UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ARMA 2022

June 29-30, 2022 Valencia, Spain

4th International Conference on Advanced Research Methods and Analytics

ADE FACULTAD DE ADMINISTRACIÓN Y DIRECCIÓN DE EMPRESAS. UPV

Departamento de Economía y Ciencias Sociales

Ivie

devstat

Congress UPV

4th International Conference on Advanced Research Methods and Analytics (CARMA 2022)

The contents of this publication have been evaluated by the Program Committee according to the procedure described in the preface. More information at http://www.carmaconf.org/

Scientific Editors

Josep Domenech
María Rosalía Vicente

Publisher

2022, Editorial Universitat Politècnica de València
www.lalibreria.upv.es / Ref.: 6106_01_01_01

Cover design by Gaia Leandri

ISBN: 978-84-1396-018-0 (print version)
Print on-demand

ISSN: 2951-9748

DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15956

Preface

Josep Domenech [1], María Rosalía Vicente [2]
[1] Dept. Economics and Social Sciences, Universitat Politècnica de València, Spain. [2] Dept. Applied Economics, Universidad de Oviedo, Spain

Abstract

Research methods in economics and social sciences are evolving with the increasing availability of Internet and Big Data sources of information. As these sources, methods, and applications become more interdisciplinary, the 4th International Conference on Advanced Research Methods and Analytics (CARMA) is an excellent forum for researchers and practitioners to exchange ideas and advances on how emerging research methods and sources are applied to different fields of social sciences as well as to discuss current and future challenges. This edition was celebrated in a hybrid format because of the uncertainties created by pandemic.

Keywords: *Big Data sources, Web scraping Social media mining, Official Statistics, Internet Econometrics, Digital transformation, global society.*

1. Preface to CARMA2022

This volume contains the selected papers of the Fourth International Conference on Advanced Research Methods and Analytics (CARMA 2022) hosted by the Universitat Politècnica de València, Spain during 29 and 30 June 2022. This fourth edition consolidates CARMA as a unique forum where Economics and Social Sciences research meets Internet and Big Data. CARMA provides researchers and practitioners with an ideal environment to exchange ideas and advances on how Internet and Big Data sources and methods contribute to overcome challenges in Economics and Social Sciences, as well as on the changes in the society after the digital transformation.

The selection of the scientific program was directed by Maria Rosalia Vicente, who led an international team of 47 scientific committee members representing institutions worldwide. Following the call for papers, the conference received 58 paper submissions from all around the globe. All submissions were reviewed by the scientific committee members under a double-blind review process. Finally, 35 papers were accepted for oral presentation during the conference, ensuring a high-quality scientific program. It covers a wide range of research topics on the Internet and Big Data, including public opinion mining, web scraping, search engine data, tourism and mobility, social behavior, data economy, or marketing and social media, among others. Additionally, 11 papers with promising work-in-progress research were selected for presentation during the conference.

The scientific program includes two keynote speakers that will review the state-of-the-art techniques and applications of the Internet and Big Data. The first keynote address is given by Giuliano Resce (University of Molise, Italy) to overview the latest digital methods for Economics and Social Sciences. The second keynote speech is delivered by Fabian Braesemann (Oxford Internet Institute, UK) and deals with the Social Data Science in the Digital Economy.

CARMA 2022 also featured two special sessions on “Big Data in Central Banks” and “Internet and Big Data in Official Statistics,” chaired by Juri Marcucci and Aidan Condron, respectively. Both sessions gave a complementary institutional perspective on how to use the Internet and Big Data sources and methods for public policy and official statistics.

The conference organizing committee would like to thank all who made this fourth edition of CARMA a great success. Specifically, thanks are indebted to the authors, scientific committee members, special session organizers, invited speakers, session chairs, reviewers, presenters, sponsors, supporters, and all the attendees. Our final words of gratitude must go to the Faculty of Business Administration and Management of the Universitat Politècnica de València for supporting CARMA 2022.

2. Organizing Committee

General chair

Josep Domènech, Universitat Politècnica de València

Scientific committee chair

María Rosalía Vicente, Universidad de Oviedo

Local organization

Eduardo Cebrián Cerdá
Pilar Malagón Selma
Héctor Martínez Cabanes
Joan Manuel Valenzuela

3. Sponsors and Supporters

Universitat Politècnica de València
Facultad de Administración y Dirección de Empresas
Departamento de Economía y Ciencias Sociales
Instituto Valenciano de Investigaciones Económicas (IVIE)
DevStat — Development of Statistics. Statistics for Development

4. Scientific committee

Anto Aasa, University of Tartu
Fernando Almeida, University of Porto & INESC TEC
José A. Álvarez-Jareño, University of Valencia
María del Pilar Ángeles, Universidad Nacional Autónoma de México
Concha Artola, Banco de España
Nikolaos Askitas, IZA – Institute of Labor Economics
Seyhmus Baloglu, University of Nevada
Catherine Beaudry, Polytechnique Montreal
Silvia Biffignandi, Consultant Economic Statistics Studies (ESS)
Ludovic Calès, European Commission, Italy
Roger H.L. Chiang, University of Cincinnati
Cihan Cobanoglu, University of South Florida
Marisol B. Correia, ESGHT-Universidade do Algarve & CiTUR
Lisa Crosato, Ca' Foscari University of Venice
Pablo de Pedraza Garcia, Italy
Giuditta de Prato, EC JRC, Spain

Index

Full papers

Abstracts

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.14936

A visual analysis of the literature on Internet neutrality

Klaudijo Klaser[1], Lucía Desamparados Pinar García[2]
[1]Università degli Studi di Trento, Dipartimento di Economia e Management, via Inama 5, 38122, Trento, Italy. [2]University of Valencia, ERI-CES and Department of Economic Analysis, Av. de los Naranjos, s/n, 46022, Valencia, Spain.

Abstract

Internet neutrality – a principle against the discrimination between Internet data packages – has been one of the most debated Internet regulation policies in the last decade. However, this debate seems to be very fragmented and there is not a global comprehension of the direction of it. In this paper we try to fill this gap circumscribing the literature on Internet neutrality. Through the open-source software VOSviewer we provide a visual analysis of the relationship between the 50 most relevant words occurring in the abstracts of scientific publications on the topic in the last 15 years.

Keywords: *Internet, Literature Analysis, Net Neutrality, VOSviewer.*

1. Introduction

Internet (or net) neutrality has been the most relevant Internet regulatory policy of the last decade (Jacobides, 2020). The principle of net neutrality requires that all Internet data packages, regardless of their content, origin, destination**,** or type of equipment used, should be equally treated. (Wu**,** 2003). In general terms we can say that the debate on the need to set the obligation of net neutrality is between: on the one hand the Content Providers with strong market power (e.g. Amazon, Google, Meta, Netflix) which defend it under the argument of maintaining Internet as an open and global network that fosters innovation; and on the other hand the Internet Service Providers (e.g. Orange, Movistar, TIM, Vodafone) which argue that net neutrality discourages investments in maintenance and extension of network capacity because free rider behavior from the Content Providers side is allowed.

The academic results about the argument are often inconclusive and the regulatory policy sometimes contradictory[1]. In other words, the literature about Internet neutrality seems to be very fragmented and there is a lack of comprehension of its global direction. Therefore, the inconsistency between the results often obtained in differents fields of literature may explain the contentious and polemic debate that still continues on whether net neutrality is necessary and how to enforce it.

In this paper, through an analysis of the words occurring in the abstracts of scientific publications, we provide a global snapshot of the academic literature on net neutrality. We show that the debate is very interdisciplinary, involving law, economics, engineering and political sciences. However, in our analysis we found two dominant macro-areas of study in the literature: the the economic debate and the legal perspective. We interpret this compartmentalization of fields as an absence of synergy between economic and legal outcomes, which adds complexity and generates confusion in the net neutrality debate.

In Section 2 we present the methodology employed to provide a visual analysis of the net neutrality literature. Section 3 presents the main results. Finally, conclusions are provided.

2. Method

On the database Web of Science[2] we looked for scientific publications which had the locution “internet neutrality” or “net neutrality” either in the title or in the abstract section.

[1] In 2015, the European Union incorporated regulations on net neutrality like the one established in the US between 2015 and 2017. In the European Union, since 2016, it is the Body of European Regulator for Economic Communications BEREC that manages the guidelines about net neutrality. In general terms, BEREC prohibits any type of quality discriminatory practice such as prioritization. However, certain financial discrimination practices such as zero-rating are allowed on a case-by-case basis. More recently, in September 2021, the Court of Justice of the European Union, by analyzing the cases of Vodafone and Deutsche Telekom in Germany, decided that zero rating offers (use of applications without data consumption) violate the net neutrality principle. This is a clear signal about how still confusing and hectic the debate is.

[2] Entered the 15th of February 2022.

This search produced a total of 368 available contributions, of which 264 articles, 52 proceeding papers and 26 book chapters. Only 296 pubblications resulted supplied with an abstract field[3]. For our analysis we focused on the co-ocurrences of words within the abstracts.

Using a minimum threshold of occurrence of 10 times – counting also if a word appeared more than once within the same abstract – we identified a total of 175 possible relevant terms. Starting from that list of words we brainstormed and selected the most important ones given the context, that is Internet neutrality, according to standard procedures in the field (Caputo et al. 2021 and Donthu et al. 2021). Thus, even if some terms occurred very frequently, we excluded from the visual analysis 92 words which had no relevance given the context or which had a too general or ambiguous meaning, in the sense that they could be easily coupled with several other words in several different ways. For instance, we excluded words such as "analysis", "content", "decision", "internet", "issue" or "paper".

In a second stage of refinement of the remaining 83 terms left we grouped the words with a coincidental or identical meaning. For example, we merged "commission" and "fcc" with "federal communications commission", "Europe" with "European Union" and "law" with "legislation". In the same way we decided to merge "costumer", "consumer", "end user", "internet user" and "user" into one single category. This further refinement left us with 48 unique words with no ambiguity or synonyms plus other 10 words derived from merging 35 similar terms.

For the visual analysis of the network between the identified words we used the open-source software VOSviewer (van Eck and Waltman 2010, 2014).

3. Results

We first start with some general indexes and then we move to the network analysis.

3.1 General bibliometric indexes

Figure 1 shows the publication and citation history of the 368 contribution identified on Web of Science starting from 2006. As it is possible to notice from the bar chart, the literature on Internet neutrality constantly grew until 2018, when we can observe the pick for both the mentioned variables. FUrthermeore, the scientific production on the theme had a significant drop during the Covid years. On the one hand this is might be a natural trend of many fields of research, because the focal center moved on the health emergency. On the other hand, this conspicuous drop is surprising given the importance that Internet had

[3] Since the number of abstracts was relatively low we decided not to exclude any publication from the following analysis.

during the lockdown periods to carry out fundamental socio-economic activities and the consequent increase in traffic volumes (Feldmann et al. 2021).

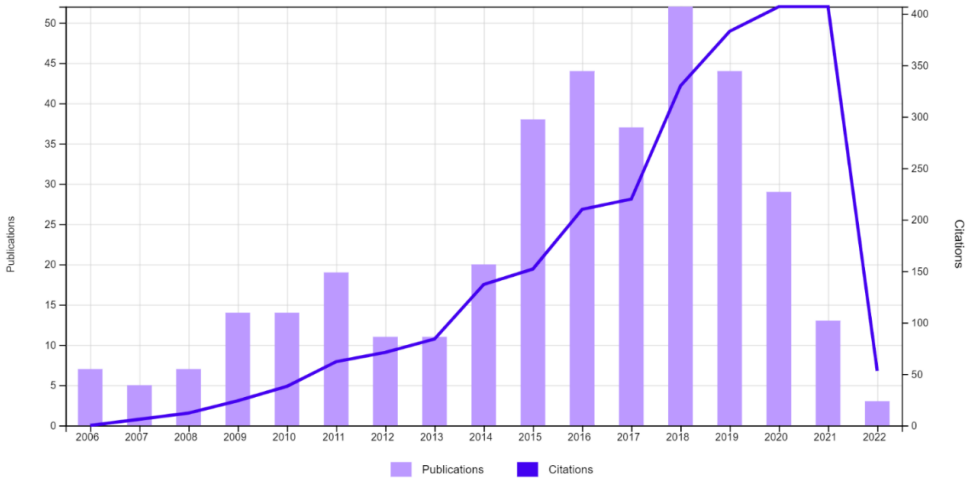


Figure 1. Publications and citations over time, 2006-2022. Source: Web of Science, Citation report

Table 1 summarizes instead the main fields of literature of the publicaitons according to the Web of Science categorization[4]. The research fields that mostly engage with the Internet neutrality debate are communication, telecommunication, information science, law, economics. Other 57 categories were present, but only less than nine contributions per category belonged to those areas. Therefore, they are not shown in Table 1.

3.2 Visual analysis of the network

Focusing on the relationship between the words present in the abstracts of the identified publications, we can notice (Fig. 2)[5] that already after the first round of refinement – thati is the elimination of non-significant words – we obtain a clear split of the 50 most relevant words in two macro-clusters (areas of research), mainly connected through the very central node represented by the term "neutrality".

[4] One publication can belong to more than one field, so there is the possibility of multiple counting.

[5] For this visualization we used a binary counting of words, which counts only the number of documents in which a term occurs. The double or triple appearance of the same word within the same abstract is counted as one. Out of the 66 total words so identified we selected the 50 most relevant.

Table 1 – Web of Science main categories of the identified contributions

Field	Times
Communication	70
Telecommunications	66
Information Science Library Science	62
Law	60
Economics	49
Engineering Electrical Electronic	42
Computer Science Information Systems	33
Computer Science Hardware Architecture	27
Computer Science Theory Methods	23
Computer Science Software Engineering	20
Management	19
Polittical Science	13

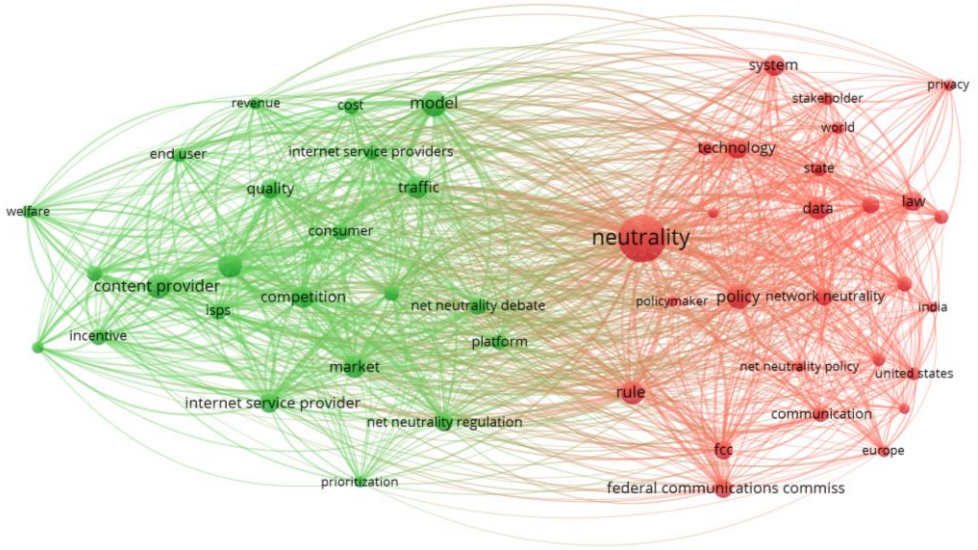


Figure 2. Network visualization of the 50 most relevant words, no merge of synonyms. Source: VOSviewer

The left cluster (green) contains words like "content provider", "internet service provider(s)", "market", "model" and "competition". This is a clear indication that one important part of the literature on Internet neutrality is basicallly focused on the normative (meant in an economic sense) configuration of the Internet market and the economic relationship between the main market actors. This is confirmed also by the co-presence of

words like “end user” or “consumer” and “welfare”, that refer to the idea of the impact of different market structures on the aggregate welfare in general and on the Internet costumers in particular and by words like “cost”, “revenue” and “incentive” that refer to a strict economic perspective. For an extensive review on the economics of net neutrality we readdress to Lee and Wu (2009), Schuett (2010), Krämer et al. (2013), Greenstain, et al. (2016), Krämer and Peitz (2018).

The right cluster (red) of Fig. 2 contains terms like “state”, “federal communication commission”, “law” and “policy”, indicating that another part of the literature has been so far focused on the legal equilibrium, that is the enforcement of the principle of Internet neutrality. However, there are still some words like “technology” or “stakeholder” that, despite having some intrinsic meaning in the Internet context, do not have a clear collocation within the network. For an extensive review on regulatory instruments and competition law linked to net neutrality we readdress to Owen (2014), Ohlhausen (2016), Maniadaki (2019) and Comeig et al. (2022).

Furthermore, analyzing the 10 strongest links within the network of Fig. 2 we observed that the central node of the network is exactly the word “neutrality”, and this mainly occurs together with very significant terms like “market” and “policy”.

After the second stage of refinement – that is the merge of synonims – we can observe from Fig. 46 how the network becomes subject to a more detailed interpretation.

As before (Fig. 3) we can still observe a clear split between the economic and the law perspective, with “service provider” and “regulation” constituting a sort of focal points of an ellipsis around which all the other words are distributed: the two focuses are indeed connected by the strongest link in the whole network. We interpret this in a straightforward way: the literature on net neutrality mainly deals with the regulation of Internet Service Providers, seen as the issue. However, compared to Fig 2, in Fig. 3 a new cluster is formed (light blue). The latter, together with some words in green and red – like “cost” or “infrastructure” –, indicates in our opinion that there is also a part of the literature focused on the technological and infrastructural dimension of Internet neutrality, probably a reflection of the research categories listed in Section 3.1.

[6] For this visualization we used a full counting method of words, includes the total number of occurrences of a term within all the documents, even if it appears multiple times within the same abstract. Out of the 58 total words that respected our criteria, we selected the 50 most relevant. We added the further filter of having a minimum cluster of five words.

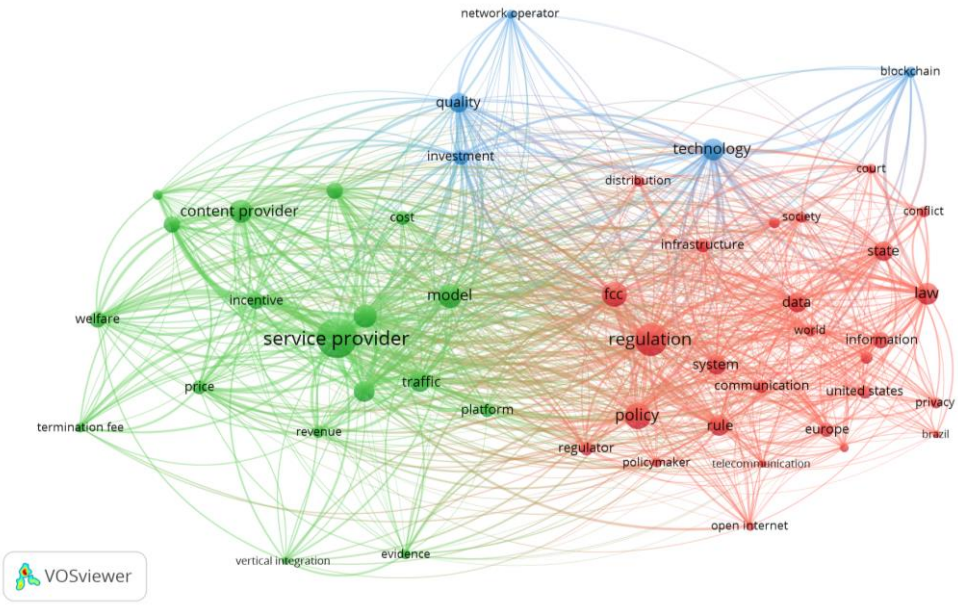


Figure 3. Network visualization of the 50 most important words. Source: VOSviewer

4. Conclusions

By using the VOSviwer software we obtained a visual analysis of the academic literature on Internet neutrality. The network outputs disclosed two macro-clusters of words representing two main areas of research with little interconnections: the economic studies and the legal perspective. The third identified cluster, representing the results related to the more technical fields (telecommunications, electronic engineering or information systems), stays somehow in between and draws upon the former two.

In conclusion, with our paper, we pointed out that there seems to be no convergence in the debate on Internet neutrality, with the economic and legal perspectives still focused on very different elements, even if both sides agree on the need to regulate the market through the regulation of Internet Service Providers. Policymakers should take into account this divide and try to include the results of these two macro-areas in a transversal way into the regulation. Indeed, only an overarching approach can guarantee an ergonomic regulation for such a technically and economically dynamic sector as the Internet market.

References

Caputo, A., Pizzi, S., Pellegrini, M. M., & Dabić, M. (2021). Digitalization and business models: Where are we going? A science map of the field. Journal of Business Research, 123, 489-501. https://doi.org/10.1016/j.jbusres.2020.09.053

Comeig, I., Klaser, K., & Pinar, L. D. (2022). The paradox of (Inter) net neutrality: An experiment on ex-ante antitrust regulation. Technological Forecasting and Social Change, 175, 121405. DOI https://doi.org/10.1016/j.techfore.2021.121405

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. Journal of Business Research, 133, 285-296. https://doi.org/10.1016/j.jbusres.2021.04.070

Faulhaber, G. R. (2011). The economics of network neutrality. Regulation, 34, 18.

Feldmann, A., Gasser, O., Lichtblau, F., Pujol, E., Poese, I., Dietzel, C., Wagner, D., Wichtlhuber, M., Tapiador, J., Vallina-Rodriguez, N., Hohlfeld, O. & Smaragdakis, G. (2021). A year in lockdown: how the waves of COVID-19 impact internet traffic. Communications of the ACM, 64(7), 101-108. DOI: https://doi.org/10.1145/3465212

Greenstein, S., Peitz, M., Valletti, T., 2016. Net neutrality: A fast lane to understanding the trade-offs. Journal of Economic Perspectives, 30 (2), 127-50. DOI: 10.1257/jep.30.2.127

Jacobides, M. G. (2020). Regulating Big Tech in Europe: Why, so what, and how understanding their business models and ecosystems can make a difference, White paper. DOI http://dx.doi.org/10.2139/ssrn.3765324

Krämer, J., Peitz, M., 2018,. A fresh look at zero-rating. Telecommunications Policy, 42 (7), 501-513. https://doi.org/10.1016/j.telpol.2018.06.005

Lee, R. S., Wu, T., 2009. Subsidizing creativity through network design: Zero- pricing and net neutrality. Journal of Economic Perspectives, 23, 61–76. DOI: 10.1257/jep.23.3.61

Ohlhausen, M. K. (2016). Antitrust over Net Neutrality: Why We Should Take Competition in Braodband Seriously. *Colo. Tech. LJ*, *15*, 119.

Owen, B. M. (2014). Net Neutrality: Is Antitrust Law More Effective than Regulation in Protecting Consumers and Innovation?. *Available at SSRN 2463823*.

Schuett, F., 2010. Network neutrality: A survey of the economic literature. Review of Network Economics, 9 (2). https://doi.org/10.2202/1446-9022.1224

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3

van Eck N.J., Waltman L. (2014) Visualizing Bibliometric Networks. In: Ding Y., Rousseau R., Wolfram D. (eds) Measuring Scholarly Impact. Springer, Cham. https://doi.org/10.1007/978-3-319-10377-8_13

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15013

Ethical Behavior and Legal Regulations in Artificial Intelligence

Thomas Hauer
Department of Social Sciences, Technical University of Ostrava (VSB), Czech Republic.

Abstract

The European Union is also working intensively on the ethical development and use of artificial intelligence. The European Parliament, which commented on the issue of intelligent autonomous robots in January 2017, expressed the need to supplement the existing legal framework with ethical principles and an "effective ethical framework for the development, manufacture, use and modification of robots". This ethical framework should be based on the principles of expediency, harmlessness, autonomy and justice. The study analyzes the interconnection of ethical and legal rules in the field of AI and shows possible directions of development

Keywords: *machine ethics; values; ethical behavior; legal regulations; AI autonomous machines and platforms; moral philosophy.*

1. Introduction: Moral machine

The current state of affairs regarding the topic of Machine Ethics and the ethical autonomy of AI algorithms can be summarized as follows. Machines and platforms equipped with advanced AI and machine learning algorithms may differ in what their purpose is, even coming up with surprising solutions, plans and designs, but only in serving the goals that we set for them. The algorithm whose programmed goal is to "prepare a good dinner" may decide to serve steak, lasagne, or even a tasty new dish it creates by itself, but cannot decide to assassinate its owner, take his car and go to Iceland to rescue penguins because it became a vegetarian. Similarly, algorithms that are part of weapons systems, drones and unmanned aircraft, which choose their own targets without human intervention, fulfil their mission, adapt and respond to unforeseen circumstances with minimal human oversight, but cannot change or cancel their own mission if they had any moral reservations. It seems most rational to consider the ethical issues that may arise from these technologies before the technology is widely disseminated and deployed in practice (Allen et al., 2005). Moral machines capable of autonomous ethical reasoning and decision-making without any human oversight will necessarily emerge in the future. However, recent approaches to machine ethics have shown that researchers and programmers need to seek advice from philosophers and ethics to avoid novices' mistakes and to understand deep-rooted methodological problems in ethics better. If they fail to address these problems properly, their efforts to build adequate moral machines will be severely hampered. How to decide what steps are morally right is one of the most difficult questions in our lives. Understanding the ethical pitfalls and challenges associated with these decisions is essential to building intelligent, moral machines.

The first area of AI ethics, research deals with creating and applying ethical rules and standards. This area formulates recommendations that should respect fundamental rights, applicable regulations, and guiding principles and values, ensuring the ethical purpose of AI while ensuring their technical robustness and reliability. Ethical requirements and rules should be included in the various steps of the AI creation process, from research, through data collection, the initial design phase, testing the system to its deployment and practical application. Thus, this area of AI ethics mainly addresses questions about how developers should behave to minimize the ethical damage that may occur in AI, whether due to poor (unethical) design, inappropriate use, or misuse. This branch is commonly referred to as robotic ethics and has already led to the formulation of many declarations (Montreal Declaration for a Responsible Development of Artificial Intelligence), to postulating the main ethical principles and rules (Boddington, 2017; Boden 2016), formulating standards for producers and developers (International Organization for Standardization 2016), and designing best practices for developing and manufacturing platforms with AI (IEEE Standards Association 2017).

2. Machine ethics and ethical autonomy of AI

The most important and rapidly developing area in terms of AI ethics is the layer of ethics of autonomous intelligent systems and AI platforms evolving over time through self-learning from Machine Ethics data. The second branch of AI ethics research deals with how robots and AI platforms can behave ethically autonomously (Allen et al., 2005). This area of AI ethics research is referred to as Machine Ethics. The main aims and assumptions of this branch of machine ethics have been formulated by the authors and participants of the AAAI Fall 2005 symposium[1]. On this basis, authors W. Wallach and C. Allen have developed the term - artificial moral agents (AMAs), which is now used in this area of AI ethics. AMAs research can be considered a scientific endeavour to answer the question of whether, in principle, it is possible to model moral behaviour – whether ethical rules are convertible to algorithms autonomously (Anderson & Anderson, 2010). In many areas, it is impractical to wait for human decision-making because the amount of data, the speed of response, and waiting for human intervention make the decision impractical. In recent years, the interdisciplinary field of machine ethics – how to use machine learning to create algorithms with ethical rules to become either implicit or explicit moral factors – has become extremely important due to current and expected technological developments in computer science, artificial intelligence (AI) and Robotics (Gunkel, 2014; Lin et al., 2012). On the basis of great technological advances in AI, the emergence of fully autonomous, human-like, intelligent robots capable of ethical reasoning and decision-making seems inevitable. “Robots with moral decision making will become a technological necessity (Wallach, 2007). “Artificial Moral Agents are necessary and, in a weak sense, inevitable!”

I use the term “AMAs” to refer only to algorithms, platforms, and robots that are explicit ethical agents (those who have an explicit set of normative principles that they can use in decision making). AMAs come into play only when the task is fully automated (Moor 2006). Regarding autonomous technology, the transfer of moral roles to machines is not a matter of specific choice. Conversely, such delegation depends on the general characteristics of the automated task. If we choose to automate a task that does not, in any way, require the exercise of moral authority when performed by humans, there is no need for moral machines (AMAs). And vice versa, if a task requires some form of moral authority when it is performed by humans, then delegating the same task to autonomous machines necessarily means transferring a moral role. However, if autonomous machines are deployed to perform tasks that, when performed by human beings, show a moral aspect, then we either decide not to address that side, or we need to find a way to implement some form of morally relevant data processing and action selection into the machines themselves.

[1] https://www.aaai.org/Library/Symposia/Fall/fs05-06.php

"As systems get sophisticated even more, and their ability to function autonomously in different contexts and environments expands, it will become more important for them to have 'ethical subroutines' of their own" (Allen et al. 2006, p. 14).

This, of course, would not mean that the algorithm would "become" human. Instead, it would provide us with tools for cooperation that would not offend our moral sensitivity and satisfy our moral expectations. The general objective of AMAs research is complex. Researchers want to create machines with autonomous ethical decision making. Thus, it seems that the moral issues arising from the autonomous functioning of the machine need to be specifically addressed (Allen et al. 2005). This is what Machine Ethics is trying to do: to build machines that work not only efficiently, not only safely, but also in a morally satisfactory way – that is, in a way that would ideally prevent moral harm and agree to confirm moral goodness. In the long run, AI will be ubiquitous, and it should, because in many areas, it can do better work than humans. Not only will their intellectual prowess exceed ours, but their moral judgment may be better.

3. Human Well-being with Autonomous and Intelligent Systems

We seem to be in an intermediate period before the mass diffusion of a new and fundamental technology, which is advanced AI algorithms (Anderson & Anderson, 2007; Allen et al., 2006; Boden, 2016; Moor, 2006). As a strategic technology, AI is now rapidly developed and used around the world. However, it also brings with it new risks for the future of jobs and raises major legal and ethical questions (Lin et al., 2012). AI technologies should be developed, deployed and used with an ethical purpose and based on respect for fundamental rights, taking into account societal values and ethical principles of beneficence, non-maleficence, human autonomy, justice and explainability (Moor, 2006; Wallach, 2007). It is a prerequisite for ensuring the credibility of AI. In order to address the ethical risks and make the most of the opportunities that AI brings, the European Commission has published a European strategy on the Ethics of AI. It puts humans at the centre of AI development and defines so-called Human Centric Artificial Intelligence (HCAI).

4. Artificial Intelligence for Europe

At the European level, the European Commission's Communication Artificial Intelligence for Europe[2] and the Coordinated Plan on Artificial Intelligence "Made in Europe"[3] issued by the European Commission in December 2018 are the starting documents in the field of

[2] https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe

[3] https://ec.europa.eu/digital-single-market/en/artificial-intelligence#Coordinated-EU-Plan-on-Artificial-Intelligence

AI. This Coordinated Plan sets out the European Union's strategic objectives and priorities in the field of artificial intelligence. It is the overarching European strategy for AI, which was developed in collaboration with the Member States and calls on the Member States at the national level to -implement the Coordinated Plan. The Member States are thus required to submit national AI strategies by the end of 2019 at the latest, including setting investment measures and implementation plans. In April 2018, the European Commission published a Communication on Artificial Intelligence for Europe, proposing a comprehensive and integrated European approach to AI. According to this document, the EU should respond to the current developments in AI and create a pan-European initiative focusing on three pillars:

- increasing technological and industrial capacity and deploying artificial intelligence across the economy,
- focus on socio-economic issues arising in the context of artificial intelligence (AI)
- providing an ethical and legal framework for AI technology.

The third pillar of EC Communication deals with legal and ethical issues related to AI. The European Commission has committed to developing ethical standards and guidelines for the use of AI. In this context, the High-Level Expert Group on Artificial Intelligence[4], which brings together AI experts, has been established to develop guidelines and recommendations on AI ethics. As part of the Communication, the EC also initiated the creation of the so-called European Artificial Intelligence Alliance, a broad discussion platform for various interest groups. The main strategic documents on AI ethical issues, which also provide a framework for this area, are:

1. Draft Ethics guidelines for trustworthy AI – published on 18 December 2018[5]
2. Communication: Building Trust in Human Centric Artificial Intelligence – published on 8 April 2019[6]

The third pillar, focusing on the ethical and legal context of AI development, is based on the above-mentioned strategic documents of the European Commission and formulates the main objective based on them. Credible human-centred AI has two components:

1. it should respect fundamental rights, applicable regulations, and the guiding principles and values shared in the EU, thereby ensuring the "ethical purpose" of AI
2. it should be technically robust and reliable because even without intentional malice, AI technologies can cause unintended harm or damage.

[4] https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

[5] https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top

[6] https://ec.europa.eu/digital-single-market/en/news/communication-building-trust-human-centric-artificial-intelligence

5. Conclusion: Trustworthy AI

The ethical requirements for trustworthy AI should be incorporated into every step of the AI algorithm development process, from research, data collection, initial design phases, system testing, and deployment and use in practice. And how things really are? Do we really consider the benefits of AI versus the possible risks? Do we currently emphasize the ethical dimension of the development and implementation of new innovations in robotics and artificial intelligence?

Acknowledgements

Published with the support of the Technology Agency of the Czech Republic (TA CR), project number TL01000299.

References

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. AI Magazine, 28(4), pp. 15-26

Anderson, M., & Anderson, S. L. (2010). Robot be good: A call for ethical autonomous machines. Scientific American, 303(4), pp. 15–24

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. Ethics and Information Technology, 7(3), pp. 149–155. https://doi.org/10.1007/s10676-006-0004-4.

Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics? IEEE Intelligent Systems, 21(4), 12–17. https://doi.org/10.1109/MIS.2006.83.

Boddington, P. (2017). Towards a Code of Ethics for Artificial Intelligence (Artificial Intelligence: Foundations, Theory, and Algorithms), Springer; 1st ed

Boden, A. M. (2016). AI: Its Nature and Future, 1st Edition, Oxford University Press

Bryson, J. (2008). Robots should be slaves. In Y. Wilks (Ed.), Close Engagements with artificial companions: Key social, psychological, ethical and design issue (pp. 63–74). Amsterdam: John Benjamins Publishing.

Dignum, V. 2017. Responsible Artificial intelligence: Designing AI for human values, ITU Journal: ICT Discoveries, Special Issue No. 1, 25 Sept. 2017,

Gunkel, D. J. (2014). A vindication of the rights of machines. Philosophy & Technology, 27(1), pp. 113–132. https://link.springer.com/article/10.1007/s13347-013-0121-z

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. IEEE Intelligent Systems, 21(4), pp. 18–21. https://doi.org/10.1109/MIS.2006.80

Lin, P., Abney, K., Bekey, G. (2012). Robot ethics: the ethical and social implications of robotics. MIT Press, Cambridge, MA

Wallach, W. (2007). Implementing moral decision making faculties in computers and robots. AI & Society, 22(4), pp. 463–475. https://doi.org/10.1007/s0014 6-007-0093-6.

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15016

The effects of the e-tailer's reputation, the e-tailer's familiarity, and the relevance of the e-tailer's social media communication on impulse buying

Yosra Akrimi
Sfax University, Tunisia.

Abstract

In this paper, we study how the interaction with consumers on social media impacts impulse buying using the data of 396 questionnaires. The results confirm that the e-tailer's reputation, familiarity, and the relevance of his social media communication positively impact trust and impulse buying. We have also found that social distance moderates the effect of the e-tailer's reputation and the perceived relevance of the e-tailer's social media communication on impulse buying. Knowing how social media communication influences impulse buying enables companies to strengthen synergy between social media presence and the online store.

Keywords: *Social media communication, impulse buying, online retailer, familiarity, reputation.*

1. Introduction

For more than a decade, social media has aroused a huge enthusiasm among economic players. Social media served as a communication medium between the user and his circle of friends and acquaintances (Husain et al, 2016). Their use quickly diversified. Consumers now use social networks to assess their consumption experience, compare offers, discover new products, recommend brands, or lead a boycott campaign (Anderson et al, 2011; Aragoncillo and Orus, 2018). Social networks are social platforms that brands and retailers use to interact with their customers. E-tailers and brands diversify the content they publish on social networks to engage and retain the consumers (Schivinski and Dabrowski, 2014). Thereby, social networks have become an efficient tool for managing and developing customer relationships and gaining customer trust (Getry et al, 2018).

Social media has brought about a change in the power balance between the brand and the consumer, which acts on the brand's communication almost instantly. The brand or the online retailer is therefore forced to manage the flow of content created by consumers by enhancing it if it is in its favor and rationalizing it if it is more critical. Brand pages on social media can serve as an effective tool for building a trusting relationship with fans by instantly answering their questions, engaging in conversations with them, and taking their suggestions and complaints into account (McClure and Seock, 2020). With the tremendous development of the Internet and social media, and therefore the ascent of multi-channel distribution, customers are exposed to marketing stimuli that promote impulse buying (Dawson and Kim, 2009). The web's flexibleness and accessibility have also augmented the tendency to buy online impulsively (Wu et al, 2016). Previous research has studied the effect of the synergy between communication on social networks and traditional communication (TV, email) on sales and purchase decisions (Kumar et al, 2016; Tarabieh, 2017). However, to our knowledge, a limited amount of research has focused on studying the link between the presence of the e-tailer on social networks and its website and its effect on impulse buying. In Marketing and information systems literature, a wide effort has been dedicated to establishing the determinants (consumer response, store cues, situational stimuli, and product characteristics) that influence impulse buying (Chan et al, 2017). However, there is a lack of research on the role of brand communication on social media in impulse buying. Chen et al (2018) state that several research studies have analyzed the effect of social media on planned buying but little research has tried to explore the facilitating role of social media in impulse buying.

This research aims to understand the synergy between the online retailer's presence on social media and his official website and its effect on impulse buying. Analyzing this synergy helps identify the underlying mechanisms which push customers to buy on a whim. Therefore, our research question is: how the non-technical aspects of the online retailer, for instance,

familiarity and reputation, and the relevance of the online retailer's official communication on social media affect impulse buying on his website?

To meet the research's objectives, first, we present the theoretical framework and the research model. Next, we discuss the results in light of previous research. Finally, we explain the research limits and future avenues of research.

2. Theoretical framework and hypothesis development

2.1. Impulse buying

Many definitions have tried to capture the extent and complexity of online impulse buying. Verhagen and Van Dolen (2011) conceptualize online impulse buying as an immediate and spontaneous purchase decision without prior reflection as in the planned purchase. Bayley and Nancarrow (1998) outline impulse buying as an unexpected, compelling shopping behavior within which the velocity of an impulse decision process precludes deliberate thinking of other information and selections. Wolny and Charoensuksai (2014) propose the concept of the consumer journey that refers to the multiple contacts with the product. The consumer encounters brands through websites, physical stores, and social networks. This myriad of stimuli can trigger impulse buying.

2.2. Perceived familiarity

Brand familiarity describes the consumer's experience or knowledge of the brand (Alba and Hutchinson 1987). The experience may result from a direct exchange following the purchase of the brand or may result from an indirect experience in the street (urban display, street marketing) or in an advertisement. Consumers can also hear about the brand from those around them, so they get to know the brand by word of mouth (Park and Stoel, 2005). The purchase of familiar brands seems to be automatic since the consumer spends less time buying familiar brands. In this case, the consumer uses brand familiarity to facilitate the decision-making process (Ha and Perks, 2005). According to Benedicktus et al (2010), brand familiarity induces trust and purchase intention. Ha and Perks (2005) argue that website familiarity is a prerequisite for consumer trust and satisfaction.

H1a: E-tailer’s familiarity is positively related to trust .

H1b: E-tailer’s familiarity is positively related to impulse buying.

2.3. Online retailer’s reputation on social media

Brand reputation can be built through brand marketing strategy as well as word of mouth. Good brand reputation assumes reliability, integrity, and quality (Creed and Miles, 1996).

Pauwels et al (2016) identify two types of media that combine to constitute the brand informational capital. The official media owned by the brand is called "owned media", like his website. which presents all the information and offers necessary to enable the purchase. The second type is unofficial media or "earned media" that are created or managed by consumers independently or in collaboration with the e-tailer mainly on social networks (community, fan page) (Sinclaire & Vogus, 2011).

H2a: Online retailer's reputation on social media is positively related to perceived familiarity.

H2b: Online retailer's reputation on social media is positively related to consumer trust.

H2c: Online retailer's reputation on social media is positively related to impulse buying on its website.

2.4. The relevance of the e-tailer's content on social media

Sperber and Wilson (1995) developed the relevance theory which states that communication is not just about sending and receiving a message, but it is about its relevance. Content is relevant if its interpretation is helpful. Hence, the message's process improves the receiver's knowledge or corrects his errors. (Xu and Zhou, 2013; Wilson and Sperber, 2002). However, an irrelevant message does not capture the receiver's attention and consequently will not be interpreted. Thus, relevant content is attractive and non-intrusive (Cook, 1992, Pérez, 2000). Ahn and Beilenson (2011) argue that if the advertisement is considered relevant, it will have a better chance of generating positive emotional and behavioral responses such as the purchase decision. Voorvled (2019) states that the messages posted by the brand on social networks are called "content" because they are mixed with the content generated by the users. This mix can make the content more relevant and engaging for the consumer. Social Media communication reduces the uncertainty that prevents the establishment of a trust bond between the consumer and the brand. Building a trust bond facilitates engagement in purchasing behavior (Tatar and ErenErdogmus, 2016; Ebrahim, 2019).

H3a: The relevance of the brand's communication on social networks is positively related to the perceived familiarity.

H3b: The relevance of the brand's communication on social networks is positively related to the e-tailer's reputation.

H3c: The relevance of the brand's communication on social networks is positively related to consumer trust

H3d: The relevance of the brand's communication on social networks is positively related to impulse buying.

2.5. Consumer trust and impulse buying

Consumer trust in an online retailer refers to consumer beliefs about the potential behavior of the e-tailer. Hence, trust refers to the consumer's expectations about the e-tailer's respect for his promises when conducting the transaction (Ou and Sia, 2010). Trust is a significant determinant of purchase intention in general and even more in online exchange settings. Trust is important in online shopping due to the vendor opportunistic behavior can behave opportunistically and the intangibility of products (Gefen, 2000). Trust is of particular importance in online transactions because it conditions the consumer's willingness to buy online or not (Yoon and Occena, 2015). We hypothesize the following:

H4: Trust in the e-tailer is positively related to impulse buying on its website.

2.6. Perceived social distance on social media

Social distance is an individual perception of closeness or intimacy between oneself and another individual or group (Magee and Smith, 2013). Constant interaction reduces social distance. The greater the desire for affiliation or belonging to the group, the smaller the distance between its members (Magee, 2020). The advent of online social networks has redefined the perception of social distance. By abolishing physical and time constraints, forming friendships is just a click away (Pappalardo et al, 2012). Online social networks have multiplied the possibilities for networking. These social networks are a source of non-redundant information. They facilitate the creation and the maintenance of close relationships (Zhang et al, 2011; Grabner-Kräuter and Bitter, 2015).

Therefore, we suggest:

H5a: Perceived **s**ocial distance moderates the relationship between the e-tailer's reputation on social media and impulse buying on its website.

H5b: Perceived social distance moderates the relationship between the relevance of the e-tailer's communication on social media and impulse buying on its website.

3. Method

3.1. Sample and data collection

The final sample contains 396 respondents who have purchased from an online store. 280 respondents confirm that they made an impulse purchase on this online store. This e-tailer develops content on Facebook and encourages its fans to share posted content to receive gifts or discount vouchers. Our sample is made of 41 % males and 59 % females. 70% of respondents are in the age category of (20-45) years

3.2. Measures

The constructs we used are well established in the literature and have good validity. To measure the constructs, we used Likert-type scales with a five-point format.

3.3. Convergent and discriminant validity

The convergent validity of each construct is above the required threshold (alphas: 0.80; AVEs: 0.50) (Table 2). To verify the convergent validity of each construct, its variance Extracted (AVE) must be greater than 0.5 according to the criterion of Fornell-Larcker (1981). The discriminant validity which refers to the sensitivity of the measurement scales is established for all the constructs (Zait and Bertea, 2011).

Table 1: Reliability and convergent validity statistics

Construct	Cronbach's alpha	AVE
Familiarity	0.988	0.944
Reputation	0.982	0.936
Relevance of communication	0.977	0.916
Trust	0.956	0.807
Impulse buying	0.966	0.875
Social distance	0.976	0.909

4. Results

4.1. Measurement model results

The measurement model indices satisfy the required thresholds (Hoe, 2008). A chi-square of 1044.185 with 583 degrees of freedom. Other goodness-of-fit indices indicate an acceptable fit [CFI] =0,984; NFI= 0,965; RMSEA = 0.043.

4.2. Structural equation modeling

The structural model describes the links between constructs (Das, 2014). Our structural model establishes the links between the e-tailer's familiarity, e-tailer's reputation, the relevance of the e-tailer's communication on social media, consumer trust, and impulse buying. The linkages (as stated in H1–H5) were tested with Maximum Likelihood. Maximum likelihood is used to assess the hypothesis and to evaluate the structural model.

Our results indicated a good fit with the data (Ding et al, 1995). (x2 = 655,369 p < 0.000; CMIN/DF = 1.757; NFI = 0.974 ; IFI = 0.986; TLI = 0.986; CFI = 0.988; RMSEA = 0.042).

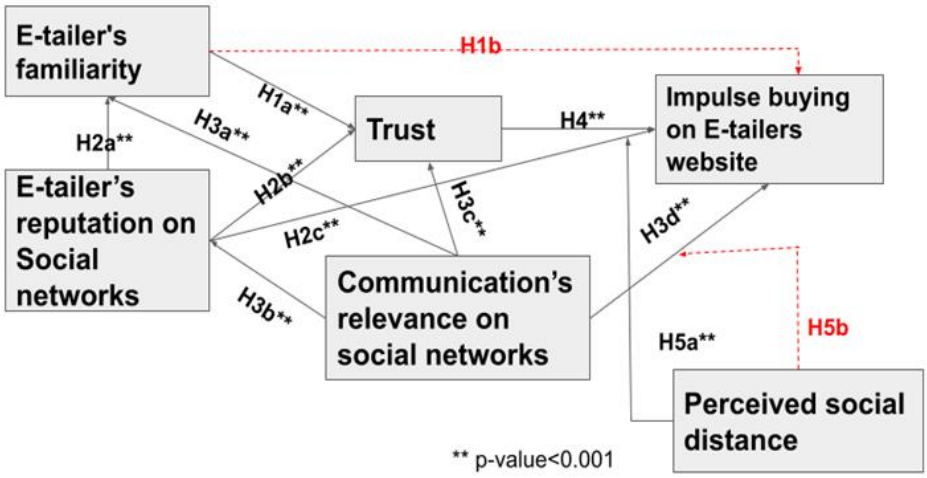


Figure 1. Research Model

5. Discussion

We have found that familiarity with the e-tailer positively and significantly impacts consumer trust (H1a: ß=0,055; p value<0.001). However, we do not find a direct effect (H1b is rejected) of familiarity on impulse buying. Likewise, H2a (ß=0,030; p value<0.001) is validated. E-tailer's reputation on social networks has a positive impact on familiarity. Familiarity captures the direct and the indirect experience with a brand within a consumer's memory. The more the consumer interacts with the online retailer, the more it is familiar to him (Campbell and Keller, 2003). H2b (ß=0.053; p value<0.001) and H2c (ß=0,046; p value=0.001) are also supported. Furthermore, online retailer's reputation consolidates trust and facilitates the purchase decision (Li, 2014; Josang et al, 2007). The hypotheses H3a (ß=0,032; p value=0.036), H3b (ß=0,050; p value<0.001), H3c (ß=0,046; p-value <0.001) and H3d (ß=0,026; p value<0.001) are supported. Kim and Jonson (2016) suggest that social media experience improves brand image and induces impulse buying. Brands use social media to maintain a rich and constant interaction with the consumer. Relevant content on social media forges a closeness and thus establishes a trusting relationship. (Khadim et al., 2018) . As we have predicted, H4 (ß=0,044; p-value <0.001) is supported. Consumer trust varies enormously depending on the context. Internet transactions are sensitive for the consumer because the risk of e-tailer misconduct is high (non-compliance with delivery deadlines, non-compliance of the delivered product with the ordered product, the opacity of the product

return policy, faulty management of complaints). We have found that perceived social distance strengthens the link between the online retailer's reputation on social media and impulse buying (H5a, ß3=0.711; p value <0.001;(H5 b, ß3=1.028; p value=<0.001). According to Chen et al. (2018), when consumers consider that they are close or similar to other consumers on social media, the perceived social distance is minimal. Accordingly, consumers can trust each other, especially since the experiences posted on social networks are well supported. Chen et al. (2016) found that the number of likes on social media commercial content is a good indicator of consumer impulsivity.

6. Conclusion and Implications

In this study, we focused on the online retailer's familiarity and reputation, which are valuable assets not directly linked to its website's technical attributes but forged by its communication strategy. We also tried to analyze the effects of perceived relevance of the content shared by the e-tailer on social networks on online retailer familiarity and reputation and impulse buying on its online store. The interweaving of the message shared "officially" by the online retailer and the users' content combine to form its reputation and increase its perceived familiarity. Sharing "stories" and experiences and the desire to maintain a connection with other users through interaction is the backbone of social networks. Consumers who easily migrate from the retailer's official page on social networks to its website to make an impulse purchase may constitute a privileged target to which it is necessary to personalize the offer. The synergy between the social media presence of the retailer and its website reinforces trust and facilitates impulse buying.The online retailer's familiarity, reputation, and the relevance of its social media content help build this synergy. Each interface will constitute the relay to other canals. For example sharing content from the e-tailer website on social networks to recommend products or buying or ordering the product by going from the brand's page on social networks to its website and vice versa.

6. Limitations and future research

It is necessary to test the model by increasing the number of online retailers who can enjoy different levels of familiarity or reputation to enhance the generalization of the results. Second, Future research may include control variables such as money and time availability, product category, gender, and age to increase the research's representativeness.

References

Anderson, MSims, JPrice, J. and Brusa, J. (2011), "Turning 'like' to 'buy' social media emerges as a commerce channel", Booz and Company Inc, available at: http://pwc.to/2kxna3V.

Aragoncillo, Laura, and Carlos Orus (2018) "Impulse buying behavior: an online-offline comparative and the impact of social media." *Spanish Journal of Marketing-ESIC*.

Campbell, M. C., & Keller, K. L. (2003). Brand familiarity and advertising repetition effects. *Journal of consumer research*, *30*(2), 292-304.

Chen, C. C., & Yao, J. Y. (2018). What drives impulse buying behaviors in a mobile auction? The perspective of the Stimulus-Organism-Response model. *Telematics and Informatics*, *35*(5), 1249-1262.

Gretry, A., Horváth, C., Belei, N., & van Riel, A. C. (2017). "Don't pretend to be my friend!" When an informal brand communication style backfires on social media. *Journal of Business Research*, *74*, 77-89.

Kaya, B., Behravesh, E., Abubakar, A. M., Kaya, O. S., & Orús, C. (2019). The moderating role of website familiarity in the relationships between e-service quality, e-satisfaction, and e-loyalty. *Journal of Internet Commerce*, *18*(4), 369-394.

Kim, D. J., Ferrin, D. L., & Rao, H. R. (2008). A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. *Decision support systems*, *44*(2), 544-564.

Magee, Joe C., and Pamela K. Smith (2013). "The social distance theory of power." *Personality and social psychology review* 17.2, 158-186.

Magee, J. C. (2020). Power and social distance. *Current opinion in psychology*, *33*, 33-37.

Ou, C. X., & Sia, C. L. (2010). Consumer trust and distrust: An issue of website design. *International Journal of Human-Computer Studies*, *68*(12), 913-934.

Pappalardo, Luca, Giulio Rossetti, and Dino Pedreschi (2012). "How Well Do We Know Each Other?" Detecting Tie Strength in Multidimensional Social Networks." *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE.

Verhagen, T., & Van Dolen, W. (2011). The influence of online store beliefs on consumer online impulse buying: A model and empirical application. *Information & Management*, *48*(8), 320-327.

Xu, Zhaohui, and Yanchun Zhou (2013). "Relevance Theory and Its Application to Advertising Interpretation." *Theory & Practice in Language Studies* 3.3.

Zaiţ, Adriana, and P. S. P. E. Bertea. (2011) "Methods for testing discriminant validity." *Management & Marketing Journal* 9.2, 217-224.

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15022

Exploring Redditors' Topics with Natural Language Processing

Yilang Zhao
Department of Curriculum and Instruction, University of Wisconsin–Madison, USA

Abstract

This paper examines how people in Reddit develop topics across threads in a given subreddit and how discussions concentrate on the topic in given threads with natural language processing (NLP) methods. By implementing an LDA topic model and TF-IDF models, this paper discovers people's aggregated concerns are related to real-world issues and their discussions are concentrative considering the topics they discuss.

Keywords: *Reddit, online discussion community, online discussion topics, natural language processing, LDA, TF-IDF*

1. Introduction and Theoretical Backgrounds

Reddit is a popular online discussion community among the young generation today, in which a lot of online discussions take place. Since topics of those discussions significantly vary, the Reddit community is divided into user-created subreddits that only allow posts with certain themes. Subreddit is regarded as real communities in virtual space in which people have their rules, languages, rules, etc. (Medvedev et al., 2017). Thus, it is worth investigating how people build up their community by exploring what they care about beyond the topic of a specific thread in a subreddit channel.

Since people's discourses are likely to focus on topic-relevant content in a given thread, scrutinizing the topics that come from a couple of threads can be constructive to understand the core interest of a subreddit community. In addition, as subreddits are self-organized spaces, its information spread yields to a more dynamic way compared to direct follower social media such as Facebook or Twitter (Medvedev et al., 2017). That is to say, the posts on Reddit may vary a lot, even in one single thread with a fixed topic. Fang et al. (2016) discover that jokes and funny comments are much more underpredicted than controversial comments, which they think can to some extent develop the discussion. Most jokes and funny comments, however, only entertain the discussion participants and may not be helpful with the discussion, yet more concentrated discussions may promote more constructive conversations on Reddit. Therefore, we are also interested in the level of concentration in a specific subreddit channel. The contribution of this research is from two aspects: one is about the community itself, and the other is about its practical application. Better understanding the topics in a subreddit channel can improve the user experience and further elevate the engagement level of participating in the discussion in online communities. As for the potential practical contribution, Park and Conway (2017) discover that public health relevant discussion on Reddit can predict the trending public interest in some public health issues and serve as an information source for certain user groups. Thus, enhancing the quality of the discussion will make some Reddit discussions more reliable information providers. Therefore, in this study, our research questions are RQ 1. what do people concern about across topics in a subreddit? and RQ 2. how do discussions in a subreddit concentrate on topics?

2. Methods

2.1. Data Collection

To answer the research questions, we collected comments from a subreddit, r/science, which is the eighth most popular subreddit channel and has over 25 million subscribers. It is an online community where people can share the latest science news and discuss. The threads we target are the top 10 hot threads of the year 2020 (see Appendix), which can be ranked by

Reddit built-in filter. These 10 topics have 20,273 comments in total when we used PRAW (Python Reddit API Wrapper) to retrieve on December 10, 2020. We stored retrieved data in a pandas data frame in which each comment is a row and has a tag of whether it is a submission (the first comment of a thread) or a following comment.

2.2. Data Analysis

As for our data analysis methods, we have two ways to analyze the data collected. In response to RQ 1, we used LDA (Latent Dirichlet Allocation) modeling. LDA is a probabilistic model that uses the Bayesian model to infer topics with their underlying probabilities and provides a representation of the document (Blei et al., 2003). We mixed all the comments and inputted them into the LDA model to discover the top 10 most probabilistic topics across the threads. To answer RQ2, we have done a two-step analysis. The first step is to implement TF-IDF modeling. TF-IDF measures the term-frequency that is the times that a term occurs in the given document and the inversed document frequency which indicates how common and rare that term is across all documents (Luhn, 1957; Jones, 1972). TF-IDF ensures that common words such as "this" are filtered out when we are looking at key information of a document as their IDF value is 0 which makes their TF-IDF value is 0. A high TF-IDF value indicates that the term is important to the given document and possibly represents key information of that document. We did TF-IDF modeling for each thread in a tri-gram way (three words as a term) to inspect if the key information extracted by TF-IDF modeling aligned with the topic. The second step is to investigate the relevance between the keywords of a thread and the title of that thread. A uni-gram (one word as a term) TF-IDF model was utilized to produce each thread's keywords. We then count the keywords that appeared in the title of a post and implement a regression model to examine the relationship between the number of keywords (explanatory variable) and the counts of its occurrences in the title (response variable).

3. Results

3.1. LDA Modeling

The below figure (see Figure 1) demonstrates the results of our LDA modeling. Each line indicates a probable topic that is consisted of some words and their probabilities. For example, the first line is the topic that contains the words "people," "money," "government," "better," "language," "federal," "asset," and "month" and their probabilities, 0.019, 0.016, 0.011, 0.011, 0.011, 0.008, 0.008, and 0.008. As all the words are tokenized in the LDA topic model, the results of LDA topic modeling yield to the researcher's interpretations. In other words, LDA, as a probabilistic model, will not provide a certain conclusion of what the exact topics are in the document but offer terms with probabilities for inference.

```
(0, '0.019*"people" + 0.016*"money" + 0.011*"government" + 0.011*"better" + 0.011*"language" + 0.008*"federal" + 0.008*"asset" + 0.008*"month"')
(1, '0.012*"virus" + 0.012*"people" + 0.009*"think" + 0.009*"marijuana" + 0.009*"neuron" + 0.006*"right" + 0.006*"better" + 0.006*"decision"')
(2, '0.045*"remove" + 0.024*"percent" + 0.010*"people" + 0.010*"adult" + 0.010*"medical" + 0.008*"problem" + 0.008*"american" + 0.008*"deductible"')
(3, '0.021*"would" + 0.013*"people" + 0.013*"economy" + 0.013*"cheap" + 0.009*"government" + 0.009*"money" + 0.009*"could" + 0.009*"billion"')
(4, '0.027*"would" + 0.011*"something" + 0.008*"placebo" + 0.006*"testing" + 0.006*"antibody" + 0.006*"people" + 0.006*"pretty" + 0.006*"assume"')
(5, '0.014*"people" + 0.013*"would" + 0.011*"state" + 0.010*"immune" + 0.010*"pathogen" + 0.008*"large" + 0.008*"school" + 0.008*"teacher"')
(6, '0.021*"neuron" + 0.017*"people" + 0.017*"leadership" + 0.012*"brain" + 0.010*"elephant" + 0.007*"involve" + 0.007*"years" + 0.007*"intelligence"')
(7, '0.016*"pressure" + 0.010*"would" + 0.010*"think" + 0.007*"people" + 0.007*"brain" + 0.007*"something" + 0.007*"human" + 0.007*"create"')
(8, '0.019*"study" + 0.010*"save" + 0.010*"patient" + 0.010*"someone" + 0.010*"better" + 0.007*"private" + 0.007*"emergency" + 0.007*"design"')
(9, '0.024*"people" + 0.009*"still" + 0.009*"vaccine" + 0.009*"community" + 0.009*"delete" + 0.009*"physical" + 0.007*"voice" + 0.007*"country"')
```

Figure 1. LDA Topic Modeling Results

3.2. Thread TF-IDF

The below figures (see Figure 2 and Figure 3) show the results of tri-gram TF-IDF modeling for each thread. From the model for each topic, we can find some key information of that thread based on the tri-grams. The TF-IDF value indicates the relevance of the term to a document. In this context, a higher TF-IDF score means that the given term is important to the comments in that thread. Therefore, the main topic of a thread can be inferred from the TF-IDF model. However, in each model, the amount of key tri-grams varies so that some topics are easier to infer based on those key terms, e.g., topic 1, whereas others like topic 7 are much more difficult to make an inference.

Topic 1	TF-IDF
juvenile incarceration place	0.198829
determine sentence natural	0.198829
state typically pay	0.198829
typically pay prison	0.198829
county led stark	0.198829
natural experiment whereby	0.198829
drop incarceration suggest	0.198829
stark drop incarceration	0.198829
experiment whereby cost	0.198829
whereby cost burden	0.198829
led stark drop	0.198829
prison county determine	0.198829
us state typically	0.198829
incarceration place county	0.198829
pay prison county	0.198829
place county led	0.198829
sentence natural experiment	0.198829
incarceration suggest mass	0.198829
county determine sentence	0.188275
cost burden juvenile	0.188275
burden juvenile incarceration	0.188275
suggest mass incarceration	0.180786
incarceration us part	0.180786
us part due	0.174978
mass incarceration us	0.174978
part due misalign	0.170232
due misalign incentive	0.162744

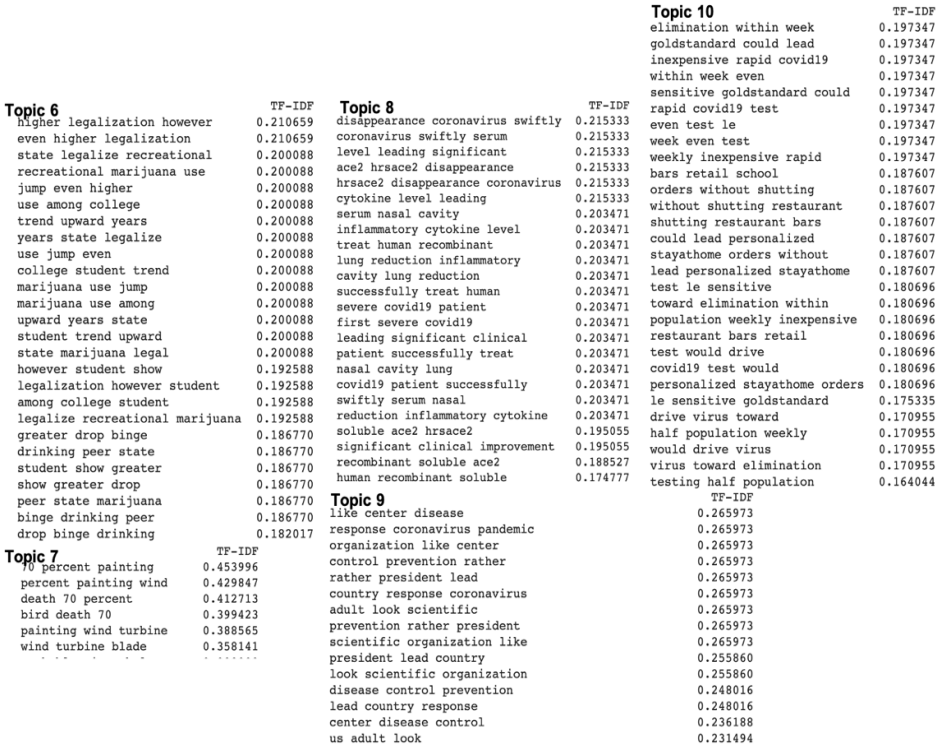
Topic 2	TF-IDF
immune response these	0.262191
these result represent	0.262191
find safe welltolerated	0.262191
result represent important	0.262191
vaccine find safe	0.262191
induce rapid immune	0.262191
safe welltolerated induce	0.262191
represent important milestone	0.262191
welltolerated induce rapid	0.262191
covid19 vaccine find	0.262191
trial covid19 vaccine	0.262191
response these result	0.262191
human trial covid19	0.252087
first human trial	0.244249
rapid immune response	0.227742

Topic 3	TF-IDF
lancet team yale	0.266959
team yale epidemiologist	0.266959
find medicare would	0.266959
yale epidemiologist find	0.266959
new study lancet	0.266959
would save 68000	0.266959
study lancet team	0.266959
epidemiologist find medicare	0.266959
annually well 450	0.254312
well 450 billion	0.254312
medicare would save	0.254312
life annually well	0.254312
68000 life annually	0.245339
save 68000 life	0.238379
450 billion cost	0.232692

Topic 4	TF-IDF
contest mind manifestation	0.242472
manifestation higher intelligence	0.242472
know know ponder	0.242472
humans higher mammal	0.242472
find crow know	0.242472
thought long believe	0.242472
research find crow	0.242472
long believe sole	0.242472
know ponder content	0.242472
analytical thought long	0.242472
mind manifestation higher	0.242472
province humans higher	0.242472
intelligence analytical thought	0.230166
believe sole province	0.230166
higher intelligence analytical	0.230166
ponder content mind	0.221435
sole province humans	0.221435
crow know know	0.193626

Topic 5	TF-IDF
cancer cell find	0.201752
find new australian	0.201752
tumour growth mouse	0.201752
cell find new	0.201752
also found venom	0.190307
component combine exist	0.190307
new australian research	0.190307
study also found	0.190307
exist chemotherapy drug	0.190307
drug extremely efficient	0.190307
reducing tumour growth	0.190307
chemotherapy drug extremely	0.190307
main component combine	0.190307
found venom main	0.190307
venom main component	0.190307
combine exist chemotherapy	0.190307
research study also	0.190307
australian research study	0.190307
efficient reducing tumour	0.190307
extremely efficient reducing	0.190307
kill aggressive hardtotreat	0.182186
found rapidly kill	0.182186
hardtotreat breast cancer	0.182186
rapidly kill aggressive	0.182186
venom honeybee found	0.182186
honeybee found rapidly	0.182186
aggressive hardtotreat breast	0.182186
breast cancer cell	0.159296

Figure 2. TF-IDF Tri-gram Keywords (topics 1-5)



Topic 6	TF-IDF
higher legalization however	0.210659
even higher legalization	0.210659
state legalize recreational	0.200088
recreational marijuana use	0.200088
jump even higher	0.200088
use among college	0.200088
trend upward years	0.200088
years state legalize	0.200088
use jump even	0.200088
college student trend	0.200088
marijuana use jump	0.200088
marijuana use among	0.200088
upward years state	0.200088
student trend upward	0.200088
state marijuana legal	0.200088
however student show	0.192588
legalization however student	0.192588
among college student	0.192588
legalize recreational marijuana	0.192588
greater drop binge	0.186770
drinking peer state	0.186770
student show greater	0.186770
show greater drop	0.186770
peer state marijuana	0.186770
binge drinking peer	0.186770
drop binge drinking	0.182017

Topic 7	TF-IDF
70 percent painting	0.453996
percent painting wind	0.429847
death 70 percent	0.412713
bird death 70	0.399423
painting wind turbine	0.388565
wind turbine blade	0.358141

Topic 8	TF-IDF
disappearance coronavirus swiftly	0.215333
coronavirus swiftly serum	0.215333
level leading significant	0.215333
ace2 hrsace2 disappearance	0.215333
hrsace2 disappearance coronavirus	0.215333
cytokine level leading	0.215333
serum nasal cavity	0.203471
inflammatory cytokine level	0.203471
treat human recombinant	0.203471
lung reduction inflammatory	0.203471
cavity lung reduction	0.203471
successfully treat human	0.203471
severe covid19 patient	0.203471
first severe covid19	0.203471
leading significant clinical	0.203471
patient successfully treat	0.203471
nasal cavity lung	0.203471
covid19 patient successfully	0.203471
swiftly serum nasal	0.203471
reduction inflammatory cytokine	0.203471
soluble ace2 hrsace2	0.195055
significant clinical improvement	0.195055
recombinant soluble ace2	0.188527
human recombinant soluble	0.174777

Topic 9	TF-IDF
like center disease	0.265973
response coronavirus pandemic	0.265973
organization like center	0.265973
control prevention rather	0.265973
rather president lead	0.265973
country response coronavirus	0.265973
adult look scientific	0.265973
prevention rather president	0.265973
scientific organization like	0.265973
president lead country	0.255860
look scientific organization	0.255860
disease control prevention	0.248016
lead country response	0.248016
center disease control	0.236188
us adult look	0.231494

Topic 10	TF-IDF
elimination within week	0.197347
goldstandard could lead	0.197347
inexpensive rapid covid19	0.197347
within week even	0.197347
sensitive goldstandard could	0.197347
rapid covid19 test	0.197347
even test le	0.197347
week even test	0.197347
weekly inexpensive rapid	0.197347
bars retail school	0.187607
orders without shutting	0.187607
without shutting restaurant	0.187607
shutting restaurant bars	0.187607
could lead personalized	0.187607
stayathome orders without	0.187607
lead personalized stayathome	0.187607
test le sensitive	0.180696
toward elimination within	0.180696
population weekly inexpensive	0.180696
restaurant bars retail	0.180696
test would drive	0.180696
covid19 test would	0.180696
personalized stayathome orders	0.180696
le sensitive goldstandard	0.175335
drive virus toward	0.170955
half population weekly	0.170955
would drive virus	0.170955
virus toward elimination	0.170955
testing half population	0.164044

Figure 3. TF-IDF Tri-gram Keywords (topics 6 –10)

3.3. Keywords and title relevance

The below table (see Table 1) shows the keywords extracted from the uni-gram TF-IDF modeling of each topic. The column *TF-IDF (>0)* indicates the number of keywords that have a TF-IDF value greater than 0 in a given thread. *InTitleCount* column refers to the counts of those keywords that are also in the submission (the first comment of the thread). *Percentage* is the ratio of those keywords in the title over the total keywords.

Our regression model tests whether the number of keywords from TF-IDF modeling in a thread predicts its occurrences in the title. The results of the regression indicates that the predictor, the number of keywords, explains 62.2% of the variance ($R^2 = .622$, $F(1,8) = 13.14$, $p < 0.01$). The result reveals that the number of keywords in a given thread statistically significantly predict the occurrences of the keywords in the title of that thread.

Table 1. Keywords and Title Analysis Results

Topic	TF-IDF (>0)	InTitleCount	NotInTitleCount	Percentage
1	25	8	17	32%
2	17	9	8	53%
3	17	8	9	47%
4	18	6	12	33%
5	28	11	17	39%
6	23	12	11	52%
7	8	3	5	38%
8	26	8	18	31%
9	17	9	8	53%
10	30	13	17	43%

4. Discussion

There are two findings from the LDA modeling analysis. The first finding is that although the overarching topics of the selected threads are different, there are aggregated concerns across those threads. In our LDA modeling results, terms that are related to people, government, and public health indicate that across these 10 threads, Redditors concern about the impact brought by the COVID-19 pandemic. This can also be inferred from the probabilistic topics 2,3,4,5,8 and 9 (see Figure 1). Redditors mention the terms of virus, vaccine, testing, placebo, economy, etc., which is reasonable as the pandemic affected everyone's life in the year 2020. The second finding is LDA generated topics may not cover all the threads. For instance, topic 7 has a theme of the bird's death and contains 1,503 rows, which takes up 7.4% of the total comments. However, the top 10 LDA topics have no evidence of this topic. Therefore, although probability-based LDA topics have the limitation of failing to represent all the concerns, it can be inferred that people's overarching concern in the subreddit channel r/science is the COVID-19 pandemic and the impacts on the public health system. Thus, in a subreddit channel that distributes the latest news, Redditors plausibly concern about what is trending in the real world.

To better understand every thread, it is necessary to inspect each of them. From Figure 2 and Figure 3, key phrases produced by tri-gram TF-IDF models in each topic make it feasible to infer what the main topic of a given thread is. For example, from the TF-IDF key terms of topic 2 in Figure 2, it can be deduced that this thread is about the COVID-19 vaccine and its trial on humans because terms like safety, immune response are key to people's discussion

in this thread based on the TF-IDF values. Therefore, to answer RQ2, Redditors' discussion in the subreddit r/science is relatively concentrated, which can also be validated in our third analysis.

The third analysis is consisted of uni-gram TF-IDF key terms generating and a follow-up linear regression modeling process. As demonstrated in the previous section, the counts of TF-IDF of each topic can predict the number of matched terms in the title of the tread in a statistically significant way. This result indicates that Redditors' comments, which are the data source of those TF-IDF terms, are relevant to the discussion topic as they frequently refer to the terms in the title of each post. This might be because the subreddit channel selected is somewhat serious so that there are fewer distracting comments. In some other subreddits such as r/todayilearned, although people are learning fun facts, they tend to respond to others in a more entertaining way and engage in discussion with more memes and jokes.

Our study has two limitations. The first is that the discussion style and language features may significantly vary across subreddits. There might even not be any subreddit-wise concerns, and an extreme example can even be that a subreddit channel might be created for amusement, and people never concentrate on any specific topics there. Therefore, whether we should care about what people are concerning depends on the purposes that people create and join a subreddit. Another is about the sample size of our data. The PRAW only allows us to send 1 request per second, which prevents us from retrieving a large volume of posts. Therefore, the generalizability of our results needs further study to test.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Fang, H., Cheng, H., & Ostendorf, M. (2016, November). Learning latent local conversation modes for predicting comment endorsement in online discussions. In Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media (pp. 55-64).

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. IBM Journal of research and development, 1(4), 309-317.

Medvedev, A. N., Lambiotte, R., & Delvenne, J. C. (2017, June). The anatomy of Reddit: An overview of academic research. In Dynamics on and of Complex Networks (pp. 183-204). Springer, Cham.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 28(1), 11-21. https://dx.doi.org/10.1108/eb026526

Park, A., & Conway, M. (2017). Tracking health related discussions on Reddit for public health applications. In AMIA Annual Symposium Proceedings (Vol. 2017, p. 1362). American Medical Informatics Association.

Appendix

2020 top 10 threads in r/science:

1.(social science)
https://www.reddit.com/r/science/comments/k1ofcu/in_the_us_states_typically_pay_for_prison_while/

2.(medicine)
https://www.reddit.com/r/science/comments/gp2hdt/the_first_human_trial_of_a_covid19_vaccine_finds/

3.(health)
https://www.reddit.com/r/science/comments/f4998k/a_new_study_in_the_lancet_by_a_team_of_yale/

4.(psychology)
https://www.reddit.com/r/science/comments/izbj3r/research_finds_that_crows_know_what_they_know_and/

5.(cancer)
https://www.reddit.com/r/science/comments/ikivxq/venom_from_honeybees_has_been_found_to_rapidly/

6.(health)
https://www.reddit.com/r/science/comments/eoomwz/marijuana_use_among_college_students_has_been/

7.(environment)
https://www.reddit.com/r/science/comments/igmtvw/bird_deaths_down_70_percent_after_painting_wind/

8.(medicine)
https://www.reddit.com/r/science/comments/jp3w7w/the_first_severe_covid19_patient_successfully/

9. (social science)
https://www.reddit.com/r/science/comments/gl1nvf/us_adults_look_to_scientific_organizations_like/

10. (epidemiology)
https://www.reddit.com/r/science/comments/jy8knh/testing_half_the_population_weekly_with/

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15023

Exploring Redditors' Communication Style

Yilang Zhao
Department of Curriculum and Instruction, University of Wisconsin–Madison, USA.

Abstract

This paper explores the communication style of parent and child posts in a popular Reddit channel, r/todayilearn. By implementing epistemic network analysis (ENA), this paper discovers that parent posts on this subreddit channel tend to have more opinions, yet child posts are more likely to present evidence or external supports to back up their statements. Understanding how people communicate in online communities may help establish a better community atmosphere for more productive discussion.

Keywords: *Reddit, online community, communication style, epistemic network analysis.*

1. Introduction

Reddit, as a popular online discussion community, has a massive amount of discussion participants (over 2 billion comments). It can be interesting to learn how people communicate with each other on Reddit, as learning scientists find that people in informal learning settings may have more effective learning (Miyake & Kirschner, 2014). Reddit has the potential to be a platform where people gain informal learning experiences from communicating with others on various topics. Thus, it is necessary to explore Redditors' communication styles to gain insights about how to establish an online environment that promotes active discussion and possibly informal learning?

2. Theoretical Background

While browsing Reddit threads, people may notice that Redditors' posts have different patterns in terms of their communication styles. Some of them prefer to refer to factual knowledge or post external links, while others are more in favor of expressing their own opinions or explaining something in their own words. What are the differences between facts and opinions?

Corvino (2014) states that the fact/opinion distinction is philosophically ambiguous, and there are three general distinctions that confuse people. First, people think facts are reality and opinions are beliefs that represent reality, but both facts and opinions can represent reality either successfully or unsuccessfully. For example, the sentence that "there is beer in the refrigerator" can either be a statement of fact or an opinion claimed by someone. Second, many people believe that facts are objective, and opinions are subjective. The problem of this distinction resides on the definition of the subjective statement. Subjectivity is mind dependent. For instance, the statement that "God created the earth" can be an object matter to God believers but seems to be a subjective assertation made by God believers to atheists. Finally, people hold the idea that facts are descriptive, but opinions are normative. Nonetheless, not all the opinions are normative. The statement that "a Democrat will win the presidency in 2016" is an opinion but it is not normative. Therefore, Corvino (2014) propose the following definitions to differentiate facts and opinions: A statement of fact is one that has objective content and is well-supported by the available evidence, and a statement of opinion is one whose content is either subjective or else not well supported by the available evidence.

Besides the fact/opinion distinction, it is also essential to know the structure of a Reddit thread. On Reddit, a piece of information can spread to a large number of people in a very short time through a person-to-person process, which is defined as a viral structure (Goel, Anderson, Hofman & Watts, 2015). However, the viral structure is extremely complicated and hard to analyze. This study only concentrates on the separate parent and child posts.

Another reason for choosing separate parent and child posts for analysis is that in Miyake and Kirschner's (2014) collaborative learning model, they claim that interdependency and task cohesion are two elements that contribute to constructive collaborative learning. They state that achieving a learning goal requires the inter-connections among the completion of sub-tasks and those independent sub-tasks have shared commitments and collaborative efforts of a group. On Reddit, the sub-tasks can be the parent and child posts which share a goal of adding new information or discussing the existing information. Thus, it is intriguing to explore the connections between parent and child posts.

Therefore, the research question of this study is: what are the communication styles of the parent and child posts on Reddit?

3. Methods

3.1. Data Collecting and Processing

The dataset was collected from the subreddit channel, r/todayilearn. The original data was the top 50 hottest (most commented) threads on March 6th, 2019, including all the posts of those threads. The total number of posts was 14,050. Half of them were child posts and the other half were their direct parent posts. Since a child post could be a parent post of other posts, there was a considerable number of overlapped lines. Since the analytical tool has a data amount limit, with all the duplicates are removed, 1706 unique parent posts and 5217 unique child posts were kept for later analysis.

3.2. Coding

Six codes are included (see Table 1) in this study, which are Link, Questioning, Explanation, Opinion, Science, and Politics. Codes are created by examining the top 3 hottest threads which have more comments than any other rest threads and then conceptually similar codes are aggregated. Since the dataset in this study has a large volume of data that is impossible for human coders to code, an automatic coding tool, nCoder (www.n-coder.org), is used to code the dataset and all the codes are statistically valid (kappa > 0.65, rho < 0.05).

Table 1. Coding Scheme

Name	Definition	Example	Kappa	Rho
Link	External references that people will request or post to share information.	Well... I stand kinda corrected. Someone conceived of the idea first. However [Tesla was the innovator](https://www.pbs.org/tesla/ins/lab_imlight.html) and deserves most of the credit and should have applied for a patent.	1.00	0.01
Questioning	Confused by the parent post and want additional information.	Still not sure if advertisement for Amtrak	1.00	0.04
Explanation	Suggesting reasons, giving details, and clarifying vagueness.	I remember a similar story, where a university student came to math class late and their professor had an unsolvable or unsolved equation/proof/theorem on the board.	0.89	0.02
Opinion	Expressing personal ideas, feelings, judgements; no matter they are new or not.	I don't get it. Why would anyone think frosted glass would be impossible?	1.00	0.03
Science	Using general scientific knowledge/facts/stories to respond to a parent post.	I mean, radiation does have its place in the medical field...we use it to treat cancers. We use it for x Ray's. We use it to trace through your blood stream for tests....toothpaste and radioactive spam though?	1.00	0.02
Politics	Responses to the parent post that contain political terms/issues.	The government has been doing this for a while but to search for a bomb instead of using it as one	1.00	0.03

3.3. Data Analysis

In this study, Epistemic Network Analysis (ENA) is applied as the analytical tool (www.epistemicnetwork.org). ENA is the tool that measures relationships among codes by quantifying the co-occurrence of codes in discourse (Shaffer, 2017). For this analysis, every single parent and child post is segmented as a conversation. The minimal unit of analysis is a single post. The ENA algorithm in this analysis uses the whole conversation as a stanza.

In the ENA model of this study, networks are visualized using network graphs where nodes corresponded to the codes, and edges reflect the relative frequency of co-occurrence, or connection, between two codes. The results are two coordinated representations for each unit of analysis: (1) a plotted point, which represents the location of that unit's network in the low-dimensional projected space, and (2) a weighted network graph. The positions of the network graph nodes are fixed, and those positions are determined by an optimization routine that minimizes the difference between the plotted points and their corresponding network centroids. Because of this co-registration of network graphs and projected space, the positions of the network graph nodes—and the connections they define—can be used to interpret the dimensions of the projected space and explain the positions of plotted points in the space. Additionally, the weights of those edges represent the degree of connection between two nodes.

4. Results

4.1. Qualitative Analysis

The below is a qualitative excerpt of a parent post and a child post.

Parent post:

There was a Carl Weathers for Governor skit in SNL once, but the only YouTube link I found had been taken down. I think his slogan was: "Carl Weathers, because I was also in Predator."

Child post:

so the ultraviolet light from your headlights is super weird to them. I would not expect that the typical incandescent car headlight would emit much UV light -- the filaments just don't get hot enough. [LEDs also do not typically emit much UV](https://sciencing.com/light-bulbs-not-emit-uv-radiation-15925.html). [HID lights do emit significant amounts of UV](https://en.wikipedia.org/wiki/High-intensity_discharge_lamp), however we then add filters to absorb this, as this UV would be a danger to us. That said, a bright light at night that suddenly appears would indeed be "super weird" to them -- at least until they've experienced a few cars. No need to blame this on UV.

From this qualitative example, the parent post attempts to express an opinion, which is subjective and not well-supported by evidence, though this person tries to use a YouTube video to support his/her claim. However, the selected child post is not only expressing and explaining personal ideas but also providing external links and using scientific terms to justify his statement, which includes both opinions and facts. Thus, the communication style of parent posts is more on the opinion side, while the communication style of child posts is more comprehensive, which usually consists of both facts and opinions.

4.2. Quantitative Analysis

From Figure 1, the positions of the codes along the X-axis are depending on their closeness to a fact or an opinion. *Questioning* uses existing posts as evidence and *Science* has scientific terms, which make them more like facts. *Explanation* and *Opinion* are more subjective and possibly lack of supportive evidence so that they are on the opinion side according to Corvino's (2014) definitions. *Link* and *Politics* can be used for either presenting facts or expressing opinions. That is why they are close to the Y-axis. In addition, the child post (green) has strong connections between the fact codes, and the opinion code. *Explanation* is also tightly linked with the left fact codes. However, the parent post has a different pattern. It is completely based on the opinion codes, *Explanation* and *Opinion*, and there is no connections with the fact codes, *Questioning* and *Science*. A two-sample t-test with the ENA web tool was conducted to examine a generalized result. Along the X axis, the parent posts (mean=0.01, SD=0.22) was statistically significantly different at the alpha= 0.05 level from the child posts (mean=0.00, SD=0.18) with a p-value equal to 0.00 and Cohen's d equal to 0.09.

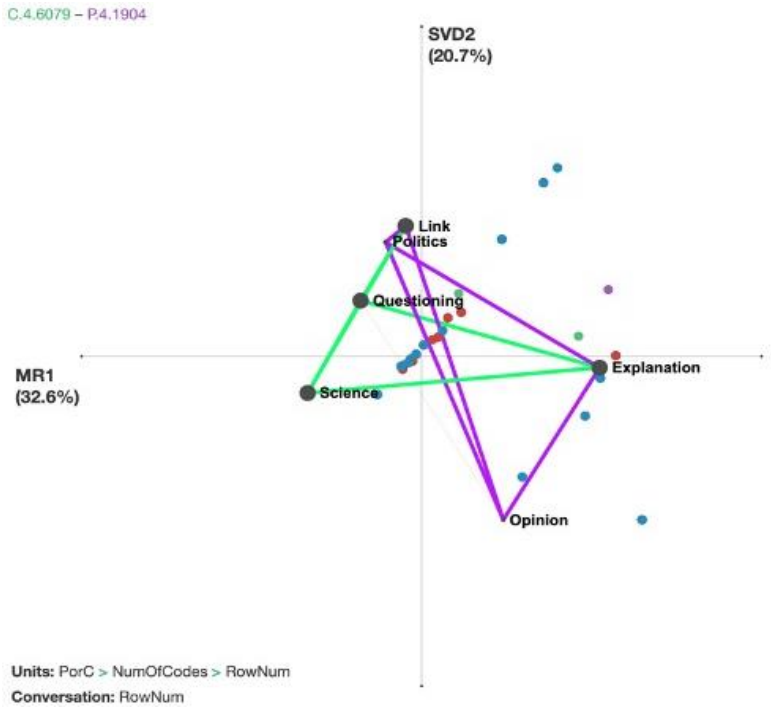


Figure 1. ENA Comparison Plot

5. Discussion

5.1. Finding

The above qualitative and quantitative results reveal that parent posts and child posts on Reddit have different communication styles. The parent posts are often simpler and express an opinion. However, in response to those opinion-based parent posts, the child posts are usually more comprehensive, including both facts and opinions. People tend to have discussions with others by using child posts that are well-supported by sound evidence and contain the replier's own ideas.

The theoretical contribution of this study is to extend understanding of the pattern with which people communicate in discussion communities to promote further collaborative online learning. As for the practical contribution, online discussion participants may understand how to post a more constructive response that provides either facts or opinions with a goal of contributing to a better informal learning environment.

5.2. Limitation

One of the major problems of this study is the Redditors' use of language. The discussion atmosphere of some Reddit channels is extensively informal so that Redditors tend to use casual languages and include many memes which cannot be accurately identified by nCoder. Additionally, the topics on Reddit vary a lot every day. It is impossible to create a coding scheme that covers all the topics. The codes selected in this study may only be effective for analyzing the data collected in the certain subreddit on that day.

References

Corvino, J. (2014). The fact/opinion distinction. The Philosophers' Magazine, (65), 57-61.

Goel, S., Anderson, A., Hofman, J., & Watts, D. (2015). The Structural Virality of Online Diffusion. Management Science, 150722112809007.

Miyake, N., & Kirschner, P. (2014). The Social and Interactive Dimensions of Collaborative Learning. In R. Sawyer (Ed.), The Cambridge Handbook of the Learning Sciences (Cambridge Handbooks in Psychology, pp. 418-438). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139519526.026

Shaffer, D. W. (2017). Quantitative ethnography. Madison, WI: Cathcart Press.

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15030

Google Trends Search Information Related to Breastfeeding in the U.S.

Richard A. Fabes, Denise Ann Bodman, Bethany Bustamante Van Vleet, Carol L Martin
Arizona State University, USA.

Abstract

Given the importance of breastfeeding to maternal and infant health, we employed Google Trends to examine search engine use for information related to breastfeeding in the U.S. We conducted an analysis of the use of the Google search engine related to the broad topic of 'breastfeeding," as well as patterns for more specific terms related to breastfeeding. Given the significant role that breastfeeding pain plays in influencing breastfeeding persistence, we examined patterns in the use of Google to seek information related to breastfeeding pain and how that compares to other breastfeeding topics. We also examined diurnal patterns in these searches as well as U.S. state-level characteristics that predict search intensity. We found that search intensity related to breastfeeding has increased over time and that searches related to breastfeeding pain were the most common. Searches tended to occur late at night and were more likely to occur in relatively unpopulated states and for states with lower income. The findings illustrate how Google Trends can be analyzed to highlight the concerns of new mothers in real-time and how such data can reveal how mothers and those who support them use the internet to seek out help, guidance, and support for issues related to breastfeeding.

Keywords: *Breastfeeding, Google Trends, pain, time of day*

1. Introduction

The health benefits of breastfeeding for mothers and infants are well established (U.S. Department of Health and Human Services, 2011). These benefits apply to mothers and children in developed nations such as the United States as well as to those in developing countries. Breastfeeding provides numerous emotional and physical benefits as human milk is uniquely suited to the human infant's nutritional needs and has unparalleled immunological and anti-inflammatory properties that protect against a host of illnesses and diseases for both mothers and children. Thus, both mothers and their infants benefit from breastfeeding. Despite this knowledge, only slightly more than half (58%) of mothers in the U.S. are likely to be breastfeeding their infants at 6 months of age and only 25% are doing so exclusively (CDC, 2020).

One of the principal reasons for breastfeeding cessation is the experience of pain during breastfeeding (McClellan et al., 2012). In addition to the discomfort, breastfeeding pain can also cause psychological distress and interfere with general activity, mood, sleep, and bonding between mother and infant (Amir et al., 1996). However, the most effective means of helping mothers establish comfortable and painless breastfeeding to promote continued breastfeeding as long as they wish has yet to be established (Kent et al., 2015).

Given the importance of breastfeeding to maternal and infant health, we used Google Trends to examine mothers' use of the Google search engine to acquire information related to breastfeeding (we assume the users are mostly mothers but acknowledge that there are others who seek such information). We began with a general analysis of the use of the Google search engine related to the broad topic of 'breastfeeding." Once we established the pattern of use, we then examined more specific terms related to breastfeeding to determine what specifically mothers were searching for. Given the significant role that breastfeeding pain plays in influencing breastfeeding duration and persistence, we examined patterns in mothers' use of Google to seek information related to breastfeeding pain and how that compares to other breastfeeding search terms. We also examined diurnal patterns in searches related to breastfeeding pain to determine if there were specific times of the day that mothers were more likely to seek such information and how these time periods may relate to issues associated with breastfeeding pain. Finally, we examined U.S. state-level predictors of Google searches related to breastfeeding pain to ascertain what demographic factors predict the relative likelihood of mothers relying on the internet for such information. The findings of this research have the potential to elucidate how we might help mothers in their search for answers to questions related to breastfeeding, and breastfeeding pain, to help mothers and their infants sustain breastfeeding as long as possible. The findings highlight how Big Data such as Google Trends can provide insight into important health and developmental processes that have significant short- and long-term outcomes for mothers and their infants, as for our society as well.

2. Methods and Results

Data on search engine use were obtained by using Google Trends (https://trends.google.com/trends/). Google Trends data are an unbiased sample of Google searches, and it has become the most popular tool for examining online behavior and interest (Mavragani & Ochoa, 2019). The data are anonymized, categorized, and aggregated. This allows for the assessment of interest in a particular topic across searches for a given time period and/or for a given region of the world. It offers a reflection of the needs, wants, and interests of its users.

Google Trends produces a real-time index of the volume of Google searches by category and geography. Google Trends does not report the absolute number of queries for a search term(s) but instead reports a *search intensity index* that reflects the fraction of a given area's Google searches devoted to that term or topic. The index reflects the total search volume for a term in a given geographic region divided by the total number of searches in that region at a particular time. The resulting numbers are then normalized between 0 and 100 and are available worldwide and for individual countries.

2.1. Breastfeeding Google Search Intensity

Our first step in assessing interest in issues related to breastfeeding was to examine the pattern of use of the search term "breastfeeding" in the U.S. and neighboring countries across the period of time available in Google Trends (2004 to the present).

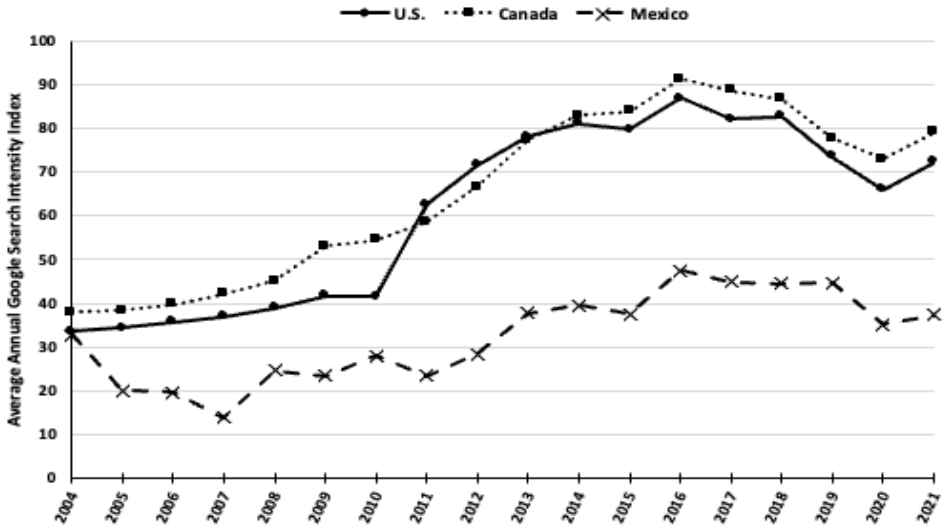


Figure 1. Google Search Intensities for "Breastfeeding" from 2004 to 2021.

The results of this search are presented in Figure 1 and reveal a positive trend in search intensity in all three countries related to breastfeeding. The figure also reveals an uptick in intensity in the U.S. beginning in 2011. Google notes an improvement in geographical assignment in 2010 that may partially account for this increase. However, this spike in 2011 did not take place in the data for either Canada or Mexico. It is interesting to note that in 2011 the U.S. Surgeon General issued a "Call to Action to Support Breastfeeding" (U.S. Department of Health and Human Services, 2011). This increase in the U.S. may reflect the impact of this society-wide emphasis on the need to support mother and their babies who are breastfeeding.

2.2. Breastfeeding Google Search Intensity Patterns Over Time

To examine the specific types of searches that may account for the increases in interest in the U.S. reflected in Figure 1, we conducted a Google Trends search using popular search topics related to breastfeeding. Specifically, we separately entered the search teams "breastfeeding pain," "benefits of breastfeeding," "breastfeeding latch," "breastfeeding help," and "breastfeeding tips." The results of these searches are presented in Figure 2, along with the slope for each of the terms. Figure 2 reveals some clear patterns. First, although each of the lines show an increase in 2011, the increases in search intensity vary in strength.

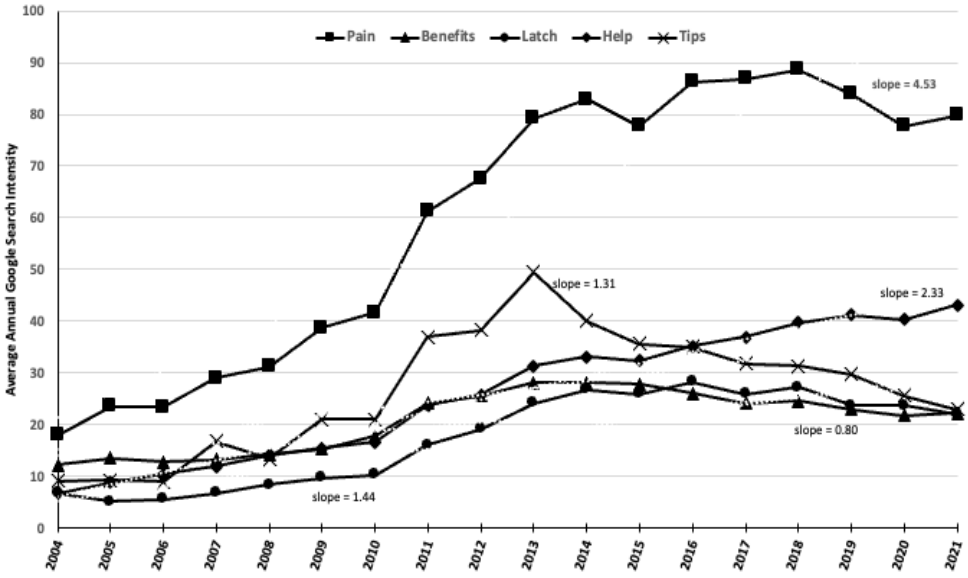


Figure 2. Google Search Intensities for popular breastfeeding terms in U.S. from 2004 to 2021.

Second, search intensity related to breastfeeding pain showed a pattern that was similar to the pattern seen for the U.S. data in Figure 1. None of the other terms showed this similarity with the general pattern found for U.S. data in Figure 1. The largest slopes were found for

"breastfeeding pain" and "breastfeeding help" – both reflecting searches for information that likely reflect mothers' concerns related to breastfeeding and an effort to find information that may help them overcome problems or issues they may be having.

2.3. Diurnal Variation in Google Search Intensity for "Breastfeeding Pain" Searches

The use of the internet as a source of information about breastfeeding pain opens up options for mothers who can access this information any time of day and from almost any location. Such a conclusion is consistent with research that has shown that new mothers often use the internet to seek information after the birth of their infants with the most common online topics searched including information about establishing breastfeeding and dealing with lactation issues (e.g., Alianmoghaddam et al., 2019). But when do mothers search for this information and are there consistent daily patterns in U.S. mothers' searches for information related to breastfeeding pain? What can daily search patterns reveal about mothers' efforts to obtain information related to breastfeeding pain?

We addressed these questions by entering "breastfeeding pain" into a Google Trends search for the U.S. over a week of data Feb 28-March 7, 2022). This search option in Google Trends provides users with hourly data across the week, up to the date of entry. The results are presented in Figure 3 and revealed a consistent pattern in Google searches. Over this week, mothers were most likely to be searching for information related to breastfeeding pain at night, typically peaking each day around midnight with lows during the daytime (results from other weeks show a similar pattern).

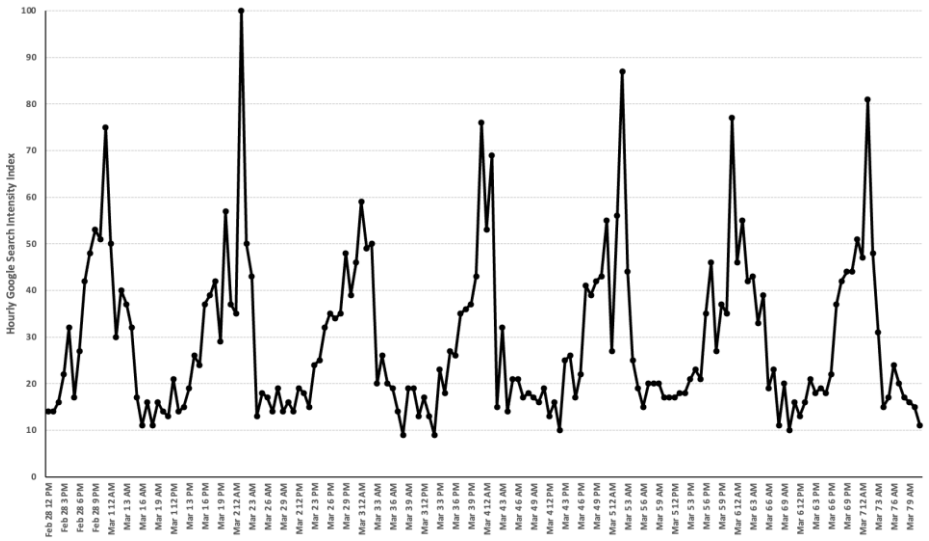


Figure 3. Google Search Intensity for "Breastfeeding Pain" by Time of Day (Feb 28 – March 7, 2022)

It has been well documented that women experience significant sleep changes after the birth of a baby (Richter et al., 2019). Mothers have been found to sleep less at night and more during the day (Gay et al., 2004). Given this, it is not surprising that mothers were searching for information on breastfeeding pain late at night while they were up and less during the day when they may have other sources to turn to, or may be sleeping, dealing with family, working and have less free time to spend on the internet. Up at night, frequently by themselves with their infants, these data support the conclusion that mothers appear to use this time to seek help related to breastfeeding pain via the internet and/or in seeking or providing virtual social support.

2.4. U.S. State-level Predictors of Google Search Intensity for "Breastfeeding Pain"

To better understand the predictors of mothers' use of Google to obtain information related to breastfeeding pain, we conducted a multiple regression in which we predicted the Google search intensity index for "breastfeeding pain" for 38 of the U.S. states that had data to calculate a search intensity index for the past 4 years (before and after the onset of COVID-19; March 2018 to March 2022). Using 2020 U.S. Census data, we entered state-level variables of total population, per capita income, percent of population that was white, and had a bachelor's degree. We also entered data on the proportion of mothers who were breastfeeding at 6 months (CDC, 2020). The results revealed that the population of a state and per capita income were inversely related to Google search intensity for "breastfeeding pain" whereas percent of mothers who were breastfeeding at 6 months was positively related (*Bs* = -.34, -.47, and .74, *ps* < .017, .0.005, and .001, respectively; R^2 = .59). Google search intensity related to "breastfeeding pain" was relatively greater in states that were less populated, had lower per capita incomes, and had more mothers who were breastfeeding at 6 months. Although we do not have data that explicate these patterns, they highlight regional differences that affect the use of Google for information about pain during breastfeeding. Such findings may suggest that when there are limitations in access to health care providers in states, either due to a more rural region or due to income limitations, mothers are relatively likely to turn to the internet for information related to breastfeeding pain.

3. Considerations and Implications

Despite much preparatory effort, numerous studies find that new mothers report feeling overwhelmed and unprepared for breastfeeding (Lansinoh, 2012). The U.S. Surgeon General's *Call to Action to Support Breastfeeding* (2011) listed several barriers to breastfeeding in the U.S., including lack of knowledge, social norms, poor family support, embarrassment, and problems with access to health services. Online breastfeeding information and support may help women meet their breastfeeding needs, particularly when they have limited access to health care providers or peer support systems.

As with all studies, there are important limitations that need to be considered. First, Google Trends data are aggregated, and we cannot disentangle the qualities of various users and any variation in search engine use for individual mothers. Second, the focus of the present research was on data from the U.S. It will be important for future research to explore these trends across different countries to compare how cultural and technological differences impact the use of the internet for information related to breastfeeding. Third, caution must be used when examining long-term trends in Google search intensity as the population of users changes over time. Moreover, Google changed its methodology over this period of time. Fourth, Google search intensity depends on the number of queries in a location. As a result, trends over time can sometimes be misleading as an increasing pattern can be found while the total number of searches declines (or vice versa). Finally, it will be important for future research to focus on how search engine information is used and how this use is related to important outcomes related to breastfeeding. For example, using a convenience sample, one study found that when online breastfeeding resources were deemed to be helpful to first-time mothers, they were more likely to continue to breastfeed and reduce their use of formula with their infants at 6 months of age (Newby et al., 2015). How patterns of use of Google search engine related to breastfeeding outcomes is an important step in understanding the impact the internet has on mothers' breastfeeding decisions.

Despite these limitations, the findings of the present study were the first to use Google Trends to highlight how such data can be used to better understand issues related to breastfeeding and how mothers use the internet to seek out help, guidance, and support. Although health care professionals do an excellent job of trying to help new mothers with issues related to breastfeeding and lactation, it is not possible for them to provide such support and guidance 24/7. As the diurnal data show, online searches for information related to "breastfeeding pain" took place late at night when women are likely alone with their infants. It is during these times of stress, emotional strain, and solitude that many women decide to give up on breastfeeding (Bennett, 2018). The U.S. state-level data presented in this paper suggest that this may be particularly true for mothers who live in relatively unpopulated states and in states with less overall per capita income. Thus, the internet may be an even more important source of information for these women and their infants. The research also confirms that a careful analysis of Google Trends may help health professionals develop timely interventions that help new mothers find appropriate and accurate information and support online.

References

Alianmoghaddam, N., Phibbs, S., & Benn, C. (2019). "I did a lot of Googling": A qualitative study of exclusive breastfeeding support through social media. *Women and Birth*, *32*(2), 147–156. https://doi.org/10.1016/j.wombi.2018.05.008

Amir, L. H., Dennerstein, L., Garland, S. M., Fisher, J., & Farish, S. J. (1996). Psychological aspects of nipple pain in lactating women. *Journal of Psychosomatic Obstetrics and Gynecology*, *17*(1), 53–58. https://doi.org/10.3109/01674829609025664

Bennett, V. (2018). Could artificial intelligence assist mothers with breastfeeding? *British Journal of Midwifery*, *26*, 212–213. https://doi.org/10.12968/bjom.2018.26.4.212

CDC. (2020). *Breastfeeding report card United States 2020*. CDC. https://www.cdc.gov/breastfeeding/pdf/2020-Breastfeeding-Report-Card-H.pdf

Gay, C. L., Lee, K. A., & Lee, S.-Y. (2004). Sleep patterns and fatigue in new mothers and fathers. *Biological Research for Nursing*, *5*(4), 311–318. https://doi.org/10.1177/1099800403262142

Kent, J. C., Ashton, E., Hardwick, C. M., Rowan, M. K., Chia, E. S., Fairclough, K. A., Menon, L. L., Scott, C., Mather-McCaw, G., Navarro, K., & Geddes, D. T. (2015). Nipple pain in breastfeeding mothers: Incidence, causes and treatments. *International Journal of Environmental Research and Public Health*, *12*(10), 12247-12263. https://doi.org/10.3390/ijerph121012247

Lansinoh. (2012). *Moms feel unprepared for and unsupported during postpartum*. https://lansinoh.com/blogs/birth-prep-recovery/moms-feel-unprepared-for-and-unsupported-during-postpartum

Mavragani, A., & Ochoa, G. (2019). Google Trends in infodemiology and infoveillance: Methodology framework. *JMIR Public Health Surveillance, 5*(2), e13439. https://doi.org/10.2196/13439

McClellan, H. L., Hepworth, A. R., Garbin, C. P., Rowan, M. K., Deacon, J., Hartmann, P. E., & Geddes, D. T. (2012). Nipple pain during breastfeeding with or without visible trauma. *Journal of Human Lactation*, *28*(4), 511–521. https://doi.org/10.1177/0890334412444464

Newby, R., Brodribb, W., Ware, R. S., & Davies, P. S. W. (2015). Internet use by first-time mothers for infant feeding support. *Journal of Human Lactation*, *31*(3), 416–424. https://doi.org/10.1177/0890334415584319

Richter, D., Krämer, M. D., Tang, N. K. Y., Montgomery-Downs, H. E., & Lemola, S. (2019). Long-term effects of pregnancy and childbirth on sleep satisfaction and duration of first-time and experienced mothers and fathers. *Sleep*, *42*(4), zsz015. https://doi.org/10.1093/sleep/zsz015

U.S. Department of Health and Human Services. (2011). *The Surgeon General's call to action to support breastfeeding*. U.S. Department of Health and Human Services. https://www.ncbi.nlm.nih.gov/books/NBK52682/pdf/Bookshelf_NBK52682.pdf

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15040

News versus Corporate Reputation: Measuring through Sentiment and financial analysis

Naiara Pikatza-Gorrotxategi, Izaskun Alvarez-Meaza, Rosa María Río-Belver, Enara Zarrabeitia-Bilbao
Universidad del País Vasco UPV-EHU, Spain.

Abstract

Today's companies cannot overlook their reputation if they want to continue to survive. One way to measure that reputation is through two factors: sentiment analysis of news stories in the press about those companies and the financial data of those companies. In this research, the sentiment analysis of news stories about several Euro Stoxx 50 companies for the years 2016 and 2019 has been carried out. For this purpose, the lexicon-based tools VADER and Hu Liu have been used. Then the trends of the results obtained for this four-year period have been analyzed and compared with the trends in their operating results in the same time period. The results obtained indicate that there is a high correlation between the sentiments reflected in the news and their operating results, i.e., when news sentiment about a company improves, its reputation also improves, and this causes its sales to increase. The same is true in the opposite direction.

Keywords: *Sentiment Analysis; Corporate Reputation; VADER SA tool, Hu Liu SA tool*

1. Introduction

In today's society, companies have increasingly more data at their disposal. This data may contain strategic information for companies, however, it is so voluminous that it is not easy to analyse it in the traditional way, making the use of artificial intelligence and data mining indispensable (Agarwal, 2020). In this context, sentiment analysis is a sub-discipline that falls under the umbrella of data mining and computational semantics. According to Gilbert and Hutto (Hutto & Gilbert, 2014), sentiment analysis, or opinion mining, is an active area of study in the field of Natural Language Processing (NLP) that analyses people's opinions, feelings, evaluations, attitudes and emotions by computationally processing subjectivity in text. It refers to the understanding of collected data obtained from sentiment-rich sources such as news, social media sites, reviews, etc. (Agarwal, 2020). Therefore sentiment analysis is concerned with extracting sentiment, opinions and emotions from text (Ravi & Ravi, 2015) and has applications in a wide range of domains, from customer satisfaction to political opinions (Medhat et al., 2014) (Mäntylä et al., 2018) (Ravi & Ravi, 2015).

Another aspect that companies cannot overlook is their reputation, as it affects, among other factors, consumer satisfaction. (Chun, 2005). According to Raithel in his article "The value-relevance of corporate reputation during the financial crisis" (Raithel et al., 2010), corporate reputation can be measured through 2 indicators: the sympathy felt towards the company, and the competence of that company. Consumers, when deciding on a company's reputation, rely on data received through word of mouth, news, advertising, etc. (Kossovsky, 2012). Therefore, one way to measure this sympathy for the company can be through sentiment analysis of news stories about those corporations. If the consumer perceives that the news has a positive tone about the companies, they will have more sympathy towards them and the company's reputation will increase.

The Python tool, VADER (Valence Aware Dictionary and Sentiment Reasoner), a sentiment analysis framework, uses a lexicon-based approach to determine the sentiment values of a sentence. This is used in conjunction with sentiment values explicitly assigned to keywords commonly found among news headlines, or in individual emails (Agarwal, 2020) (Borg & Boldt, 2020). This sentiment extraction typically results in a score that can be translated into positive, neutral or negative (Hutto & Gilbert, 2014). Another frequently used instrument, the Hu & Liu lexicon, was developed for sentiment analysis of customer reviews. The resulting categories (lexicon-based) are Sentiment (an overall measure of positivity), Positive and Negative (good classification metrics in machine learning tasks). This tool has been chosen because it has almost exclusively been used in studies that do not focus on textual production in the social media. (Mayor & Bietti, 2021)
A substantial number of sentiment analysis approaches rely greatly on an underlying sentiment (or opinion) lexicon. A sentiment lexicon is a list of lexical features (e.g., words) which are generally labeled according to their semantic orientation as either positive or

negative (B. Liu, 2010). With Hu Liu, words are categorized into binary classes (i.e., either positive or negative) according to their context free semantic orientation. (Hutto & Gilbert, 2014). Hu and Liu present a natural language-based approach for providing feature-based summaries of customer reviews. The approach uses a part-of-speech tagger to divide words into lexical categories, as only the semantic orientation of adjectives is considered by the algorithm. The use of different instruments for the automatic coding of the same dataset is essential to assess the robustness of results across tools (Mayor & Bietti, 2021). As there are 2 suitable tools, this research will measure the reputation of companies through sentiment analysis, measured with VADER and Hu Liu. Therefore the aim of the article is to analyse the possible correlation between the sentiment analysis of news about companies and their reputation.

2. Methodology

The methodology followed to obtain and analyze the data was as follows:

STEP 1.- Choice of database: The objective is to carry out a sentiment analysis of the news on the 10 highest dividend yielding companies in the Euro Stoxx 50 as of May 2021. This database was chosen because of consistent data for these companies. These companies are: Axa, Eni, Total Energies, Intesa Sanpaolo, ING, Engie, BNP Paribas, Basf, Allianz and Daimler (*El 26% Del Euro Stoxx 50 Paga Una Rentabilidad Por Dividendo Superior Al 4% | Mercados | Cinco Días*, n.d.).

STEP 2.- Data extraction: having selected the companies, their most relevant news items according to different databases were downloaded. To do so, we went to the original source and downloaded the 500 most relevant news items by company and year from the main media. It was decided to analyse the years 2016 and 2019. In the event that a company did not reach 500 news items per year, all of them were downloaded. The total number of news items per company per year is as follows:

Table 1. Extracted news items per year.

Year	2016	2019	TOTAL
News Items	3,994	3,838	**7,832**

For each of the selected companies 1,000 news items (500 per year) have been extracted, with 3 exceptions: *Intesa San Paolo* had only 482 news items in total, ING 250 news items in total, and Total Energies, due to its numerous name changes over the years, produced very little news, so it is not counted. Therefore the total number of news items is 7,832.

STEP 3.- Cleaning and classification of extracted data for SA: Having downloaded all of the news items in txt format, they are imported into the data mining software Vantage Point (W. Liu & Liao, 2017), through which the raw data can be structured for subsequent export in xlx-csv format.

STEP 4.- Conducting Sentiment Analysis: The news items are ready to be exported to Orange, a machine learning and data mining suite for data analysis through Python scripting (Demšar et al., 2013). Now the sentiment analysis of each of the news items will be carried out using the VADER and Hu Liu tools.

STEP 5.- Analyzing the correlation between the Sentiment Analysis of the news and the operating results, by company: The possible correlation between the trend between the Sentiment Analysis with VADER and Hu Liu and the operating profits of each company is analyzed.

3. Results

Once the Sentiment Analysis of the extracted news has been carried out, the results obtained, classified by company and tool used, are as follows:

Table 2. Results of the Sentiment Analysis of the news

	VADER		**HU LIU**	
	2016	**2019**	**2016**	**2019**
AXA	0.5534	0.6067	0.5560	0.9739
ENI	0.4778	0.2920	0.0806	0.0826
INTESA SANPAOLO	0.2188	0.2491	-0.6068	-0.0290
ING	0.7067	0.4582	0.0721	-0.0911
ENGIE	0.6582	0.668075	0.4966	1.063754
BNP Paribas	0.2944	0.226955	-0.2662	-0.5982
BASF	0.7113	0.4003	0.3850	0.1311
ALLIANZ	0.5568	0.5682	0.07709	0.1541
DAIMLER	0.7358	0.5607	1.3256	1.03161

One way to study the data is to analyze their trend over time, and see whether they are improving or worsening. In this way it will be possible to check the trend of the sentiments and opinions reflected in the news about each company, and, since the evolution of sympathy towards these companies is being measured, to analyze whether its reputation could improve or not. An improvement in a company's reputation will, in principle, lead to an increase in sales. Consequently, the evolution in the sentiment analysis has been compared with the evolution in the operating result of each company. The data obtained are as follows in Table 3:

Table 3. Comparison of trends in news Sentiment Analysis and operating profit, by company

	VADER	HU LIU	PROFIT		VADER	HU LIU	PROFIT
AXA				**ENI**			
2016	0.5534	0.5560	7,641		0.4778	0.0806	2,315
2019	0.6067	0.9739	8,427		0.2920	0.0802	8,597
TREND	UP	UP	UP		DOWN	DOWN	UP
INTESA				**ING**			
2016	0.2188	-0.606	8,273		0.7067	0.0721	5,903
2019	0.24914	-0.029	8,760		0.4582	-0.0911	6,834
TREND	UP	UP	UP		DOWN	DOWN	UP
ENGIE				**BNP**			
2016	0.6582	0.4966	9,491		0.2944	-0.2662	10,771
2019	0.6680	1.0637	10,366		0.2269	-0.5982	10,057
TREND	UP	UP	UP		DOWN	DOWN	DOWN
BASF				**ALLIANZ**			
2016	0.7113	0.3851	5,330		0.5568	0.0770	11,056
2019	0.4003	0.1311	4,631		0.5682	0.1541	11,855
TREND	DOWN	DOWN	DOWN		UP	UP	UP
DAIMLER							
2016	0.7358	1.3256	31,963				
2019	0.5607	1.0316	29,165				
TREND	DOWN	DOWN	DOWN				

**Operating profit is shown in millions of euros*

Having obtained the data on the sympathy generated by the companies, it is time to measure the reputation of these companies. In order to measure the reputation trend of companies, two factors should be taken into account, which are also intercorrelated; the likeability that these companies induce and their financial consistency (Raithel et al., 2010). One way to measure this sympathy can be through sentiment analysis of press releases. If those sentiment analyses improve, that will mean an improvement in the company's reputation. This will lead to an improvement in sales and therefore in the operating profit. In turn, an improvement in its financial results will cause the company's reputation to improve, completing the cycle. Figure 1 shows this correlation between the trend in sympathy towards the company and the trend in operating income.

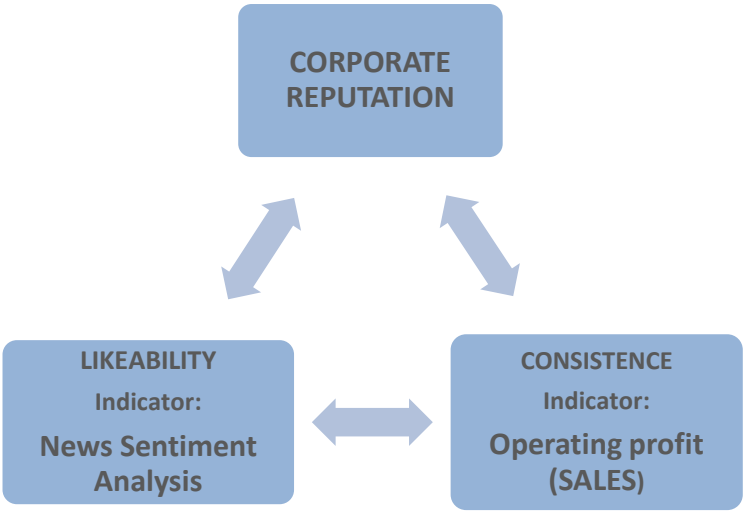


Figure 1: Corporate reputation trend measurement model

4. Conclusions

Based on the obtained data, it can be concluded that the initial thesis is correct. On the one hand, it can be seen that the VADER and Hu Liu data coincide in terms of trend. If we analyze the trend between 2016 and 2019, the trends between these two tools coincide in all cases. In 4 of the cases the trend in the sentiment analysis of the 2 tools is upward with both VADER and Hu Liu (Axa, Intesa, Engie and Allianz), and in the other 5 companies the trend is downward (Eni, ING, Bnp, Basf and Daimler). This data may be an indicator that the 2 tools coincide in their sentiment analysis measurements. If we compare these sentiment analysis trends with the trend in operating results over the same time period, we

can see that they also coincide in almost all cases, i.e., companies that have had a positive trend in their news sentiment analysis increase their operating results and vice versa. This occurs in all cases except for Eni and ING, which increase their profits within that period but lower their scores in news sentiment analysis. The reason for this discordance in the data in the case of ING may be the low number of news items analyzed with respect to the other companies (250 news items in the case of ING, and 1,000 news items in the others). In any case, the correlation between sentiment analysis and operating results is positive in 78% of the cases. Therefore, when sentiment analysis shows a positive trend, operating results increase, i.e., sales of that product increase. When the trend is negative, sales decrease.

References

Agarwal, A. (2020). Sentiment Analysis of Financial News. *Proceedings - 2020 12th International Conference on Computational Intelligence and Communication Networks, CICN 2020*, 312–315. https://doi.org/10.1109/CICN49253.2020.9242579

Borg, A., & Boldt, M. (2020). Using VADER sentiment and SVM for predicting customer response sentiment. *Expert Systems with Applications*, *162*, 113746. https://doi.org/10.1016/J.ESWA.2020.113746

Chun, R. (2005). Corporate reputation: Meaning and measurement. *International Journal of Management Reviews*, *7*(2), 91–109. https://doi.org/10.1111/j.1468-2370.2005.00109.x

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., & Zupan, B. (2013). Orange: Data mining toolbox in python. *Journal of Machine Learning Research*, *14*(August), 2349–2353.

El 26% del Euro Stoxx 50 paga una rentabilidad por dividendo superior al 4% | Mercados | Cinco Días. (n.d.). Retrieved December 2, 2021, from https://cincodias.elpais.com/cincodias/2021/05/18/mercados/1621332461_948065.html

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, 216–225.

Kossovsky, N. (2012). Reputation, stock price, and you: Why the market rewards some companies and punishes others. In *Reputation, Stock Price, and You: Why the Market Rewards Some Companies and Punishes Others* (Vol. 9781430248). https://doi.org/10.1007/978-1-4302-4891-0

Liu, B. (2010). Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition* (pp. 627–666).

Liu, W., & Liao, H. (2017). A Bibliometric Analysis of Fuzzy Decision Research During 1970–2015. *International Journal of Fuzzy Systems*, *19*(1). https://doi.org/10.1007/s40815-016-0272-z

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*,

27, 16–32. https://doi.org/10.1016/j.cosrev.2017.10.002

Mayor, E., & Bietti, L. M. (2021). Twitter, time and emotions. *Royal Society Open Science*, *8*(5). https://doi.org/10.1098/rsos.201900

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Raithel, S., Wilczynski, P., Schloderer, M. P., & Schwaiger, M. (2010). The value/relevance of corporate reputation during the financial crisis. *Journal of Product & Brand Management*, *19*(6), 389–400. https://doi.org/10.1108/10610421011085703

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, *89*, 14–46. https://doi.org/10.1016/j.knosys.2015.06.015

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15041

Cape Town road traffic accident analysis: Utilising supervised learning techniques and discussing their effectiveness

Christo du Toit, Sulaiman Salau, Sebnem Er
Statistical Sciences Department, University of Cape Town, South Africa.

Abstract

Road traffic accidents (RTA) are a major cause of death and injury around the world. The use of Supervised Learning (SL) methods to understand the frequency and injury-severity of RTAs are of utmost importance in designing appropriate interventions. Data on RTAs that occurred in the city of Cape Town during 2015-2017 are used for this study. The data contain the injury-severity (no injury, slight, serious and fatal injury) of the RTAs as well as several accident-related variables. Additional locational and situational variables were added to the dataset. Four training datasets were analysed: the original imbalanced data, data with the minority class over-sampled, data with the majority class under-sampled and data with synthetically created observations. The performance of different SL methods were compared using accuracy, recall, precision and F1 score evaluation metrics and based on the average recall the ANN was selected as the best performing model on the validation data.

Keywords: *Road traffic accidents; supervised learning methods; imbalanced data..*

1. Introduction

A road traffic accident (RTA) can be defined as a rare, random, multi-factor event always preceded by a situation in which one or more road users fail to cope with the road environment (Rospa, 2002). In 2018, there were 12,921 fatalities recorded in South Africa as a result of RTAs. In addition to the social cost, RTAs also have significant economic costs for South Africa. In order to effectively reduce the number and injury-severity of RTAs in South Africa, a better understanding of the relationship between RTA injury-severity and accident-related factors is needed. The use of supervised learning (SL) methods can be very useful for informing future road safety campaigns and potentially reducing the frequency and injury-severity of RTAs.

Several statistical models such as logistic regression, classification and regression trees (CART), random forests (RF) and artificial neural networks (ANNs) have been effectively employed in previous research in different countries including Saudi Arabia, the United States, Italy and Canada, (Al-Ghamdi, 2002; Kong & Yang, 2010; Chang & Wang, 2006; Montella, *et al.*, 2012; Akın & Akbaç, 2010; Chong, *et al.*, 2005; Olutayo & Eludire, 2014) to predict injury-severity of RTAs. These SL methods were shown to regularly outperform logistic regression methods.

There are very few studies conducted on RTAs in South Africa predicting RTA injury-severity. South African literature mostly consists of identifying significant contributors to RTAs. The literature suggests that the quality of South African RTA data is generally poor due to issues such as underreporting, duplication as well as missing values in the data. There are limited studies modeling RTA injury-severity using SL methods, with the focus area mainly in the province of Gauteng (Govender, *et al.*, 2020; Mokoatle, *et al.*, 2019; Twala, 2013; Saar-Tsechansk & Provost, 2007; Moyana & Chibira, 2016). There is a definite need for more research focusing on South African RTA injury-severity prediction especially in the Western Cape province, which is one of the few provinces in South Africa with comprehensive and easily accessible RTA data. This research aims to contribute to the literature by including variables generated from external resources (ie. weather-related variables and geolocation related variables) in addition to the variables obtained from the provincial database. Additionally, it is aimed to highlight the best SL modeling and data sampling approaches for addressing the issue of class imbalance in RTA data.

2. Data and Methods

2.1. Data

The dataset used for this study contains records of more than 82,000 RTAs that occurred during the 2015-2017 period in Cape Town. The data were sourced from the City of Cape

Town, one of the major cities in South Africa and located in the Western Cape Province. The city has a well developed and managed road network and provides researchers with access to its comprehensive RTA database. The dataset contains several variables related to the accident such as street name, crash date, weekday, time of day, alleged cause, crash type, vehicle type, number of vehicles, number of passengers, number of pedestrians involved in the accident as well as the worst injury-severity sustained during the accident. Additionally, the data was enriched with weather-related variables such as temperature, precipitation, wind speed, visibility and cloud cover. Several other variables such as those relating to whether an accident occurred on a public holiday, on a weekend, the season the accident occurred, the number of vehicles involved, whether the accident occurred during peak traffic times as well as whether an accident occurred at an intersection or non-intersection were also added. The location of an accident (amongst other variables) was geocoded in order to obtain geographical coordinates for each accident. After inspecting that valid coordinates were returned for each accident's street address, the longitude and latitude coordinates were added as variables to the dataset.

A common issue with RTA datasets is that the classification categories are imbalanced. The target variable consists of four injury classes, namely: "fatal" (0.27%), "serious" (2.54%), "slight" (10.33%) and "no injury" (86.86%) and is severely imbalanced in Cape Town for the period of 2015-2017 (N= 82,363). Imbalanced data can negatively affect the performance of certain classification methods, especially with regards to predicting the minority class (Weiss & Provost, 2001). This is an issue since the minority class is often the class researchers are most interested in predicting correctly.

Three common data sampling approaches used by researchers to address imbalanced data are utilised in this study, namely (i) undersampling of the majority class, (ii) oversampling of the minority class and (iii) Synthetic Minority Oversampling Technique (SMOTE). SMOTE is a popular over-sampling method developed by Chawla, *et al.* (2002), that creates artificial data examples of the minority class in order to improve the imbalanced distribution of the target variable. While random over-sampling methods simply duplicate existing minority class examples, SMOTE creates artificial minority examples by extrapolating between existing minority examples by finding the k-nearest neighbours of the minority class for each minority example and then generating artificial examples in the feature space of the nearest neighbours. The artificial examples cause the classifier to create larger and less specific decision boundaries resulting in decision region for the minority class to become more general (Chawla, *et al.,* 2002).

The original imbalanced data as well as the data sets generated under different sampling schemes are analysed using multinomial logistic regression, CT, RF, Gradient Boosted Machine (GBM) and ANN methods to predict the target variable, the worst injury-severity resulting from a RTA. The next section briefly discusses the methods, therefore the authors

recommend reading the resources such as Hastie, *et al.*, (2009) and Gareth, *et al.,* (2013) for further details of the various methods.

2.2. Methods

The multinomial logistic regression (MLR) model calculates the probability of an RTA belonging to each injury-severity category relative to a reference category, "no injury" in this case (Yasmin & Eluru, 2013). Classification trees (CT) are a non-parametric classification method that do not require any pre-defined underlying relationship between the predictor variables and the target variables to be specified. CTs use a tree-like structure in order to model the relationship between the predictor variables and the target variable. While a CT might be easily interpretable, it comes at the expense of predictive accuracy as well as high sampling variability (Chang & Wang, 2006). Random forests (RFs) can be used to reduce the variance of a SL method such as CTs (Hastie, *et al.*, 2009). This method is built on the idea that averaging a set of predictions reduces the variance. RF is an ensemble learning method, meaning that many CTs are combined/ensembled into one, better model. An RF model is built by growing a multiple number of trees, $B > 0$, on bootstrapped samples (random sub-samples of data with replacement). Gradient boosting machines (GBM), similar to RF, is an ensemble learning method. Unlike RF, which grows trees independently, GBM grows trees sequentially. This means each tree can learn from the errors made by the previous trees. Artificial neural networks (ANN) are another SL method that can be used for both regression and classification problems. ANNs are especially useful when one does not need interpretable results and when there are non-linear relationships present in the data (Hastie, *et al.*, 2009).

2.3. Evaluation Metrics

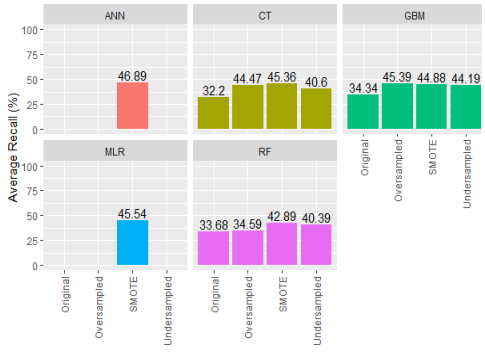
To identify the "best" performing model with regards to predicting RTA injury-severity, evaluation metrics are needed to compare the performance of the different models (multinomial logistic regression, CT, RF, GBM and ANN). While accuracy might be the most simple metric to understand and calculate, it can be misleading especially when dealing with imbalanced data. A model could classify all observations as the majority class and still achieve a relatively high accuracy despite it failing to identify any observations belonging to the minority class. For this reason, the validation data performance of the models are evaluated using the average (i) accuracy, (ii) recall, (iii) precision and (iv) F1 score of each model (Gareth, *et al.,* 2013, p. 149, Grandini, *et al.*, 2020).

3. Results

The analyses were conducted using R (R Team, 2013) and all plots were created using the "ggplot2" package (Wickham, et al., 2016) and models were built with the "caret" (Kuhn,

2008) package. Hyperparameter tuning for RF, GBM and ANN models was performed under University of Cape Town's ICTS High Performance Computing cluster: hpc.uct.ac.za. Different hyperparameters were tested with cross validation. For RFs, the number of predictors at each split considered were (2, 3 and 7), and the number of trees were (500 and 1000); in the case of the GBM models, the number of trees considered were (50, 100, 150, 500), the different shrinkage parameter or learning rates were (0.005, 0.01, and 0.05), and the number of splits in each tree were (2, 10, and 20); for NNs, three hidden layers with varying number of neurons (layer1: 30, 40, 55, layer2: 0, 15, 25, 35, layer3: 0, 5, 15, 25) were tested with different weight decay (0.0001, 0.001, 0.1) settings.

The average recall of the different SL methods are compared in order to identify the best model based on its performance on the validation data. The comparisons of the average recall of all SL methods are shown in Figure 1 (a).



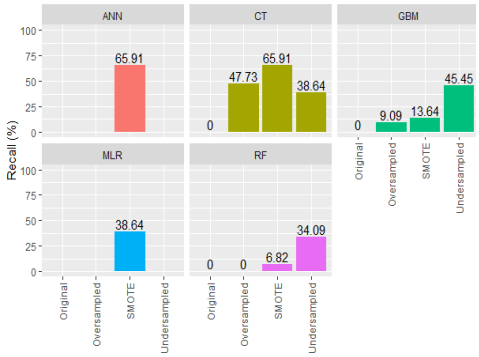


Figure 1. (a) Comparison of average recall *(b) Comparison of recall for "fatal" RTAs*

There is a noticeable difference in average recall between the different SL methods. As shown in Figure 1 (a), the model with the highest average recall is the ANN trained on the SMOTE training data (46.89%). The multinomial logistic regression model trained on the SMOTE data achieved the second highest average recall. The GBM model trained on the oversampled data and the CT trained on the SMOTE data have the next highest average recall respectively with RF model trained achieving the lowest value. It is important to note that some models, ie. ANN in the original dataset, failed to predict the true positive cases (fatal class), hence resulting in evaluation metrics of 0%.

Since RTAs that result in "fatal" or "serious" injuries carry the highest social and economic impact, the misclassification of these classes is undoubtedly the most important issue in predictions. Therefore, a comparison of the recall for the "fatal" RTAs of the different SL methods is shown in Figure 1 (b). The results show that ANN and CT trained on the SMOTE data achieved the highest recall for "fatal" RTAs compared to the other models. The results show that the ANN and CT models achieved the highest recall for "fatal" RTAs overall,

followed by the GBM, multinomial logistic regression and finally the RF models respectively. The CT, RF and GBM models trained on the original data could not identify any "fatal" RTAs. The RF model trained on the oversampled data also failed to identify any "fatal" RTAs.

The ANN model trained on the SMOTE data was selected as the "best" performing model and its performance on the test data, also known as "unseen" data, is assessed. The confusion matrix of the ANN's performance on the test data is shown in Table 1, while the evaluation metrics are shown in Table 2.

Table 1. Confusion matrix of ANN on test data

		Actual Category			
		Fatal	**No Injury**	**Serious**	**Slight**
Predicted Category	**Fatal**	38	1011	223	574
	No Injury	2	12257	64	698
	Serious	1	132	80	157
	Slight	3	908	52	272
					Overall Accuracy: 76.78%

Table 2. Evaluation metrics of ANN on test data by class

Class	Recall (%)	Precision (%)	F1 (%)
Fatal	86.36	2.06	4.02
No Injury	85.67	94.13	89.70
Serious	19.09	21.62	20.28
Slight	15.99	22.02	18.53
Average	**51.78**	**34.96**	**33.13**

The model has an overall accuracy of 76.78% while being able to correctly identify some RTAs belonging to each of the four different injury-severity categories. The model has a higher average recall (51.78%) than any of the SL methods manage to achieve on the validation data. As shown in Table 1, the model also managed to correctly identify a large number of "fatal" and "no injury" RTAs, in contrast to fewer correctly identified "slight" and "serious" RTAs. Table 2 also shows that the ANN model has a very high recall for both "fatal" (86.36%) and "no injury" (85.67%) RTAs and a comparatively low recall for "slight" (15.99%) and "serious" (19.09%) RTAs. This is consistent with the findings of Chong, *et al.* (2005), who found that several SL methods applied in their study predicted "no injury" and "fatal" RTAs most accurately out of all the injury-severity categories. The model also has a high precision for "no injury" RTAs, indicating that it is very precise at correctly identifying "no injury" RTAs. This is in contrast with "fatal" RTAs, for which the model has a low

precision score. This suggests that although the model correctly identifies the vast majority of "fatal" RTAs, it also results in a large number of false positives for "fatal" RTAs.

4. Recommendations and Future Work

Imbalanced data is a common issue found with RTA data. The comparison of the CT, RF, GBM and ANN models trained on the four different training datasets indicate that the best data sampling technique to address class imbalance in RTA datasets is the SMOTE technique with regards to maximising average recall. It is therefore recommended that future researchers use the SMOTE technique to address imbalanced RTA datasets when predicting RTA injury-severity.

It is recommended that the City of Cape Town expand and improve the quality of their RTA data. This study added several new predictor variables to the dataset obtained from the City of Cape Town, several of which were found to be significantly associated with RTA injury-severity. This will allow future researchers to analyse and model RTA injury-severity more comprehensively and identify the most comprehensive set of predictors that will help reduce the frequency and injury-severity of RTAs.

The data used for this study was sourced from the City of Cape Town, who collected and processed the data from the SAPS. The data contained several accident-related variables along with the RTA injury-severity. However, compared to RTA data used in similar international studies, the data used for this study contains relatively few accident-related variables. This can negatively affect the performance of the SL methods as the models are trained on data that is potentially missing some important accident-related variables.

The geographical coordinates of an RTA were added as predictor variables to the data. The use of models that explicitly take the spatio-temporal nature of RTA data into account could be beneficial since this study determined that the geographical location of an accident is significantly associated with RTA injury-severity. Reducing the cardinality of predictor variables in RTA data may also result in more interpretable models.

References

Akın, D. & Akbaç, B. (2010). A neural network (NN) model to predict intersection crashes based upon driver, vehicle and roadway surface characteristics. *Scientific Research and Essays*, 5(19), 2837-2847.

Al-Ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis & Prevention*, 34(6), 729-741.

Chang, L.Y. & Wang, H. W. (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis & Prevention*, 38(5), 1019-1027.

Chawla, N.V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.

Chong, M., Abraham, A. & Paprzycki, M. (2005). Traffic accident analysis using machine learning paradigms. *Informatica*, 29, 89-98.

Gareth, J., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R*, Springer.

Govender, R., Sukhai, A. & van Niekerk, A. (2020). Driver intoxication and fatal crashes. Road Traffic Management Corporation Research and Development.

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning*, Springer.

Kong, C. & Yang, J. (2010). Logistic regression analysis of pedestrian casualty risk in passenger vehicle collisions in China. *Accident Analysis & Prevention*, 42(4), 987-993.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(5), 1-26. doi:http://dx.doi.org/10.18637/jss.v028.i05

Mokoatle, M., Marivate, V. & Bukohwo, M. E. (2019). Predicting road traffic accident severity using accident report data in South Africa. Proceedings of the 20th Annual International Conference on Digital Government Research, 11-17.

Montella, A., Aria, M., D'Ambrosio, A. & Mauriello, F. (2012). Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. *Accident Analysis & Prevention*, 49, 58-72.

Moyana, H. & Chibira, E. (2016). Improving safety in the road transport sector through road user behaviour changing interventions: a look at challenges and prospects. *Proceedings of the 35th Southern African Transport Conference (SATC 2016),* 516-528.

Olutayo, V. A. & Eludire, A. A. (2014). Traffic Accident Analysis Using Decision Trees and Neural Networks. International Journal of Information Technology and Computer Science, 6(2), 22-28.

Rospa. (2002). The Royal Society for Prevention of Accidents (ROSPA) Road Safety Engineering Manual. Retrieved from hhttps://trid.trb.org/view/730321 (2021/07/19)

Saar-Tsechansky, M. & Provost, F. (2007). Handling missing values when applying classification models. Journal of Machine Learning Research, 8, 1625-1657.

Team, R. C. (2013). R: A language and environment for statistical computing.

Twala, B. (2013). Extracting grey relational systems from incomplete road traffic accidents data: the case of Gauteng Province in South Africa. *Expert Systems*, 31(3), 220-231.

Weiss, G. M. & Provost, F. (2001). The effect of class distribution on classifier learning: an empirical study. DOI: https://doi.org/10.7282/t3-vpfw-sf95

Wickham, H., Chang, W. & Wickham, M. H. (2016). Package 'ggplot2'. Create elegant data visualisations using the grammar of graphics. Version, 2(1), 1-189.

Yasmin, S. & Eluru, N. (2013). Evaluating alternate discrete outcome frameworks for modeling crash injury severity. Accident Analysis and Prevention, 59, 506-521.

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15045

Can unlisted firms benefit from market information? A data-driven approach

Alessandro Bitetto[1], Stefano Filomeni[2], Michele Modina[3]
[1]Department of Economics and Management, University of Pavia, Italy, [2]Essex Business School, Finance Group, University of Essex, United Kingdom, [3]Department of Economics and Management, University of Molise, Italy.

Abstract

We employ a sample of 10,136 Italian micro-, small-, and mid-sized enterprises (MSMEs) that borrow from 113 cooperative banks to examine whether market pricing of public firms adds additional information to accounting measures in predicting default of private firms. Specifically, we first match the asset prices of listed firms following a data-driven clustering by means of Neural Networks Autoencoder so to evaluate the firm-wise probability of default (PD) of MSMEs. Then, we adopt three statistical techniques, namely linear models, multivariate adaptive regression spline, and random forest to assess the performance of the models and to explain the relevance of each predictor. Our results provide novel evidence that market information represents a crucial indicator in predicting corporate default of unlisted firms. Indeed, we show a significant improvement of the model performance, both on class-specific (F1-score for defaulted class) and overall metrics (AUC) when using market information in credit risk assessment, in addition to accounting information. Moreover, by taking advantage of global and local variable importance technique we prove that the increase in performance is effectively attributable to market information, highlighting its relevant effect in predicting corporate default.

Keywords: *credit risk, distance to default, machine Learning, market information, probability of default.*

1. Introduction

The aim of banks' core business is to perform accurate assessment of borrowers' capability to repay their debt by collecting information about a given borrower from different sources. The type of information a bank should use when assessing credit risk has been a matter of concern for policy makers since inaccurate credit risk measurement could threaten the stability of the banking sector. In this regard, banks' need to implement reliable credit risk models to timely and precisely forecast business failure is imperative to reach appropriate lending decisions and, eventually, to engage in corrective action.

When focusing on the predictions of default risk of micro-, small- and mid-sized enterprises (MSMEs), a credit risk assessment model should take into account their peculiarities which are not similar to those of larger firms. MSMEs exhibit higher default risk and greater information opacity. Given their importance for market economies, it is imperative to implement credit assessment models specifically addressed to MSMEs with the objective to minimize expected and unexpected losses as accurately as possible.

In this paper, we develop a credit risk model for MSMEs that considers, in addition to accounting measures, market information obtained from comparable publicly listed companies adopting three statistical techniques, namely linear models, multivariate adaptive regression spline, and random forest. Assembling a comprehensive dataset that includes 10,136 unlisted Italian MSMEs, we estimate multivariate forecasting models on the incidence of corporate default by using both market and accounting information employing several advanced statistical techniques. Given the nature of our dataset, we estimate the Merton's Probability of Default (PD) based on market information obtained from listed companies and deemed as comparable by a data-driven clustering approach, avoiding any a-priori assumption of mapping by size, industry and number of employees.

The paper contributes to the literature along two dimensions. The first one involves the implementation of predictive models and their explainability. Our work contributes to a new stream of research (usually called eXplainable Artificial Intelligence) by implementing both a non-linear parametric and non-parametric ML algorithms. Specifically, we go beyond the forecasting of corporate defaults and implement an advanced methodology that involves the use of two cutting-edge techniques to evaluate the importance of variables on forecasts: Permutation Feature Importance (Fisher et al., 2018) explains the overall variables' relevance, whereas Shapley Additive Explanations (Lundberg et al., 2020) provide the contribution of each variable's values to the predicted probability of default for a single observations. In addition, we implement a sophisticated clustering technique that, to the best of our knowledge, is the first application of Artificial Neural Networks to compress the information of financial ratios so to map each unlisted MSMEs to a pool of listed ones. Secondly, our hybrid credit scoring models, which use a combination of market

and accounting information, provide better default predictions for unlisted firms when compared with the respective predictive power of models which only use accounting or market information. We demonstrate that the estimated Merton default probability (PD) measure has incremental predictive power over corporate default when added to a multivariate predictive regression model that already includes accounting information.

One policy implication resulting from our findings is that banks can potentially integrate their hybrid credit scoring methodologies with market information for credit risk assessments, with the purpose of increasing the accuracy of forecasting corporate defaults for unlisted firms. This would allow banks to expand the spectrum of information used in credit risk measurements helping them to enhance their internal hybrid credit scoring by including both accounting and market information on the credit quality of a given borrower. Thus, results reported in this paper could be very helpful for forward-looking financial risk management frameworks (Rodriguez Gonzalez et al., 2018).

The remainder of this paper is organized as follows. Section 2 discusses the data and Section 3 presents the econometric methodology. Section 4 illustrates the empirical results.

2. Data

We use two sources of information for our analysis: a proprietary one, consisting of granular information of 10,136 Italian unlisted MSMEs, and a public one, comprising data on comparable publicly listed companies, i.e., peers.

2.1. MSMEs data

We exploit a unique and disaggregated dataset on an unbalanced panel sample of 10,136 firms and 113 cooperative credit banks, for a total of 19,743 firm-year observations over the period 2012–2014. Specifically, we consider firms with less than 250 employees and revenue at most of 50 million. We selected a subset of 22 financial ratios out of 30 removing the ones showing high partial correlation with many other ratios. Therefore, some ratios with mild correlation with at most one other ratio are still kept because the models we use for the predictions are robust to multicollinearity.

2.2. Peers Data

We select a panel of 40 Italian listed firms, evenly distributed in manufacturing and services sector. We collect accounting figures from Orbis database, developed by Bureau Van Dijk (a Moody's analytics company), by matching the VAT code for each given peer firm . The accounting figures are used to reconstruct and match or proxy the 22 financial ratios of the MSMEs dataset. Moreover, daily stock prices are collected from Refinitiv

Eikon database and are used to compute the annual assets volatility of comparable publicly-listed companies.

3. Methodology

The aim of this paper is to assess the impact of market information, i.e., the Merton's probability of default (PD), in predicting corporate default risk of unlisted firms, in addition to accounting based measures. Our analysis can be summarized into three steps. Firstly, we match each MSME to one or a group of peers and evaluate its firm-wise PD. Section 3.1 recalls how the PD is evaluated following the Merton's model and Section 3.2 describes the peers-to-firm matching procedure, consisting of a low dimensional representation of the 22 variables space and its subsequent clustering. Secondly, we predict corporate default by calibrating different classification models, both using financial ratios as predictors (baseline) and including the PD (extended). Section 3.3 shows the calibration of the models and the differences of models' performance between the baseline and extended cases. Lastly, we investigate which predictor contributes the most to predict corporate default, by means of feature importance techniques. Section 3.4 reports the estimation of the contribution of each variable to the predicted class (default or non-default) for both the baseline and extended cases.

3.1. Estimation of the Merton model

We estimate the Merton model of corporate default risk for our sample of MSMEs. According to the Merton model, the corporate default takes place when the company is unable to pay off its debts, or when the current market of assets falls below the market value of liabilities. For this reason, the market value of equity of the MSME is treated as a call option on the asset value of the MSME with strike price equal to the market value of debt . The MSME asset value process follows a Geometric Brownian motion as shown in Equation (1) below:

$$dA_t = rA_t dt + \sigma_A A_t dz \tag{1}$$

where At is the firms market value of assets and σ_A is the volatility of assets, r is one-year maturity risk-free rate of return, which we choose to be the yield of the 1-year maturity domestic government bond with 1-year maturity .

3.2. Matching unlisted firms with peers

Since there are no market data available for our sample of unlisted MSMEs, we proxy the market volatility of assets of unlisted MSMEs with those of their comparable publicly-listed companies. As for the latter, the market value of assets is computed as the daily

product of their share price multiplied by the number of shares outstanding. Our implicit assumption made for the estimation of the Merton's Probability of Default (PD) and Distance-to-Default (DD) is that those MSMEs which operate in the same industry sectors and have similar balance sheet behaviour with our Italian peers share the same risk profile and belong to the same (market) risk class of the latter . In order to render the matching procedure as accurate as possible, we opt for a clustering approach: we find the optimal number of clusters in the MSME dataset and then we assign each peer to the most similar cluster by minimizing the average distance from all firms in the cluster.

3.3. Prediction of default

After assigning the PD to all our unlisted MSMEs, we calibrate three different models to predict the binary target, (1) for defaulted firm and (0) otherwise. Each model is calibrated with the set of 22 variables (baseline) and with the addition of the PD (extended). First, we inspect the distribution of each input variable with respect to the target variable. Second, we opt for a non-linear and piecewise model, the Multivariate Adaptive Regression Spline (MARS), that estimates multiple polynomial relationships in different partition intervals of each input variable. So, the model can be seen as an ensemble of sub-models that are estimated in each combination of partitions in which input variables can be divided. As MARS is a parametric algorithm, meaning that we have to define a structure of each estimation function, e.g. polynomial, we test also a non-parametric model, the Random Forest (RF).

3.4. Importance of variables

We explore which input variable contributes the most in each model predictions, focusing on the changes when the PD is added. For this reason, we evaluate the predictive power of the variables using two state-of-the-art techniques for feature importance: Permutation Feature Importance (PFI) and Shapley Additive Explanations (SHAP). PFI evaluates the importance of the j-th variable by comparing the performance, e.g. F1-score, of the model that predicts the observations used for the calibration against the performance of the model that predicts the same observations where the values of the j-th column are shuffled. In this way the correlation between the j-th variable and all the others is broken thus removing the influence of that variable on the model predictions. If the change in performance is negligible, the j-th variable is not important for the model. SHAP is based on Shapley values, a method from coalitional game theory which provides a way to fairly distribute the payout among the players by computing average marginal contribution of each player across all possible coalitions. SHAP, uses Shapley values to evaluate the difference of the predicted value of a single observation, comparing the prediction of all possible combinations of variables that include the j-th variable against the ones that do not. The differences are then averaged and the positive or negative change in the prediction is used

as variable importance. For example, if the model predicts the probability of default, SHAP evaluates, for a single observation, which variable contributed most to increase/decrease the final probability. In this way, by exploiting the additive property of Shapley values, it is possible to estimate the impact of all variables on the final predicted value, for each single observation. PFI provides a global measure of importance by assessing the impact of all observations together. Moreover, it measures the changes of a global performance. SHAP, on the other way, provides a local measure of importance, measuring the impact of variables for every single observation. However, taking the average of the absolute values of each observation's SHAP, it is still possible the get a global measure of the average importance of the variables. Instead, taking the average of the Shapley values rather than their absolute value, provides an average effect of each variable on the predictions.

4. Results

As described in Section 3.2, we firstly find the embedding that minimizes the Reconstruction Error. Table 1 reports the optimal embedding dimension k, the reconstruction error of the different algorithms and the R^2. In our context, in analogy with the classical R^2, we compute the RSS term as the Reconstruction Error given by the embedding and the TSS term as the total variance contained in the original data and represents a proxy of how much intrinsic information within the data is preserved in the transformation.

Table 1. Results of dimensionality reduction

Input level	Rows	Columns	Method	Input Dimension	Embedding Dimension	Reconstruction Error (% of Avg Abs Input)	R^2
Firm-year	Firm-year pairs	Variables	AE	19,743 x 22	19,743 x 6	0.1418 (20%)	98%
			RobPCA	19,743 x 22	19,743 x 9	0.2033 (30.6%)	95.70%
Firm (batch of years)	Firms	Variables	AE-LSTM	10,136 x 22	10,136 x 10	0.2138 (31.8%)	94.60%
Firm	Firms	Variables-year pairs	AE	10,136 x 66	10,136 x 32	0.2391 (35.9%)	91.30%
			RobPCA	10,136 x 66	10,136 x 15	0.3857 (58%)	84.80%

Source: our elaboration.

The embedding resulting from AE (AutoEncoder) with the firm-year level approach performed best showing the lowest reconstruction error and the highest R^2. Methods evaluated with firm level approach performed worst and won't be included in the following analysis. Then, we look for the optimal number C of clusters. We select $C = 5$ clusters identified on the AE embedding. Moreover, we apply the UMAP algorithm to visualize the

clusters into a 3-dimensional space. Figure 1 depicts the five clusters for all observations (small points) as well as the matched peers (bold spheres), showing a good separation, even if there is small overlapping between the yellow and green cluster and few blue peers are mapped close to the red ones. We recall that the embedding function *f* is estimated only on the MSMEs dataset and then the peers' embedding is evaluated by applying *f*. Being the PD assigned, we calibrate the prediction models. The following results refer to the PDs evaluated with the pointwise-PD approach because it performed better than the average-PD one, although the findings described below still hold robust. We tune the parameters of each model with the Stratified Cross-Validation and we calibrate the models with the optimal parameters on the entire dataset, so to have a single model to be used for feature importance evaluation. In Table 2 we report the performance on the entire dataset and the average performance on validation folds for each model as well a comparison between the models trained with the 22 ratios only and the ones with the addition of PD. Random Forest is the only model with good performances, being able to capture the different local separation of the data, as discussed in Section 4.3. Nevertheless, all models show an improvement on class-specific performance, i.e. F1-score for the defaulted class, and on the AUC when the PD is included as predictor.

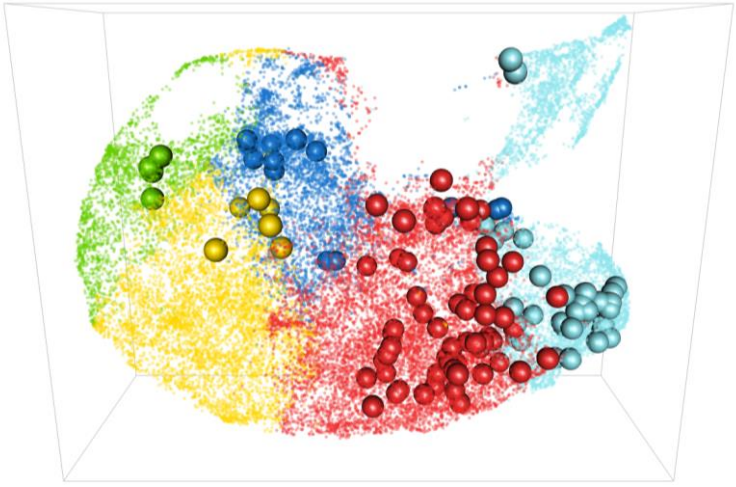


Figure 1. 3D visualization of five clusters for the 6-dimensional AE embedding. Source:our elaboration.

Finally, we explore the feature importance for all models. PFI and SHAP are evaluated on model calibrated with input variables and with the addition of PD. Figure 2 shows the PFI of Random Forest model, where the changes of F1-score are normalized. PD is the second most important variable, slightly below the financial interest on revenues.

Table 2. F1-score and AUC for Elastic-Net, MARS and Random Forest calibrated on dataset with input variables only and with the addition of PD

	F1 (Cross-Val)		AUC (Cross-Val)	
Algorithm	Baseline	With PD	Baseline	With PD
Elastic-Net	30.7% (30.1±1.7%)	35.1% (35.1±1.5%)	79.8% (79.6±0.6%)	82% (81.7±0.8%)
MARS	36% (33.8±1.4%)	40% (37.5±0.6%)	82.5% (81.7±0.6%)	84.2% (82.8±0.8%)
Random Forest	89.5% (85.1±1.7%)	95.8% (91.4±1.2%)	89.8% (85.4±1.1%)	96.1% (91.7±0.7%)

Source: our elaboration.

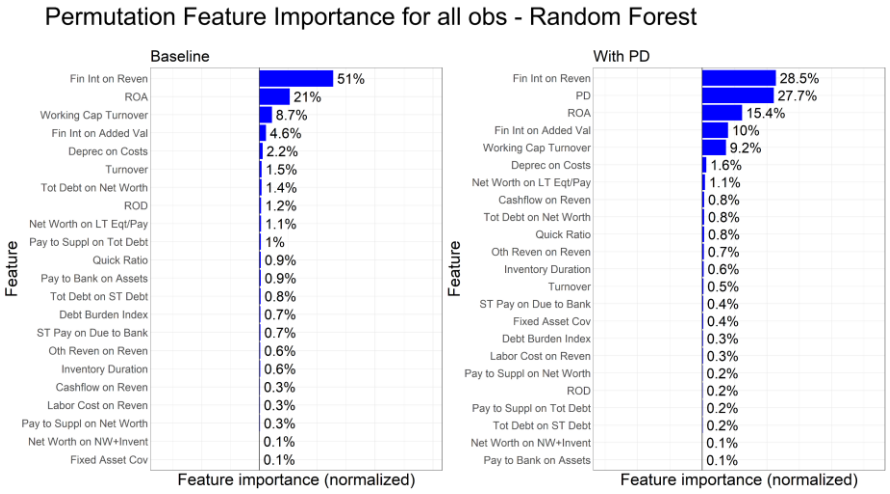


Figure 2. Permutation Feature Importance for Random Forest model. Source:our elaboration.

References

Fisher, A., Rudin, C., and Dominici, F., 2018. odel class reliance: Variable importance measures for any machine learning model class, from the 'rashomon' perspective. URL http: //arxiv.org/abs/1801.01489.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I., 2020. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2:2522–5839.

Rodriguez Gonzalez, M., Basse, T., and Kunze, F., 2018. Early warning indicator systems for real estate investments: Empirical evidence and some thoughts from the perspective of financial risk management. *ZVersWiss*, 107:387–403.

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15052

Analysis of Wellness Experiences in a Tourist Destination

Lourdes Cauzo-Bottala[1], Francisco Javier Quirós-Tomás[1], Myriam González-Limón[2], María del Rocío Martínez-Torres[1]

[1]Departamento de Administración de Empresas y Marketing, Universidad de Sevilla, Spain,
[2]Departamento de Análisis Económico y Economía Política, Universidad de Sevilla, Spain.

Abstract

Wellness tourism has experienced rapid growth in recent years. This has attracted the interest of both researchers and industry representatives. However, experiential tourism has not been investigated in depth through user generated content (UGC) dimensions to create the tourism destination image.

The aim of this paper is to analyse UGC published on Airbnb Experiences in eight Spanish tourist destinations (Barcelona, Islas Canarias, Granada, Madrid, Málaga, Mallorca, Seville and Valencia) to identify the dimensions of Wellness and their relationship with the tourism destination image.

Keywords: *Wellness tourism; UGC (User Generated Content); Tourism Destination Image; consumer behaviour; eWOM; Social media*

1. Introduction

Wellness tourism is considered one of the ten key sectors of the wellness economy, empowering tourists to incorporate wellness behaviours, activities, and life habits into their lives (Global Wellness Institute [GWI], 2018). A new kind of tourism experience is emerging, in which the host/hostess is an essential part of the wellness perceived by the tourist. The image of the tourism destination can be investigated through User Generated Content (UGC) on the tourism of experience.

The aim of this paper is to analyse the UGC published on Airbnb Experiences to identify the dimensions of wellness and their relationship with the image of the tourist destination.

2. Review of the literature

Destination image has been one of the most researched constructs in the tourism literature since the first studies were published in the early 1970s. The use of social media has become so important that the image reflected in the UGC by tourists who share their experiences can influence the perceived image of potential ones, being a good basis for analysing the image of a destination from a demand perspective (Rodríguez-Rangel & Sánchez-Rivero, 2021).

Dunn (1959) originally introduced the concept of "holistic wellness" that included the physical, the mind, the spirit, and the environment dimensions. Wellness tourism is a subset of health tourism (GWI, 2018). However, the mechanisms by which a wellness tourism experience provides avenues to support overall wellness are unclear (Smith & Diekmann, 2017). Exploring this gap is important to better understand the mechanisms through which different types of wellness are achieved while traveling. Thus, in our research, we try to identify the elements of the dimensions of holistic wellness through the UGCs of tourism experiences.

3. Research Methodology

The qualitative thematic classification of travel reviews provides a significant and in-depth understanding of the experience of wellness tourism through the use of netnography analysis in combination with framework analysis. Netnography is defined as ‘a qualitative research methodology that adapts ethnographic research techniques to study the cultures and communities that are emerging through computer-mediated communications ’(Kozinets, 2002, p. 62). Framework analysis utilizes a well-defined process where collected data are selected, listed, and organised in line with key issues and emergent themes discovered through the data. Figure 1 illustrates the general outline of the methodology used in this study, which is detailed in the following subsections.

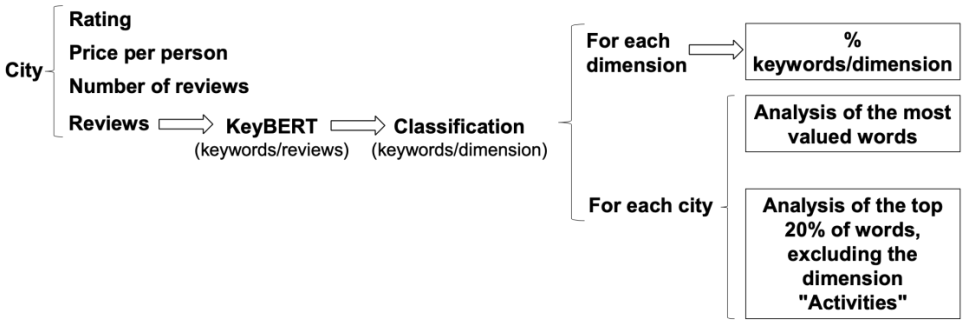


Figure 1. General scheme of the methodology

2.1. Dataset

The data have been extracted from the experiences shared by Airbnb users in the main tourist destinations in Spain: Barcelona, Granada, Islas Canarias, Madrid, Málaga, Mallorca, Sevilla, and Valencia. For each city, the 18 most relevant experiences according to Airbnb have been selected and, for each one of them, not only all the reviews shared by users have been extracted, but also their rating, number of reviews, and price. All reviews were written in English.

Table 1. Number of reviews collected per city

City	Reviews	City	Reviews	City	Reviews
Madrid	8.254	Málaga	1.992	Valencia	3.529
Barcelona	10.774	Gran Canaria	3.013	Granada	2.793
Seville	6.353	Palma de Mallorca	2.452		

All reviews captured for each city have been merged into a single document for further processing.

2.2. Keyword extraction

Keyword extraction relies on the publicly available keyword extraction approach keyBERT (Grootendorst, 2020), which is a deep learning model used to extract keywords from statements or documents. The main idea of keyBERT is that it uses BERT embeddings and cosine similarity to find the words in a document that are most similar to the document itself (Yoo et al., 2021).

2.3. Data Analysis

2.3.1. Classification

To address the dimensions of wellness tourism, the 500 keywords with the highest value according to keyBERT were taken from each destination. These keywords were first classified by two of the researchers and then given to the other members of the research team to identify and discuss any differences or disagreements to ensure the reliability standards.

The dimensions of the experience of wellness tourism were taken according to Dunn's (1959) holistic concept: *Spirit, Environment, Mind* and *Body*. Within each dimension, different elements can be distinguished. Spiritual wellness is the birth place of emotions (*Spirit*), and they are not located in any physical location. It is related to a spiritual connection with the host and immersion in the community. Likewise, *Environment* is related to the physical space where the experience is taken place. On the other hand, *Mind* and mindfulness are an essential part of the holistic concept of wellness (Dunn, 1959), and they refer to experiences that enable a tourist to be aware and conscious of his/her thoughts (Dilette, Douglas and Andrzejewski, 2021), looking for experiencing a profound effect of relaxation and rejuvenation (Voigt, Brown and Howat, 2011). Finally, *Body* is related to the physical reality in contrast to the spirit, and so includes interpretative elements such as activities and services.

2.3.2. Vertical Analysis

A vertical analysis was carried out to see what percentage of each element that makes up each of the dimensions was present in the cities analysed. It was calculated according to the following formula:

$$\frac{\sum Keywords_{ij}}{\sum keywords_{i}}$$

where $\sum Keywords_{ij}$ is the sum of the keywords referring to the element *i* in the city *j* and $\sum Keywords_{i}$ is the sum of the keywords referring to the element *i*.

2.3.3. Horizontal analysis

Taking each of the cities separately, a qualitative analysis was made of the keywords for which keyBERT had the highest value for each city. Having captured data that in most cases were about activities, a second analysis was also carried out with the 20% of the highest value words for each city, but without taking into account those words which previously had been classified as "Activities".

3. Findings

3.1. Experiences features

Information has been collected about three characteristics of the experiences: rating, number of reviews, and price. The rating goes from 1 to 5 stars, with the average of the punctuation of the reviews of the activities of 4.93, with many of them rated 5 stars. Only Sevilla has no activity rated 5 stars, although one of them has a rating of 4.99. The rating of the activities is very high, being the lowest of the experiences analyzed, the 'Pub Crawl Madrid Experience' in Madrid, with 'only' a rating of 4.5 stars.

The number of reviews varies from place to place. The city with more reviews is Barcelona with more than 10.000 of them, while the one with less is Mallorca with 1,727.

The price of experiences varies between cities and within them. The highest price experience is 'Journey into the heart of Gran Canaria' with a cost of 315€, while several of them cost just 1€ such as 'Free tour of Madrid on foot' or 'Sagrada Familia-Symbolism, Architecture, and Gaudi's vision'. The average price ranges from 65.56€ for Islas Canarias to 35.44€ for Granada.

3.2. Dimensions of the Wellness Tourism Experience

Related to the four dimensions, 15 interpretative elements have been attached to them, 3 to *Spirit*, 6 to *Environment*, 2 to *Mind,* and 4 to *Body* (Figure 2). These are based on Dilette et al. (2021). These authors identify and describe 14 of them. Most of them are used in this paper. However, 2 of them have been combined, as well as 2 new ones have been stated: *landscape* and *reception*, both related to *Environment*. Another main change is the interpretative element *Host/Hostess*. It is similar to the one stated by Dilette et al. (2021) as *Staff*. In 'Airbnb experiences', the features of the hosts are essential, since this person is the basic in the selection of the experience by the tourist, as opposed to the staff in the case of a typical tourist company. He/she is who designs the activities, which are specifically aimed at going one step beyond the activity itself, with the intention of providing a 'Memorable Tourism Experience' (MTE).

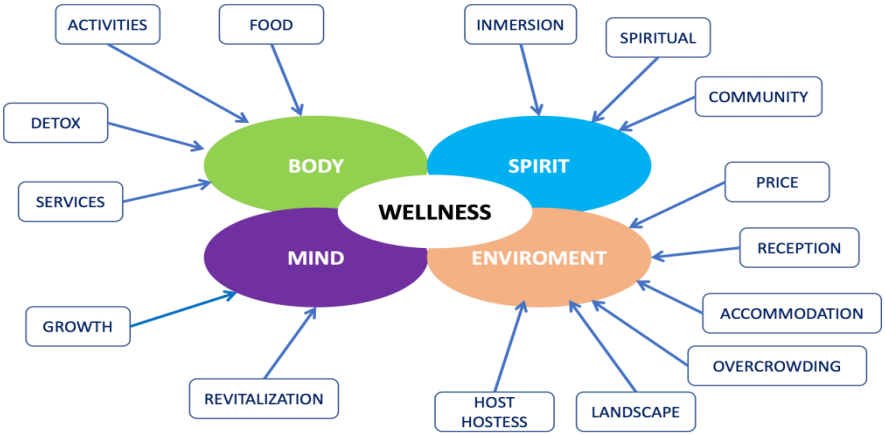


Figure 2. Dimensions and elements of the Experience of Wellness Tourism

The connection of interpretative elements with the feeling of wellness by tourists can be seen as well as pathways as barriers to wellness. In most cases, the reviews analysed show a positive sentiment towards wellness, while several of them can show a clear barrier to its successful achievement.

3.3. Tourism destination image according to wellness experiences

A first approximation to the image of tourist destinations can be found by relating the dimensions of wellness with their key words (Figure 3). Looking at each of the dimensions, we can see how each destination stands out in the different elements that make up each dimension. Among other things, it stands out how the *Mind* dimension is more present in destinations such as Mallorca and Valencia, *Spirit* in Granada, Málaga, Valencia, and Barcelona, *Body* in both islands, as well as in Málaga and Granada, and *Environment* in all destinations except Mallorca.

A second analysis takes into account how each keyword is related to its tourist destination. It can be seen that in all of them except Mallorca, the "Activities" element of the *Body* dimension predominates. As mentioned above, this is logical because Airbnb experiences tells stories about activities. Going deeper, we can see the different activities that stand out in each destination. For example, Barcelona has sea activities ('sailing'); Islas Canarias with visits to 'vineyards'; Granada, Seville and Valencia with excursions ('hike', 'tourguide', 'gotovalenciadaytour'); Madrid and Málaga with activities related to the idiosyncrasy of the country ('flamenco', 'carnival'). Finally, for Mallorca, the element "Food", also included in the dimension *Body*, stands out.

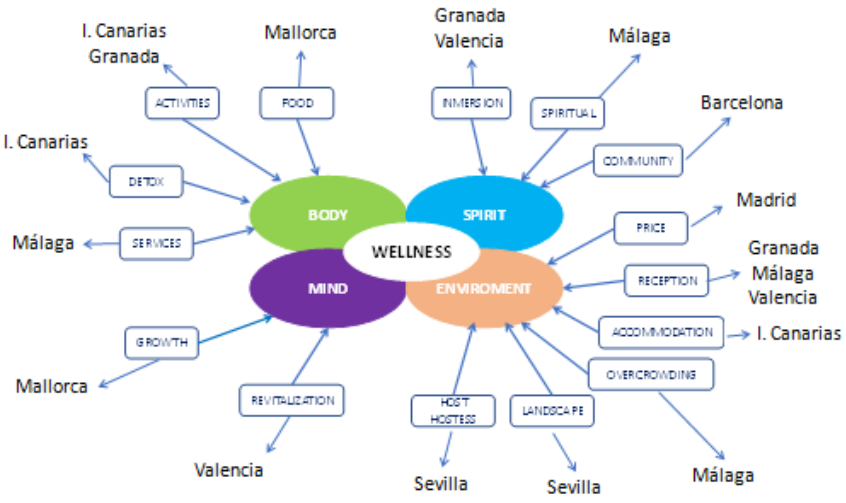


Figure 3. Wellness experiences in a tourist destination

Finally, if we take into account the 20% of the keywords most related to each destination without taking into account those corresponding to "Activities", we can see that in the islands the experience of well being is obtained through "Food", in the inland cities this is more related to a "Spiritual" sensation, and in the coastal cities there is not so much unanimity, as Barcelona and Málaga would be more "Spiritual" and Valencia "Food".

4. Conclusions

The purpose of this study was to identify the dimensions of wellness and their relationship with the image of the tourism destination. These dimensions have been analysed according to the Dunn holistic concept in 8 Spanish tourist cities. Some elements of these dimensions have been adjusted with respect to those proposed by Dilette et al. (2021).

In all the tourist destinations analysed, the wellness experience is achieved through the dimension of *Body*, being the predominant element of this dimension "Activities" in all cities except Mallorca, which was "Food". Without taking into account the wellness generated through "Activities", this study first supports Dilette et al. (2021) in uncovering the importance of culinary experiences (interpretative element "Food") to achieve holistic wellness, as can be seen in Valencia, the Islas Canarias and Mallorca, and second, identifies the importance of those experiences that provide a sense of spirituality, as can be seen in the other cities analysed and refer to the dimension *Spirit*.

Some limitations to this study are, first, the data collected for this study are limited to only those wellness travelers who choose to review their experience online and through Airbnb

experiences, assuming that all provided reviews are honest; second, the study focuses on the 18 better experiences appointed to Airbnb of each destiny analysed, which cannot necessarily be considered representative of the entire population of experiences.

To strengthen the findings of this study, future research on wellness tourism experiences should incorporate other methods of qualitative analysis (e.g. in-depth interviews, focus groups, case studies, etc.) as well as quantitative techniques.

References

Dillette, A.K., Douglas, A.C. & Andrzejewski, C. (2021). Dimensions of holistic wellness as a result of international wellness tourism experiences. *Current Issues in Tourism*, 24(6), 794-810. doi: 10.1080/13683500.2020.1746247

Dunn, H. L. (1959). High-level wellness for man and society. *American Journal of Public Health and the Nations Health*, 49(6), 786-792. doi: 10.2105/AJPH.49.6.786

Grootendorst, M. (2020). KeyBERT: Minimal keyword extraction with BERT. *Zenodo.* doi: 10.5281/zenodo.4461265

Global Wellness Institute (2018). The global wellness tourism economy report. Retrieved from https://globalwellnessinstitute.org/industry-research/globalwellness-tourism-economy/

Kozinets, R.V. (2002). The field behind the screen: Using netnography for marketing research in online communities. *Journal of Marketing Research*, 39(1), 61-72. https://doi.org/10.1509/jmkr.39.1.61.18935

Rodríguez-Rangel, M.C. & Sánchez-Rivero, M. (2021). Qualitative analysis of the online tourist image of Zafra (Spain) through the comments in Tripadvisor. *Investigaciones Turísticas*, 128-151.

Smith, M.K. & Diekmann, A. (2017). Tourism and wellbeing. *Annals of Tourism Research*, 66, 1-13. doi: 10.1016/j.annals.2017.05.006

Voigt, C., Brown, G. & Howat, G. (2011). Wellness tourists: In search of transformation. *Tourism Review*, 66(1/2), 16-30. doi: 10.1108/16605371111127206

Yoo, Y., Lim, D., & Kim, K. (2021). Artificial Intelligence Technology analysis using Artificial Intelligence patent through Deep Learning model and vector space model. *arXiv preprint arXiv:2111.11295*.

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15057

The effect on purchase intention of social media influencers recommendations

Miguel Gonzalez-Mohino, L. Javier Cabeza-Ramirez
Department of Business Organization, University of Cordoba, Spain.

Abstract

The present study aims to examine the impact of involvement (measured through fashion conciousness), perceived authenticity of the message, and perceived risk on purchase recommendations made by influencers. Furthermore, the relationship between these variables is investigated as a risk mitigator in the purchase intention, being induced by influencers in their followers. The global rise of social media has created a new context in which the figure of influencers has become a strategic communication tool that makes the product more familiar, acceptable and desirable to the audience. However, the negative aspects that could influence the purchase intention, such as the risk perceived by the audience, have not yet been studied in depth. To fill this gap, we present a structural equation model using the SmartPLS tool on 948 influencer followers. The results obtained suggest the remarkable influence of involvement with the product, the authenticity of the message and the presence of risk derived from the recommendations; as well as a strong impact of the authenticity of the message as the main mitigating factor of the perceived risk.

Keywords: *Influencer marketing; Purchase intent; Social media; Perceived Risk; Involvement.*

1. Introduction

In these fast-paced and changing times, the global rise of the digital age and the Internet economy have established social media marketing as a pervasive activity for society and an essential part of almost every company's promotional strategies. (Dumitriu et al., 2019). The massive dissemination of all kinds of content through consolidated social platforms such as Facebook, Twitter, Instagram or YouTube (Arora et al., 2019), the progressive arrival of increasingly dynamic and versatile new media such as Twitch, or Tik Tok (Cabeza-Ramirez et al., 2021), their unprecedented integration into people's daily lives (Tafesse and Wood, 2021), as well as the increase in their popularity have given companies and organizations new opportunities to spread brand awareness, attract customers and improve their relationships in a way that had not been done before (Lou and Yuan, 2019). In the field of marketing, it is striking to observe how the figure of the influencer fits perfectly into the definition of marketing, becoming a strategic communication and persuasion tool that makes the product more familiar, acceptable and desirable for the audience (Enke and Borchers, 2019). The use of influencers is linked to the informal communication process that arouses the interest of the potential client (Schwemmer and Ziewiecki, 2018), aligning it with the new paradigm that represents the user of social media as an independent brand ambassador (Boerman, 2020).

However, the emerging literature on influencer marketing has not yet deepened the understanding of the most negative aspects that could influence the purchase intention (Enke and Borchers, 2019; Hudders et al., 2021), particularly those related to the risk perceived by the audience on the recommendations received. This issue has perhaps been overlooked when most approaches consider the absence of risk, and assume that the recommendations are always received as reliable. This probably happens under the premise of an audience that presupposes the influencer as a person who is close, reliable and highly knowledgeable about the product (Casalo et al., 2020).

2. Objectives and hypotheses

The present study seeks to fill this research gap and aims to examine the impact of involvement (measured through fashion conciousness), perceived message authenticity, and perceived risk in purchase recommendations made by influencers. Several mitigating factors of risk are studied in depth, such as the involvement of the follower with the type of sponsored product (Mou et al., 2020), as well as the perception of authenticity of the message transmitted by the influencer (Hudders et al., 2021; Martínez- Lopez et al., 2020). To fill this gap, this research letter proposes the exploration of the model represented in Figure 1, and the following hypotheses:

-H_1: The perception of message authenticity has a positive effect on the purchase intention of the product or service recommended by the influencers.

-H_2: The perception of the sponsored message authenticity, decreases the perceived risk of the recommendations made by influencers.

-H_3: The perception of risk in influencers' suggestions has a negative impact on purchase intentions.

-H_4: Involvement with the product through fashion awareness positively influences the perception of authenticity of the influencers' message.

-H_5: Involvement with the product through fashion awareness has a positive impact on the purchase intention of the products suggested by the influencers.

-H_6: Greater involvement with the product (measured through awareness of fashion), will mean an increase in the perceived risk of the product recommended by the influencers.

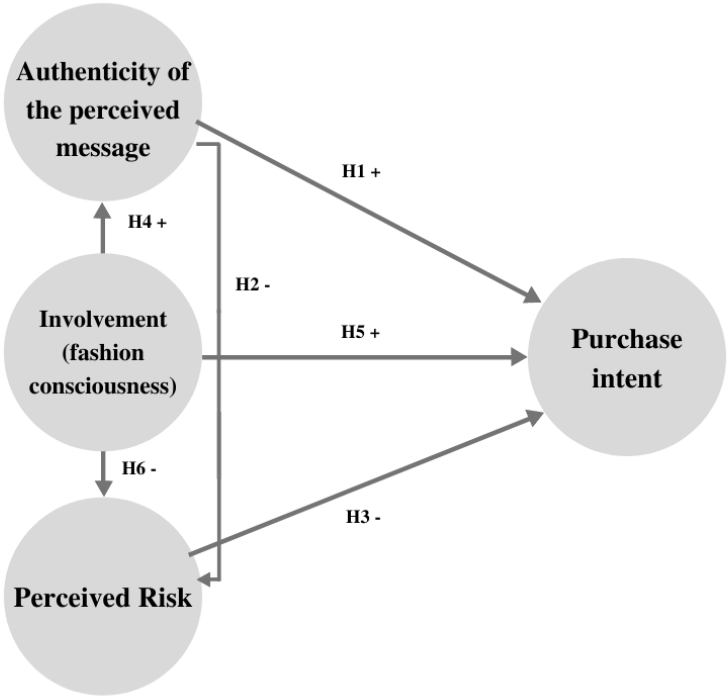


Figure 1. Hypothesis model

3. Methodology and Results

A measurement instrument was designed based on previous studies and specialized literature on the health passport composed of 4 blocks: (a) Perceived Risk (PR) (3 items, (Alalwan et al., 2018; Singh et al., 2021)) ; (b) Authenticity of the Perceived Message (APM) (3 items, (Martínez-López et al., 2020)); (c) Involvement, measured through fashion conciusness, (Inv) (4 items (Lertwannawit & Mandhachitara, 2012)); (d) Purchase intention (Int) (3 items (Hwang and Kim, 2020)). The final sample collection was performed using the SurveyMonkey platform. A link to collect answers and a QR code were generated. The link or the code was distributed through social networks, and through the Moodle learning management tool in different university centers in the south of Spain. The questionnaire remained open during the months of April and May 2021.

The survey was answered by 948 Spanish participants aged 18 or over, followers of an influencer in the field of fashion. Is a exploratory survey, the sample is self-selected no probability based. The sociodemographic data reflect a sample made up of a higher percentage of women (66.7%). Regarding age, we mostly found people between the ages of 18 and 30 (80.1%), with a university or higher educational level (54.85%). Table 1 shows the results of the confirmatory factor analysis (CFA). The items were evaluated according to the values suggested by Hair et al. (2014), ranged from 0 to 1. External loads exceeded the cutoff value of 0.707 suggested by Carmines and Zeller (1979), so no load had to be removed. The validity and reliability criteria of the construct were measured with the criteria proposed by Fornell and Larcker (1981), the composite reliability coefficient (CR) and Cronbach's alpha were above 0.7. In addition, all the AVEs (Average Variance Extracted), as presented in Table 1, range from 0.662 to 0.822, which exceed the recommended 0.50 threshold (Hair, et al. 2017).

The proposed theoretical model was estimated using the SEM structural equation modeling technique through Partial Least Squares (PLS). The SmartPLS 3.3.7 software, developed by Ringle,Wende, and Becker (2015), was used to analyze the relationships proposed in the hypotheses (see figure 1). Table 2 shows the results and the results in relation to the proposed hypotheses. Five of the six proposed relationships are supported. Hypothesis 6, which relates involvement with the product (measured through fashion conciousness), and the perceived risk of recommendations made by influencers (H_6), is rejected as it is not significant ($\beta= 0.059$, $p \geq 0.1$).

Table 1: Measurement model. Factor loadings

Constructs	Items	Factor Loads	Mean (SD)	α Cronbach	CR	AVE
Perceived risk (PR)	PR1. It is risky to buy products recommended/promoted by influencers.	0.925	3.99 (1.77)	0.818	0.882	0.715
	PR2. Buying products recommended/promoted by influencers adds uncertainty about the results I will get from buying the product.	0.82	4.05 (1.83)			
	PR3. Influencers' recommendations expose me to a general risk about the outcome of the product.	0.785	3.82 (1.83)			
Authenticity perceived message (APM)	APM1. I perceive that the influencers' fashion suggestions are authentic.	0.875	3.19 (1.69)	0.864	0.917	0.787
	APM2. Online influencers' fashion posts look real to me.	0.909	3.25 (1.70)			
	APM3. The opinions of influencers on fashion are reliable.	0.877	3.20 (1.67)			
Involvement (Inv)	Inv1. I usually have one or more items of clothing that are in the latest fashion.	0.748	4.56 (2.14)	0.829	0.887	0.662
	Inv2. When it comes to choosing between two outfits, I go by what is in fashion rather than comfort.	0.779	3.43 (2.00)			
	Inv3. Dressing fashionably is important to me.	0.876	3.46 (2.01)			
	Inv4. It is important to me that my clothes are as trendy as possible.	0.846	3.40 (1.99)			
Purchase intent (Int)	Int1. I intend to buy fashion products recommended by influencers	0.899	2.54 (1.77)	0.892	0.933	0.822
	Int2. In the future I will try to buy products sponsored by influencers	0.928	2.50 (1.70)			
	Int3. I will make an effort to buy fashion products recommended by influencers	0.891	2.12 (1.61)			

Source: own elaboration

Table 2. Structural model. Path coefficients and result of the hypotheses

Hypothesis	Independent variable	Dependent variable	Path coefficients (p-values)	Results
H_1	APM →	Int	0.465 (0.000)***	Support
H_2	APM →	PR	-0.141 (0.000)***	Support
H_3	PR →	Int	-0.058 (0.055)*	Support
H_4	Inv →	APM	0.435 (0.000)***	Support
H_5	Inv →	Int	0.195 (0.000)***	Support
H_6	Inv →	PR	0.059 (0.134)	Not Support

***p<0.001; **p<0.05; *p<0.01

Source: own elaboration

4. Discussion and Conclusions

To the best of our knowledge, our research is one of the few that addresses the effects of product involvement on perceived authenticity of the sponsored message as mitigators of perceived risk in influencer-sponsored recommendations. Therefore, it proposes a valuable approach in gaining insight into the presence of general risk associated with influencer endorsements. In the first place, we explore the perception of the authenticity of the influencer's message from a double perspective: as a determinant of the consumer's intention to follow the recommendations towards the product (H1), and as a mitigator of the risk perceived by the audience (H2). The results obtained from hypothesis 1 add to the abundant previous literature that points to trust, authenticity and credibility of the transmitted message as the main causes of the impacts on attitudes and purchase intention in different contexts (Yoon and Kim, 2016; Chakraborty and Bhat, 2018). Regarding the incidence of the authenticity of the message on the perceived risk, the relationship was also verified, in line with the findings of Kim and Lennon (2013) who showed how credibility through reputation has a determining effect on the emotions of consumers and a significant negative effect on perceived risk (Hussain et al. 2017). The third relevant finding have a bearing on the incidence of risk perception on purchase intention (H3). In a way, this finding highlights the need to consider the construct in future research, since it could be a mistake to assume that endorsements generated by influencers are always received as trustworthy (Casalo et al., 2020). Concerning the direct effect between the implication with the product and the perception of the authenticity of the message (H4), it complements the findings reported by Xue and Zhou (2010), which indicate that the greater the implication with the product, the greater the trust towards the source. On the other hand, the significant and positive influence of involvement on purchase intention (H5) stands out, in line with the effects previously identified by Huang et al. (2010) in their work on the involvement of travel blog followers and their purchase intention. Finally, hypothesis 6, which relates

involvement with the product and the perceived risk to the recommendations made by influencers (H6), is rejected. This result could be explained based on the work of Chu and Chen (2019), and Liao et al., (2021) who pointed out that when online consumers have the need to buy products with high perceived risk, they are more active in gathering information, and more receptive to the opinions of others. This research is subject to limitations, among which we can find: (1) the exploratory nature of the model; (2) the concept of risk used, measured as a general perception; (3) a large convenience sample, in a particular context, after the Covid-19 lockdown, and for a specific product (fashion). Consequently, future analyses should consider these limitations as new opportunities for future work.

References

Alalwan, A. A., Dwivedi, Y. K., Rana, N. P., & Algharabat, R. (2018). Examining factors influencing Jordanian customers' intentions and adoption of internet banking: Extending UTAUT2 with risk. *Journal of Retailing and Consumer Services, 40*, 125-138.

Arora, A., Bansal, S., Kandpal, C., Aswani, R., & Dwivedi, Y. (2019). Measuring social media influencer index- insights from facebook, Twitter and Instagram. *Journal of Retailing and Consumer Services, 49*, 86-101.

Boerman, S. C. (2020). The effects of the standardized instagram disclosure for micro-and meso-influencers. *Computers in Human Behavior, 103*, 199-207.

Cabeza-Ramirez, L. J., Fuentes-Garcia, F. J., & Munoz-Fernandez, G. A. (2021). Exploring the Emerging Domain of Research on Video Game Live Streaming in Web of Science: State of the Art, Changes and Trends. *International Journal of Environmental Research and Public Health, 18*(6), 27.

Carmines, E. G., & Zeller, R. A. (1979). *Quantitative Applications in the Social Sciences: Reliability and Validity Assessment* (S. Publications, Ed.)

Casalo, L. V., Flavian, C., & Ibanez-Sanchez, S. (2020). Influencers on Instagram: Antecedents and consequences of opinion leadership. *Journal of Business Research, 117*, 510-519.

Chakraborty, U., & Bhat, S. (2018). The Effects of Credible Online Reviews on Brand Equity Dimensions and Its Consequence on Consumer Behavior. *Journal of Promotion Management, 24*(1), 57-82.

Chu, S.-C., & Chen, H.-T. (2019). Impact of consumers' corporate social responsibility-related activities in social media on brand attitude, electronic word-of-mouth intention, and purchase intention: a study of Chinese consumer behavior. J. Consum. Behav. 18 (6), 453–462.

Dumitriu, D., Militaru, G., Deselnicu, D. C., Niculescu, A., & Popescu, M. A. M. (2019). A perspective over modern SMEs: Managing brand equity, growth and sustainability through digital marketing tools and techniques. *Sustainability*, *11*(7), 2111, 1-24.

Enke, N., & Borchers, N. S. (2019). Social Media Influencers in Strategic Communication: A Conceptual Framework for Strategic Social Media Influencer Communication. *International Journal of Strategic Communication, 13*(4), 261-277.

Fornell, C. & D. F. Larcker (1981). "Structural Equation Models with Unobservable Variables and Measurement Error: Algebra and Statistics." Journal of Marketing Research 18(3): 382-388.

Hair, J.F.J., Hult, G.T.M., Ringle, C.M. & Sarstedt, M. (2017), A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM), 2nd ed., SAGE Publications, Thousand Oaks, CA, available at: https://uk.sagepub.com/en-gb/eur/a-primer-on-partial-least-squares-structuralequation-modeling-pls-sem/book244583

Hair, J. F., Sarstedt, M., Hopkins, L., & Kuppelwieser, V. G. (2014). Partial least squares structural equation modeling (PLS-SEM) An emerging tool in business research. *European Business Review, 26*(2), 106.

Hudders, L., De Jans, S., & De Veirman, M. (2021). The commercialization of social media stars: a literature review and conceptual f(ramework on the strategic use of social media influencers. *International Journal of Advertising, 40*(3), 327-375.

Hussain, S., Ahmed, W., Jafar, R. M. S., Rabnawaz, A., Jianzhou, Y. (2017). eWOM source credibility, perceived risk and food product customer's information adoption. *Computers in Human Behavior, 66*, 96-102.

Hwang, C., & Kim, T. H. (2020). Muslim Women's Purchasing Behaviors Toward Modest Activewear in the United States. *Clothing and Textiles Research Journal, 39*(3), 175-189.

Kim, J., & Lennon, S. J. (2013). Effects of reputation and website quality on online consumers' emotion, perceived risk and purchase intention. *Journal of Research in Interactive Marketing, 7*(1), 33-56.

Lertwannawit, A., & Mandhachitara, R. (2012). Interpersonal effects on fashion consciousness and status consumption moderated by materialism in metropolitan men. *Journal of Business Research, 65*(10), 1408-1416.

Liao, S. H., Hu, D. C., Chung, Y. C., & Huang, A. P. (2021). Risk and opportunity for online purchase intention–A moderated mediation model investigation. *Telematics and Informatics*, *62*.

Lou, C., & Yuan, S. (2019). Influencer Marketing: How Message Value and Credibility Affect Consumer Trust of Branded Content on Social Media. *Journal of Interactive Advertising, 19*(1), 58-73.

Martínez-López, F.J., Anaya-Sánchez, R., Esteban-Millat, I., Torrez-Meruvia, H., D'Alessandro, S., & Miles, M. (2020). Influencer marketing: brand control, commercial orientation and post credibility. *Journal of Marketing Management, 36*(17-18), 1805-1831.

Mou, J., Zhu, W., & Benyoucef, M. (2020). Impact of product description and involvement on purchase intention in cross-border e-commerce. *Industrial Management y Data Systems, 120*(3), 567-586.

Ringle, C. M., Wende, S. & Becker, J. M. (2015). "SmartPLS 3." Boenningstedt: SmartPLS GmbH, http://www.smartpls.com.

Schwemmer, C., & Ziewiecki, S. (2018). Social Media Sellout: The Increasing Role of Product Promotion on YouTube. *Social Media + Society, 4*(3), 2056305118786720.

Tafesse, W., & Wood, B. P. (2021). Followers' engagement with instagram influencers: The role of influencers' content and engagement strategy. *Journal of Retailing and Consumer Services, 58*, 102303.

Xue, F., & Zhou, P. (2010). The Effects of Product Involvement and Prior Experience on Chinese Consumers' Responses to Online Word of Mouth. *Journal of International Consumer Marketing, 23*(1), 45-58.

Yoon, D., & Kim, Y.-K. (2016). Effects of Self-Congruity and Source Credibility on Consumer Responses to Coffeehouse Advertising. *Journal of Hospitality Marketing y Management, 25*(2), 167-196.

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15059

What are Gen Z's and Millennials' opinions on Masculinity in Advertising: a Qualitative Research Study

Toms Kreicbergs, Deniss Ščeulovs
Faculty of Engineering Economics and Management of Riga Technical University, Latvia.

Abstract

The aim of the research is to explore young audiences such as Generation Z's and millennials' opinions on traditional and modern masculinity in advertising. The researchers used the YouTube platform for opinion mining on several advertisements selected to find out what themes emerge from these discourses. By using Nvivo 11 qualitative data analysis software researchers conducted qualitative content analysis, sentiment analysis, and discourse analysis. The results showed that masculinity in advertising gets a lot of Gen Zers' and millennials' attention while the product discourse does not get any noteworthy importance in the discussions about the advertisements. In addition, the research found that when commenting on the advertisements consumers take into consideration the entire context of masculinity and the contemporary notions of it in society, media, popular culture, and competitor's advertisements. The study also concluded that that consumers are more emotionally expressive and opinionated when viewing modern masculinity advertisements than traditional.

Keywords: *Generation Z; millennials, advertising; masculinity, gender.*

1. Introduction

In recent years, there has been a renewed interest in the changing notions about masculinity in advertising, largely due to Gillette's controversial ad aiming against toxic masculinity and promoting inclusivity. Consumer perception of changing notions of masculinity is particularly important for marketing professionals because advertising is created based on assumptions about society (Daechun & Kim, 2007), what does the society appreciate and want to have including material possessions and preferable version of themselves. This question is particularly relevant when advertisers and researchers want to better understand younger audiences such as millennials and Generation Z because many scholars now suggest that the millennial generation has created a much more inclusive culture (McCormack 2012; Thurnell-Reid 2012; Robinson et al., 2019). Nevertheless, marketing to Gen Zers and millennials presents special challenges. While traditional media is still important to these groups, it cannot be compared to the importance of social media and YouTube, which are the main channels to reach Generation Z (Kotler & Armstrong, 2018). That is why this research used YouTube as a platform to gather data from consumers which according to marketing professionals and analysis consist mainly of younger audiences. The data was in a form of YouTube comments on advertisements where masculinity is presented as the key concept. The research question is: What are the current Generation Z's and millennials' opinions on masculinity in advertising and what themes emerge from these discourses?

2. Literature review

After conducting an extensive literature review, the authors concluded that researchers mainly distinguish two very different types of masculinity: traditional masculinity, and modern masculinity. Traditional masculinity is most commonly associated with physical strength (Pollack, 2017), bravery (Smith, 2012), patriotism and emotional stoicism (Ging, 2013), wealth (Zayer et al., 2020), dominance, and a sense of entitlement (Connell, 2014), decisiveness and risk-seeking (Jaffe, 1990), and being a breadwinner, in other words providing for the family (Kimmel, 1996). In contrast, modern masculinity is most commonly associated with progressive thinking (Ging, 2013), being emotionally expressive (Ging, 2019), open-minded (Kimmel, 1996), being sensitive and compassionate (Lalancette & Cormack, 2018).

The characteristics of traditional and modern masculinity go hand in hand with two theoretical concepts that helped authors distinguish brand archetypes and masculinity archetypes in the advertisements selected for this research. Masculinity archetype theory (King, Lover, Magician, and Warrior) is developed by Carl Jung to classify masculinity archetypes and their key characteristics (Moore & Gillette, 1990). Similarly, brand

archetype theory (Mark & Pearson, 2011), which includes twelve different archetypes with their distinctive characteristics and features helped authors notice the differences in selected advertisements' main features and their behaviors.

3. Methods

Since this research is based on qualitative methods, the focus, therefore, was on analyzing consumer engagement, consumer feedback, sentiment, and discourse, by using qualitative content analysis, sentiment analysis, and discourse analysis.

3.1. Data collection

Based on the reason that YouTube comments provide a certain level of authenticity (Tolson, 2010) the authors decided to use online data collection by extracting YouTube comments as data. YouTube is a key site where the discourses of participatory culture and the emergence of the creative, empowered consumer have been played out (Benson, 2016). What is more, researchers argue for the academic value of using YouTube comments as data, saying that YouTube has attracted academic interest in emerging literature that tends to view YouTube as a technological, media, or cultural phenomenon (Jones et al., 2015). On YouTube consumers such as Generation Zers and millennials willingly give their opinions on specific ads where masculinity is at the centre of the advertisement. The data collection was done using a YouTube comments downloader tool. This tool nor any other cannot help understand the researchers what age are the commentators nor any other information about them. The researchers assumed that majority of the commentators would be Generations Zers and millennials based on common sense and data about the average YouTube user. There were six advertisements selected from different brands (Gillette, Barbasol, Old Spice, Axe, National Football League, and Dos Equis), selling different products, such as shaving products, deodorants, shampoos, beer, and so on. The reason for choosing these specific advertisements were because they had a lot of comments for the analysis purposes; they had a variety of masculinity archetypes and brand archetypes; and they provided researchers with different young audiences and their unique feedback about the advertisements due to the audiences the advertisers were communicating to. Three of the advertisements had traditional and the other three had modern masculinity at the core of the ad. The authors of this research took a sample of 400 YouTube comments from each advertisement, making a total of 2400 comments.

3.2. Data analysis

At the beginning of sentiment analysis and discourse analysis, there was qualitative content analysis which was conducted using Nvivo 11 qualitative data analysis software to help with the process of organizing, analyzing, and finding relevant insights in the YouTube

comments. The authors chose to have a mixed content analysis of conventional and direct content analysis. That means that some codes were defined before the analysis of the data based on the theoretical framework and some codes were defined during the analysis of data making it a partially open and partially preconceived coding. After the coding process was done with 198 different codes, the codes were sorted into larger categories based on how they are related and what they reveal. These larger categories were formed into themes that emerged from the YouTube comments. Themes were later merged into larger discourses. Discourse analysis was conducted with the intention to get a deeper understanding of what consumers are experiencing when viewing these ads and to get valuable insights in consumer perception of the masculinity depicted in the particular advertisements. Discourse analysis helped the researchers to pinpoint the key characteristics, behaviors and opinions of consumers. Discourse analysis consisted of larger categories where multiple themes were combined into, but at the same time there was a greater focus on what specifically consumers appreciate about the advertisements in question and what consumers dislike about them.

4. Results

4.1. Results of qualitative content analysis

The results of the qualitative content analysis showed the themes that emerged from each of the selected advertisements, showing how dominant was the masculinity theme and how much did the consumers approve or disapprove of the advertisement and its depiction of masculinity. An example of the qualitative content analysis end result can be seen in Figure 1. Gillette's "We Believe: The Best Men Can Be" advertisement's comment section presented researchers with a variety of themes, where many of them were negative towards the brand. It suggests the researchers that the advertisement and its depiction of modern masculinity and its caregiver brand archetype were not appreciated at all by the young consumers that commented on the advertisement. When combined all the negative themes together, it makes an astounding disapproval rate of 65%, which cannot even be compared with the score of other advertisements analysed in this research, which normally received a 3% or 5% disapproval rate. To make matters worse for Gillette, the theme of Ad appreciation was evident only in 4% of the comments. The combined percentage of masculinity being involved in the advertisement's comment section is 28%, while product discussion did not get any noteworthy attention.

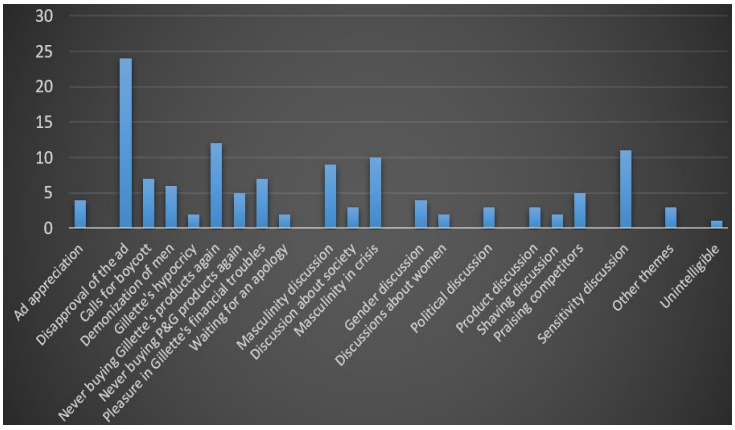


Figure 1. Themes from Gillette's "We Believe: The Best Men Can Be" advertisement's comment section, in % (Source: Authors' original work based on YouTube comments)

However, it has to be pointed out that Gillette's advertisement was an anomaly in the research. Qualitative content analysis of the remaining five advertisements showed very different results, suggesting greater importance in the main character of the advertisement and the displayed masculine characteristics and a greater appreciation of the advertisements where masculinity was at the core of the ad.

Another advertisement as an example of qualitative content analysis for this research is Axe's "Is it ok for guys" (Figure 2), which unlike Gillette received very positive consumer engagement and feedback with 41% ad appreciation.

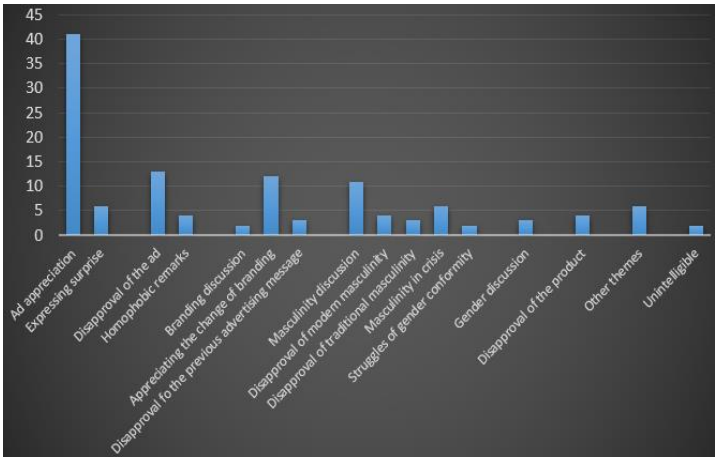


Figure 2. Themes from Axe's "Is it ok for guys?" advertisement's comment section, in % (Source: Authors' original work based on YouTube comments)

The advertisement had a total masculinity discourse of 24% with only 4% of the comments discussing the actual product. A more detailed summary of all six advertisements' qualitative content analysis results are provided in the Table 1. The table shows the top 10 most common themes in all of the six advertisements. As evident by the table the most consistently common theme is Ad appreciation and humour appreciation, while other themes have highly inconsistent frequency in the YouTube comment sections.

Table 1. The results of qualitative content analysis

Theme	Barbasol	Old Spice	Dos Equis	Gillette	NFL	Axe
Ad appreciation	32%	31%	21%	4%	38%	41%
Humour Appreciation	13%	16%	16%	0%	20%	0%
Disapproval	3%	1%	5%	65%	8%	17%
Competitor discourse	31%	0%	2%	5%	0%	1%
Masculinity discourse	15%	2%	3%	19%	3%	24%
Main character discourse	2%	7%	31%	0%	4%	0%
Product discourse	14%	1%	7%	10%	7%	4%
Branding discussion	1%	0%	2%	2%	1%	17%
Satire	1%	24%	29%	1%	8%	0%
Popular culture	0%	22%	6%	0%	5%	0%

Source: Authors' original work based on YouTube comments

4.1. Results of sentiment analysis

The sentiment analysis measuring likeability or how positive, negative or neutral were each advertisement's comment section showed that of the selected ads traditional masculinity's advertisements comment sections were on average more positive than modern masculinity's, with an average of 56% positivity rate to 46% positivity rate (Figure 3). However, that might be due to the significant discrepancy between Gillette's positivity rate and NFL's and Axe's. For instance, only 8% of Gillette's "We Believe: The best man can be" comments were positive, with 78% being negative. While on the other hand the rest of the comment sections, especially NFL's "Touchdown celebrations" (70% positivity rate) had a high level of positivity rate, despite what people mostly associate internet comment sections with. Due to Gillette's "The best man can be" advertisement's high negativity rate, the average negativity rate of the selected modern masculinity ads (39%) is notably higher than the average negativity rate of traditional masculinity ads (8%).

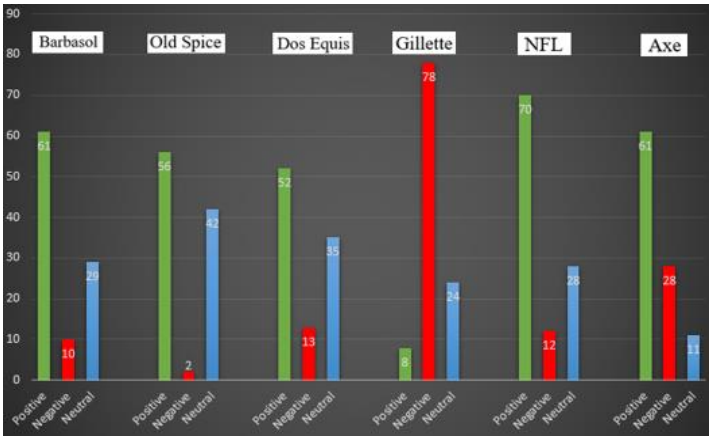


Figure 3. Sentiment analysis results of selected advertisements (Authors' original work)

When it comes to the neutrality of comments, suggesting consumers' lack of emotional involvement, the traditional masculinity ads had much higher neutrality (35%) in their comments than modern masculinity ads (21%). This indicates that consumers are more emotionally expressive and opinionated when viewing modern masculinity content than traditional.

4.1. Results of discourse analysis

Discourse analysis also provided authors with valuable insights into what consumers think about the advertisements in question and their versions of masculinity depicted in both, modern and traditional ways. The masculinity discourse was more evident in Axe's and Barbasol's advertisements. In the case of Barbasol, the discourse analysis revealed how much consumers liked this advertisement and expressed their opposition to Barbasol's competitor Gillette and their advertisement against toxic masculinity. Gillette's comment section had a very negative discourse with consumers showing their distaste for the advertisement and calling for a boycott of the brand. Dos Equis advertisement's discourse analysis of the YouTube comment section revealed a significant debate about the main character of the ad displaying traditional masculinity characteristics. Consumers expressed their appreciation for the original character and distaste for the new, who is not displaying so much traditional masculinity as the original character, also signalling consumer preferences on the matter.

5. Discussion

First, on one hand, the qualitative content analysis, as well as the discourse analysis, showed that masculinity in advertising gets a lot of attention from the young consumers who are the primary commentators on YouTube videos, in this case, advertisements.

However, it also showed very little attention paid to the actual product being advertised. Discourse analysis showed that consumers perceive the advertisement in a larger context of masculinity that is influenced by the society, media, and other competitors of the brand as well as indicating the importance of the main character and his masculinity characteristics in an advertisement. Second, on one hand, the sentiment analysis showed that contrary to a popular belief the discourse in internet comments concerning masculinity is more positive rather than negative. Nevertheless, it has to be pointed out that these advertisements are considered to be quite effective by research done earlier on this matter, determining the effectiveness of advertising. Finally, after analyzing masculinity and brand archetypes in advertising as well as combining the research with traditional and modern masculinity characteristics, the researchers found that traditional masculinity in advertising is rooted in somewhat old-fashioned stereotypes about men and masculinity that perpetuates the idea of gender conformity. While modern masculinity, on the other hand, in advertising is rooted in equality, inclusiveness, opposition to masculine stereotypes.

6. Conclusion

This study aimed to understand the current consumer perceptions of modern and traditional masculinity in advertising and how should advertisers depict masculinity in an effective way so it can resonate with consumers. The study found that masculinity in advertising has such a significant interest for the Gen Z and millennials, that the product discourse does not get any noteworthy importance, suggesting that consumers might be too distracted on the main character and depiction of masculinity to pay attention to the product. The study also concluded that when presenting traditional masculinity in advertising, the main character is notably important as evident by the qualitative content and discourse analyses. And finally, the research found that when commenting on the advertisements consumers take into consideration the entire context of masculinity and the contemporary notions of it in society, media, popular culture, and competitor's advertisements. Further research will expand the methods for assessing consumer perceptions of masculinity in advertising by conducting surveys and focus groups as well as interviews with the experts of the advertising industry.

References

Benson, P. (2016). The Discourse of YouTube: Multimodal Text in a Global Context. Routledge.

Connell, R. (2014). The study of masculinities. Qualitative Research Journal, 14 (1), 5-15.

Daechun, A., Sanghoon, K. (2007). Relating Hofstede's masculinity dimension to gender role portrayals in advertising: A cross-cultural comparison of web advertisements. International Marketing Review, 24 (2), 181-207.

Ging, D. (2013). Men and Masculinities in Irish Cinema. Palgrave Macmillan.

Ging, D. (2019). Alphas, Betas, and Incels: Theorizing the Masculinities of the Manosphere. Men and Masculinities, 22, 638-657.

Jaffe, L.J. (1990). The Effect of Positioning on the Purchase Probability of Financial Services Among Women with Varying Sex-Role Identities. Northeastern University, 17, 874- 881.

Jones, R.H., Chik, A., Hafner, C.A. (2015). Discourse and Digital practices: Doing Discourse Analysis in the Digital Age. Routledge.

Kimmel, M. (1996). Manhood in America: A Cultural History. The Free Press.

Lalancette, M., Cormack, P. (2018). Justin Trudeau and the play of celebrity in the 2015 Canadian federal election campaign. Celebrity Studies, open access.

Mark, M., Pearson, C. (2001). Building Extraordinary Brands Through the Power of Archetypes. McGraw Hill.

McCormack, Mark. 2011. "Hierarchy without Hegemony: Locating Boys in an Inclusive School Setting." Sociological Perspectives 54, 83–101.

Robinson, S., White, A., Anderson, E., Privileging the Bromance: A Critical Appraisal of Romantic and Bromantic Relationships; Men and Masculinities 2019, Vol. 22(5) 850-871

Smith, J. (2012). The Thrill Makers: Celebrity, Masculinity, and Stunt Performance. University of California Press LTD.

Thurnell-Read, T (2012). "What Happens on Tour: The Premarital Stag Tour, Homosocial Bonding, and Male Friendship." Men and Masculinities, 15 (2), 49–70.

Tolson, A. (2010). A new authenticity? Communicative practices on YouTube. Critical Discourse Studies, 7 (4), 277-289.

YouTube Video of Axe advertisement (2017). Is it ok for guys... https://www.youtube.com/watch?v=0WySfa7x5q0

YouTube Video of Barbasol advertisement (2013). Shave Like a Man- War Hero. https://www.youtube.com/watch?v=CzC47F1DTO8

YouTube Video of Dos Equis advertisement (2014). The most interesting man in the world. https://www.youtube.com/watch?v=dYde7LbQrG4&t=23s

YouTube Video of Gillette advertisement (2019). We Believe: The Best Men Can Be. https://www.youtube.com/watch?v=koPmuEyP3a0&t=2s

YouTube Video of National Football League advertisement (2018). Touchdown celebrations. Authors' reuploaded version (due to the original being made private) https://www.youtube.com/watch?v=_tCRjWxWk_o

YouTube Video of Old Spice advertisement (2010). The Man Your Man Could Smell Like. https://www.youtube.com/watch?v=owGykVbfgUE

Zayer, L.T., McGrath, M.A., Castro-González, P. (2020). Men and masculinities in a changing world: (de)legitimizing gender ideals in advertising. European Journal of Marketing, 54 (1), 238-260.

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15067

Influence of popularity on the transfer fees of football players

Pilar Malagón-Selma[1], Ana Debón[1], Josep Domenech[2]
[1]Centro de Gestión de la Calidad y del Cambio, Universitat Politècnica de València, Spain,
[2]Departamento de Economía y Ciencias Sociales, Universitat Politècnica de València, Spain.

Abstract

Search popularity, as reported by Google Trends, has previously been demonstrated to be useful when studying many time series. However, its use in cross-section studies is not straightforward because search popularity is not provided in absolute terms but as a normalized index that impedes comparisons. This paper proposes a novel methodology for calculating popularity indicators obtained from Google Trends to improve the prediction of football players' transfer fees. The database is formed by 1428 players who competed in LaLiga, Premier League, Bundesliga, Serie A, and Ligue 1 on the 2018-2019 season. Random forest algorithm and multiple linear regression are used to study the popularity indicators' importance and significativity, respectively. Results showed that the proposed popularity indicators provide significant information to predict players' transfer fees, as models including such popularity indicators had lower prediction error than those without them. This study's developed method could be used not only for analysts specialized in sports data analysis but for researchers of other fields.

Keywords: *Popularity Indicators; Google Trends; Transfer fees*

1. Introduction

With 158 years of history, football is not only the King sport of today's society but one of the most profitable businesses in the world. According to Ajadi et al. (2021), the combined turnover of the top 20 clubs was €8.2 billion in 2019/20. However, such amounts of income are accompanied by significant expenses. In 2017, Paris Saint-Germain F.C. carried out the most expensive transfer in history, paying €222 million to F.C. Barcelona for Neymar Jr. A year later, this same team bought Kylian Mbappé for €180 million, becoming the second most expensive transfer fee[1] in the history of this sport (Trujillo, 2021). These expenses can only be understood by considering that the main assets of football teams are the players. Thus, given the impact of transfer fees on the economy of football clubs, academics, managers, and other experts have tried to find their main determinants. Factors affecting the transfer fees include the players' performance, position (forward, midfielder, defender, or goalkeeper), the club they play for and, physical characteristics (height, age, etc.) (Garcia-del-Barrio & Pujol, 2007; Herm, Callsen-Bracker, & Kreis, 2014; Müller, Simons, & Weinmann, 2017).

Furthermore, football players are brands themselves, and they have been benefited from the emergence of social networks such as Instagram or Twitter. So, it seems reasonable to study their online popularity and how it impacts the transfer fees, especially if this information is open and easy to access. Previous research has already used popularity measures, such as the followers on social media (Müller et al., 2017; Hofmann, Schnittka, Johnen, & Kottemann, 2019) and their exposition in Google, measured as the number of hits (Garcia-del-Barrio & Pujol, 2007; Herm et al., 2014; Hofmann et al., 2019) to predict the football player transfer fees. Müller et al. (2017) also incorporate Reddit posts, Wikipedia views, YouTube videos, and a Google Trends search index[2]. In this regard, Garcia-del-Barrio and Pujol (2007) and Herm et al. (2014) have found the number of hits in Google results is statistically significant in predicting players' transfer fees, while Hofmann et al. (2019) did the same for the number of followers on social media. Similarly, Müller et al. (2017) found all popularity variables to be statistically significant except the Google Trends search index. This could be because GT does not provide time series of absolute searches but term-dependent normalized indexes from 0 to 100, so they cannot be directly used to compare different players.

This article proposes novel ways to use GT to measure player popularity by requesting several terms (i.e., player names) simultaneously. To demonstrate its usefulness, this

[1] Actual prices paid on the market (Müller et al., 2017).

[2] Google Trends is a tool that allows users to measure the interest that a topic or a person arouses in the world over time according to the number of searches in Google Search Engine (Rogers, 2016).

methodology has been applied to help predict the transfer fees of the players sold during the summer market of the 2018-2019 season.

The rest of the paper is organized as follows. The second section is devoted to explaining how the proposed popularity indicators have been calculated. The third section describes the database and the statistical methods used for carrying out the analysis. The fourth and fifth sections introduce the results obtained and the conclusions achieved, respectively.

2. Popularity indicators with Google Trends

The time series provided by Google Trends (Rogers, 2016) contain a relative index of term popularity, normalized from 0 to 100, which takes value 100 in the period with the highest number of searches. This normalization makes it difficult to compare player popularities since the corresponding series are individually normalized. That is, all series are rescaled considering their maximum. According to Rogers (2016), one way to put the search interest into perspective is to add additional terms. Thus, using two terms (each one representing a player) simultaneously, the results of both series are jointly normalized, i.e., with respect to the highest popularity of any of the terms. Therefore, both series are on the same scale, and it is possible to compare them.

Unfortunately, GT series are reported as whole numbers instead of real numbers. Thus, if a famous player is compared to an unpopular one, GT reports a search index of 0 for the latter, making it difficult to compare the popularity of less searched players. To deal with this issue, we propose to use different reference players according to the relative popularity and position, since the notoriety of a player depends on his position.

In this study, three popularity layers were defined ("High" for the most popular players, "Middle" for the relatively popular, and "Low" for the less popular), each one with a specific reference player. The reference player of the first layer was the one who, compared to the rest, had the highest average search index (in the case of the forwards, Cristiano Ronaldo). The reference player for the second layer was a player whose average popularity index was 1 when put together with the reference player in the first layer. Among all those satisfying this criterium, the least popular one was selected as the reference for the second layer. Less popular players (that is, receiving an average search index of 0 when compared to the layer-1 reference player) were then compared to the reference player of the second layer. This process was repeated in the three levels of the three considered player positions (defender, midfielder and forward). After that, all series were rescaled to account for the different reference players used.

Once the time series of the weekly popularity of the players are on the same scale, their information is summarized in six popularity indicators that can be used in cross-sectional studies: First Principal Component (CP1), mean, median, maximum, minimum and variance.

3. Methodology

The following section presents the database used to carry out the study and the statistical methods used in the predictive analysis. Free R software was used for the analysis (R Core Team, 2019).

3.1. Models

In order to know if the proposed indicators have a significant impact on the transfer fees prediction, two different models were considered. In Model 1, considered as the baseline, the transfer fee for each player *i* is explained by his characteristics[3] and his performance[4].

$$\text{Transfer fee}_i = f(characteristics_i, performance_i) \quad (1)$$

Model 2 extends Model 1 by including the popularity indicators described in Section 2.

$$Transfer\ fee_i = f(characteristics_i, performance_i, popularity_i) \quad (2)$$

3.2. Data

The database used to carry out the analysis was formed by 1428 players who competed in LaLiga, Premier League, Bundesliga, Serie A and Ligue 1 on the 2018-2019 season with 36 explanatory variables related to player characteristics, performance, and six popularity indicators[5]. To train the models, the estimated market value[6] of 1235 players not sold where used. The model error was assessed using the transfer fees of the 193 players sold during the summer market after that season.

3.3. Methods

Random Forest algorithm (RF) (Breiman, 2001) and Multiple Linear Regression (MLR) (Berry, Feldman, & Stanley Feldman, 1985) where used to fit the models. Before carrying

[3] Player characteristics: Position, age, height, and contract.

[4] Player performance: Playing time, aerial duels accuracy, tackles accuracy, interceptions, shots intercepted, fouls, yellow cards, red cards, goals, shots, shots accuracy, assists, dribbles, crosses, corners, passing accuracy, short passes accuracy, long passes accuracy, key passes, progressive passes, deep passes, penalty area, last half quarter, and free kicks.

[5] Popularity indicators were calculated using values for the time period from 17 May 2018 to 26 May 2019 (popularity per week).

[6] Amount of money that a club would be willing to pay for an athlete to sign a contract, regardless of an actual transaction (Herm et al., 2014). Source: www.transfermarkt.com

out the MLR and for alleviating the multicollinearity, variance inflation factors (VIF) were obtained using the `vif_function` (Thompson, 2013), removing those variables with VIF >5 only for MLR. Later, the most relevant variables were selected in the fitted linear model according to the Akaike information criterion (AIC) (Akaike, 1974) using the `MASS` R-package (Venables & Ripley, 2002).

In addition, the repeated k-fold cross-validation technique (in this case, k=5 and repetitions=5) was used to optimise the hyperparameters of the training set for both methods, RF and MLR, using the `caret` R-package (Kuhn, 2020). Finally, the model's performance was obtained on transfer fees of 193 players, who had not been used to build the model.

4. Results

After applying the methodology Table 1 shows the performance for each method and model measured through the root mean square error (RMSE).

Table 1. Summary of model performance measured by means of the RMSE (EUR).

	RF	MLR
Model 1 (baseline)	16,583,803	17,045,285
Model 2 (popularity)	12,083,185	15,338,117

Source: Own calculations

According to Table 1, using the popularity indicators, the RMSE decreased by €1,707,168 and €4,500,618 for the MLR and RF methods, respectively.

Table 2. Variables selected by the Multiple Linear Regression after applying the AIC in model 2

Type of variables	Variables
Player characteristics	Position, age, and contract
Player performance	Playing time, aerial duel accuracy, fouls, goals, shots, assists, dribbles, short passes accuracy, passes in the last quarter of the opponent half, deep passes, free kicks, and corners
Popularity indicators	Variance, minimum, median

Source: Own calculations

Table 2 shows variables selected by the MLR after applying the AIC in model 2. Note that, after applying the *vif_function* the variance, minimum, and median were the only popularity indicators that remained in the model. Thus, the MLR selected these three popularity indicators included in the model.

RF algorithm allows knowing the importance of the variables in the regression model. Liaw and Wiener (2002) incorporated, in the randomForest R package, the calculation of the average increase of the mean squared error (IncMSE%) in the out-of-bag when one variable's values are permuted in the training dataset while the others remain unchanged (the greater the prediction error, the greater the importance of the variable). Figure 1 shows the importance of the variables according to the IncMSE% in the model 2.

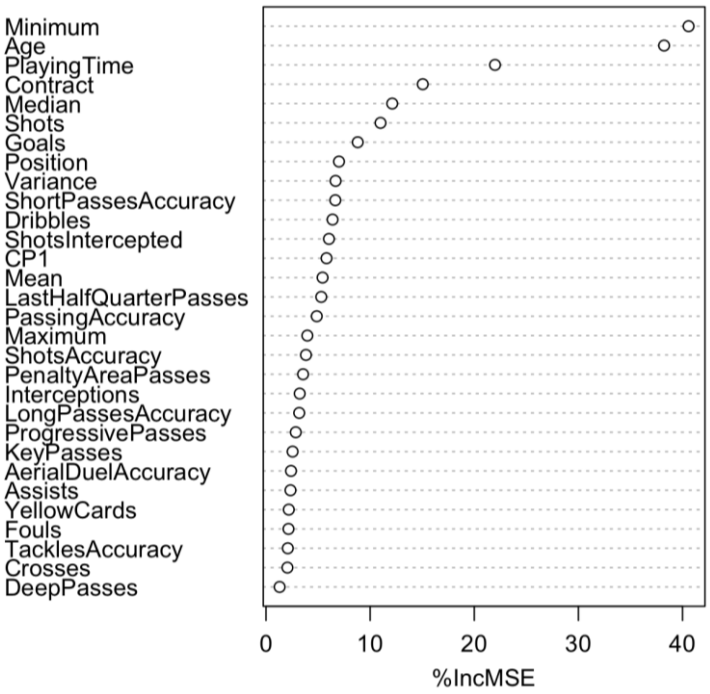


Figure 1. Importance of variables of Random Forest algorithm for the model 2. Source:Own calculations.

Figure 1 shows that in the case of RF algorithm, the most important variable is the "Minimum" popularity indicator, which stores information about the week in which players were least searched. Additionally, in the same way as MLR, the variables "Median" and "Variance" take a relevant position.

4. Conclusion

This work proposed new ways to use GT data to measure player popularity for predicting his corresponding transfer fee. First, because the time series given by GT is individually normalized, it should not be used directly to measure player popularity. Thus, this document recommends using reference players classified by popularity levels as a possible solution. Second, the results (Table 2 and Figure 1) show that the popularity indicators calculated through the proposed methodology (see section 2) improve the prediction of transfer fees. This information may be helpful to analysts who might add these indicators to their models to improve transfer fees prediction.

Acknowledgments

This work was partially supported by grants PID2019-107765RB-I00 and funded by MCIN/AEI/10.13039/501100011033.

References

Ajadi, T.; Bridge, T.; Hanson, C.; Hammond, T.; Udwadia, Z. (2021). Deloitte Football Money League 202. *Deloitte Sports Business Group,* 2-58. https://www2.deloitte.com/uk/en/pages/sports-business-group/articles/deloitte-football-money-league.html.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19(6), 716-723.*

Berry, W., Feldman, S., & Stanley Feldman, D. (1985). *Multiple regression in practice.* Thousand Oaks, CA: Sage.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Garcia-del-Barrio, P., & Pujol, F. (2007). Hidden monopsony rents in winner-take-all markets?sport and economic contribution of Spanish soccer players. *Managerial and Decision Economics, 28*(1), 57-70.

Herm, S., Callsen-Bracker, H.-M., & Kreis, H. (2014). When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community. *Sport Management Review, 17*(4), 484-492.

Hofmann, J., Schnittka, O., Johnen, M., & Kottemann, P. (2019). Talent or popularity: What drives market value and brand image for human brands? *Journal of Business Research, 124*, 748-758.

Kuhn, M. (2020). *Caret: Classification and regression training.R package version 6.0-86*.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News, 2*(3), 18-22. http://CRAN.R-project.org/doc/Rnews/.

Müller, O., Simons, A., & Weinmann, M. (2017). Beyond crowd judgments: Data-driven estimation of market value in association football. *European Journal of Operational Research, 263*(2), 611-624.

R Core Team. (2019). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

Rogers, S. (2016, September 1). *What is Google Trends data — and what does it mean?* (Google News Lab) Retrieved December 12, 2021, from GoogleNews2016: https://medium.com/google-news-lab/what-is-google-trends-data-and-what-does-it-mean-b48f07342ee8.

Thompson, S. (2013, February 5). *Collinearity and stepwise VIF selection*. Retrieved October 15, 2020, from http://beckmw.wordpress.com/2013/02/05/collinearity-and-stepwise-vif-selection/.

Trujillo, I. (2021, August 27). *¿Qué lugar ocupará Mbappé entre los fichajes más caros de la historia del fútbol?*. Retreived March 9, 2021, from LaRazón: https://www.larazon.es/deportes/futbol/realmadrid/20210827/fz345lazmreqjfsn5wrq4u5ogy.html.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S.* New York: Springer.

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15068

Study of e-commerce trends based on customer characteristics in Latvia

Igors Babics[1,2], Rosita Zvirgzdina[3]
[1]Chairman of The Board at DEVNRISE Web Agency (devnrise.com), [2]Business Administration, Turiba University, Latvia, [3]Business Administration, Turiba University, Latvia.

Abstract

The current topics of e-commerce studies in Latvia are examined and basic directions of research are highlighted. The key trends of e-commerce development processes in the country are analyzed, based on the study of the main characteristics and preferences of e-costumers. The main problems in the development of e-commerce in Latvia and further steps to address them are substantiated.

This article aims to investigate trends in Internet commerce in Latvia based on the characteristics of e-customers and determine the prospects and ways in which Latvian businesses can take advantage of the opportunities offered by the Internet.

Based on the research results, an author can note that there is significant e-commerce development potential in Latvia and, in particular, by local businesses.

Keywords: *Internet commerce, trends, e-customers, e-commerce, consumer characteristics, Latvia*

1. Introduction

The steady growth of the digitalization of modern society, significantly accelerated in recent years by the impact of the global pandemic and its economic consequences, has brought to a new level the challenge of improving business processes and, in particular, profit-making processes through the Internet and digital technology. It has become virtually impossible for modern businesses to compete successfully in the marketplace without using the tools and capabilities of the global network and e-commerce. At the same time, the presentation of a company's products for sale on various Internet services is not in itself a prerequisite for successful sales. The effectiveness of e-commerce in today's world is driven by several factors, not the least of which is a clear understanding of your target audience and their needs. In this context, the study of e-commerce through the prism of consumer preferences is of particular relevance.

2. Literature review

During the last decade, the topic of e-commerce has been on the radar screen of a large number of researchers all over the world. In Latvia's case, several research areas prevail in the research field. The first group aims at finding ways to develop small and medium-sized businesses using e-commerce. For example, it is worth mentioning a joint research paper by several Latvian and Lithuanian researchers on the problems of the e-commerce segment in the Baltic states (Rivza et al. 2020), one of its key conclusions being that there is a lack of specialists with socio-technical knowledge, which prevents the domestic e-commerce market in Lithuania and Latvia from reaching the level of most EU member states. In other words, there is a lack of business understanding of the needs of e-customers, and a failure to recognize and study their target audience, including in terms of choosing effective e-marketing tools.

The second line of research is based on identifying the factors that contribute to the development of e-commerce in the Baltic States at the macro level. Particularly, Gudele and Rivza (2015) and Gudele and Jekabsone (2020) conclude that the development of e-commerce in the country depends, among other things, on the general level of education and digital literacy of the population, stressing the point that Latvian businesses do not take full advantage of the existing potential of e-commerce.

The third group of researchers focuses specifically on e-commerce management, attempting to justify innovative ways of developing Internet commerce through customer interaction analysis (Pollack et al 2021).

The fourth group of researchers concentrates on the study of the characteristics and factors of successful performance of firms in the field of Internet marketing. In particular, we can

note the work of Gulevičiūtė, Išoraitė, and Sohail (2019), The author have analyzed the functioning and performance of digital marketing companies in the Baltic States, and have been able to justify different employment and profitability models for companies in the digital marketing sector.

A particular area of research is the area of evaluating the effectiveness of e-commerce activities and digital marketing channels. So, Sceulovs and Lorencs (2017), based on a study of the main characteristics of the digital marketing sector in Latvia and an expert survey, a list of indicators for evaluating the effectiveness of applied digital marketing channels was formed. In turn, Kotane, Znotina, and Hushko (2019) conducted research on key trends in the use of digital marketing tools to identify the most effective strategies for local businesses.

Thus, we can say that two main directions prevail in the Latvian scientific field: the study of e-commerce processes itself and the study of the sphere of digital marketing as a factor in its development. At the same time, there is a great need for further research into the prospects and ways of effective e-commerce development in Latvia based on the characteristics of online customers, which would allow local businesses to form online sales strategies and thus effectively compete in the market not only in Latvia but also at least in other EU countries.

3. Aim, Scientific novelty, Theoretical significance and Methods

3.1. The aim of the article

This article aims to investigate trends in Internet commerce in Latvia based on the characteristics of e-customers and determine the prospects and ways in which Latvian businesses can take advantage of the opportunities offered by the Internet.

3.2. The scientific novelty

The scientific novelty of the study consists in highlighting the characteristics and trends of e-commerce processes in Latvia during the last decade and substantiating the ways of development of online trade by national businesses.

3.3. The theoretical significance

The theoretical significance of the study lies in deepening the understanding of the Latvia's place in comparison to other European Union countries, in terms of e-commerce share in a gross domestic product, which can be further applied to form state programs of online business sector development and to form private online commerce strategies of companies.

3.4. Methods

The methodological basis of the study is general scientific and special methods of economic theory. In particular, in the process of work on the study the following methods were used:

comparative analysis and synthesis - to detail the object of research; economic and mathematical - to analyze the behaviour of businesses and users on the Internet; graphic - to illustrate and chart the subject of research; abstract-logical - to justify objectives, generalizations, and formulation of conclusions.

4. Research results

The progressive development of e-commerce has given a significant boost to the number of online consumers of products and services due to COVID-19 worldwide, not excluding Latvia, where, according to US Office of International Trade estimates (International Trade Administration, 2022), as of summer 2021, 90.8% of adults were using the Internet daily. This is a huge potential and partially active e-commerce audience, as 85% of those surveyed had made at least one online purchase during the year (International Trade Administration, 2022).

For its part, the European Digital Trade Association estimates (The European Digital Commerce Association) that the growth of e-commerce in Latvia in 2020 is 27%, but e-commerce only accounts for 1.03% of Latvia's gross domestic product (Ecommerce Europe, 2021).

Based on the above results, an author can note that there is significant e-commerce development potential in Latvia and, in particular, by local businesses, which is still passively used nowadays in comparison to EU countries. To understand the key factors of this situation, let us compare the indicators of enterprises in Latvia with e-commerce with the number of real e-customers (Fig. 1).

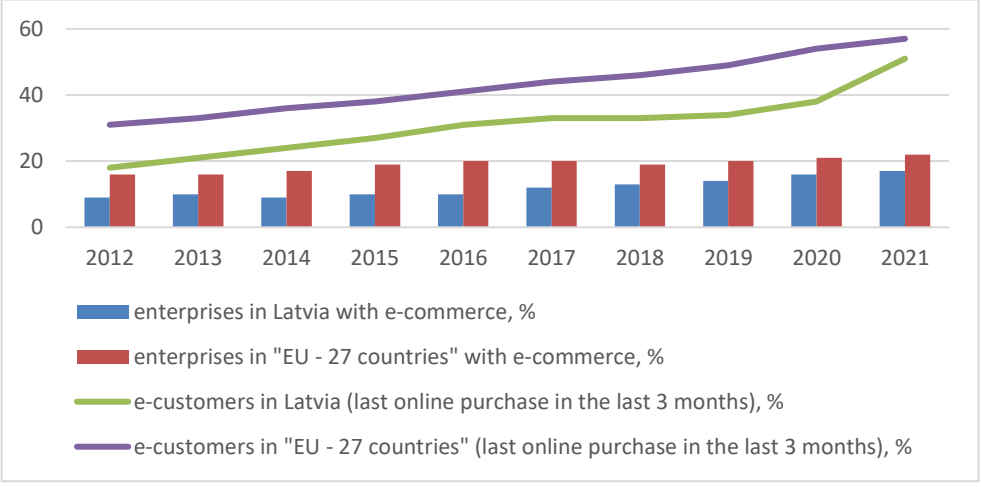


Figure 1. Comparison of the share of companies using e-commerce tools and the share of real e-customers in Latvia in 2012-2021, %. * formed on the basis of data from Eurostat (2022).*

These data allow conclusions to be drawn, first of all, about the reasons why the role of internet commerce in Latvia's gross domestic product is so insignificant. Thus, as of 2021, only 17% of companies were using e-commerce means in their business activities. However, until 2017, the share of companies operating in the field of e-commerce did not exceed 10%. This is a low figure if we compare it with all the EU countries, where, according to Eurostat (2022), 17% of companies were actively using e-commerce opportunities in 2014, and in 2021 the proportion was already 22%.

Thus, we can say that Latvia's Internet business has only now reached the level of Europe in 2014. At the same time, there is a reduction in the gap between Latvia and all European countries in the use of e-commerce by businesses, which allows a conclusion about the growing e-commerce potential in Latvia.

The logical explanation for this situation could be a lack of e-customers in Latvia who made at least one online purchase in the last 3 months. During the period under study, the share of online users in Latvia who have made at least one online purchase increased from 18% in2012 to 38% in 2020, which means 20% growth only, it is 3% less in comparison to All European countries. Although the growth rate of the share of businesses using e-commerce tools in Latvia during the survey period is quite significant, it is not enough to fully meet the demand of e-shoppers.

However, the data obtained show quite significant growth rate of e-customers in Latvia who made at least one online purchase in the last 3 months during the 2021 when it increased by a significant 13% in 2021. An author expects it can be strongly reflected in the e-commerce gross domestic product percentage in Latvia in 2021.

This is supported by the fact that no more than 40% of online purchases are made by national businesses (Figure 2). At the same time, such high rates of online shopping by national businesses in Latvia were observed only in 2021, largely due to multiple quarantine restrictions and fear of individuals stimulating them to change their habitual lifestyle and, among other things, to switch to an increased list of goods bought online.

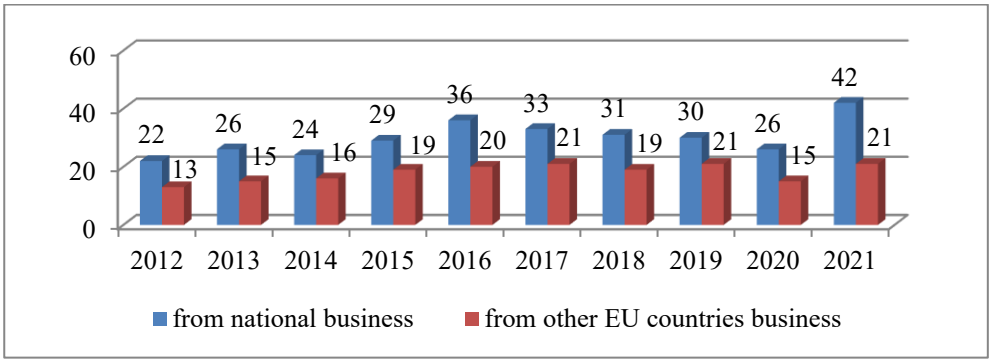


Figure 2. Share of online purchases in Latvia depending on the seller's territorial affiliation in 2012-2021, %. * formed on the basis of data from Eurostat (2022).*

And it is only in 2021 that there is a significant gap between the share of online purchases from local businesses and companies from other EU countries. In fact, for the first time in ten years, the share of online purchases by users from Latvian businesses was twice as high as the share of online purchases by Latvians from businesses from other EU countries. A dynamic analysis of the age structure of Internet users who have never made an online purchase (Figure 3) suggests that the last decade has seen a very significant change in the purchasing behaviour of users under the age of 45. In particular, if the results of the study in 2012 showed that more than 40% of users in this age group have never made an online purchase, by the end of 2020 there will be no more than 11% in each of the selected age groups. This situation confirms the global trend of growth in the active use of online shopping opportunities among the younger population.

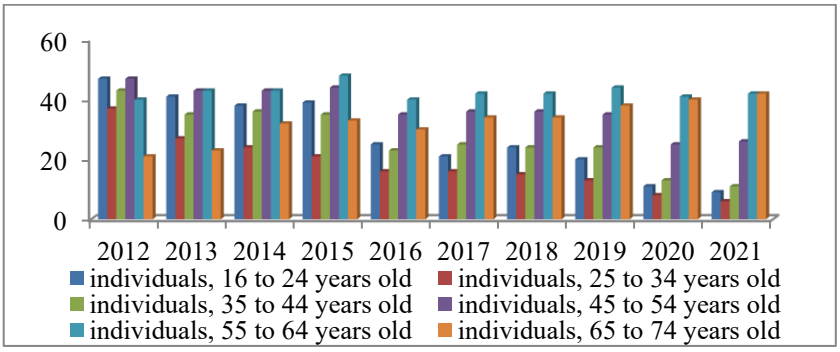


*Figure 3. Dynamics of the share of Latvian Internet users by age category who have never made an online purchase in 2012-2021, %. * formed on the basis of data from Eurostat (2022).*

If we consider individuals aged 45-54, more than 25% of them have never made an online purchase in Latvia as of 2021, and the share of users over 55 not buying online is over 40%. If we add the education criterion to the age criterion, we get the following data (Fig. 4).

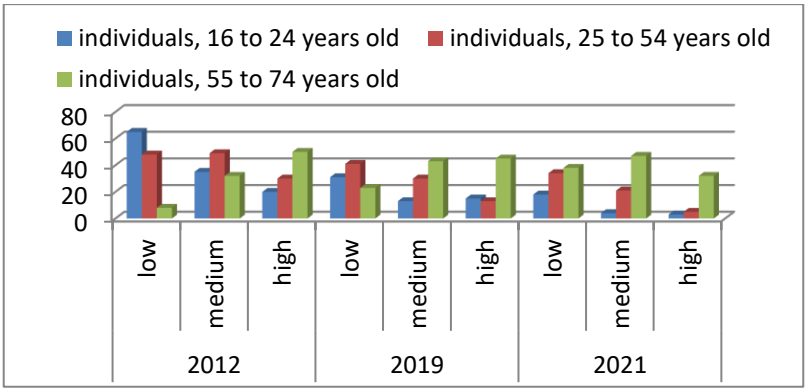


*Figure 4. Dynamics of the share of Latvian users who have never made an online purchase, by age and education, %. * formed on the basis of data from Eurostat (2022).*

The results show that the criterion of users' education has a greater weight the younger the individual is. Thus, there are significant differences in the proportion of young people - from 16 to 24 years old - who have never made an online purchase, depending on their level of education. In 2012, 65% of young Internet users with a low level of education did not use the opportunities of e-commerce, while their peers with a high level of education had only 20%.

While until 2019 the share of Latvian Internet users who have never made an online purchase was decreasing gradually, the last two years have seen significant shifts. At the same time, the importance of the educational criterion has become quite clear in the group of users aged 25-54 years old - thus, according to a survey in 2021, only 5% of users in this age category with a high level of education have never made an online purchase in Latvia.

On the other hand, the 55+ age group in Latvia is the least dependent on educational criteria in the context of online shopping. When targeting this group of consumers, it is worth looking for other factors that contribute to their involvement in e-commerce processes.

Another aspect of studying e-commerce trends in Latvia is an analysis of the preferences of e-shoppers, especially in the context of the impact of the global pandemic (Fig. 5).

For this purpose, we have chosen 2019 as the year before the pandemic and 2021 as the period which is a vivid reflection of the impact of COVID-19 both on the Latvian economy as a whole and the consumer preferences of online shopping users.

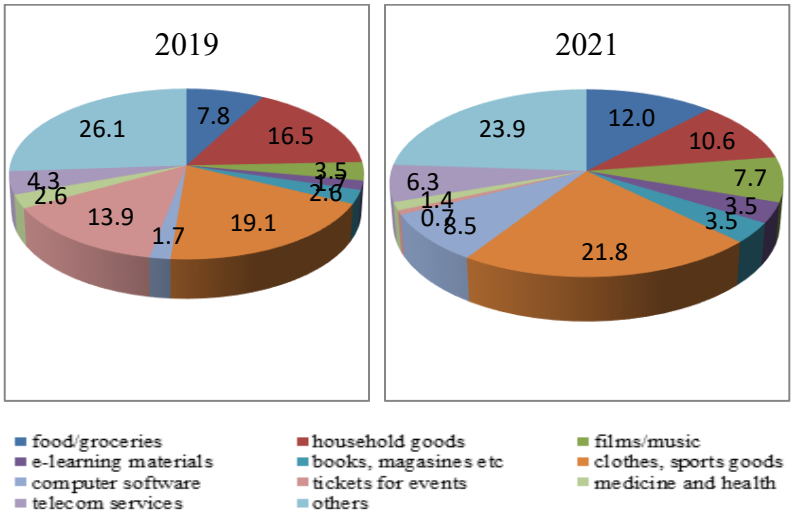


Figure 5. Structure of online purchases by Latvian Internet users by product type in 2019 and 2021, %.
** calculated on the basis of data from Eurostat (2022)*

The analysis of the above data allows us to conclude that significant changes in purchasing preferences in Latvia during the study period occurred. Thus, if before the pandemic, 13.9% of online purchases by the country's users were for tickets to various events, in 2021 it was only 0.7%. Instead, users increased the number of online purchases of software by a factor of five, movies and music by a factor of two. In addition, the number of online purchases of food, educational materials, clothing and sports goods, and telecommunications services has also increased.

Thus, we can identify several key characteristics of e-shoppers in Latvia that should be taken into account by the central authorities responsible for e-commerce development in the country and by businesses themselves when shaping their e-commerce strategy:

The first is the willingness and availability of users to shop online - 75% of the country's online users have experienced online shopping, and more than 51% of them have done so in the past three months. At the same time, only 17% of Latvian companies use e-commerce tools in their operations, which is still relatively less nowadays in comparison to EU countries.

Secondly, there is still an insufficient interest of Latvian e-costumers in national businesses' products and online services. Thus, in the study period, no more than 35% of online purchases by Latvians were purchased from representatives of local businesses. The exception was 2016 with 36% and 2021 with 42% of purchases.

Third, there is a fairly high correlation between a user's willingness to shop online and their level of education. Users between 16 and 54 years of age with a medium to high level of education are more open to online shopping.

Fourth, the analysis of the structure of online purchases by Latvian Internet users by product type in 2019 and 2021 shows significant changes in online purchasing preferences during the study period.

All the above-mentioned allow us to state the necessity of e-commerce sector development in Latvia, as at the moment e-customers to a greater extent finance the formation of the gross domestic product of other countries. At the same time, a special role should be given to working with businesses themselves, so the key areas for further research should be to identify the problematic aspects of online sales tools application by national businesses in the context of different business areas, to build a profile of e-shopper for different product categories, and to determine the main obstacles for users to make online purchases.

5. Conclusion

Significant growth in the digitalization of the modern economic space leads to the increasing dependence of businesses on the level of application of information technologies, including in communications with consumers and the sale of products. In this context, most Latvian businesses have not yet adapted to new aspects and continue to base their operations on twentieth-century approaches and concepts. At the same time, users in the country becomes more active in online shopping, and the situation with the global pandemic has made this number especially high. Thus, there is a need for active development of e-commerce in Latvia, primarily through the application of its tools by businesses, which requires further research into the reasons holding companies back from entering the online segment and identifying incentives for its activation.

References

Ecommerce Europe (2021). European E-commerce Report 2021. Amsterdam/Brussels: Amsterdam University of Applied Sciences & Ecommerce Europe. Retrieved from: https://ecommerce-europe.eu/wp-content/uploads/2021/09/2021-European-E-commerce-Report-LIGHT-VERSION.pdf

Eurostat (2022). Digital Economy and Society database. Retrieved from: https://ec.europa.eu/eurostat/web/digital-economy-and-society/data/database

Gudele I., and Rivza B. (2015) Factors influencing e-commerce development in Baltic rural areas. *Nordic view to sustainable rural development*, 496–500.

Gudele, I., and Jekabsone, I. (2020). Factors Contributing to the Development of E-Commerce by the Degree of use in Latvia. *European Integration Studies*, (14), 207-216.

Gulevičiūtė G., Išoraitė M. and Sohail M. (2019) Effectiveness and possibilities of digital marketing: a case study of Baltic countries. *Annals of Marketing Management & Economics*. Vol. 5. No 1-2, 37–45.

International Trade Administration (2022). Latvia - Country Commercial Guide. Retrieved from: https://www.trade.gov/country-commercial-guides/latvia-ecommerce

Kotane I., Znotina D. and Hushko S. (2019) Assessment of trends in the application of digital marketing. Periodyk Naukowy Akademii Polonijnej. Vol. 33. No 2, 28–35.

Pollák, F., Konečný, M., and Ščeulovs, D. (2021). Innovations in the management of E-commerce: analysis of customer interactions during the COVID-19 pandemic. *Sustainability*, *13*(14), 7986.

Rivza B., Kruzmetra M., Rivza P., Miceikiene A., Balezentis A., Jasaitis J. (2020) E-commerce as a Consequence of Innovation and the Cause of New Innovations for SMEs: the Perspectives of Latvia and Lithuania. *Comparative Economic Research. Central and Eastern Europe* 23 (3), 7–20.

Sceulovs D. and Lorencs E. (2017) How to measure the efficiency of the digital marketing channels? *Proceedings of the 21st world multi-conference on systemics, cybernetics and informatics (WMSCI 2017)*, 62–68.

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15069

Social Desirability and the Willingness to Provide Social Media Accounts in Surveys. The Case of Environmental Attitudes

Beate Klösch[1], Markus Hadler[1], Markus Reiter-Haas[2], Elisabeth Lex[2]
[1] Department of Sociology, University of Graz, Austria, [2] Institute of Interactive Systems and Data Science, Graz University of Technology, Graz, Austria.

Abstract

This paper contributes to the research on combining public opinion surveys and social media data by a) analyzing the effects of social desirability on the willingness to provide social media account information in surveys, and b) evaluating the congruence of opinions expressed in the survey and on social media. We analyze these questions by considering the willingness to make a sacrifice for the environment, i.e., the willingness to pay higher taxes and higher prices. Our results show that Facebook users who oppose environmental measures are less likely to share their account information in the survey, whereas this effect could not be found among Twitter users. Considering the congruence of opinions expressed in the survey and on Twitter, we find similar tendencies both at the aggregate and the individual level.

Keywords: *Survey; Social Media; Facebook; Twitter; Sentiment Analysis; Environmental policies.*

1. Introduction

In a previous paper (Hadler et al. 2022), we showed that the likelihood of providing one's social media account information in surveys is higher among respondents who are in favor of various COVID-19 policy measures. One explanation is the occurrence of a social desirability bias. Social desirability refers to the effect that respondents tend to report behaviors and opinions that are generally assessed positively and give socially desirable answers rather than share their true thoughts if those deviate from social norms (Grimm 2010). Anonymous interviewing and ensuring confidentiality are known to reduce this bias (Larson 2019). Providing access to their social media accounts removes the respondents' anonymity, as researchers will know their actual identities. In line with the idea of social desirability, respondents who share the mainstream opinion to support COVID-19 measures more often provide their account information than respondents who oppose them.

The COVID-19 pandemic and related policy measures, however, are specific topics as they resulted in a polarized public opinion (Reiter-Haas et al. 2022). The current paper thus tests whether our findings on the effects of social desirability are also applicable to another topic – environmental attitudes, i.e., the willingness to pay higher taxes and prices to protect the environment. Finding similar effects would support the general idea of a bias in providing one's social media account towards respondents who share mainstream views. Therefore, our first research question is: *Do attitudes towards environmental protection measures influence the willingness to provide social media accounts in a public opinion survey?*

Alongside this question, we also compare opinions towards environmental measures expressed on social media to those stated in our survey. First, we compare attitudes reported in a public opinion survey with sentiments expressed on social media at the aggregate level. Second, we also take a closer look at the congruence at the individual level and check whether single individuals express the same opinions in the survey and on social media. Therefore, our second research question is: *Do the attitudes expressed in surveys match the social media sentiments at the aggregate and the individual level?*

2. Methods

Our analyses are based on three data sources, i.e., a public opinion survey, the tweets of survey respondents who shared their social media account names, and Twitter data for the same time period. The survey was conducted online in the DACH region (i.e., Germany, Austria, Switzerland) in the summer of 2020. The sample comprises a total of 2560 respondents and resembles the sociodemographics of each country. The questions included attitudes towards the COVID-19 pandemic, environmental attitudes, the use of various social media platforms, and sociodemographics. Respondents were also asked to share the name of

their Facebook and Twitter accounts. However, we were only able to access the Twitter data due to Facebook's terms and conditions.

Our dependent variable is derived from the questions on sharing one's account information and includes the following groups for Facebook and Twitter users respectively: a) respondents without an account, b) account holders who were not willing to share their account name, and c) account holders who shared their account name. Independent variables include the sociodemographic variables gender, age, and education, as well as attitudes towards environmental measures ('How willing would you be to a) pay higher prices, and b) pay higher taxes in order to protect the environment?'). Responses to this item are measured on a five-point scale with 1 = no acceptance and 5 = high acceptance. We also computed an index, displaying the mean score of the two attitudes. An overview of the survey variables is provided in Table 1.

Table 1. Sample characteristics

Variables	Mean (SD) or %
Acceptance of environmental measures (1 = no acceptance, 5 = high acceptance)	
How willing would you be to…	
a) pay higher prices in order to protect the environment	2.90 (1.17)
b) pay higher taxes in order to protect the environment	2.40 (1.17)
Index (mean score of previous variables)	2.65 (1.20)
Sociodemographic variables	
Female	50.4%
Age	44.34 (13.90)
Education	
Compulsory school	35.2%
Vocational training	11.6%
High school degree	23.9%
University degree	29.3%

For the Twitter data, we used the Twitter Search API[1] and matched the time period to the survey data (i.e., from July 30th, 2020, to August 10th, 2020). The search query was restricted to the German language containing terms related to the environment[2], which resulted in a total of 16,780 tweets. Furthermore, we considered two smaller subsets of these tweets containing the German and English terms for prices (i.e., 'preis' and 'price' = 275 tweets) and taxes (i.e., 'steuer' and 'tax' = 470 tweets). Subsequently, we conducted a sentiment

[1] We used twarc2 with the full-archive search using the academic research product track (https://twarc-project.readthedocs.io/en/latest/twarc2_en_us/#search).

[2] The list contains 'environment', 'climate', their German counterparts 'umwelt' and 'klima', as well as their corresponding hashtags (e.g., #environment).

analysis using the TextBlob library with the German language extension, which includes a sentiment polarity lexicon that we use for sentiment extraction. We chose this approach as it applies well to the analysis of short texts, such as tweets. After extracting the sentiment, we removed tweets that express no sentiment to exclude purely objective statements which would otherwise dominate the resulting distribution. The extracted sentiments are on a scale from −1 for negative to +1 for positive sentiment.

For the analysis at the individual level, we manually annotated all tweets mentioning the keywords 'environment'/'umwelt', 'climate'/'klima', or the accompanying hashtags from those respondents, who provided us with their Twitter handle and compared their overall opinion with their survey answers. This resulted in 60 tweets by 9 individuals, as only this small number of respondents who provided their account information used the selected keywords or hashtags in their tweets or retweets. Since they also shared only a few tweets, we expanded our time period from the 12 days of the survey to the entire year 2020. Our research required specific ethical considerations (Sloan et al. 2020). We informed the survey respondents about the content of our research, the voluntary nature of their participation, as well as the confidential treatment of their data. Hence, the archived dataset (Hadler et al. 2021) does not include any information that would allow the identification of individuals, including Facebook and Twitter account names or tweets.

3. Results

3.1. Social Media use and the willingness to provide account information

As shown in Table 2, our results indicate that almost 70% of the survey respondents use Facebook, whereas only 16% use Twitter. Among these, a similar proportion describes themselves as active users in each case (40% vs. 35%), as well as a similar number of users of both platforms have shared their account information (35% vs. 30%) which is in line with the numbers reported in previous studies (Al Baghal et al. 2020). However, Facebook's terms and conditions do not allow access to their data, which is why we could only analyze the Twitter data of our respondents. At last, we were able to access 79 Twitter accounts as 40 people (accidentally) provided a false account name or protected account.

Table 2. Overview on Social Media use and the willingness to provide account information

	Facebook	Twitter
Account holders[3]	1774 (69.3%)	404 (15.8%)
Active users[4] (% of account holders)	700 (39.5%)	141 (34.9%)
Account provided[5] (% of account holders)	617 (34.8%)	119 (29.5%)
Successfully accessed[6] (% of account holders)	N.A.*	79 (19.6%)
Total sample	2560	2560

*Account access is restricted by Facebook's terms and conditions.

We conducted two multinomial logistic regression models to estimate the influence of opinions on environmental measures and sociodemographics on the willingness to provide one's account information. The reference groups in Table 3 are respondents who provided their account information. Hence, the regression models show the results for survey respondents who do not have a Facebook or Twitter account, and for those who do have an account but did not share them.

Table 3. Regression analysis on Social Media use and the willingness to provide account information

	Facebook account (ref.: account shared) b-values		Twitter account (ref.: account shared) b-values	
	no account	**account but not provided**	**no account**	**account but not provided**
Intercept	-.629	.666*	.864	1.210
Acceptance of env. measures (index)	-.115*	-.154**	-.103	-.029
Gender (Female)	-.138	.162	.770***	-.052
Age	.031***	.004	.025**	-.002
Compulsory School	-.177	-.201	.275	-.178
Vocational training	-.225	-.152	.319	-.401
High school degree	.334*	.157	-.017	-.159
Cox-Snell	.040		.044	
Nagelkerke	.046		.067	
X^2 *(df)*	103.621***		112.674***	
n	2516		2516	

Note: Acceptance of env. measures low value = low acceptance; Gender 1 = male, 2 = female; Age numeric; Education reference category = University degree. * = $p < 0.05$, ** = $p < 0.01$, *** $p < 0.001$.

[3] 'Do you have a private Facebook/Twitter account?' (yes/no)

[4] 'How would you describe the way you use Facebook/Twitter?' (actively, i.e., posting/passively, i.e. reading etc.)

[5] 'We would like to find out who is using Facebook/Twitter and for which purposes. If you provide access, we will keep all your information confidential. Do you agree to provide us with your personal Facebook/Twitter username as well as access to your data in order to link it to this survey data?' (yes/no)

[6] 'What is your username?'

As for Facebook, the results indicate that respondents without an account tend to be older and have a high school degree rather than a university degree. In terms of environmental willingness, they are more likely against the surveyed environmental measures. Similarly, respondents who do have a Facebook account but did not share it in the survey are more likely opposed to environmental measures. In comparison, opinions on environmental measures have no significant influence on the willingness to share one's Twitter account. Only some sociodemographic variables show significant effects in the sample, i.e., that women and older individuals are less likely to have a Twitter account. To evaluate the soundness of our results, we additionally performed binary logistic regression models (account shared vs. account not shared; account shared vs. account not shared and no account), whose results are very similar to those of the multinomial models.

3.2. Comparing public opinion and sentiments expressed on Twitter

We need to limit our second research question – do the opinions regarding environmental measures shared in the survey match those expressed on social media – to the Twitter data, as we do not have access to the Facebook data. The results regarding the overall Twitter data show a positive sentiment in tweets using the keywords or hashtags 'environment' (median = 0.7, mean = 0.21) or 'climate' (median = 0.7, mean = 0.27). Both keywords show a similar distribution, whereby the sentiment regarding the term 'environment' is slightly more dispersed. For the matching at the aggregate level, we first compare the survey responses to the sentiment analysis considering all tweets related to environmental measures 'pay higher prices' and 'pay higher taxes'. Second, we turn towards the individual comparison of those survey respondents that provided their Twitter handles.

Figure 1 shows the opinions on the two environmental measures expressed on Twitter (top) as well by all survey respondents (middle) and survey respondents who have a Twitter account (bottom). The Twitter boxplots (top) show that price (median = 0.7; mean = 0.32) receives way more positive sentiments than tax (median = -0.7; mean = -0.37). Regarding price, the opinions are more diverse as the wider quartile range indicates, yet show the most positive sentiment within all data sources.

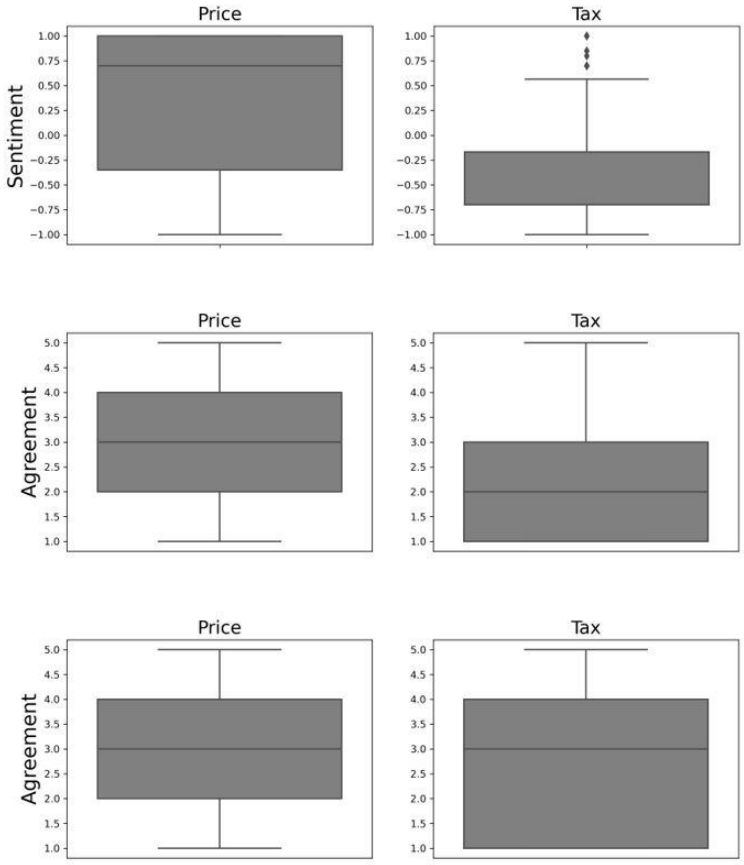


Figure 1. Opinions towards environmental measures expressed on Twitter and in the survey. Top: Twitter sentiments. Middle: All survey respondents. Bottom: Survey respondents with a Twitter account.

Note: Polarization in the Twitter data in terms of sentiment regarding the two environmental measures during the survey period. The sentiments are measured per tweet on a range from −1 for the maximum negative sentiment to +1 for the maximum positive sentiment. Tweets with neutral sentiment are excluded. n (price)=275, n (tax) = 470. Polarization in the survey data based on the question: 'How willing would you be to a) pay higher prices, b) pay higher taxes in order to protect the environment?'; 1=no acceptance and 5=high acceptance. n_{all} *(price)=2512,* n_{all} *(tax)=2511;* $n_{account}$ *(price)=398,* $n_{account}$ *(tax)=397.*

As for the survey respondents (middle and bottom boxplots), opinions regarding paying higher prices to protect the environment are distributed similarly ($median_{all}$ = 3, $mean_{all}$ = 2.90; $median_{account}$ = 3, $mean_{account}$ = 2.98), regardless of whether they have a Twitter account or not. Opinions regarding paying higher taxes, however, tend to be more positive in respondents holding a Twitter account ($median_{account}$ = 3, $mean_{account}$ = 2.55), compared to all survey respondents ($median_{all}$ = 2, $mean_{all}$ = 2.40), as the median is one value higher, and the dispersion grows towards agreement. Overall, the measure 'pay higher prices' receives more support in all datasets than 'pay higher taxes to protect the environment', and the comparison

of survey opinions and Twitter sentiments at the aggregate level shows a relatively good overlap concerning the tendency of opinions. However, on Twitter people can express their opinions using multiple tweets whereas in the survey there is only one answer per respondent and item. Therefore, we also take a look at the individual level and compare the survey responses of those who provided their Twitter accounts with their actual tweets regarding environmental measures.

Considering the match at the individual level, we were able to assess 60 tweets and retweets by 9 survey respondents using the keywords or hashtags 'environment' or 'climate'. The agreement with environmental measures was assigned manually, using the same scale (1 = no acceptance, 5 = high acceptance) as in the survey, and compared in terms of congruence to the opinions towards environmental measures shared by these respondents in the survey. Overall, the results show a relatively high congruence within the two data sources by these 9 people, although no assessment could be made for two respondents' opinions on Twitter due to differing or unclear content (such as using #climate in working atmosphere content).

4. Conclusion

To summarize our findings regarding our first research question, we find divergent results regarding the willingness to share one's social media account depending on the selected social media platform. As for Facebook, opinions regarding environmental measures have a significant effect on the willingness to share one's account information. Respondents who oppose those measures are more likely to refuse sharing their social media accounts. However, this effect cannot be proven for Twitter. These results are in line with our findings regarding COVID-19 measures (Hadler et al. 2022). Hence, we assume that Facebook is potentially a more polarized platform, whose users are more cautious about sharing their data and that the social desirability bias is stronger on Facebook than on Twitter, regardless of the investigated topic. This may be attributable to the fact that different social media platforms follow diverse agendas and have different aims, and users may differ accordingly. For Twitter, we assume, based on our findings, that users tend to be more scientifically minded and therefore agree with current policies as well as with making their data available for research purposes.

Regarding our second research question, we find a relatively high congruence between the opinions shared in the survey and those posted on Twitter, both at the aggregate and the individual level. Our findings differ in this respect from previous studies, such as Pasek et al. (2020), or Amaya et al. (2020), who both showed that social media content concerning political issues on Twitter or Reddit differs from public opinion, i.e., in the European Social Survey. This could be due to the fact that opinions in our data regarding environmental measures are less subject to social desirability bias than other, mainly political, topics.

However, our study faces some limitations. First, social media users are not representative of the overall population in terms of sociodemographics, as they tend to be younger. Second, comparisons between social media data and survey data at the aggregate level always face the bias of an unequal number of responses per individual, as social media users can express their opinions in multiple tweets or postings whereas there is only one answer per item per survey respondent. Finally, our congruence analysis at the individual level is based on a very small number of subjects, as individuals who provide their social media account must also tweet about the particular topic in order to be included in the analysis. Hence, the findings might not be suitable for generalization but can provide a first insight into the field and serve as a basis for further investigations.

References

Al Baghal, T., Sloan, L., Jessop, C., Williams, M. L., & Burnap, P. (2020). Linking Twitter and Survey Data: The Impact of Survey Mode and Demographics on Consent Rates Across Three UK Studies. *Social Science Computer Review* 38 (5), 517-532. 10.1177/0894439319828011.

Amaya, A., Bach, R., Kreuter, F., & Keusch, F. (2020). Measuring the Strength of Attitudes in Social Media Data. In: Hill, C. A., Biemer, P. P., Buskirk, T. D., Japec, L., Kirchner, A., Kolenikov, S., Lyberg, L. E. (eds.) *Big Data Meets Survey Science: A Collection of Innovative Methods*, 163-192. 10.1002/9781118976357.ch5.

Grimm, P. (2010). Social desirability bias. *Wiley international encyclopedia of marketing, Part 2 Marketing Research*. 10.1002/9781444316568.wiem02057.

Larson, R. B. (2019). Controlling social desirability bias. *International Journal of Market Research* 61 (5), 534-547. 10.1177/1470785318805305.

Hadler, M., Klösch, B., Lex, E., & Reiter-Haas, M. (2021). *Polarization in public opinion: Combining social surveys and big data analyses of Twitter* (SUF Edition). 10.11587/OVHKTR, AUSSDA, V1, UNF:6:jPjxWXqS6RVg4uYo3Zplcw== [fileUNF].

Hadler, M., Klösch, B., Reiter-Haas, M., & Lex, E. (2022). Respondents' opinion on COVID-19 and their willingness to provide their social media info in a survey. Paper presented at the *Annual Meeting of the American Sociological Association*, Los Angeles, August 2022.

Pasek, J., McClain, C., Newport, F., & Marken, S. (2020). Who's tweeting about the president? What big survey data can tell us about digital traces? *Social Science Computer Review* 38 (5), 633-650. 10.1177/0894439318822007.

Reiter-Haas, M., Klösch, B., Hadler, M., & Lex, E. (2022). Polarization of Opinions on COVID-19 Measures: Integrating Twitter and Survey Data. *Social Science Computer Review*. 10.1177/08944393221087662.

Sloan, L., Jessop, C., Al Baghal, T., & Williams, M. (2020). Linking survey and Twitter data: informed consent, disclosure, security, and archiving. *Journal of Empirical Research on Human Research Ethics* 15 (1-2), 63-76. 10.1177/1556264619853447.

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15071

Evaluation of the use of influencers for the development of consumer satisfaction in the Baltic consumer goods market

Iveta Linina[1], Velga Vevere[2], Rosita Zvirgzdina[3]
[1]Department commerce Turiba University, Latvia; [2]EKA University of Applied Sciences, Latvia, [3] Department commerce Turiba University, Latvia.

Abstract

A business focused on the consumer and its satisfaction is an important factor in ensuring the company's competitiveness. The process of attracting new customers always involves more money, time and energy. In order for a company to retain existing customers and gain only new ones, one of the main tasks is to know the factors that make them happy. Within the framework of this work, the authors want to study the theoretical foundations of consumer satisfaction, to understand the peculiarities of the development of consumer satisfaction using digital content creators - influencers. The use of influencers is an integral part of today's business development, enabling companies to operate successfully in a competitive environment. This study identifies factors that influence consumer satisfaction with the use of influencers to enable companies to improve their use and become more competitive. The study uses both secondary data analysis and expert interviews and consumer surveys. The study describes the situation in the field of influencer marketing use in the Baltic States. The study finds that influencers provide a higher level of consumer satisfaction. In order to achieve the goal of the research, three tasks were set: 1) to analyse theoretical basis of consumer satisfaction and the use of influencers; 2) describe the use of influence agents and their contribution to consumer relations; 3) to study consumer evaluation of influencer activities. A monographic or descriptive method was used to analyse the theoretical aspects of the use of influencers, an analysis of secondary data was used to describe the situation, and a consumer survey was conducted to examine consumer perceptions of influencer activity and its contribution to consumer satisfaction.

Keywords: *Influencers, consumer, consumer satisfaction.*

1. Introduction

The development of the digital age and the increase in the number of social networking sites have led to changes in consumer behaviour. This transformation has created more and more opportunities and challenges. The growing importance of digital influencers has been recognized by both practitioners and academics. However, given its contemporaneity, the academic literature on the subject faces some limitations. This study looks at digital content creators or influencers. With the development of influencers, their power over brand and company perception has developed significantly, so it can greatly affect both the company's operations and its reputation (Vodas, Novysedlák, Čakanová & Pekár, 2019). These new opportunities for providers need communication professionals who are constantly working with target customers through a variety of social media channels. In turn, consumer satisfaction and its management has become the basis of the company's competitiveness and an integral part of the business. It is important for businesses to ensure and promote consumer satisfaction and to develop a system that makes them want to stay in business. Based on the research, it has to be concluded that attracting a new consumer is 5-10 times more expensive than selling to an existing consumer, and the existing consumer spends 67% more money than new consumers (Anderson, Jolly, Fairhurst, 2007). So working with consumers and building relationships in the long run is an essential foundation for a successful business. By gaining an understanding of the factors that make up consumer satisfaction and using them skilfully using digital influencers, the company gains more customer confidence and significantly increases its competitiveness. Consumer relationship management is the company's business strategy to attract, serve and retain consumers through understanding and meeting their needs, developing long-term cooperation. In order to achieve the goal of the study, three tasks were set:

1. To analyse the theoretical basis of consumer satisfaction and the use of influencers;

2. Describe the use of influence agents and their contribution to consumer relations;

3. To find out the consumer's assessment of the activity of influencers.

A monographic or descriptive method was used to analyse the theoretical aspects of influencer marketing and their use, secondary data analysis was used to describe the situation, and a consumer survey was conducted to study consumer attitudes towards influencers.

The research period is from June 1, 2021 to January 1, 2022. As a result of the study, it was found that the selection of appropriate influencer agents and the development of guidelines for cooperation with them can be used by companies to ensure consumer satisfaction and, as a result, increase competitiveness. After reading the scientific literature, the authors

conclude that there are many conceptual Different definitions of consumer satisfaction for different practical purposes.

2. Literature Review

Attempts to summarize these multifaceted definitions of consumer satisfaction have been made by J.L. Giese and J.A. Cote (2000), summarizing the definitions of consumer satisfaction offered in the scientific literature over thirty years and comparing their wording with respondents' perceptions of the essence of satisfaction. As a result of the research, a three-dimensional framework has been created, the aim of which is not to offer a general definition of satisfaction, but to crystallize the basic components of consumer satisfaction:

- A total emotional reaction, the intensity of which may vary;
- Satisfaction related to product selection, purchase / purchase and consumption;
- Depending on the situation, the time taken to carry out the assessment is limited, with each of the components being adaptable to the specific situation and the range of consumers (Giese, Cote 2000).

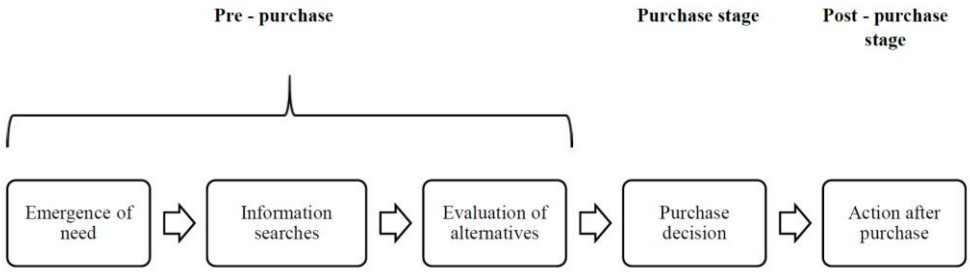


Figure 1. Stages in the process of consumer behaviour in the market. Source: (Elliot & Percy, 2007, 6)

It is in the first stage of pre-purchase that a relationship with the consumer is established, which will be the basis for successful relationship building and satisfaction in general. As there is no direct contact with the consumer, it is important to address it correctly and build a successful communication. Communication is an interactive communication process in which participants realize their goals and interests by influencing each other. Effective communication pays attention not only to the process of information transfer, but also to the full reception and processing of information. The goal of effective communication is to present the message in a way that is not only understandable, but also able to be remembered, analysed and used by the recipient.

Entrepreneurs perceive influencers as an advertising channel with a directly reachable potential audience, so they pay for advertising on their social accounts. Significantly, at a

time when various ad-blocking tools are popular in the digital environment, ads on social networking accounts are able to bypass them because they are unable, at least for the time being, to filter out social networking ads from influencers.

Influencers can be defined as a social networking person who, by creating his or her original content in one of the fields, has gained thousands of followers and advertises the products or services of various companies, institutions and organizations to his audience for a fee, a commission or a product. The need for product experimentation through the exchange of experiences and views of digital influencers is an important tool for building relationships with consumers. Credibility in influencers makes followers communicate with them and trust their views and serves as a means of communication, as they can shape brand messages in their stories, incorporate them into their daily lives in a natural and authentic way and reach a variety of audiences, as shown by several studies (Djafarova & Rushworth, 2017; Piskorski & Brooks, 2017; Veirman et al., 2017). It is the consumer's choice and following the particular influencer that is the basis for building a long-term relationship between the business and the consumer, where the influencer acts as an intermediate in this relationship. Studies also indicate that influencers are given creative freedom to communicate with the consumer (Casaló, Flávian & Ibáñez-Sánchez, 2018). However, industry experts point to the importance of existing guidelines and the approval of content before publication (Piskorski & Brooks, 2017), which will enable companies to use influencers as an important factor in building relationships with consumers that will further ensure consumer satisfaction. Today's consumer has the ability to block corporate advertising in a variety of ways, but influencers address a circle of followers who trust him. It is the best method to gain consumer trust, and the established relationship between the consumer and the influencer is almost impossible for the brand itself (Hall, 2016). Given the scale and speed of the Internet, influencers can quickly attract mass audiences and gain "fame," by accumulating the cultural capital and making a company competitive (McQuarie, Miller, & Phillips 2013).

3. Research and discussion

One of the most pressing issues is how to measure the influencer's outcomes. Quantitative indicators of success (such as the number of Likes) are generally readily available and are mainly used by both stakeholders and companies. However, it is still unclear what the value of these metrics is for influencer marketing and whether they are an appropriate substitute for content quality. This is of particular interest as companies have only limited control over the content published by influencers. In general, professionals consider the reach of the influencer and the number of their interactions to be the most important indicators of success. Contrary to that, when professionals face a trade-off between multiple indicators,

they rely primarily on the mood of user comments as a basis for determining consumer satisfaction (Grave & Greff, 2018).

A study was conducted to further explore consumers' attitudes toward influencer activity. Consumer satisfaction is seen as a phenomenon that looks at consumer and process perception in the context of a particular study, so the study is considered analytical and the research paradigm is positivism (Kumar & Thondikulam, 2005). The current study has a mixed methods approach. First, it is quantitative because it aims to characterize consumers' attitudes towards the influencer activity in the digital environment as a phenomenon based on an assessment of consumers' current supply (Kristapsone, Kamerāde et al. 2011, 49-81).

The population sampling was carried out by the purposeful snowball method (Kristapsone, Kamerāde et al. 2011, 71), using the personal contacts of the study authors, the questionnaire was sent via e-mail to the respondents, who further shared this link. The questionnaires were filled out by 1448 respondents, they all were recognized as valid for the research. In 2020, the population of all three Baltic States was taken as a general population. At the 95% confidence level and the 5% margin of error, the minimum sample size in each country was calculated to be 1155 respondents (Arhipova & Bāliņa, 2006, 98–104). The research questionnaire consisted of closed, open-ended questions, and in order to evaluate the consumer experience, questions with a Likert scale of 5 points were created, where 1 is very bad and 5 is very good. The responses were processed with SPSS software. The characteristics of the respondents are summarized in Table 1.

Table 1. Socio-demographic indicators of the survey respondents.

Nr.	Socio – demographic indicators of respondents		Number of
1.	Gender		
		Women	842
		Men	606
		In Total	1448
2.	Age		
		0-25	205
		26-40	607
		41-55	332
		56-63	101
		64>	3
		In total	1448

Source: Created by the authors.

Respondents use from one to three or more social networks. When it comes to how many content creators they follow, more than half, that is 73%, failed to answer this question. Although 57% of respondents stated that the follow-up was not spontaneous, it was a balanced decision. When asked whether they are more attracted to high-quality photos or video content, the overwhelming majority indicated that 87% referred to video content. Although in the open question about video, respondents indicate that they prefer short, meaningful video clips.

Table 2. Evaluation of Influencers Performance from the Consumer Perspective.

Assessing the performance of influencers from a consumer perspective	**Arithmetic mean**	**Standard Error of mean**	**Median**	**Moda**	**Standard deviation**	**Variation**
Consumer confidence in influencers	4.52	0.04	4	4	0.82	4.00
Assortment of advertised products	3.55	0.05	4	4	0.99	3.00
Quality of inflator content creation	3.28	0.05	4	5	1.09	3.00
Influence of the influencers on the positive attitude towards the product manufacturer	4.15	0.05	5	5	1.07	4.00
Influence of the influencers on the positive evaluation of the product	3.99	0.05	4	4	0.96	2.00

Source: Created by the authors.

As can be seen from the Table 2, consumer confidence in influencers is very high - 4.52 out of 5 possible ($\bar{X}$ = 4.52; Me = 5.00; Mo = 5.00). Respondents rated the range of products offered by influencers much lower (3.55; 4.00; 4.00), while the quality of influencer content was rated even lower (3.28; 4.00; 5.00). Respondents rated the ability of the influencer with the content of its profile towards the product to be high (4.15; 5.00; 5.00), but against the manufacturer of the product only slightly lower (3.99; 4.00; 4.00).

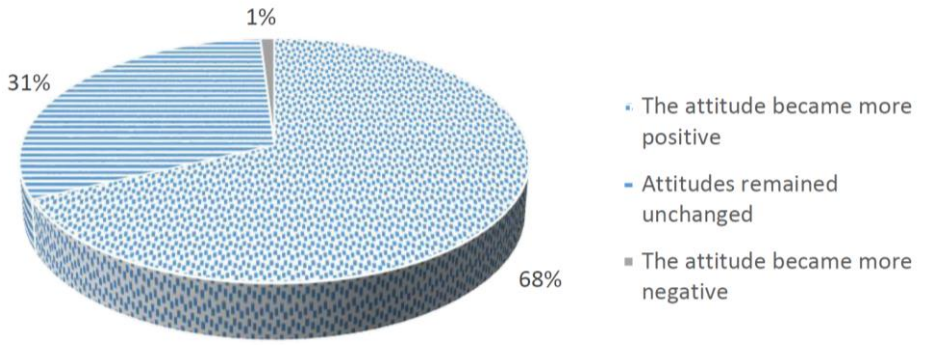


Figure 2. Respondent's opinion on the change of attitude under the influence of influencers. Source: created by the authors.

Respondents were also asked about the impact of influenza on his opinion and how it changed. In this question, it was found out that 68% of respondents' opinion about goods / services as well as about the company itself changes in a positive direction due to inflation, only 31% does not change it, but in the case of 1% it may change to a negative side (see Figure 2).

The study also calculated correlations based on demographic factors.

Table 3. Correlation between respondent's age and influencer's impact assessment.

Evaluation of influencer operation	Age of the respondent
Pearson Corelation	0.870

Source: Created by the authors

It was concluded that the correlation between the respondent's age and more positive evaluations of influencer activity is 0.87, which is very significant (see Table 3). No correlation was observed between the impact assessment and the gender of the respondents.

4. Conclusions

Consumer satisfaction is the result of both the cognitive and emotional response of the consumer, and consumer satisfaction can be seen as a process whose final outcome is influenced by certain factors, a comparison of alternatives, and so on. The process also involves a satisfaction assessment, which involves five stages, which can be summarized in three stages: the pre-purchase stage, the purchase stage and the post-purchase stage. The pre-purchase stage consists of the emergence of a need, the search for information and the evaluation of alternatives. It is at this stage that the relationship with the consumer is

initiated and ensures the further development of this relationship and can be the basis for ensuring consumer awareness and building relationships. Consumer relationship management is the company's business strategy to attract, serve and retain consumers through understanding and meeting their needs, developing long-term cooperation.

In the digital age, influencers are a new type of independent third party that uses a variety of content creation tools to shape consumer attitudes towards business and the brand, using social media as a communication channel. These activities can be very diverse, involving both the expression of opinions, such as product reviews, video tips and tricks, the organization of competitions, and the publication of images containing products or services (Bernitter, Verlegh, & Smit, 2016).

Influencers are one of the key factors in building relationships with consumers and play an important role in the development of the brand and the attitude towards the company as a whole. As a result of the research, it was proved that it is important for companies to work together with influencers to develop guidelines for cooperation in order to target the target audience in a more targeted way.

The study also showed that in 68% of cases, the successful choice of influenza and its effects on consumers can lead to a change in the attitude towards the goods / services and the company itself, which can be the basis for building a relationship with the consumer.

The study found a positive correlation between respondents' age and trust and satisfaction with the brand and the company, which further strengthens the role of influencers in communicating with consumers in the future.

Based on the contribution of influencers to the creation of digital content, which provides an increase in the number of followers, the company can communicate with consumers to ensure trust in the company and its product / service and can build long-term relationships that increase business value and competitiveness.

References

Anderson, Joan L.; Jolly, Laura D.; Fairhurst, Ann E. (2007) Customer relationchip management in retailing: A content analysis of retail trade journals, Journal of Retailing and consummer services 14, no.6.

Arhipova, I., Bāliņa, S. (2006). Statistika ekonomikā un biznesā. Datorzinību centrs, Rīga, 98. –233.lpp.

Casaló, L. V., Flávian, C., & Ibáñez-Sánchez, S. (2018). Influencers on Instagram: Antecedents and consequences of opinion leadership. Journal of Business Research, 117, 510-519. https://doi.org/10.1016/j.jbusres.2018.07.005

Djafarova, E., & Rushworth, C. (2017). Exploring the credibility of online celebrities' Instagram profiles in influencing the purchase decisions of young female users. Computers in Human Behaviour, 68, 1-7.

Elliot, R., Percy, L. (2007). Strategic Brand Management. Bath Press, p. 6.

Evans N. J., Phua J., Lim J. & Jun H. (2017). Disclosing Instagram Influencer Advertising: The Effects of Disclosure Language on Advertising Recognition, Attitudes, and Behavioral Intent. Retrieved 09.11.2021 https://web.a.ebscohost.com/bsi/detail/detail?vid=2&sid=2cf32453-5587-4bb2-878c-86d266a4aab3%40sessionmgr4006&bdata=JnNpdGU9YnNpLWxpdmU%3d#AN=127728010&db=bsu

Giese J.L., Cote J.A. (2000) Defining Consumer Satisfaction// Academy of Marketing Science Review, Vol.No.1, p.15.

Grave, J.F.; Greff, A.(2018). Good KPI, Good Influencer? Evaluating Success Metrics for Social Media Influencers. SMSOCIETY'18: Proceedings Of The 9th International Conference On Social Media And Society, Page 291-295, DOI: 10.1145/3217804.3217931

Hall, J. (2016). 'The Influencer Marketing Gold Rush Is Coming: Are You Prepared?' http://www.forbes.com/sites/johnhall/2016/04/17/the-influencer-marketing-gold-rushiscoming-are-you-prepared/ #26a8f05f2964

Kristapsone, S., Kamerāde, D.u.c. (2011). Ievads pētniecībā: stratēģijas, dizaini, metode. RaKa, Rīga, 284 lpp.

Kumar, S., Thondikulam, G. (2005). Knowledge management in a collaborative business network. Information Knowledge Systems Management, 5, pp. 171-187.

McQuarie, E. F., Miller, J., and Phillips, B.,J. (2013). 'The Megaphone Effect: Taste and Audience in Fashion Blogging.' Journal of Consumer Research 40 (1): 136–58.

Piskorski, M., & Brooks, G. (2017). Online broadcasters: How do they maintain influence, when audiences know they are paid to influence. Proceedings of the 2017 Winter AMA, 28, D70- D80.

Veirman, M., Cauberghe, V., & Hudders, L. (2017). Measuring through Instagram influencers: The impact of number of followers and product divergence on brand attitude. International Journal of Advertising, 36(5), 798-828.

Vodák, J., Novysedlák, M., Čakanová, L., Pekár, M. (2019). Influencer Marketing as a Modern Phenomenon in Reputation Management. Retrieved 01.03.2022. https://www.hippocampus.si/ISSN/1854-6935/17.211-220.pdf

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15074

Collaborate for what: a structural topic model analysis on CDP data

Camilla Salvatore[1], Alice Madonna[2], Annamaria Bianchi[3], Albachiara Boffelli[2], Matteo Kalchschmidt[2]

[1]Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy, [2]Department of Management, Information and Production Engineering, University of Bergamo, Bergamo, Italy, [3]Department of Management, Economics and Quantitative Methods, University of Bergamo, Bergamo, Italy

Abstract

This paper aims to understand why firms engage with their suppliers to collaborate for sustainability. For this purpose, we use the Carbon Disclosure Project (CDP) Supply Chain dataset and apply the Structural Topic Model to: 1) identify the topics discussed in an open-ended question related to climate-related supplier engagement and, 2) estimate the differences in the discussion of such topics between CDP members and non-members, respectively focal firms and first-tier suppliers. The analysis highlights that the two prominent reasons why firms engage with their suppliers relate to several aspects of the supply chain management, and the services and good transportation efficiency. It is further noted that first-tier suppliers do not possess established capabilities and, therefore, are still improving their processes. On the contrary, focal firms have more structured capabilities so to manage supplier engagement for information collection. This study demonstrates how big data and machine learning methods can be applied to analyse unstructured textual data from traditional surveys.

Keywords: *sustainable supply chain management; carbon disclosure project; supplier collaboration; structural topic model; text mining*

1. Introduction

In recent years, environmental disclosure programs, in which firms communicate how they manage their impact on climate change, are gaining more and more traction. While at the beginning of the 2010s, these programs were deemed to provide a competitive advantage, today, they are almost mandatory in a supplier selection procedure (Serafeim, 2020). Previous studies have utilised data from these programs to understand their impact on the firm's performances (Madonna, Boffelli and Kalchschmidt, 2021), but they have failed to understand the reasoning behind different behaviours. Particularly, distinguishing between different tiers along the SC is crucial, as their approach to sustainability could have happened in different time frames and for different reasons (Schmidt et al., 2017). Thus, this study aims to fill this gap by trying to answer the following research question:

"Why do firms collaborate for sustainability along their supply chain?"

Taking on this goal, the Carbon Disclosure Project (CDP) Supply Chain (SC) dataset has been considered suitable due to the depth of information provided and the availability for the respondent to describe the engagement strategies through open-ended questions (CDP, 2018). The availability of open-ended survey questions allows us to deeply investigate the behaviour of businesses with respect to standard closed-ended questions. However, texts are unstructured data and Machine Learning (ML) approaches are fundamental to extracting information from such data. To this purpose, we apply the Structural Topic Model (STM) technique, which allows us to discover the latent topics discussed and to estimate the effect of relevant metadata (being a Member of CDP) on the discussion proportion of topics. The main reasons for joining CDP concern enhancing the firm's image and reputation and receiving insights into one's suppliers. The data are gathered thanks to CDP members who request their suppliers to fill in a questionnaire to report information about climate change management, after filling in the questionnaire themselves. We expect the comments to highlight a divergence in the reasoning behind the engagement from the firms that, leveraging the data collection procedure, we can allocate into different tiers in the supply chain. In particular, we assign to CDP Members the role of *focal firms*[1] and the Non-Members the role of *first-tier suppliers*.

The novelty of the work is to be found firstly in the methodology, which is approached in the field of Sustainable Supply Chain Management (SSCM) for the first time. Indeed, in the field of sustainability, STM has been applied only to study open-ended questions about climate change (Tvinnereim, & Fløttum, 2015) and CSR disclosure in tweets (Salvatore, Biffignandi & Bianchi 2020). The second novelty introduced by this work relies on the analysis's

[1] Focal firms are those firms considered the leaders and the power fulcrum of their supply chain. The distinction between focal firms and first-tier suppliers has been done by leveraging the data collection procedure.

perspective. The "business-as-usual" of SSCM research is to observe how business decisions impact firms' performance. This study has taken on the challenge to invert the viewpoint, considering that sustainability actions are required, mandatory at times (Serafeim, 2020), even though they do not necessarily influence firms' performances (Pagell & Shevchenko, 2014).

The remainder of this article is organised as follows. Section 2 introduces the model. In Section 3, the data and the model selection strategy are presented. The results are discussed in Section 4. The main conclusions are drawn in Section 5.

2. The Structural Topic Model (STM)

Topic modelling (TM) is an unsupervised learning technique that allows studying the underlying properties of a text to discover the topics discussed and get signals from the data (Vayansky & Kumar, 2020). Among the different algorithms to implement TM, we select the STM, which was originally designed to analyse open-ended survey questions, and which is becoming increasingly popular due to the possibility of estimating models including document-level metadata, thus characterising the relationship between topics and metadata (Roberts, Stewart, & Airoldi, 2016).

In the following, we briefly introduce the STM algorithm. Please refer to Roberts, Stewart, & Airoldi (2016) for more details. STM is based on the bag of words assumption, which means that each document is represented as a vector of words without considering the order in which they appear. A topic is defined as a mixture of words, and a document as a mixture of topics. In STM, document-metadata influences two components of the model, the topical prevalence, which is defined as the proportion of the document associated with a topic, and the topical content, which refers to the usage rate of a word in a topic. Thus, topical prevalence covariates affect the discussion proportion of the topic (θ), while topical content covariates affect the rate of word usage within a topic (β). Here we focus only on topical prevalence covariates. The model can be represented in plate notation as in Figure 1.

The first step in estimating the model is to specify the algorithm initialisation strategy and the number of topics. Usually, the output is very sensitive to the specified initialisation. In this respect, the suggestion is to use spectral initialisation, a deterministic algorithm based on the method of moments, due to its stability (Roberts, Stewart, & Tingley, 2019). With respect to the specification of the number of topics, it is worth noticing that there is no true number of topics, and the suggestion is to test different numbers of topics by comparing some metrics and manually evaluating the results. Roberts, Stewart, & Tingley (2019) argue that four metrics should be compared: held-out likelihood, residual dispersion, semantic coherence and exclusivity. The held-out likelihood is a measure of the predictive power. The higher the held-out likelihood, the higher the model's predictive power. Residual dispersion is equal to

one when the model is well specified. This is a very strict requirement, and for practical purposes, the analyst should prefer models with low residuals and evaluate residuals in combination with the other metrics. Semantic coherence measures the co-appearance rate of the most probable words in that topic, so the higher this metric is, the better a topic is defined. However, semantic coherence decreases as the number of topics increases, i.e., if the number of topics is small, it is likely that they will be composed of the same words. Thus, practitioners should also look at exclusivity, which measures whether the top words for that topic do not appear as top words in other topics (exclusivity of words to a topic).

After the initialization and the number of topics is specified, model estimation and inference are based on an appropriate variational E-M algorithm, which returns as output the discussion proportion of the topics for each document, the rate of word usage within each topic, and the effect of covariates on the topical prevalence and topical content.

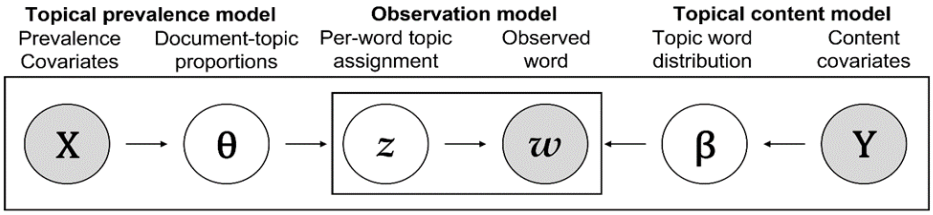


Figure. 1. Structural Topic Model. Source: Amended from Roberts et al. (2016).

For our analyses, we use R and, in particular, the `STM` package (Roberts et al., 2019) to estimate the model and the `quanteda` package (Benoit et al., 2018) to clean and prepare the data.

3. Data and model selection

The data we analyse are part of the SC dataset of the 2018 CDP questionnaire (CDP, 2018). The Carbon Disclosure Project is a non-profit organisation which encourages firms to disclose information about their climate-change-related risks and opportunities through a yearly survey. Figure 2 shows the sample refinement process, which lead to a final dataset of 314 firms. Each respondent could comment on different types of the deployed engagement strategies, namely Compliance & onboarding, Information collection, Engagement & incentivisation, Innovation & collaboration, and others. Thus, 461 short comments on the different rationale for why the 314 respondents engage their suppliers.

Before analysing the comments, it is necessary to clean the data. This involves different operations aiming at keeping only relevant words, reducing the complexity of the model and

speeding up the estimation process. In this respect, the steps we implemented are: elimination of punctuation, stop words and numbers, conversion to lowercase, and stemming. An additional step in data cleaning is the removal of infrequent terms. This allows reducing the noise in the data, making topic detection easier. The rule of thumb is to remove the terms that appear in less than 0.5-1% of the documents (Denny & Spirling, 2018). We fixed the threshold to 1%, and the final set of unique stems is composed of 728 units. These data are ready to be analysed.

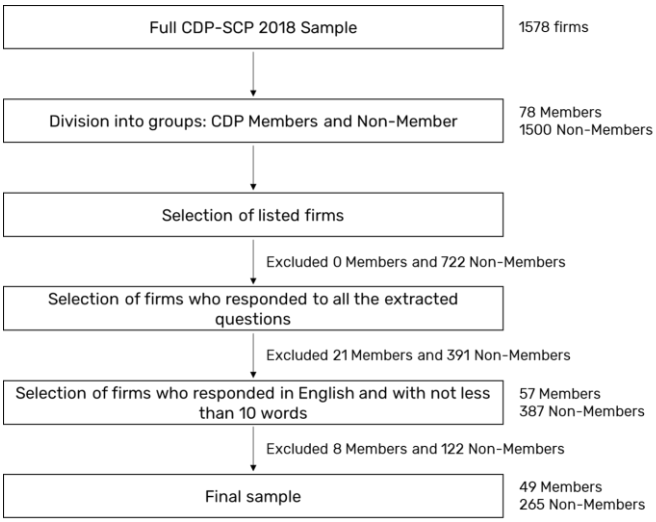


Figure 2. Sample refinement procedure.

We consider only one topic prevalence covariate: being a member of the CDP. We study the effect of this variable on the proportion of discussion of topics.

The optimal number of topics is identified by looking at the metrics described in Section 2 and represented in Figure 3. Although it is not possible to identify the *true* number of topics, this procedure helps identify a set of plausible values. The appropriate number of topics seems to be around 20 and 30, where residuals are relatively low. After a manual evaluation of the quality of topics, we selected the model with 20 topics.

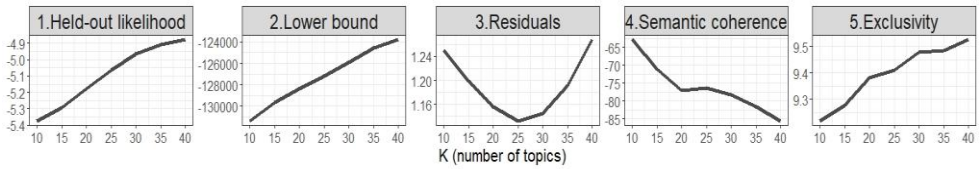


Figure 3. Evaluation metrics for choosing the number of topics

4. Results and discussion

Topics are identified by looking at the most probable stems for each topic, and labelling them consequently. Figure 4 shows the proportion of identified topics classified by macro-dimension (left panel) and how topics are correlated (right panel).

The most prevalent dimension (30% of the topics) relates to Supply Chain Management (SCM) issues, particularly suppliers' management (Topic 10, 13, 17), control (T5, 6) and accountability (T4). The second one (20%) is about Services and Materials Transportation (SMT), which addresses transport optimisation (T12), outsourcing of services (T9, 11), and transportation of sold goods (T3). The remaining dimensions relate to the measurement of GHG emissions and more globally carbon footprints (MS – T7, 8, 19), compliance with different standards (COMP – T1, 16, 18), the use of data to make informed decisions (DDE - T14, 15), and finally, activities to promote sustainability (PS – T2, 20). Figure 4 (right panel) clearly shows that these macro-topics are not independent of each other.

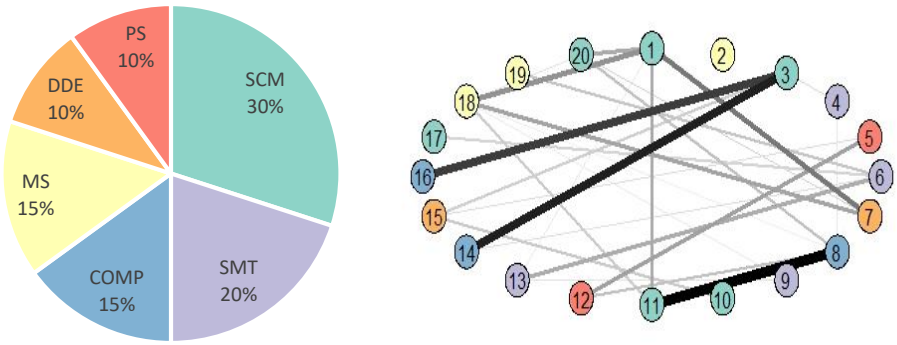


Figure 4. Left panel: Proportion of topics by macro-dimension (SCM: Supply Chain Management - green; COMP: Compliance - blue; SMT: Services and Materials Transportation - purple; MS: Measuring Sustainability - yellow; DDE: Data-Driven Evaluations – orange; PS: Promoting Sustainability - red. Right panel: Correlation plot.

Looking at the effects of being a CDP member on the discussion proportion of the topics, we estimate the changes in topic proportions shifting from firms that are CDP members and firms that are not. It turns out that differences are significant for 7 topics out of 20, as represented in Figure 5. In particular, topics 9, 11 and 12 that are prevalent for Non-Members refer to outsourcing services and optimising the firm's processes, all topics that concern an active process. Instead, Members are characterised by topics that refer to the management of other SC actors and the measurement of different parameters (topics 15, 18, 2 and 4).

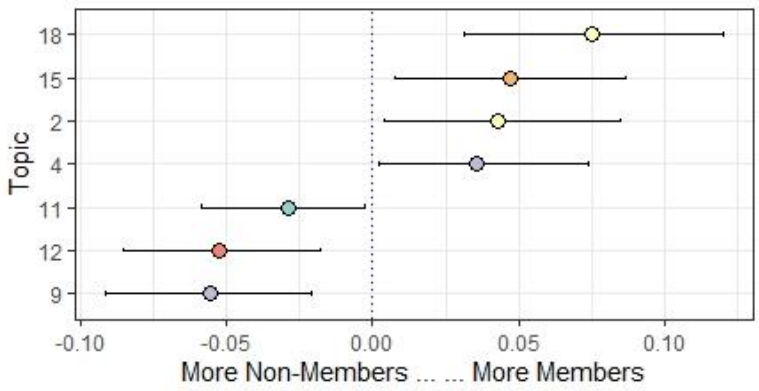


Figure 5. Estimated topic proportion difference between CDP members and non-members with 95% confidence interval.

5. Conclusions

In this work, we propose to apply the STM model for analysing one open-ended survey question about the rationale of engaging with suppliers to pursue sustainability-related goals along the supply chain. From a methodological point of view, although we do not consider a big data source, our research shows how machine learning approaches can be applied to unstructured textual data from traditional surveys to study socio-economic matters.

From a substantive point of view, implications of this work concern the enlightenment of the goals divergence for CDP members, namely focal firms, and non-members, namely first-tier suppliers, when it comes to collaborating along the supply chain for sustainability. This result supports the initial hypothesis to allocate the firms by membership into different tiers of the supply chain (focal firms and first-tier suppliers), as it supports the theoretical characteristics that belong to each category. For instance, focal firms have been classified as first movers toward the transition to sustainability, and therefore they have established resources and capabilities to confront the relevant stakeholders (Schmidt et al., 2017). On the contrary, first-tier suppliers are late entrants into the sustainability movement and, therefore, are still adapting their operations.

Future developments of the work foresee the development of an econometric model wherein the topic model will try to estimate how the discussion is reflected in the firm's value and performance. The ultimate goal will be to provide managerial implications on how these environmental disclosure programs results are perceived and acknowledged over time by external parties.

References

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., & Müller, S. M. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30), 774. doi: doi: 10.21105/joss.00774

CDP. (2018). *Closing the Gap: Scaling up sustainable supply chains*. Retrieved from https://www.cdp.net/en/research/global-reports/global-supply-chain-report-2018/

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26(2), 168-189.

Elijido-Ten, E. O. (2017). Does recognition of climate change related risks and opportunities determine sustainability performance? *Journal of Cleaner Production, 141*, 956-966. doi:10.1016/j.jclepro.2016.09.136

Madonna, A., Boffelli, A., & Kalchschmidt, M. (2021). The role of sustainable supply chain management on improving environmental performance: a longitudinal analysis of CDP data. In *Proceedings of 52nd Annual Conference: Decision Sciences Institute,* ISBN: 978-0-578-62648-2, ISSN: 2471-884X,.

Pagell, M., & Shevchenko, A. (2014). Why Research in Sustainable Supply Chain Management Should Have no Future. *Journal of Supply Chain Management, 50*(1), 44-55. doi:10.1111/jscm.12037

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association*, 111, 988-1003.

Roberts, M., Stewart, B., & Tingley, D. (2019). Stm: An R package for structural topic models. *Journal of Statistical Software*, 91(1), 1-40.

Salvatore, C., Bianchi, A., & Biffignandi, S. (2020). Communicating Corporate Social Responsibility through Twitter: a topic model analysis on selected companies. In *CARMA 2020: 3rd International Conference on Advanced Research Methods and Analytics*, pp 269-277

Schmidt, C. G., Foerstl, K., & Schaltenbrand, B. (2017). The Supply Chain Position Paradox: Green Practices and Firm Performance. *Journal of Supply Chain Management,* 53(1), 3-25. doi:10.1111/jscm.12113

Serafeim, G. (2020). Social-impact efforts that create real value. *Harvard Business Review, 98*(5), 37-48.

Tvinnereim, E., & Fløttum, K. (2015). Explaining topic prevalence in answers to open-ended survey questions about climate change. Nature Climate Change, 5(8), 744-747

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582.

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15076

Text mining methods for innovation studies: limits and future perspectives

Pietro Cruciata, Davide Pulizzotto, Catherine Beaudry
Polytechnique Montréal, Canada.

Abstract

This study offers alternative and promising approaches to word count methods, largely used to develop innovation indicators from unstructured text. We propose a method based on Information Retrieval (IR) and word-embedding models to tackle the semantic ellipsis, one of the main issues of word count methods. We test our IR model by investigating the concept of collaboration and comparing our approach with a baseline corresponding to the keyword search. To ensure the best performances, we use several ways to represent queries and documents in a vector space and three pre-trained word-embedding models. The results prove that our approach can alleviate the semantic ellipsis problem. Indeed, the IR model developed outperforms the classical keyword search in terms of F1-score and Recall. Moreover, we create a combined method that achieves the highest F1-score. These preliminary results can facilitate the creation of reliable innovation indicators from unstructured textual data substituting or complementing survey-based questionnaires.

Keywords: *Text mining, Natural Language Processing, Information Retrieval, innovation measures.*

1. Introduction

The interest to analyze and comprehend innovation dynamics increased since technological progress became the main driver of economic growth, augmenting the need for researchers of public databases and questionnaire-based surveys as a source of data for their quantitative studies. Nevertheless, these sources of information have many weaknesses. Public databases are often incomplete or not specific whereas questionnaire-based surveys (especially large-scale as the biennial European CIS or the annual MIP) lack regional granularity, coverage, timeliness, and furthermore, they are costly (Axenbeck & Breithaupt, 2021). For all of these reasons traditional innovation indicators rarely provide the full picture (Kinne & Lenz, 2021)

Alternative or complementary to these sources is web-based unstructured textual data. Indeed, the rising amount of data available in the form of digitalized text opened up new possibilities for researchers. Although it seemed difficult to measure "signals" of innovation dynamics through corporate websites or other web sources, researchers on innovation and technology management have obtained good results by building new indicators with large amounts of texts. For example, Arora et al. (2013) built five descriptive variables by analyzing a sample of small and medium-sized high technology graphene firms in the US, UK, and China based on keyword analysis of their webpages. Gök et .al. (2015) created web indicators of R&D activities by extracting the keywords from companies' websites. Their study proved that R&D activities captured through the web indicators were significantly more numerous, compared to the R&D activities documented in the other sources. Libaers, et al. (2016) harnessed the data of the companies' websites to develop a taxonomy that identified strategies used by small firms to commercialize their innovations. They analyzed the content of firms' websites to extract the keywords related to possible strategies used by companies. Blazquez & Domenech (2018) used web-based variables built with keywords to predict firm export orientation. Héroux-Vaillancourt et al. (2020) built innovation indicators of four core concepts (R&D, IP protection, collaboration and external financing) from the complete texts of 79 corporate websites of Canadian nanotechnology and advanced materials firms using keywords frequency analysis.

As highlighted in the few examples above, most of the indicators created on textual data are based on keyword search and keyword frequency with several weighting schemes, such as TF-IDF. However, these indicators have two important drawbacks when it comes to analyzing concepts in texts: polysemy of words, a semantic phenomenon that illustrates the relationship between one word and multiple meanings (e.g., river "bank" vs. "bank" as a financial institution) and semantic ellipsis, which refers to the emerging of concepts in sentences where standard words referring to it have been omitted (e.g., the concept of "collaboration" can be expressed by using a combination of words such as "joint venture",

"work with", "join forces" etc.). There are two tasks in Natural Language Processing (NLP) that apply advanced methodologies to solve these two issues. The first task is Word Sense Disambiguation (WSD) which refers to the problem of determining what is the word's meaning in a particular context. The other task is Information Retrieval (IR) whose goal is to search through documents to retrieve the best answer to a query.

In our research, we tackle the problem of the semantic ellipsis by developing a word-embedding-based IR model. Considering that collaboration is one of the main innovation indicators, we chose this concept to test our approach.

To the best of our knowledge, this is the first work that uses pretrained neural networks to preprocess unstructured text for an unsupervised approach in innovation studies. Moreover, this research provides evidence that our IR model alleviates the semantic ellipsis increasing the chances to find the collaboration concept

2. Experiments

2.1. Data

We use WordNet lexical database (Miller et al., 1990) to select the words referring to the concept of "collaboration" (Table 1). We limit ourselves to the least ambiguous words to reduce the risk of noise that keyword searching would cause. Then, to evaluate our approach we build a test dataset by partially labeling SemCor (Miller & Charles, 1991), a dataset manually annotated with the synsets from WordNet.

Table 1. List of words chosen.

• Consortium	• Alliance	• Collaborator
• Partnership	• Collaborate	• Cooperative
• Cooperation	• Cooperate	• Collaborative

2.2 IR model

An IR model comprises three key components: a query, a target corpus and an IR System. In our case, the query is represented by the list of words referring to collaboration (Table 1), the target corpus is our test dataset and the IR system is based on a cosine similarity computation between the word-embedding vectors representation of the query and the target corpus. Finally, we evaluate the performances of our IR model with the F1-score.

The key elements in our IR approach are the pre-trained word-embeddings, which aim to model the proprieties of a language in a vector space. These techniques leverage neural networks to learn the vector representation of millions of words according to the context in

which they appear (Mikolov et al., 2013). The representation of words in dense vector enables the computation of semantically related words and can be used to represent phrases and short texts, reducing the sparsity of traditional vector-space representations (Pelevina et al., 2017). During the years, several pre-trained embedding models were created. We compare the results using GloVe (Pennington et al., 2014) FastText (Bojanowski et al., 2017), and GoogleNews (Mikolov et al., 2013). The three pre-trained models differ in some aspects related to the model architecture and the training corpora used. GoogleNews was the first model to be able to compute high dimensional word vectors from large corpora due to its lower computational complexity. The model is trained on the 100 billion words of the Google news dataset. The backbone is the Skip-gram neural network that predicts the word context giving the word itself (Mikolov et al., 2013). On the other hand, GloVe trained on 840 billion tokens of Common Crawl represents a development on the previous model developed by Mikolov et al. (2013). The model leverages statistical information by training only on the nonzero elements of a word-word co-occurrence matrix (Pennington et al., 2014). Finally, FastText, trained on 600 billion Common Crowl words, improves the GoogleNews algorithm due to its capacity to represent the vector embedding of unseen words as the sum of the vector representations of its n-grams characters.

In the next two sections, we present our results firstly by comparing different settings of the IR model and secondly, by comparing the best IR model with a keyword search method – which is the baseline of this research.

2.3 IR models comparison

To achieve the best performances, we use different parameters for our IR model's three core components generating 24 different settings (with the combination of two queries, three word-embedding models, and four different corpus representations.) For the query, we built the first by transforming each word from Table 1 into a dense vector (we refer to it as W, e.g, GoogleNewsW). From the same list of dense vectors, we then created the second query vector with the arithmetic average (we refer to this one as "Semantic Field" and identify it with the acronym SF e.g., GoogleNewsSF). We justify the last query with the assumption that the SF vector represents the whole semantic field of the concept of "collaboration" which can thus mitigate the problems related to semantic ellipsis. For the target corpus, we execute a preprocessing step of the documents (morphological analysis, lemmatization, removing stop words, etc.), and then we divided each sentence of the target corpus using n-grams, i.e., contiguous word sequences in a document. Indeed, for our target corpus, we tried four different n-grams representations: 1-gram, 2-gram, 3-gram, and 4-gram to find the best setting in our IR model. Finally, for the IR system, we use the three different pre-trained word-embedding models mentioned in section 2.2. Therefore, our IR model works as follows: the IR system takes one word at a time from our list of words as the query, transform it into a dense vector, and searches through the different sentences of

the target corpus divided in n-grams, which in turn are transformed into dense vectors, to find the most similar. The similarity is measured by computing the cosine similarity between two dense vectors representing the target corpus and the query. Once the most similar n-gram dense vector is found, the whole sentence is returned as a sentence where the concept of "collaboration" emerges. Finally, since the target corpus is partially annotated, we measure the final results of the IR model by computing the F1-score of the several cosine thresholds to sort out the best ones.

Table 2 shows the F1-score of our several IR models. First, we notice that IR models with W query setting outperform IR models with SF query setting, except for GloVe^{SF} which has a higher F1-score compared to GloVe^{W}. Moreover, we can observe that IR GoogleNews^{w} models get the best performances. In particular, the best model is GoogleNews^{W} with a 3-gram setting, which obtains an F1-score of 0.8635. For the SF query setting, the best model is GoogleNews^{SF} with the 2-gram sequence representation that reaches an F1-score of 0.7926 – which is far less performant than the GoogleNews^{W} model. Additionally, we notice that the performances of the IR models decrease with the 4-gram sequence representation of the target corpus, except for GloVe^{W} which probably requires a longer sequence representation to perform better. It is important to highlight that FastText^{SF} and FastText^{W} perform better in a 1-gram setting, probably due to the subword representation typical to FastText.

Table 2. F1-score of the IR models. The superscript SF indicates that the query setting is the Semantic Field while W indicates the single word; the subscripts indicate the cosine similarity threshold

	1-gram	**2-gram**	**3-gram**	**4-gram**
GoogleNews^{W}	$0.5748_{0.545}$	$0.8540_{0.55}$	$\mathbf{0.8635}_{0.525}$	$0.8385_{0.52}$
Glove^{W}	$0.1835_{0.55}$	$0.1963_{0.55}$	$0.2207_{0.55}$	$0.2375_{0.55}$
FastText^{W}	$\mathbf{0.7532}_{0.55}$	$0.6813_{0.55}$	$0.6048_{0.54}$	$0.4979_{0.53}$
GoogleNews^{SF}	$0.56_{0.51}$	$\mathbf{0.7926}_{0.51}$	$0.7063_{0.50}$	$0.5981_{0.51}$
Glove^{SF}	$0.2870_{0.55}$	$0.3395_{0.55}$	$0.3277_{0.55}$	$0.3040_{0.55}$
FastText^{SF}	$\mathbf{0.6667}_{0.54}$	$0.6064_{0.53}$	$0.45_{0.54}$	$0.3596_{0.54}$

2.4 IR models comparison with baseline

In light of the previous findings, we select the best IR model settings to compare with the method used as the baseline for this study: the keyword search. To ensure that our baseline reaches the best F1-score we apply several pre-processing steps (e.g., lemmatization, lower case, etc.). Additionally, we test combined approaches of keyword search and IR GoogleNews models. Table 3 presents the results of the best models including the measures of Precision and Recall.

Table 3 shows that GoogleNews^{w} with the 3-gram sequence representation outperforms the keyword search in terms of F1-score and Recall. Despite the higher precision of the keyword search, this IR model has a higher F1-score due to its higher number of sentences retrieved (thus, a higher Recall) in which the "collaboration" concept emerges. On the other hand, GoogleNews^{SF} with the 2-gram sequence representation performs worse than the baseline. Among all the methods, the combination keyword search-GoogleNews^{W} with the 4-gram sequence representation yields the highest F1-score in our study. Although this combined method has less Recall than the best GoogleNews^{W} model, it has almost the same precision as the keyword search. Finally, testing the combination of the GoogleNews^{W} with GoogleNews^{SF} models (see Table 3) proved to be the best method to achieve the highest Recall. This result proves that the SF setting, despite its lower performance, significantly contributes to the improved results. In other words, the SF setting and the W setting should be used together since they retrieve different sentences in which the concept emerges.

Table 3. Comparison between IR models, keyword search and combined methods. The superscript SF indicates that the query is the Semantic Field while W indicates that is the single word. The subscripts indicate the cosine similarity threshold

	F1	Recall	Precision
$GoogleNews^{W}_{0.53}$**4-gram +**$GoogleNews^{SF}_{0.535}$**2-gram**	0.8571	**0.8483**	0.8662
$GoogleNews^{W}_{0.53}$ **4-gram+Keyword**	**0.8736**	0.7862	0.9827
$GoogleNews^{SF}_{0.535}$ **2-gram+ Keyword**	0.8560	**0.7793**	0.9495
Keyword search	**0.8412**	0.7310	0.9965
$GoogleNews^{W}_{0.545}$ **3-gram**	0.8635	0.8069	0.9286
$GoogleNews^{SF}_{0.51}$ **2-gram**	0.7925	0.7379	0.856

3. Conclusion

This research provides a solution to alleviate one of the main issues of the widespread word count methodologies, namely semantic ellipsis. To do so, we compare our baseline, a simple keyword search, with two different methods: an IR model and a combination of IR model with the keyword search. The results show that the IR models mitigate the semantic ellipsis problem outperforming the baseline. However, the combination of the IR models with the keyword search reaches the best performances showing a certain level of complementarity. In particular, the combination of GoogleNews^{w} and GoogleNews^{SF} gets the highest complementarity, thus improving its precision could lead to reach the best performances. Finally, our results suggest that using the methods developed represent an alternative to build stronger and reliable innovation indicators from unstructured text.

Nevertheless, the method developed suffers from two main limitations stemming from the pre-trained models. The first is the impossibility to disambiguate because they conflate all meanings of a word into a single vector (Pelevina et al., 2017). The second limitation is due to the domain sensitivity of the word embedding training corpus that reduce its generalisation.

For our future research directions, we plan to explore four different approaches to improve the results of this study. The first will be to integrate an advanced WSD approach to our combined method. As mentioned above, improving the precision of the IR models will lead to improve their performances. Previous researchers have already used and documented similar combinations of approaches to achieve greater results. For instance, (Rothe & Schütze, 2015) combined word embeddings based on WordNet synsets to obtain sense embeddings, whereas Pina & Johansson (2016) applied random walks on the Swedish Wordnet to generate training data for the Skip-gram model. The second approach will be to use word embedding created with state-of-the-art NLP model as Bert for their capacity to disambiguate words. Indeed, they give to the same word different vector representation based on the context solving one main issue of the pre-trained word-embedding models. The third approach will be to train a supervised model on a manually labeled dataset. This approach united with word embedding could leverage the capacity of the advanced NPL models to disambiguate the words' meanings. Finally, the fourth approach will be to create our word-embedding model based on a collection of documents on innovation studies. We believe that this pre-trained model could have the capacity to learn different concepts and words related to innovation facilitating the creation of new indicators.

References

Arora, S. K., Youtie, J., Shapira, P., Gao, L., & Ma, T. (2013). Entry strategies in an emerging technology: A pilot web-based study of graphene firms. *Scientometrics*, *95*(3), 1189–1207.

Axenbeck, J., & Breithaupt, P. (2021). Innovation indicators based on firm websites—Which website characteristics predict firm-level innovation activity? *PLOS ONE*, *16*(4), e0249583. https://doi.org/10.1371/journal.pone.0249583

Blazquez, D., & Domenech, J. (2018). Web data mining for monitoring business export orientation. *Technological and Economic Development of Economy*, *24*(2), 406–428.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, *102*(1), 653–671.

Héroux-Vaillancourt, M., Beaudry, C., & Rietsch, C. (2020). Using web content analysis to create innovation indicators—What do we really measure? *Quantitative Science Studies*, *1*(4), 1601–1637.

Kinne, J., & Lenz, D. (2021). Predicting innovative firms using web mining and deep learning. *PloS One*, *16*(4), e0249071.

Libaers, D., Hicks, D., & Porter, A. L. (2016). A taxonomy of small firm technology commercialization. *Industrial and Corporate Change*, *25*(3), 371–405.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, *3*(4), 235–244.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*(1), 1–28.

Pelevina, M., Arefyev, N., Biemann, C., & Panchenko, A. (2017). Making sense of word embeddings. *ArXiv Preprint ArXiv:1708.03390*.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Pina, L. N., & Johansson, R. (2016). Embedding senses for efficient graph-based word sense disambiguation. *Proceedings of TextGraphs-10: The Workshop on Graph-Based Methods for Natural Language Processing*, 1–5.

Rothe, S., & Schütze, H. (2015). Autoextend: Extending word embeddings to embeddings for synsets and lexemes. *ArXiv Preprint ArXiv:1507.01127*.

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15080

The demand side of information provision: Using multivariate time series clustering to construct multinational uncertainty proxies

Florian Schütze
Department of Socioeconomics, University of Hamburg, Germany

Abstract

Information demand in the modern world is met to a huge extent by information supply from the search engine Google. Humans use the search engine to gather information which shall help to reduce perceived personal uncertainty about a specific subject. Google Trends is providing insights into this information demand in a timely manner and for a variety of different countries. In this paper, multinational Google Trends data and unsupervised learning techniques are used to construct meaningful country clusters resembling the economic, geographic and political relationships of the considered countries. Additionally, these clusters are stable over time. Under the assumption that an increase in Google search requests reflect elevated uncertainty, the cluster information is used to construct economic and political uncertainty time series for 43 different countries. This uncertainty index Granger causes quarterly GDP growth in more countries compared to an existing multinational uncertainty index proofing its usefulness in the field of forecasting. Furthermore, the new index is available up to a daily frequency and can be applied to additional countries and regions.

Keywords: *Google Trends; Economic Uncertainty; Unsupervised machine learning; Forecasting; Clustering.*

1. Introduction

Economic and political uncertainty can be inferred in diverse ways. For example, by measuring the volatility of macroeconomic variables (Bloom, 2009; Jurado et al., 2015), the dispersion among forecasters (Bachmann et al., 2013) or counting the occurrence of uncertainty related keywords, like the Economic Policy Uncertainty index by Baker et al. (2016) or the World Uncertainty Index by Ahir et al. (2022). A different strand of uncertainty measurement lies in Google Trends data. In contrast to the previous methods, Google allows to measure uncertainty among the general population instead of measuring uncertainty in journalist or forecasters. These Google Trends uncertainty measurements have an influence on real economic variables like investment, consumption, industrial production, and stock market returns (Bontempi et al., 2021; Castelnuovo and Tran, 2017). The underlying assumption is that people feeling uncertain about a subject turn to Google and search for the subject to reduce said uncertainty. Therefore, a higher search request reflects elevated uncertainty.

The mentioned studies used Google Trends keywords, which are prone to language selection. In this paper, the approach uses topics instead of keywords. While the English keyword "economy" only reflects search request which contains the English word "economy", the Google Trend topic "economy" covers keywords like "economic" or "economical" and terms in different language, for example the German word "Wirtschaft". This makes this approach very applicable in a multinational context. Kupfer and Zorn (2019) demonstrated that uncertainty proxies constructed using Google Trends topics have an influence on economic activities in European countries.

The contribution of this paper is as follows: firstly, giving insights in the diverse demand side of information provision across the globe with a cluster analysis. Secondly, showing that these clusters are stable over time. And thirdly, that these insights into the demand side of information can be used to form a timely uncertainty index, which outperforms the uncertainty index by Ahir et al. (2022) when it comes to forecast performance.

The rest of the paper has the following structure: The next chapter gives an insight to the used data and the construction approach of the country clustering. In the third chapter the country clusters of the second chapter are used to form country specific uncertainty measurements to compare them to the uncertainty proxy by Ahir et al. (2022). The last chapter concludes.

2. Multivariate time series clustering

In this section the data collection and clustering approach is explained and subsequently the result of the clustering is shown.

2.1. Data

The data for the approach of this paper stems from Google Trends. Google Trends allows for various search requests, for example the coverage of keywords or topics for different country, for different regions and for a certain time span, up to daily data. The main advantage of topics compared to keywords lies in its robustness against word selection and in its applicability in a multilanguage framework.

The complete set of data contains 109 Google Trends topics for 43 different countries spanning monthly from 01/2004 (the earliest date possible with Google Trends) until 02/2022. The 109 Google Trends topics are based on 184 uncertainty keywords by Bontempi et al. (2021) and by Baker et al. (2016) which are available only in English and Italian. For example, two of the keywords are “taxation” and “taxed”. Both were inserted in the Google Trends interface and the primary suggestion by Google for the underlying topic “tax”, therefore leading to the topic “tax” being among the final 109 different Google Trends topics. This procedure was then repeated for all uncertainty keywords. The R package "gTrendsR" was used to download all topics for all countries. Additionally, the 43 countries are chosen because the complete set of 109 topics exists for each country. The names of all used countries can be seen in figure 1 in the next chapter.

2.2. Cluster construction and optimal number of clusters

A hierarchical clustering procedure was applied to the Google Trends data to obtain country clusters of similar information demand. It is assumed that the entire world can be seen as one major information demand cluster with subclusters regarding to economic, political, geographical and/or historical ties. Hierarchical clustering is used in economics for example when it comes to clustering of countries with similar tax burden (Simkova, 2015).

The similarities between the different time series were identified by using Dynamic Time Warping. In contrast to the Euclidean distance, which compares pairs of datapoints directly, Dynamic Time Warping calculates the smallest distance between all datapoints. Therefore, it allows for possible "leads" and "lags" in the data, which could be important because there might exist “Google search spillover effects” from one country to another. While dynamic time warping stems from the area of speech recognition it is slowly also applied in economic, for example to predict recessions (Raihan, 2017).

The clusters are calculated by using agglomerative Ward´s method (Miyamoto et al., 2015), starting with single clusters for each of the 43 countries. These single country clusters are then merged based on minimum within-cluster variance gain leading in the end to a single cluster containing all countries. Therefore, this approach minimizes the intra-cluster variance.

All time series are Z-Score normalized before used in the clustering process. For the clustering procedure the R-package “dtwclust” was used.

While clustering will always result in different cluster sizes it is paramount to identify the optimal number of clusters which fits the data best. For this purpose, two internal evaluation metrics are used, primarily because in contrast to an external evaluation metric no assumption about cluster size and distribution must be made. The first metric is the Silhouette index. It ranges from -1 to 1, measuring the standardized averaged distance from all points within a cluster A to the next cluster B. Here, zero means a poor fit with a lot of overlapping clusters, whereas a value of one means a perfect fit with no overlapping clusters. Therefore, a higher Silhouette index reflects a better fitting clustering. The second metric is the Davies-Bouldin index. This index measures the ratio of the within cluster separation to the between cluster separation. A lower index can reflect to things: Firstly, that the within cluster separation is better, meaning that the data is more compact within a cluster. Secondly, that the separation between clusters is better, meaning that the cluster are not overlapping.

2.3. Clustering results

In figure 1 the cluster dendrogram of the whole dataset can be seen. The nearer the countries are clustered together the bigger the similarities in information search behaviors using Google among these countries. Potential clusters can be formed wherever the dendrogram splits into subclusters.

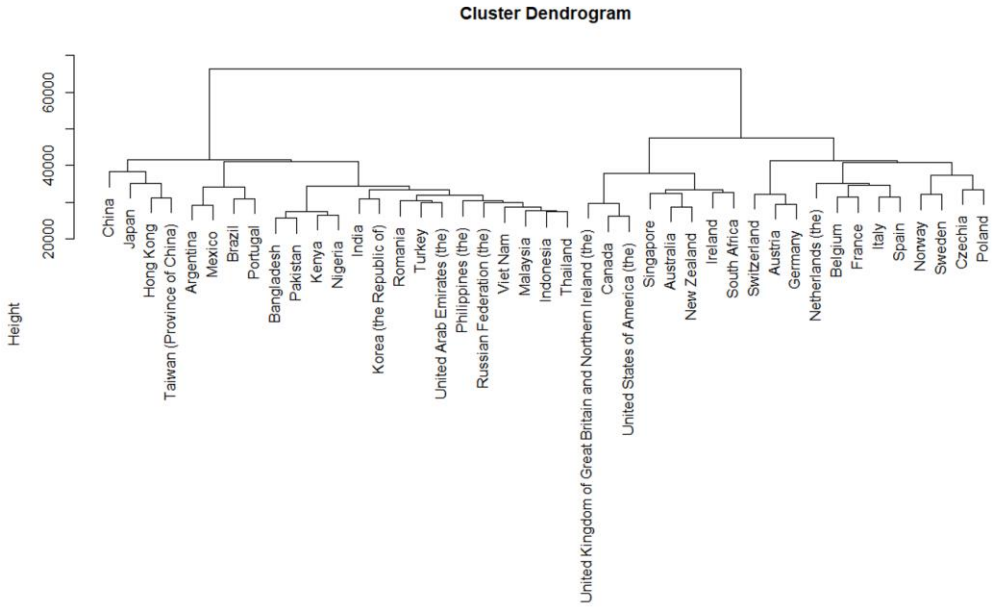


Figure 1: Clustering of Countries based on Google Trends topics data; complete set. Source: own Calculation

Starting at the top there exists a twofold split, resulting in two subclusters. The left-hand side cluster contains (mostly) the emerging economies of the world while the right-hand side

consists only of emerged economies. This alone demonstrates the real-world application of the clustering approach, because the two major branches of the dendrogram are not based on randomness but on similar economic structures resulting in similar information seeking behavior with regards to Google internet searches.

When going to a lower cluster region (or sub-clusters), the splitting is either according to geographical or economical/political reasons. Within the emerging economies cluster on the left side there exists an Asian sub-cluster including China, Japan, Hong Kong and Taiwan. The sub-clustering bordering the Asian one to the right can be interpreted as a South American cluster. The last emerging economies sub-cluster is more based on economic ties instead of geographic ones. In the case of the emerged economies cluster the split is again based on either geographical or economical/political affiliation. The left-hand side split can be interpreted as an "Anglosphere" sub-cluster consisting of mainly English-speaking countries. The right-hand side contains purely (central) European countries.

While it is possible to cluster the countries according to each branching in the dendrogram the optimal fitting number of clusters is based on objective internal evaluation metrics mentioned above. The optimal cluster number is five and was chosen based on the combination of a high silhouette and a low Davies-Boulding index (Silhouette 0.0459, Davies-Boulding index 1.6261) compared to a lower or higher number of clusters.

Table 1. The resulting five clusters with their respective countries

Cluster	**Countries**				
Cluster 1 “Emerging Economies”	Argentina	Bangladesh	Brazil	India	Indonesia
	Korea	Nigeria	Portugal	Thailand	Viet nam
	Malaysia	Pakistan	Romania	United Arab Emirates	Kenya
	Mexico	Phillippines	Russia		Turkey
Cluster 2 “Anglosphere”	Australia	Canada	Ireland	New Zealand	Singapore
	South Africa	United Kingdom	United States		
Cluster 3 “German speaking”	Austria	Germany	Switzerland		
Cluster 4 “Europe”	Belgium	Czechia	France	Italy	Netherlands
	Poland	Spain	Sweden	Norway	
Cluster 5 “Asia”	China	Hong Kong	Japan	Taiwan	

Source: Own calculation.

In table 1 the membership of countries regarding the five clusters are shown. The first cluster can be interpreted as a cluster of mostly emerging countries. The second cluster is an "Anglosphere" cluster. The third and fourth clusters are a "German speaking" and a "European" cluster, respectively. The last cluster is an "Asian" cluster. To sum up, all clusters reflect the connectedness of different countries, either due to geographical, political or economic links or a mixture out of these.

To validate if the cluster results are stable over time, the whole time span (01/2004-02/2022) was cut in half and the clustering has been applied to both subsamples. For the first 9 years the optimal number of clusters is four and the major difference compared to the complete set is the "German speaking" cluster merges with the "European" cluster. When looking at the last nine years the optimal number of clusters is back to five and the "German speaking" cluster is expanded by Czechia and Poland, two non-German speaking countries but with a close distance to Germany. The other clusters stay for the most part the same.

To sum this chapter up the approach using a hierarchical clustering procedure on multinational Google Trends topics data leads to meaningful country clusters being in line with political and geographical proximities. Furthermore, these clusters are stable over time except for the "German" clusters showing up only in the second half of the time span. With these results at hand the next step will be to research if the Google Trend queries of countries within certain subclusters can be used in an economic application context, i.e. as an uncertainty measurement of said countries.

3. Construction of country specific uncertainty indices

This chapter describes how uncertainty proxies using Google Trends topics are constructed and how they perform again an already existing uncertainty proxy.

3.1. Construction based on topic clusters

To construct uncertainty measurements for each country the next step is to identify the optimal number of topic clusters within a respective country. For this task, the 109 Google Trend topics are averaged over all countries within the corresponding cluster. The routine described in the previous chapter is then applied to this new dataset to calculate how many optimal topic clusters exists within each of the five country clusters.

In table 2 the optimal number of topic clusters is stated for all five country clusters. The optimal number is two except for the case of the "Emerging Countries" cluster where the optimal number is four, since this cluster is more diverse than the others when it comes to geography or political affiliation. When looking at the content of the two subclusters for each country, one cluster leans more to theme "economy" while the other cluster is more driven by "politics". For example, in the "Anglosphere" case the first cluster consists of 79 topics ("Tax", "Trade War", "Income tax" etc.) while the second cluster has 30 topics ("Economy", "Business", "Central Bank" etc.).

Table 2. Optimal number of topic clusters within the country clusters

	No. Of optimal topic subclusters	Distribution of topics	Silhouette	Davies-Bouldin index
Cluster 1 (”Emerging Countries”)	4	16-50-22-21	0.19	1.3
Cluster 2 (”Anglosphere”)	2	79-30	0.19	0.9
Cluster 3 (”German speaking”)	2	78-31	0.19	0.8
Cluster 4 (”Europe”)	2	55-54	0.21	1
Cluster 5 (”Asia”)	2	93-16	0.17	1

Source: Own calculation.

3.2. Comparison to the World Uncertainty Index

Published by Ahir et al. (2022) there exists a World Uncertainty Index (WUI) for 143 different countries using the Economist Intelligence Unit reports. The index is constructed counting the word "uncertainty" in the respective report for a certain country and a given time. This uncertainty measurement exists on a quarterly base which is a major disadvantage when it comes to the timely identification of elevated uncertainty regarding time.

To compare the GCUs with the WUI the frequency of the GCU must be adapted, because the GCUs exist on a monthly frequency. For comparing the Google Trends data to the WUI the months for the respective quarters where averaged, for example the first quarter of 2004 consist of the average of the first three month of the year 2004 to keep the informational content as high as possible.

An uncertainty measurement is identified as superior regarding forecast performance if it can significantly (α=0.05) Granger causes quarterly GDP growths in more countries. This was evaluated in a VAR approach using the Toda and Yamamoto (1995) procedure for Granger causality. The data for the quarterly GDP growth stems from OECD (2022), is available for 31 countries and the time span is from 01/2004 to 4/2021. All time series were seasonal adjusted and made stationary prior to the procedure. The optimal lag length for each national VAR was evaluated using the AIC.

Table 3. How often does … Granger causes the national quarterly GDP (α=0.05)

	WUI	GCU1	GCU2	GCU3	GCU4
Cluster 1 (”Emerging Countries”)	2	2	0	7	2
Cluster 2 (”Anglosphere”)	4	3	0	-	-
Cluster 3 (”German speaking”)	1	2	0	-	-
Cluster 4 (”Europe”)	3	3	1	-	-
Cluster 5 (”Asia”)	2	2	0	-	-
Total	12	12	1	7	2

Source: Own calculation.

In table 3 the results of the Granger causality procedure are shown. For all 31 considered countries the WUI Granger cause the quarterly GDP twelve times. This is comparable to the GCU performance being also twelve for the first cluster. When using the third cluster for the "Emerging Countries" instead of the first cluster the GCU Granger causes the GDP in 17 countries, which is a distinctly better result compared to the WUI.

To sum this chapter up, it was shown that the constructed multinational Google Cluster Uncertainty indices do Granger cause quarterly national GDP growth in different countries, implying valuable forecast characteristics. The constructed indices perform better doing so in comparison to an existing uncertainty index, the WUI. Furthermore, while the WUI is only available on a quarterly base, the GCU is also available on a monthly base (and even up to a daily base) making it more useful in short term forecasting.

4. Conclusion

In this paper multinational Google Trends search queries were used to show that meaningful country clusters can be formed. These clusters are stable over time and overlap with economic, geographic or political affiliation. These clusters can be used to identify relevant topics in different countries leading to a deeper understanding of the distribution of information demand around the world.

Topics within the country clusters were then clustered to construct Google Cluster Uncertainty indices for 31 different countries. On average, these indices perform better than an already existing uncertainty measurement regarding forecast ability of GDP growth.

The main advantage of the used procedure lies in its applicability and real time availability. Until now, only 43 countries are considered, but it can be applied to almost all countries when there is an a-priori decision to which cluster a new country belongs. Then, there is no need for the complete set of 109 different topics, but only for the topics in the cluster which are needed to construct the GCU. Furthermore, unlike already existing keyword-based Google

Trends uncertainty indices, which are prone to language selection, the usage of topics offers an easy multinational application.

Furthermore, the Google Trends data can be obtained even on a daily base and for subregions within countries, opening a variety of application for future research and practical forecasting considerations.

References

Ahir, H., Bloom, N., & Furceri, D. (2022). The world uncertainty index. NBER Working Paper 29763.

Bachmann, R., Elstner, S., & Sims, E. R. (2013). Uncertainty and economic activity: Evidence from business survey data. American Economic Journal: Macroeconomics, 5 (2), 217–249.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. The Quarterly Journal of Economics, 131 (4), 1593–1636.

Bloom, N. (2009). The impact of uncertainty shocks. Econometrica, 77 (3), 623–685.

Bontempi, M. E., Frigeri, M., Golinelli, R., & Squadrani, M. (2021). Eurq : A new web search–based uncertainty index. Economica, 88 (352), 969–1015.

Castelnuovo, E. & Tran, T. D. (2017). Google it up! a google trends-based uncertainty index for the United States and Australia. Economics Letters, 161, 149–153.

Jurado, K., Ludvigson, S. C., & Ng, S. (2015). Measuring uncertainty. American Economic Review, 105 (3), 1177–1216.

Kupfer, A. & Zorn, J. (2019). A language-independent measurement of economic policy uncertainty in eastern European countries. Emerging Markets Finance and Trade, 4 (1), 1–15.

Miyamoto, S., Abe, R., Endo, Y., & Takeshita, J.-i. (2015). Ward method of hierarchical clustering for non-euclidean similarity measures. In 2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR), (pp. 60–63). IEEE.

OECD (2022). Quarterly GDP (indicator). 10.1787/b86d1fc8-en (Accessed on 19 March 2022.

Raihan, T. (2017). Predicting us recessions: A dynamic time warping exercise in economics. SSRN Electronic Journal.

Simkova, N. (2015). The hierarchical clustering of tax burden in the eu27. Journal of Competitiveness, 7 (3), 95–109.

Toda, H. Y. & Yamamoto, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes. Journal of Econometrics, 66 (1-2), 225–250.

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15081

Machine Learning and MADIT methodology for the fake news identification: the persuasion index

Luisa Orrù[1], Christian Moro[1], Marco Cuccarini[2], Monia Paita[1], Marta Silvia Dalla Riva[1], Davide Bassi[1], Giovanni Da San Martino[2], Nicolò Navarin[2], Gian Piero Turchi[1]
[1]Philosophy, Sociology, Pedagogy and Applied Psychology, University of Padova, Italy,
[2]Mathematics Department, University of Padova, Italy-

Abstract

The phenomenon of fake news has grown concurrently with the rise of social networks that allow people to directly access news without the mediation of reliable sources. Recognizing news as fake is a difficult task for humans, and even tougher for a machine. This proposal aims to redesign the problem: from a check of truthfulness of news content, to the analysis of texts' persuasion level. That is how information is introduced to the reader, assuming that fake news is aimed at persuading towards the reality of sense they intend to convey. M.A.D.I.T. methodology has been chosen. It is useful to describe how texts are built, overcoming the content/structure analysis level and stressing the study of Discursive Repertories: discursive modalities of reality of sense building, classified into real and fake news categories thanks to the Machine learning application. For the dataset building 7,387 news have been analysed. The results highlight different profiles of text building between the two groups: the different and typical discursive repertories allow to validate the methodological approach as a good predictor of the persuasion level of texts, not only of news, but also of information in domains such as the economic financial one (e.g. GameStop event).

Keywords: *Fake news; Persuasion index; MADIT methodology; Machine learning; Dialogic analysis; Discursive configuration.*

1. Introduction

In recent years, the phenomenon of fake news showed its critical effects (Tandoc, et al., 2018; Tagliabue, et al., 2020). In light of this, the scientific community has worked to provide tools to help citizens identify fake news. The scientific efforts in contrasting fake news outcomes have been undertaken in two "main directions" (Lazer et al., 2018): empowering individuals in evaluating the fake news they encounter, and implementing structural changes on online platforms and algorithms, to prevent exposure of individuals. Referring to the first category of interventions, different studies focused on finding personal characteristics and cognitive processes that play a role in dealing with fake news (Pennycook, Cannon & Rand, 2018; Pennycook & Rand, 2020). About the second category of interventions, Artificial Intelligence and Natural Language Processing have become increasingly important tools to help citizens when dealing with fake news (Oshikawa, et al, 2020). Currently, several different computerized methods are available for detecting fake news (for a review, see Zhou & Zafarani, 2020; Oshikawa, et al., 2020). The plurality of available methods can be traced back to the lack of a specific and shared definition of "fake news", which currently assumes different characteristics depending on the author or the research considered (Tandoc, et al., 2018; Andersen & Søe, 2020). At the same time, Zhou e Zafarani (2020) offer valuable support, organizing all the constructs "similar to fake news (es. satire, clickbait, etc) in function of their intention and their truthfulness. However, some questions are left open: how is the truth of a certain content decided, when the news cannot find factual correspondences (Andersen & Søe, 2020)? How much of the analyzed content is false, and how to consider the news when it's partially false (Oshikawa, et al., 2020)?

The mere truthfulness of a piece of content does not offer any insight into how it is conveyed by the text: the presented work attempts to abandon the dichotomy "fake-news/real-news" in favor of the construct of *persuasion*. In doing so, we do not intend to replace content truthfulness analysis with persuasion analysis, but rather to shift the focus of the investigation: from the reader who assumes this information to be fake or true, to the degree to which the modalities used to convey the text lead the reader to assume the same narrative position, i.e. the same modalities used in and by the text itself. It is possible to anticipate how fake news conveyed through highly persuasive modes can lead readers, and more broadly the community, towards more high-risk interactive settings, e.g. of social fragmentation. This perspective is in continuity with Barron-Cedeno et al. (2019) contribution on propaganda (Da San Martino et al., 2020): the authors move beyond the fake/real news distinction, in favor of an approach that provides an index related to how much the news tries to influence the reader's opinion through automated analysis of the structure and content of the text: a persuasion index (Miller & Levine, 2019; Festinger 2001; Grandberg, 1982; Cacioppo, Cacioppo, & Petty, 2018; Druckman, 2021). O'Keefe's perspective on persuasion (2016), integrated with the analysis tools made available by the

NLP, enables the extraction of the characteristics of the arguments that allow us to effectively change the interlocutor's "position" on a given argument ("persuasion techniques"; Li, et al., 2020; Hunter et al., 2019). Following these recent works, the analysis of the ways in which language is used enables the identification of typical modes of persuasive messages. The focus of our work is the same ways of using language: the persuasiveness of a text, which is in fact associated with the rhetorical-argumentary architecture of the text itself (how it is discursively structured) and the amount of critical reading competence and attention needed to consume it. As a theoretical reference to solve this problem, we have chosen Dialogical Science (Pinto et al., 2022; Turchi, et al., 2021), which, through the formalization of Natural Language, has given value to the description of Natural Language's use, formalizing the rules of its use. The formalization of language has given value to saying, defining and grouping it, and then measuring it (Turchi and Orrù, 2014). Following the approach of Dialogical Science, we defined persuasiveness as "the possibility of a text to make the reader use discursive modes and contents similar to those of the text itself"[1]. This possibility of the text is therefore independent of the truthfulness of the contents; however, we anticipate that fake news disseminates texts that are more prone to this possibility. We then defined a novel index of persuasiveness of a text derived from the Discursive Repertories (DRs; Turchi et al., 2021; Turchi and Orrù, 2014), namely specific language use modalities as defined by M.A.D.I.T methodology (Turchi et al., 2021). There are 24 DRs, each identifying one of the possible language use modalities that can be traced in a text. Based on the specific characteristics and properties of each DR, we hypothesized that 12 DRs can be understood as the elements that constitute the DNA of a persuasive news, since facts can be manipulated through language in order to convince the reader about some version of them, creating also what we called fake news (Iswara and Bisena, 2020). This exploratory experimentation wanted to observe whether the DRs that, theoretically, should characterize persuasive texts, characterize the texts of fake news more than the one of real news. We anticipate that even real news texts could employ DRs related to the construct of persuasion, but we consider that fake news and persuasion index are related, certainly it will be the object for a future work of deepening.

2. Methods

In order to pursue the goal of this exploratory experimentation, building a dataset containing both fake news and real news was necessary. We define real all the news taken by authoritative Italians newspaper and fake all the news taken from list of blogs and web sites

[1] Rephrasing of the technical definition of Persuasion: "*Possibility of a text to generate a configuration of sense tending in turn to occupy - without ever overlapping - the same discursive space of the original text*".

that spread medical, scientific and political misinformation. The list was provided by the fact checking web site: Bufale.net. To obtain the texts of these articles automatically, we used a package of the python library which, after providing the link of the news' container site, returns the title, the text and other information of the news itself. We found an imbalance in the topics covered: fake news usually focuses on a certain set of topics, ignoring others. To overcome this issue, we eliminated from real news the topics that are rarely covered in fake ones, for example sports. The dataset is composed of 2776 real news and 4611 fake news. We built a corpus of human annotated texts and devised a machine learning approach to identify the DRs. Naive Bayes was initially employed, but its low precision rate (0.35) and recall rate (0.37) led us to use BERT. We created a model which manages to divide the inserted text into excerpts. Subsequently, BERT classifies these excerpts according to the 24 possible categories (as the DRs). The model that is closest to human performance has a precision level of 0.47 (recall=0.43; f1-score=0.43), which can be improved considering that human roles trained with a basic training have precision=0.65; recall=0.63; f1-score=0.63. We use a model previously trained for the identification of DRs, that model is a BERT. For identification of DRs, we used a dataset of 14567 excerpts, each of them belongs to one of the 24 possible categories (as the DRs). We have defined different possible models and using the cross-validation we chose the best one in terms of performance. In our case was that one is described in the Table 1. Chose the best model, we split the dataset in train (75%) and test (25%), we train it and we get the result.

Table 1. BERT model structure

Model structure	Pretrained Weights	Batch Size	Learning Rate	Freeze to
The structure of model that produce the pertrained weights.	bert-base-italian-xxl-uncased	16	1e -05	1

AdamW Eps	Max Epochs	Patience Epochs	Embed Dim	Activation Function
0.0001	200	20	768	ReLU

For the construction of the dataset with the DRs distribution, we implement a model for text division in extracts, we used the function sent_tokenise of the package NLTK. This tokenizer divides a text into a list of sentences by using an unsupervised algorithm. Following the construction of the dataset, we used the model to analyze the texts of fake and real news, to detect how the chosen model divides the various excerpts and names them, going into detail of the distribution of the DRs for fake and real news. Having for each text, real or fake, the distribution of DRs, we used a model to predict whether a text is a real news or a fake news;

this, to provide further support for the theoretical link between persuasion and biased texts, based on the distribution. We then implemented a Random Forest model, which allows us to define the importance that each repertory has had in the classification. Either in this case we spit the data in train (75%) and test (25%) for searching the best hyper parameters, we used RandomisezSearchCV of the package Sklearn. Hyper-parameters are optimized by cross validated search over parameter settings.

3. Results

Figure 1 shows the distributions of RDs according to the text categorization in fake and real news. Generally, a strong presence of characteristic repertories is observable for both the text codifications.

Table 2 addresses the Figure 1's distributions, and reports the percentage occurrences of the 12 DRs in fake and real texts. We use the distribution to characterize a persuasive text. The assumption of our research was that we would be able to trace an increased occurrence of persuasion-indicating DRs regarding the texts of fake news, more than the real news ones. The results confirm this hypothesis for seven of the twelve indicated DRs ("Anticipation", "Cause of Action", "Declaration of Aims", "Prediction", "Justification", "Certify Reality" and "Evaluation").

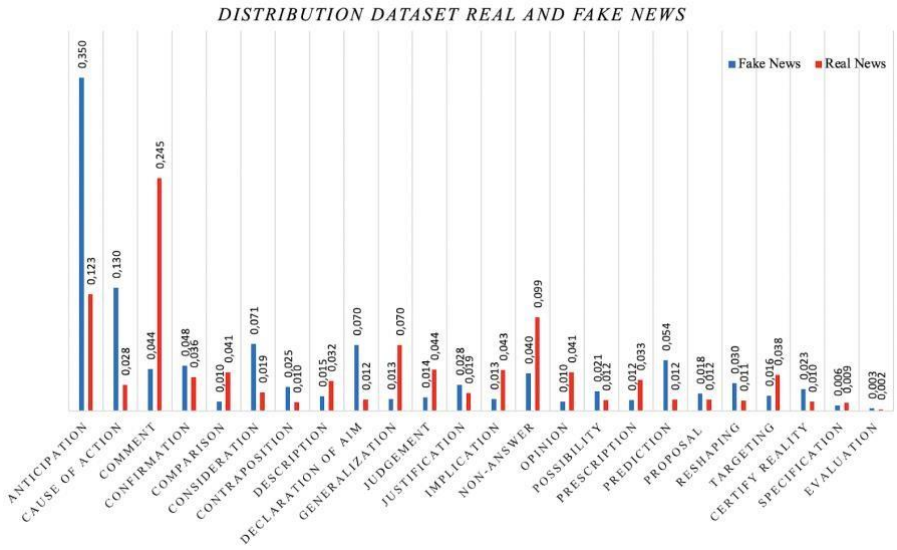


Figure 1. Distribution Dataset of real and fake news

Table 2. Distribution of indicators of persuasions between fake and real news

DR	Fake	Real	RF	DR	Fake	Real	RF
Anticipation*	0.35	0.123	0.00937	Targeting	0.016	0.038	0.01921
Cause of Action*	0.13	0.028	0.03254	Judgment	0.014	0.044	0.03176
Declaration of Aim*	0.07	0.012	0.02566	Implication	0.013	0.044	0.03111
Prediction*	0.054	0.012	0.05226	Prescription	0.012	0.033	0.03958
Justification*	0.028	0.019	0.02134	Opinion	0.01	0.041	0.02325
Certify Reality*	0.023	0.01	0.03547	Evaluation*	0.003	0.002	0.02545

When looking at the results, it is also necessary to consider the imbalance between fake and real news in the dataset, the level of accuracy of the model with respect to the naming task (0.47) and, in general, the errors brought behind in the various phases of the program, both by the model that classifies the DRs but also from the one that divides the text. The results in the prediction of a text as real or fake news are with the Random Forest' model: Precision=0.76; Recall=0.76, F1-score=0.76; which can be considered very good results due to the difficulty of the task and the margin of error of the model that predicts the DRs of each text.

4. Conclusions

In line with the work of Barron-Cedeno and colleagues (2019), a perspective was adopted that attempted to overcome the issues of the traditional 'fake-news/real-news' distinction. Therefore, a persuasion index was constructed to provide the reader with elements to evaluate the bias of a certain piece of news, based on the specific ways in which language is used in a text. The analysis of language use was carried out according to Dialogic Science (Turchi and Orrù, 2014), and implemented through an automatic process, based on the BERT transformer. The results obtained from the experiment generally support what has been argued in theory. Specifically, we observed a greater occurrence of DRs - which theoretically should generate persuasion in the reader - in fake news texts, and a greater occurrence of "less persuasive" DRs in real news texts, as was also shown by the results of the Random Forest model. Therefore, the elements of the 'DNA' of persuasive news, identified through the Dialogic Science, have been matched by experimental data. However, this kind of distribution, in support of theory, has not been found in all repertories that are supposed to promote a process of persuasion. This can be traced back to several factors: the imbalance of the dataset, the error made by the model in processing the text in excerpts, as well as in the naming process. Thus, the findings highlight certain aspects of the model used and the need

of the automated textual analysis to be specified and refined in the future, to increase the accuracy of the automated analysis. A further future work perspective concerns the construction of the dataset. As a matter of fact, the need for pre-processing work on the fake news dataset emerged: on the one hand, by maintaining a similar distribution between the fake news part and the real part; on the other hand, by refining the process of "cleaning" the text from parts that do not relate to the article in question, but are mistakenly downloaded from the Python library. Lastly, the index was applied in the economic-financial domain, to about 100 posts (on Reddit and Twitter) regarding the GameStop case (January 2021). A multitude of small investors, gathered on the Reddit page r/WallStreetBets, bought the company's shares in a mass movement that led them to rise, in less than a month, from $17.25 to $348, controversially setting up the entire financial market. It emerged that about 70% of the DRs used were indicators of persuasiveness: the posts addressed users in a uniform and coordinated way towards the goal of "opposing" Wall Street and big investors. This scenario could have been anticipated, observed and measured thanks to the persuasion index. The same index could be applied in the detection of war propaganda and in politics, since public convincing becomes the main aim of the professional roles involved, in order to gather consensus in a strategic way, avoiding risks of losing support (Durante and Zhuravskaya, 2018).

References

Andersen, J., & Søe, S.O. (2020). Communicative actions we live by: The problem with factchecking, tagging or flagging fake news – the case of Facebook. *European Journal of Communication,* 35(2), 126-139.

Barron-Cedeno, A., Jaradat, I., Da San Martino, G., & Nakov, P. (2019). Proppy: Organizing the news based on their propagandistic content. *Information Processing and Management,* 56, 1849-1864.

Cacioppo, J.T., Cacioppo, S., & Petty, R.E. (2018). The neuroscience of persuasion: A review with an emphasis on issues and opportunities. *Social Neuroscience,* 13(2), 129-172.

Da San Martino, G., Shaar, S., Zhang, Y., Yu, S., Barrón-Cedeno, A., & Nakov, P. (2020, July). Prta: A system to support the analysis of propaganda techniques in the news. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 287-293).

Druckman, J.N. (2021). A Framework for the Study of Persuasion. *Annual Review of Political Science,* 25.

Durante, R. & Zhuravskaya, E. (2018). Attack When the World Is Not Watching? US News and the Israeli-Palestinian Conflict. *Journal of Political Economy,* 126(31), 1085-1133.

Festinger, L. (2001). *Teoria della Dissonanza Cognitiva.* Franco Angeli.

Grandberg, D. (1982). Social Judgment theory. *Annals of the International Communication Association,* 6(1), 304-329.

Hunter, A., Chalaguine, L., Czernuszenko, T., Hadoux, E., & Polberg, S. (2019). Towards computational persuasion via natural language argumentation dialogues. In *Joint German/Austrian Conference on Artificial Intelligence* (Künstliche Intelligenz) (pp. 1833). Springer, Cham.

Iswara, A.A., & Bisena, K.A. (2020). Manipulation And Persuasion Through Language Features In Fake News. *RETORIKA: Jurnal Ilmu Bahasa*, 6(1), 26-32.

Lazer, D.M., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., ... & Zittrain, J.L. (2018). The science of fake news. *Science*, 359(6380), 1094-1096.

Miller, M.D., & Levine, T.R. (2019). Persuasion. In D. W. Stacks, M. B. Salwen, & K. C. Eichhorn, *An Integrated Approach to Communication Theory and Research* (p. 261 - 277). New York: Routledge.

O'Keefe, D.J. (2016). *Persuasion. Theory and Research.* London: Sage.

Oshikawa, R., Qian, J., & Wang , W.Y. (2020). A Survey on Natural Language Processing for Fake News Detection. *LREC*, 6086-6093.

Pennycook, G., Cannon, T.D., & Rand, D.G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general*, 147(12), 1865.

Pennycook, G., & Rand, D.G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of personality*, 88(2), 185-200.

Pinto, E., Alfieri, R., Orrù, L., Dalla Riva, M.S., & Turchi, G.P. (2022). Forward to a methodological proposal to support cancer patients: the Dialogic contribution for the precision care. *Medical Oncology*, 39(5), 75.

Tagliabue, F., Galassi, L., & Mariani, P. (2020). The "Pandemic" of Disinformation in COVID-19. *SN comprehensive clinical medicine,* 2(9), 1287-1289.

Tandoc, E.C., Lim, Z. W., & Ling, R. (2018). Defining “Fake News” A typology of scholarly definitions. *Digital journalism,* 6(2), 137-153.

Turchi, G.P., Dalla Riva, M.S., Ciloni, C., Moro, C., & Orrù, L. (2021). The Interactive Management of the SARS-CoV-2 Virus: The Social Cohesion Index, a Methodological Operational Proposal. *Frontiers in Psychology*, 12.

Turchi, G.P., & Orrù, L. (2014). *Metodologia per l'analisi dei dati informatizzati testuali: fondamenti di teoria della misura per la scienza dialogica*. Edises.

Zhou, X., & Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Survey,* 53(5), 40.

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15082

Monitoring the survival of subscribers in a marketing mailing list

Andrea Marletta
Department of Economics, Management and Statistics, University of Milano-Bicocca, Italy.

Abstract

The statistical literature proposed many contributions about survival analysis in medical research, in this work this approach is proposed in a business context. The aim of this paper is to control the mortality of the users belonging to an e-mail subscribers list for a company operating in the healthcare information sector. Having available the survival times for each subscriber, the choice was oriented to survival models to evaluate the abandon of the customers. A survival analysis was conducted through a Cox model considering some risk factors of the subscriber. The selected Cox model carried to the identification of risk profiles representing different situations in terms of probability of abandon.

Keywords: *Mailing list marketing; Healthcare information; Survival analysis; Cox Model*

1. Introduction

During last years, the marketing strategies experienced many changes and the introduction of new technological tools favoured innovative approaches more hinged on a direct relationship with customers. One of these approaches is based on the use of the e-mail. This tool passed from a simple communication media to a privileged channel for the direct attainment of a possible list of customers. The user could enter in a marketing mailing list after a volunteer subscription, indeed during an on-line purchase, the entering of an e-mail address in a mandatory information to complete the process. This information should be used for a faster interaction between the customer and the seller and it is inserted into a business database. But even after the accomplishment of the purchase process, the communication with the seller proceeds with the sending of the advertising material in order to make easier new purchases in the future. In other cases, the addition in the e-mail database occurs indirectly, for example after the transfer of the information by a third stakeholder, according to provided modalities by the contract. The user often accepts unconditionally these terms, above all during e-commerce purchases, without giving importance to the transfer of personal data, generating problems in terms of data privacy.

The introduction of the direct mailing among the marketing strategies generated a raising of the available data. This tool led to creation of a business databases that starting from the e-mail address, it joined personal information of the customer (name, surname, age, gender, professional role, telephone number, address,...) and other information related to the interaction business-customer (date of entering, last mail sent, number of sent mails, last click on e-mail, ...). Using this database, the companies could obtain useful information to profile the customer list, customizing the sending of personalized contents during time. Beyond the profiling, the firms could be interested to monitor the abandons, due to voluntary cancellations from the mailing list or for the closure or disuse of the e-mail address automatically reported after the shipping.

The existent literature showed recently interest for this issue focusing on how customers respond to email messages or looks at the average effect of email on transactional behaviour (Zhang, 2015) or investigating the effectiveness of triggered e-mail marketing (Goic et al., 2021). In this paper, this strategy is faced from a statistical point of view. Here, the statistical approach is focused on survival analysis, a method usually applied in medical research. This technique is well-known and described in Clark et al. (2003); Collett (2015); Cox and Oakes (2018); Hougaard (2000).

The term survival is here intended as synonym of permanence of an individual in the mailing list with the registered e-mail address. This approach is data-driven since this technique needs a survival time variable obtained as difference between the date in which the e-mail address entered the list and the date in which the last mail was sent. In this date, the customer chose

to not receive e-mails. Another similarity with the medical approach is in the concept of censored data, that is to say, the statistical unit that did not experiment the death effect. In this case, the censored unit is the customer that received the last e-mail when the data collection period is over.

Applying this technique in this context could have multiplex aims: firstly, to monitor from a descriptive point of view the effectiveness of the direct mailing; secondly to compute the abandon rates of the mailing list; finally, to detect customer profiles more at risk. The aim of this work is to verify the applicability of these models in a context different from the usual one and to provide to the companies a new tool suitable to check the reliability of this marketing strategy.

2. Survival analysis

Survival analysis contains all the techniques and statistical models designed for the description and the analysis of time events of a statistical unit. It is necessary to identify the unit exposed at risk respect to this event and the measure of the time duration and the end of these event. Survival is therefore characterised by a time variable with a start-up and an end-point. In medical research, start-up corresponds to time in which an individual has been introduced in the experimental study or a clinical treatment or the start of a particular condition for a disease. On the other hand, if the end-point is the death of the patient, data are referred to the death time. The end-point could be not necessarily the death, but also the end of a pathological state. For this work, the start-up is the date in which the customer was subscribed in the e-mail lists and the end-point is represented by the exit of the customers from the list.

Survival analysis can be treated using non-parametric, parametric or semi-parametric models. The first nonparametric approach considers the estimate of the survival function of a t time variable using the life-tables. These tables are obtained dividing the observation period in temporal intervals (Collett, 2015). Non-parametric models are very flexible but they do not guarantee consistent and precise estimates. This is why they are usually as exploratory tools. For this reason, parametric models have been proposed proposing that the time variable assumes a probability distribution depending on some parameters. Once the function probability distribution function $h(t)$ and the cumulative hazard risk function $f(t)$ is chosen, then it is possible to obtain the survival function $H(t)$. Finally, semi-parametric models were introduced by Cox (1972) and it is so defined because even if it is based on the hypothesis of proportional hazards, it makes no assumption about a probability distribution for the survival times. The Cox model assumes the hazard risk function $h(t)$ as a product of two components:

$$h(t) = h_0(t) * exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}) \quad (1)$$

The first component $h_0(t)$ is named baseline hazard risk function, the second one is the exponential of the sum of the combination terms $\beta_i x_i$ extended to all p explanatory variables. The computed model allows to identify some categories of users more at risk.

3. Application

The analysis was based on research proposed by PKE, Professional Knowledge Empowerment, a company created to manage Italian healthcare databases. Over time, the areas of expertise have expanded, thus specialising both in data management and communication. In the communication area, one of the services is the e-mail marketing. From an increasingly digital standpoint, communication strategies must also take into account the change that PKE reinterprets, by making email marketing projects available that guarantee precious and exclusive value: in-depth knowledge of the health professional and in particular of doctors.

In this paper, the dataset is composed by all the subscribers in the PKE e-mail marketing list until 2021. PKE sends over 18 million e-mails every month, this tool has allowed it to perfect communication models for promotion of drugs in launch, mature or in decline, in siding or replacing the local pharmaceutical representative. The target audience is made of pharmaceutical companies, medical device companies, certification bodies, scientific societies, patient associations, insurance, technology companies, public/private bodies of the NHS, CME providers, publishing companies, public utilities. The type of subiscription is volunteer since the user takes part to some events related to the world of healthcare information. The professional background of these users is mainly represented by professionals in the medical sector.

Several models could be obtained considering different dependent variables of the Cox model. A time variable could be computed as difference between the subscription in the list and the last received e-mail. Another time variable could be the difference between the subscription and the last time the subscriber opened or clicked the e-mail. Once these Cox models are estimated, it is possible to define some risk profiles and determine the categories of target audience more inclined to abandon the e-mail marketing strategy.

The variables considered as potential risk factors for the abandon state are referred to the personal information of the subscriber:

- gender;
- age;
- workplace;
- main professional figure;
- medical specialization.

The full dataset is composed by 651.783 mailing lists for the PKE subscribers. First of all, it could be interesting to compute the percentage of abandoned e-mails over to the total. The number of abandons is equal to 162.960. So the abandon rate could be easily computed:

$abandon\ rate$ = not-active e-mails/ $total\ email\ addresses$ = 162.960/651.783 = 25,0%

Some operations of aggregation and encoding were applied on the original variables in order to better organize the data. The age of the subscriber was categorized in three slots: young (until 35 years old), medium-age (between 35 and 65 years old) and retired (more than 65 years old). About the workplace of the subscriber, this information is available at different levels according to the Istat classification: NUTS1 (macro-regions), NUTS2 (regions) and NUTS3 (provinces). The number of professional figure considered are 55, they are grouped in medical and non-medical area. For individuals in medical area, it is available the sub-category of specialization. It was divided into 5 macro-categories: Medical area, Clinical Services, Surgery, Dentistry and Other. Specialization in general medicine groups internal and specialized medicine, psychiatry and pediatrics. Clinical services gathers radio diagnostics, anaesthesia, rehabilitation medicine and public health. Surgery specialization includes general surgery, neurosurgery and heart surgery.

In table 1, the frequency distribution for the risk factors are shown. There is equal distribution between men and women among subscribers (51,8% vs 48,2%). The majority of subscribers is in the middle age class (2 over 3), more than 1 over 4 is more than 65 years old and only 7,5% of the customers is under 35 years old. For this representation, workplace variable is represented by the higher level, NUTS 1, Northwest and Central Italy are the most present area with respectively (28,0% and 26,4%), followed by Northeast and South Italy (17,2% and 18,0%). Only 10,4% of the subscribers works in Insular Italy. Since the source of the dataset is a company specialized in healthcare information, most of the subscribers (87,3%) belongs to the medical area as main professional figure. For this customers, it is also available the information about the specialization. For subjects in medical area, 45,6% is represented by physicians operating in general medicine specializations followed by clinical services (23,3%) and surgery (22,8%), finally 6,5% of subscribers are in the dentistry area.

The frequency distribution allows to build the baseline profile, joining the typical features of the subscriber. This customer is a man of medium-age with a workplace in Central Italy belonging to medical area specialized in general medicine. Central Italy has been chosen instead of Northwest Italy as a reference level because when spatial variables are considered, it is preferable to choose a central area. To measure the abandon risk, a semi-parametric Cox model was adopted as described in the previous section. Estimated coefficients β_i, the hazard ratio relative risk $e(\beta_i)$ for all risk factors are presented in table 2. Applying this model, it is possible to create the profile of a customer with higher or lower abandon risk.

Table 1. Risk factors frequency distribution

Risk factor	Percentage (%)
Gender	
Men	51,8%
Women	48,2%
Age of the subscriber	
Young	7,5%
Medium-age	66,6%
Retired	25,8%
Workplace (NUTS1)	
Northwest	28,0%
Northeast	17,2%
Central	26,4%
South	18,0%
Insular	10,4%
Professional figure	
Medical area	87,3%
Non-medical area	12,7%
Medical Specialization	
General medicine	45,6%
Clinical services	23,3%
Surgery	22,8%
Dentistry	6,5%
Other	1,8%

In table 2, the underlined levels of risk factors represented the baseline profile, the βi coefficients could be interpreted in terms of significance and in terms of abandon risk considering the hazard ratio relative risk $exp(\beta_i)$. Last column reports the p-value, associated to the hypothesis test. Since all p-values are under the threshold of 5%, then all the explanatory variables have significant coefficients, this means that no risk factors have to be deleted from the full model.

The estimated parameters could be interpreted in terms of sign and value. The positive sign implicates an higher risk in comparison with the baseline level. The hazard ratio HR $e(\beta_i)$

indicates how much increase the risk for the subscriber with that level of the explanatory variable.

Table 2. Full semi-parametric Cox Model

Risk factor	β_i	$exp(\beta_i)$	P-value
Gender			
Men	0,000		
Women	0,074	1,077	0,000
Age of the subscriber			
Young	0,522	1,685	0,000
Medium-age	0,000		
Retired	0,238	1,269	0,000
Workplace (NUTS1)			
Northwest	0,098	1,103	0,000
Northeast	0,251	1,285	0,000
Central	0,000		
South	0,017	1,017	0,217
Insular	-0,112	0,894	0,000
Professional figure			
Medical area	0,000		
Non-medical area	0,140	1,150	0,000
Medical Specialization			
General medicine	0,000		
Clinical services	0,146	1,157	0,000
Surgery	0,090	1,094	0,000
Dentistry	-0,407	0,666	0,000
Other	-0,035	0,966	0,489

For example, the value $\beta_i = 0{,}074$ for women reports an higher risk of abandon of the mailing list for this category. The value $e(\beta_i) = 1{,}077$ means that there is a +7% ($exp(\beta_i)-1$) of risk for these users. Young and retired subscribers are more at risk of abandon compared to medium-age subscribers (+68% and + 27%). Users with workplace in Northeast of Italy are the more at risk (+28% vs the Central Italy). In the medical area, physicians specialized in

clinical services are more inclined to leave the list. Missing values for this variable are residuals, so they did not affect the consistency of the model.

4. Conclusions

The proposed approach combines the use of survival analysis in a business context for measuring abandon risk in e-mail marketing. The availability of information about time days between first and last e-mail sent suggested the use of a semi-parametric Cox model. The results for this model for abandon rate are satisfactory detecting low and high risk profiles. These results could be very useful for the company owner of the mailing list. For example, starting from the Cox Model, it is possible to compute a risk score for each profile. Using this tool, when a new user enters the list, it could be immediately classified as user more or less inclined to abandon. The age of the subscribers seems to be a significant risk factor, considering young and retired users as more at risk individuals. Some territorial differences are present using NUTS1 and specialization, a possible enhancement could regard the use of NUTS2 and NUTS 3 variables focusing the attention on a smaller area.

Future works could regard different survival analysis models, for example comparing these preliminary results with methods as Kaplan-Meier survival curves or with exponential or Weibull models. These methods are parametric and usually proposed for descriptive issues. Finally, the presented model could be enhanced using the number of emails sent during the considered period or different time variables such as time between first e-mail sent and last click.

References

Clark, T.G., Bradburn, M.J., Love, S.B., Altman, D.G., (2003). Survival analysis part i: basic concepts and first analyses. British journal of cancer 89, 232–238.

Collett, D., (2015). Modelling survival data in medical research. CRC press.

Cox, D.R., (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) 34, 187–202.

Cox, D.R., Oakes, D., (2018). Analysis of survival data. Chapman and Hall/CRC.

Goic, M., Rojas, A., Saavedra, I., (2021). The effectiveness of triggered email marketing in addressing browse abandonments. Journal of Interactive Marketing 55, 118–145.

Hougaard, P., (2000). Analysis of multivariate survival data. volume 564. Springer.

Wu, J., Li, K.J., Liu, J.S., (2018). Bayesian inference for assessing effects of email marketing campaigns. Journal of Business & Economic Statistics 36, 253–266.

Zhang, X., (2015). Managing a Profitable Interactive Email Marketing Program: Modeling and Analysis. Georgia State University.Blazquez, D., & Domenech, J. (2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, 130, 99-113.

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15083

Digital Ethnography Redux: Interpreting Drone Cultures and Microtargeting in an era of Digital Transformation

Gavin Mount[1], David Beesley[2]
[1]School of Humanities and Social Sciences, UNSW Canberra, Australia, [2]School of Media and Communciation, RMIT University, Australia.

Abstract

This paper affirms and demonstrates the application of digital ethnography methodologies to two digitally transformative phenomena that are fundamentally enmeshed in the public sphere: personal drones and microtargeting. We review recent methodological studies on digital ethnography that can be delineated into three forms: research that is online or remote by necessity because of physical distance between researcher and participants; research that uses natively digital tools to study phenomena (Rogers 2013; Fish 2019) and research focused on digital cultures (Markham 2020). Our application of digital ethnography is further informed by qualitative ethnographic research undertaken by Pink, Horst, Postill and Hjorth (Pink, et al., 2016); and Manovich's work on the application of digital ethnography to examine automation and big data (Manovich & Arielli, 2022). Beesley (forthcoming) utilises longitudinal visual ethnography as a lens to understand consumer drone cultures and disentangle the multiple narratives surrounding these disruptive technologies. Mount (2020), utilised digital ethnography to review two decades of microtargeting activities, employed by Strategic Communication Laboratories and Cambridge Analytica, to influence electoral behaviour. This methodological research will be combined with our conceptual swarm hermeneutics framework (Mount & Beesley, 2022) to develop scenario based simulations that will further evaluate interpretive schemas and behaviours.

Keywords: *Digital ethnography; drones; microtargeting; big data; culture.*

1. Introduction

This paper affirms and demonstrates the application of digital ethnography methodologies to two digitally transformative phenomena that are fundamentally enmeshed in the public sphere: personal drones and political microtargeting (PMT). In this paper, we review and refine applicable ethnographic methodologies in these digital domains to inform our current research. Building upon our conceptual *swarm hermeneutics* framework (Mount & Beesley, 2022), we aim to combine these conceptual and methodological approaches and develop scenario based simulations that will be used to test and train interpretive schemas and behaviours.

2. What is digital ethnography?

Ethnography is the systematic study of cultural phenomena from the point of view of the subject of the study and the behaviour of participants in a given social situation. *Digital* ethnography can be delineated into three forms: (i) research that is online or remote by necessity because of physical distance between researcher and participants; (ii) research that uses natively digital tools to study phenomena (Rogers 2013; Fish 2019) and; (iii) research focused on digital cultures (Markham 2020).

Digital ethnographic methods are a powerful approach to theorising, conceptualising and practising research on cultural phenomena in digital and data rich environments. Pink et al. (2016) have proposed the following set of five principles to guide digital ethnography research:

i. **Multiplicity** – There is more than one way to engage with 'the digital'. Research is unique to the research question, as well as by the needs and interests of different research partners, stakeholders and participants.
ii. **Non-digital-centric-ness** – The digital is de-centered in digital ethnography, yet it is also inseparable from the other activities, technologies, materialities and feelings through which they are used, experienced and operate.
iii. **Openness** – Digital ethnography is an open event. It is not a research method that is bounded nor is it a unit of activity or a technique with a beginning or end. Rather, it is processual and often iterative.
iv. **Reflexivity** – Digital ethnography involves reflexive practice. Ethnographers consume and produce knowledge through encounters with other people and things. Pink et al. (2016) argue that reflexive practice is necessarily an ethical practice in that it enables researchers to acknowledge the collaborative ways in which knowledge is made.
v. **Unorthodox** – Digital ethnography embraces the complexities of contemporary social contexts by encompassing a diverse set of methods that are adaptive, allowing the ability to find the best suited tool for a given situation.

The above model provides an extended framework of digital ethnographic techniques which are required to make sense of rapidly evolving enmeshed societies and emerging digital cultures. These broad set of data collection methods are not, however, what make ethnography inherently meaningful. The methodology is enriched only when it is engaged through a particular disciplinary or interdisciplinary analytical framework and used in relation to other practices and ideas within a research process (Pink, et al., 2016).

3. Why digital ethnography for Big Data?

When applied to Big Data analytics, a term which increasingly encapsulates our digitally transfomed society, digital ethnography is a fascinating and illuminating method to study data produced through human behaviors and the resulting movement and flow of information. As a method, it is equally applicable to both small- and large-scale research, from a single case, instance, individual or small group, through to exploring patterns in aggregated large datasets, and allows for the analysis of upswells or shifts of interest in events or crisis, for example, by examining how ideas flow or emerge through various groups, platforms, or networks (Markham 2020). Research groups such as RMIT's Center for Automated Decision Making and Society [ADM+S] are increasingly using ethnographic techniques to explore notions of ethics and bias in AI. For example, Graham and Thompson (2022) have used ethnography to monitor cultures of misinformation through a study of pro-Russian Twitter bots to demonstrate how they have an exponential capability to spread harmful information across limitless networks. Media theorist Lev Manovich uses ethnographic techniques in his recent studies on generative art and large datasets (Manovich & Arielli, 2022) to elicit the underlying cultural forms. Likewise, the examples discussed below employ diverse digital ethnographic methods to interpret the cultural dynamics of consumer drones and political microtargeting in digitally transformative contexts.

4. Case studies

4.1. Case Study 1 – Drone Cultures

Drones are 'uniquely transformative technologies capable of extending and elevating human and more-than-human senses to the edges of the internet and into entanglements with other forces and species' (Fish, 2019). Beesley (forthcoming) uses digital ethnographic techniques and tools to document and chronicle the social and cultural significance of the physical and virtual communities of practice surrounding the proliferation of thousands of personal drones. The techniques used to explore these emerging communities are longitudinal digital-video ethnography, field-notes, and semi-structured interviews alongside a cultural studies "Circuit of Culture" analytical framework (Du Gay, et al., 2013). Utilising digital ethnography provides a means to illuminate and obtain a visceral understanding of the drone

cultures that are emerging as humans and increasingly smart machines interact and develop new modes of co-performance, and as means to disentangle incumbent narratives.

Consumer drones are best thought of as an assemblage comprised of a human pilot interacting with the machine elements of the drone platform itself, yet the assemblage also comprises other inter-linked socio-cultural factors such as public perception and media representation, the regulatory landscape, and models of consumption and production. By de-centering the digital and utilising an ethnographic approach to document and study the communities of practice that form around these assemblages, it has been revealed that drone cultures are segregated and differentiated by where the locus of agency resides within the assemblage, or, to what degree automation and assistive technologies play a role. FPV (first person view) pilots, for example, wear head-up displays to fly bespoke camera mounted drone assemblages with low levels of automation, and as such the skills of the human pilot are paramount.

Conversely, the communities of practice centered around using drones for recreational purposes such as photography or videography, where the data stream is of more importance than the act of flight, tend to operate drones with extremely high levels of automation, increasingly sophisticated sensor suites and embedded AI. It is these communities in particular that are increasingly allowing and trusting the hardware in the assemblage to take ownership of both agency and appropriateness. Agency in the sense that many basic operations of flight, including take-off, landing, returning to home, collision avoidance and other areas of potential human error are now handled by the drone hardware and software itself; and appropriateness in the sense that the drone through geo-location and referencing regularly updated software databases of no-fly zones and other airspace restrictions, will limit its flight operations accordingly. This is in part an acknowledgment that the hardware and software elements have greater contextual and situational awareness than the human elements of the assemblage, with the human relegated to an almost secondary role accordingly.

Longitudinal ethnographic video studies clearly highlighted how these communities of practice form, function and evolve as the drone technologies themselves mature. By engaging with both the physical communities of practice 'in the field' and through participation with the many virtual communities of practice in the digital realm via blogs and forums, one gains a unique perspective and first-hand insights into the complexities and actualities surrounding human interactions with increasingly autonomous, intelligent and data driven machines, and the practices and discourses – or cultural activity – that circulate around them. As Fish (2019) observed, to come to a more realistic notion of what the drone is, does and why it matters, one needs a synthesis of ethnographic, epistemological and ontological scholarly perspectives.

Considering the use of personal drone assemblages as a socially and culturally embedded, skilful practice, performed at a particular moment in time and situated in a specific physical

context, and by paying attention to the surrounding social and material interactions, requires an ethnography of a technology in use, or technography. Through adopting a technographic perspective one assumes that outcomes in a given situation will be emergent properties, determined by the specific context; temporal, institutional, geo-spatial and social. This means incumbent narratives about what a technology is for, or how it should be configured, need to be set aside and instead attention is paid to the discourses and practices of all the people and organisations involved in the activity as all have some degree of influence over how the technological assemblage works. Ethnographic methods reworked for hybrid digital communities support a reflexive style of conception and analysis with the researcher becoming fully engaged with the physical and digital lives of the participants and thereby achieves an understanding that is inaccessible to those who insist on remaining neutral and distant.

4.2. Case Study 2: Political Microtargeting

This case study combines conventional ethnographic methodology with an analysis of the emerging, and intensely digital, phenomenon of political microtargeting [PMT]. 'Culture' is central to the analysis because identifying, reinforcing and amplifying identity conflict (Kreiss, 2017) and racial prejudice (Shaw, 2019) have been central elements of contemporary PMT strategies.

A recent study on political microtargeting (Mount, 2000) revealed how ethnopolitics can be leveraged to achieve electoral influence. Ethnic minorities may become disenchanted with the electoral system if they experience prolonged 'invidious treatment' (Gurr, 1989) and ineffective power sharing (Horowitz, 1985). Democracy privileges the will of majority which *can* systematically marginalise political and cultural minorities. Capturing the votes of ethnic demographics ('Latino', 'Jewish', 'Afro-American', 'Afro-Carribean', 'Asian') has come to form a key strategic element of contemporary US and UK electoral campaigns. Conversely, strategies that actively target feelings of displacement and resentment against 'other' or 'foreign' communities in behalf of a besieged 'White' status quo has also emerged as a powerful electoral strategy. It was this later strategy that characterised the successful campaigns of Brexit and Trump (see Kreiss, 2017; Haynes, 2019; Shaw, 2019).

Bennet (2015) has identified four Big Data trends in contemporary democratic politics that have accelerated the process of digitising the electoral process. (i) Voter databases have become integrated into interactive voter management platforms; (ii) Election campaigns have also shifted from mass-messaging to tailored micro-targeting employing profiled data from commercial data brokerage firms; (iii) Enhanced social media analytics allow for messaging to respond to trends in real timeand; (iv) Data analysis become decentralised and mobile allowing campaigns to adopt hyper local electoral strategies.

The International Institute for Democracy and Electoral Assistance (IDEA) produced a useful report on how Big Data has significantly transformed not only the amount and type of data that is gathered, but also how it is utilised.

Table 1. How big data has changed traditional targeting into microtargeting (IDEA, 2018)

Traditional targeting	Digital microtargeting
Collecting data	Increased availability of big datasets: collected by parties themselves, government agencies, polling agencies, voter files, as well as consumer data purchased from commercial market research firms Data can be collected more easily: citizens' personal information can be reached more readily online, as can their digital footprint Data can be stored more easily through larger servers. For example, US President Donald Trump's election campaign had 'more than 300 terabytes of data' (Halpern 2017)
Dividing voters into segments based on characteristics such as personality traits, interests, background, or previous voting behaviour	'Predictive analytics': patterns can be recognized more easily with the use of complex algorithms 'Psychological targeting': squaring voter data collected by political parties with consumer data purchased from commercial market research firms; this helps to build a more detailed profile: what people buy, eat or watch in some cases can help to predict how they vote. The impact of psychological targeting is being debated.
Designing personalized political content for each segment	'A/B testing': sending out hundreds of thousands of slightly different versions of the same message to different population segments to test patterns in their responses, such as how quickly they click, how long they stay on a page, what font and colour layout they like
Using communication channels to reach the targeted voter segment with tailor-made messages	Pairing voter profiles with social media user data to reach the right people with the right message

Political microtargeting was effectively used in both of Obama's Presidential campaigns and had a decisive impact on the UK Brexit Referendum and US Trump Presidential victory in 2016. Notoriously, Cambridge Analytica was heavily involved in both campaigns and it was later revealed that they had improperly harvested over tens of millions of Facebook accounts to build and target voter profiles (Confessore, 2018). In their website, Cambridge Analytica promised depth of experience with a new "pyschographic" methodology of voter profiling:

> We bring together 25 years' experience in behavioral change, pioneering data science, and cutting-edge technology to offer unparalleled audience insight and engagement services and products (Cambridge Analytica , 2016).

A number of recent studies are now tentatively concluding that PMT may be less influential in changing voter behaviour (Dobber, 2017; Zarouali, 2020). The reasons for this are not

clear. It could be that the demos has become inoculated from these tactics; or that the techniques are only effective among certain subsets of the electorate such as "angry", "fearful" or "swinging" voters.

Applying Pink's et.al. (2016) five criteria of digital ethnography could help to understand and interpret these discrepancies by analysing the complex evolving and adaptive influence of political microtargeting in contemporary democracies. Microtargeting needs to be understood through a *multiplicity* of disciplinary lenses and from different points of view. Big data and the analytic tools utilised on them has given micro-targeting a distinctive character; but it was used as a 'trigger' to agitate cultural bias, fears, suspicions, resentment and broader sociopolitical anxiety for the purpose of enhancing reactionary, popularist, and nativist political ideologies. Microtargeting is certainly an *open* problem. Microtargeting is *reflexive* because monitoring of electoral beliefs, defines the scope and scale of public discourse and thereby actively transforms the electoral landscape.

The study of political microtargeting in a cultural cointext requires new interdisciplinary theories and methodologies. Sociological concepts will need to be adapted in unorthodox ways to explore a new 'logic of accumulation' in big data ecosystems. Shoshana Zuboff's notion of the "Big Other" requires analysis of 'often illegible mechanisms of extraction, commodification, and control that effectively exile persons from their own behavior' (2015: 75). Strategies designed to manipulate and distort need to be carefully scrutinised with these innovative and advanced methodologies.

5. Conclusion

By adapting well established ethnographic techniques to these examples, it is apparant that the enmeshment of subject-observer-participant becomes even further entwined by commonality of the increasingly digital technologies and methods that are shared by both researchers and informants alike. In effect, it is the digital transformation and application of ethnographic techniques that further dissolves boundaries between researcher and subject. By necessity, ethnographic methods must continually adapt in order to interrogate our digitally transfomed and transforming society to the extent – to paraphrase Fish – that our digital methods become entangled with the technologies, landscapes, research subjects, data and practices being studied and analysed.

References

Beesley, D. (*forthcoming*) *Head in the Clouds: documenting the rise of personal drone cultures*. Doctoral Dissertation. RMIT Melbourne.

Bennett, C. (2015) 'Trends in Voter Surveillance in Western Societies: Privacy Intrusions and Democratic Implications.' *Surveillance & Society*, no. 3/4, pp. 370-389

Cambridge Analytica, (2016) "About Us", *Internet Archive,* Captured 16 Feb 2016:

Confessore, N. (2018, 4 April) "Cambridge Analytica and Facebook: The Scandal and the Fallout So Far", *New York Times*

Dobber, T., et.al. (2017). Two crates of beer and 40 pizzas: The adoption of innovative political behavioural targeting techniques. *Internet Policy Review,* 6(4), 1–25

Du Gay, P., Hall, S., Janes, L., Madsen, A. K., McKay, H., Negus, K., & Open University. (2013). *Doing cultural studies: the story of the Sony Walkman* (Second ed.). London: Sage Publications.

Fish, A. R. (2019). Drones: Visual Anthropology from the Air. En *Handbook of Ethnographic Film and Video* (págs. 247-255). Routledge.

Graham, Timothy and Jay Daniel Thompson (15 March 2022) "Russian government accounts are using a Twitter loophole to spread disinformation", *The Conversation:*

Gurr, T. R., & Scarritt, J. R. (1989). Minorities Rights at Risk: A Global Survey. *Human Rights Quarterly*, 11(3), 375–405.

Haynes, T., et.al. (2019). *Brexit : The Uncivil War*. BBC.

Horowitz, D. L. (2014). Ethnic Power Sharing: Three Big Problems. *Journal of Democrac*y, 25(2), 5–20.

International Institute for Democracy and Electoral Assistance (2018), *Digital Microtargeting: Political Party Innovation Primer 1*

Kreiss, D. "Micro-targeting, the quantified persuasion". *Internet Policy Review,* 6 (4), 2017

Manovich, L., & Arielli, E. (18 January 2022). *Who is an artist in "software" era?*

Markham, A. N. (8 December 2020). *Doing Ethnographic Research in the Digital Age.*

Mount, G. (2020). Microtargeting and Ethnic Tension. *Oceanic Conference on International Studies.* Canberra.

Mount, G., & Beesley, D. (2022). Swarm Hermeneutics: A preliminary review. *(forthcoming).*

Pink, S., Horst, H., Postill, J., Hjorth, L., Lewis, T., & Tacchi, J. (2016). *Digital Ethnography: Principles and Practice.* London: SAGE Publications Ltd.

Ricoy-Casas, R. M. (2022) "Use of Technological Means and Personal Data in Electoral Activities: Persuasive Voters", in Á. Rocha et.al. (eds) *Communication and Smart Technologies*. Springer,

Rogers, R. (2013). *Digital Methods.* Cambridge: MIT Press.

Shaw, M, (9 January 2019) "Vote Leave relied on racism. Brexit: The uncivil war disguises that", *The Guardian*

Strategic Communications Laboratories (SCL), (2018) "Home", Internet Archive, 20 Mar 2018

Zarouali B, et.al. (2020) "Using a Personality-Profiling Algorithm to Investigate Political Microtargeting: Assessing the Persuasion Effects of Personality-Tailored Ads on Social Media". *Communication Research*. October 2020

Zuboff, S. (2015) "Big Other: Surveillance Capitalism and the Prospects of an Information Civilization", *Journal of Information Technology* 30, pp. 75-89

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15085

Applying NLP techniques to characterize what makes an online review trustworthy

José Carlos Romero[1], María Olmedilla[2], María Rocío Martínez-Torres[3], Sergio Toral[4]
[1]Department of Computer Architecture, University of Malaga, Spain, [2]SKEMA Business School, France, [3]Facultad de Ciencias Económicas y Empresariales, University of Seville, Sevilla, Spain [4]E. S. Ingenieros, University of Seville, Sevilla, Spain.

Abstract

Users spend a significantly amount of time reading and exchanging reviews online in e-commerce and eWOM communities that help them with their purchase decisions. Source credibility theory is gaining more importance as some online reviews are currently being damaged by those fake reviews that promote an untruthful image not only of the products but also of those online websites. Thus, trustworthiness of online reviews is a key aspect not only for the users that want to make more informed decisions regarding the products, but also for the websites whose credibility might be affected. In this regard, this study proposes a classification system using two Natural Language Processing (NLP) models that can predict trustworthy online reviews (helpful and truthful) applied to the product category "Cell phones & accessories" of Amazon. After using a keyword extractor among those trustworthy online reviews we can characterize their most important features. The results reveal that those features are related to brands, physical and technical features and the UX of the mobile phones.

Keywords: *Source credibility; trustworthiness; helpfulness; online reviews; classifier; Natural Language Processing.*

1. Introduction

The importance of source credibility has been extensively investigated for decades. In this regard, Joshua et al. (1986) investigated the relationship between the constructs of trust and credibility. Their experiments were based on the automobile example from Wart and McGinnies (1980) and the trustworthiness manipulation was based on the Kelly's (1973) ideas. The authors used the level of expertise of the source to show that a more trustworthy source is more credible than a less trustworthy.

Currently, the widespread of the Internet has changed the way messages are exchanged and along with it is the electronic word-of-mouth (eWOM) communication, which has changed the way in which users evaluate the credibility of information (Utz et al., 2012). Thus, most studies have focused on how to cope with source credibility in the context of online reviews (Chakraborty, 2019; Hsieh & Li, 2020). Besides, since the rise of social media, assessing perceptions of source credibility among online reviews has become more important (Metzger et al. 2010). Because online reviews are still playing an important role as information source in the consumer decision making process (Shan, 2016). However, given the decontextualization and anonymity of eWOM the concept of source credibility might seem unclear to consumers (Xie et al. 2011). Likewise, much information exchanged across social media is suspicious or sometimes malicious (Zhang & Ghorbani, 2020). In this respect, recent research works show that AI and ML techniques are playing a vital role in detecting online malicious content (Ahmed et al., 2018; Kaliyar et al., 2020; Ozbay and Alatas, 2020).

Consequently, this paper aims at investigating the relationship between the constructs of trust and credibility applying AI techniques to online reviews. First, we statistically model the online review helpfulness as a component of the construct of source credibility. Then, we make an experiment in which the trustworthiness of the online reviews is characterized by combining two NLP classifiers. One of the NLP classifiers predicts the online helpful reviews, and the other one predicts the truthful online reviews. The common online reviews of both classifiers are called trustworthy reviews. We use a keyword extractor to further characterize the features that make an online review trustworthy.

This paper is structured as follows. Section 2 presents the performed methodology for this work including the followed steps and the data analysis. Section 3 describes the obtained results summarized in three main steps. Finally, Section 5 finishes with the conclusions of the study.

2. Methodology

Figure 1 shows the schematic of the developed methodology.

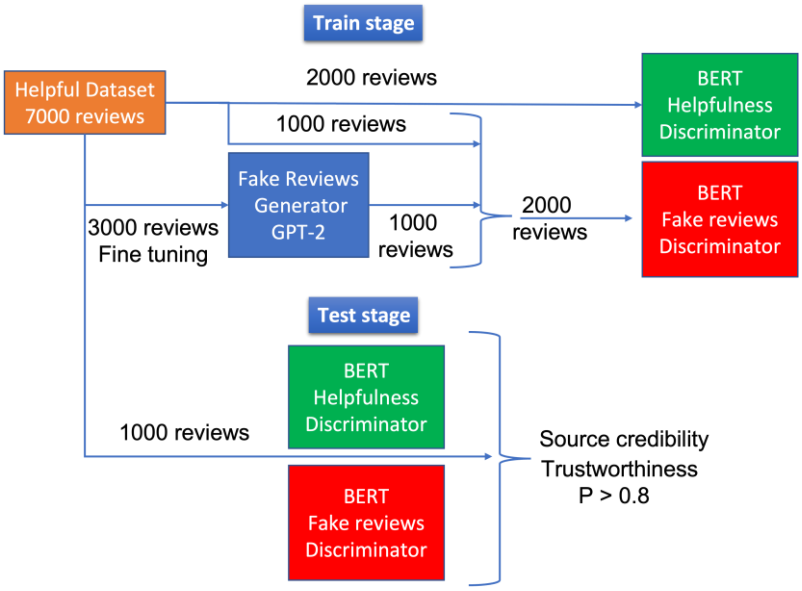


Figure 1. Scheme of the methodology

2.1. Data extraction and preprocessing

The original dataset has been extracted from Ni et al. (2019), which provides many reviews from different types of products. For this work we have selected a specific dataset, *"Cell phones and accessories*" (1,128,437 reviews). The dataset has been preprocessed to filter and to keep only the useful reviews. Firstly, the reviews with no votes have been deleted. Secondly, after analysing the dataset, we found that the reviews with a date prior to 2012 generally had lower number of votes. Thus, to have a variety of values among the votes, we have filtered and kept only the reviews with a date after 2012. Finally, we have filtered again and kept only the reviews with a verified user to ensure the trustworthiness of the review. After this preprocessing, the dataset has been ordered in quartiles according to the number of votes. The reviews with two votes or below have been considered as *not helpful*. The reviews with nine votes or more have been considered as *helpful*. The final *"helpful dataset"* has been generated by extracting 3500 *not helpful* reviews and 3500 *helpful* reviews. The dataset has been used for two purposes: (1) to provide an input of *"truthful dataset"* for the fake reviews generator and (2) to provide a *"helpful dataset"* for the BERT Helpfulness Discriminator.

2.2. Fake reviews generation using GPT-2

After the preprocess of the original dataset, the fake reviews generator has been trained and used to generate a *"fake dataset"*. We have used the GPT-2 transformer model developed by

Radford et al. (2019). The GPT-2 model has been pre-trained to generate realistic texts. The model only needs to be fine-tuned to generate texts on a specific topic. We have fine-tuned the model using 1000 steps with an average training loss of 0.92 at the end of the execution. 3000 reviews from the *"helpful dataset"* that were generated in the previous step have been used as training data. Based on all of this, our customized model is able to generate fake reviews on the product category "Cell phones and accessories".

Using this model, we have generated 1000 fake reviews. Afterwards, we have created a new dataset made of the newly generated fake reviews and 1000 truthful reviews from the *"helpful dataset"*. Then, we have added a new column assigning a 0 if the review was truthful or 1 if it was fake. This dataset called *"fake-truthful dataset"* will be further used for the classifier step.

2.3. Fake and Helpfulness reviews classifiers

Once the *"helpful"* and *"fake-truthful"* datasets have been created, the next step was to train the classifiers to predict the helpful and the truthful reviews. The classifiers used in this work are based on the BERT model developed by Devlin et al. (2018). We have trained two different classifiers: one to predict helpful reviews (BERT-Helpful) and the other one to predict truthful reviews (BERT-Truthful). BERT-Helpful has been trained using 2000 reviews from the *"helpful dataset"*, which were truthful reviews combining helpful and not helpful reviews. BERT-Truthful has been trained using 2000 reviews from the *"fake-truthful"* dataset.

Once both models have been trained and validated, we have generated predictions using the same dataset for both models in the test stage. The dataset contained 1000 fake and truthful reviews, which were not used for training the classifiers. The objective was to combine both classifiers to determine the reviews that were at the same time helpful and truthful, the so-called trustworthy reviews. BERT-Helpful predicts a specific range of reviews as helpful. BERT-Truthful also predicts another range of reviews as truthful. Our goal is to merge both ranges of reviews, helpful and truthful, to find the ones contained in there.

Finally, the pool of reviews has been analyzed, characterizing the features that make a review trustworthy. This characterization has been made using *KeyBert*, developed by Grootendorst (2020). This tool allows the extraction of keywords for BERT models. It can be customised to extract keywords of different lengths starting in 1 (keywords or keyphrases) or to set the diversity of keywords we want to obtain.

3. Results

3.1 Experimental setup

The experimental evaluation has been conducted using the free version of *Google Collab*. The resources provided by the free version are a single node with one GPU Tesla K80 and 64 GB of RAM. Within this version, the availability of resources are not guaranteed and limited, and sometimes the usage limits fluctuates depending on the demand. The code has been developed using *Python3* and pandas library for the datasets management and processing. The GPT-2 transformer model has been applied with the *gpt-2-simple* library[1], which uses *TensorFlow* and wraps existing model fine-tuning and generates texts for GPT-2. It uses specifically the "small" 124M and "medium" 355M hyperparameter versions. We have used the "small" 124M version of the model for this work. Thus, the model can be handled by our system. The BERT classifier has been used with a version developed in *PyTorch*[2].

3.2 Classifiers performance

BERT-Helpful has been trained in 20 epochs with a training loss of 0.384 and a validation loss of 0.268. BERT-Truthful has been trained in 20 epochs with a training loss of 0.378 and a validation loss of 0.283. The testing for the classifiers has been made using the same dataset combining 1000 fake and truthful reviews. The accuracy is 0.9 for BERT-Helpful classifier and 0.91 for BERT-Truthful.

3.3. Trustworthy reviews extraction and analysis

Once the predictions have been generated, the next step is to merge results and extract the common reviews that both classifiers have to discover the trustworthy reviews. The default classifier uses a threshold of 0.5 to classify the data: whether it is helpful or not, or whether it is truthful or not. We wanted to ensure that our data was truly helpful and truthful. So the threshold needs to be adjusted.

Figure 2 explores different thresholds applied to the classifiers and the number of trustworthy reviews we got. It can be observed that as the threshold increases, the number of trustworthy reviews decreases proportionally. The higher the threshold, the higher the number of trustworthy reviews. We have fixed the threshold at 0.8, since at 0.9 the number of reviews decreases too much.

[1] https://github.com/minimaxir/gpt-2-simple

[2] https://github.com/prateekjoshi565/Fine-Tuning-BERT

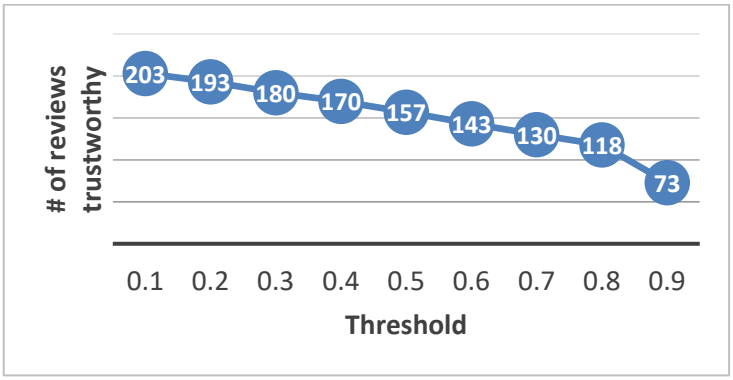


Figure 2. Exploration for different # of reviews trustworthy per threshold applied

Finally, in order to facilitate a deeper interpretability of results, we have gotten the main keywords that are most similar to a document from the 118 trustworthy reviews using the tool *KeyBert.* After doing a qualitative analysis of those extracted keywords we have identified that the content of trustworthy reviews has to do with mobile phones brands (e.g., Samsung Galaxy, iPhone), accessories brands (e.g. otterbox), the physical features of the mobile phone (e.g., long, size, sleek, weight, etc.), the physical features of the mobile phone accessories (e.g., quality of glass screen protector, adaptability of cases to the mobile phones), the technical features of the mobile phone (e.g., sound quality, mobile network coverage/signal, hardware performance) and about the UX (e.g., screen touch responsiveness).

4. Conclusions

In this work we have developed a methodology to detect and characterize trustworthy reviews within the product category "Cell phones and accessories". We have generated fake reviews using the GPT-2 model to train a classifier to detect truthful reviews. Likewise, a second classifier has been trained to detect helpful reviews. We obtain a higher accuracy, around 0.9, for both classifiers and 118 trustworthy reviews from a dataset of 1000 reviews. Finally, the keywords allow us to characterize some of the trustworthy reviews' common patterns for this product category, thus understand the content of those reviews.

One limitation of this work was the limited resources available, thus it was necessary to reduce the number of data used to generate our results. Future work could incorporate more powerful systems able to analyse much larger datasets from different product categories. Moreover, the detection and analysis of trustworthy reviews could be improved.

References

Ahmed, H., Traore, I., & Saad, S. (2018). Detecting opinion spams and fake news using text classification. Security and Privacy, 1(1), e9.

Chakraborty, U. (2019). The impact of source credible online reviews on purchase intention: The mediating roles of brand equity dimensions. Journal of Research in Interactive Marketing.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Grootendorst, M. (2020). KeyBERT: Minimal keyword extraction with BERT(Version v0.3.0). Version v0.3.0. doi:10.5281/zenodo.4461265

Hsieh, J. K., & Li, Y. J. (2020). Will you ever trust the review website again? The importance of source credibility. International Journal of Electronic Commerce, 24(2), 255-275.

Joshua L. Wiener and John C. Mowen (1986) ,"Source Credibility: on the Independent Effects of Trust and Expertise", in NA - Advances in Consumer Research Volume 13, eds. Richard J. Lutz, Provo, UT : Association for Consumer Research, Pages: 306-310.

Kaliyar, R. K., Goswami, A., Narang, P., & Sinha, S. (2020). FNDNet–a deep convolutional neural network for fake news detection. Cognitive Systems Research, 61, 32-44.

Kelly, H. (1973), "The Process of Causal Attribution," American Psychologist, 28, 107-128.

McGinnes, E. and C. Ward (1980), "Better Liked Than Right: Trustworthiness and Expertise in Credibility,-Personality and Social Psychology Bulletin, 6, 467-472.

Metzger, M. J., Flanagin, A. J., and Medders, R. B. Social and heuristic approaches to credibility evaluation online. Journal of Communication, 60, 3, 2010, 413–439.

Ni, J., Li, J., & McAuley, J. (2019, November). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 188-197).

Ozbay, F. A., & Alatas, B. (2020). Fake news detection within online social media using supervised artificial intelligence algorithms. Physica A: Statistical Mechanics and its Applications, 540, 123174.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Xie, H. J., Miao, L., Kuo, P. J., & Lee, B. Y. (2011). Consumers' responses to ambivalent online hotel reviews: The role of perceived source credibility and pre-decisional disposition. International Journal of Hospitality Management, 30(1), 178-183.

Shan, Y. (2016). How credible are online product reviews? The effects of self-generated and system-generated cues on source credibility evaluation. Computers in Human Behavior, 55, 633-641.

Utz, S., Kerkhof, P., & Van Den Bos, J. (2012). Consumers rule: How consumer reviews influence perceived trustworthiness of online stores. Electronic Commerce Research and Applications, 11(1), 49-58.

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management, 57(2), 102025.

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15087

Emergency Calls in the City of Vaughan (Canada) During the COVID-19 Pandemic: A Spatiotemporal Analysis

Ali Asgary[1], Adriano O. Solis[2], Nawar Khan[1], Janithra Wimaladasa[1], Maryam S. Sabet[3]
[1]Disaster & Emergency Management Area, School of Administrative Studies, York University, Canada, [2]Decision Sciences Area, School of Administrative Studies, York University, Canada, [3]Fleming College – Sutherland Campus, Canada.

Abstract

The COVID-19 pandemic has required governments to introduce various public health measures in order to contain and manage the pandemic's unprecedented impacts in terms of illnesses and deaths. This study analyzes the spatiotemporal distribution of emergency incidents in Vaughan, a medium-sized city in the Canadian province of Ontario, comparing occurrences prior to and during the pandemic. Emergency calls received and responded to by the Vaughan Fire and Rescue Service were examined using spatial density and emerging hotspot analysis based on 11 periods of various public health measures and restrictions set in place from 17 March 2020 to 15 July 2021, as compared with corresponding pre-pandemic periods in the preceding three years (2017-2019). The resulting analyses show significant spatiotemporal changes in emergency incident patterns, particularly during periods of more stringent public health measures such as 'stay at home' orders or lockdowns of nonessential business establishments. Results of the study could provide useful insights for managing emergency service resources and operations during public health emergencies.

Keywords: *Spatiotemporal analysis, COVID-19 pandemic, emergency calls, kernel density, emerging hot spot analysis, City of Vaughan.*

1. Introduction

Since the beginning of the COVID-19 pandemic in Canada in March 2020, regions and cities/municipalities in the Province of Ontario have gone through several stages of public health measures, including closures of public and sports facilities, schools, places of worship, and nonessential businesses. Using standard temporal analysis methods, Solis et al. (2022) observed dramatic downward shifts in frequencies of certain types of emergency calls received by the fire and rescue service of Ontario's City of Vaughan during the first ten months of the COVID-19 pandemic. The drops in emergency incidents have appeared to be consistent with public health measures set in place by government authorities at different stages of the pandemic.

Spatiotemporal methods have been rapidly developing in the field of data analytics, and have been applied in the analysis of emergency incidents for some time. With advances in technology, various new methods have been developed and used (Špatenková and Virrantaus, 2013; Yao and Zhang, 2016; Shafiei Sabet et al., 2019; Yao et al., 2019). Emerging hotspot analysis is among new methods added to GIS-based analyses and its usage in spatiotemporal analysis has been growing (Adepeju et al., 2016; Gudes et al., 2017; Rabiei-Dastjerdi and McArdle, 2020; Hart, 2021). Recent studies have sought to examine spatial patterns, some including emerging hotspot analysis, of COVID-19 cases in various contexts (Andersen et al., 2020; Mollalo et al., 2020; Mylona et al., 2020; de Cos et al., 2021; Purwanto et al., 2021).

Understanding how the pandemic impacts spatial and spatiotemporal distributions of emergency incidents during different pandemic stages may provide useful insights for fire and rescue service decision makers in managing resources as public health measures evolve over time. The current study is aimed at examining the spatiotemporal distribution, using spatial density and emerging hotspot analysis, of emergency calls in Vaughan during different COVID-19 phases and corresponding public health measures. To the best of our knowledge, this is the first study to examine the spatiotemporal patterns of emergency call variations during different phases of the pandemic.

2. City of Vaughan and Dataset

The City of Vaughan is one of nine municipalities in the Regional Municipality of York (also known as York Region) of the Canadian province of Ontario. Vaughan, which has a land area of 273.56 square kilometers, currently has an estimated population of close to 341,000 (City of Vaughan, 2022). It is situated just north of the City of Toronto, which is the provincial capital of Ontario and the largest Candian city in terms of population.

As of January 1, 2019, the Vaughan Fire and Rescue Service (VFRS) operated with ten fire districts and corresponding fire stations 7-1, 7-2, …, 7-9, and 7-10 (Figure 1). VFRS

maintains a Standard Incident Report dataset in which every single emergency call is recorded using a unique incident ID number and with specific attributes including incident type, longitude and latitude coordinates of the incident location, alarm date/time, responding station, dispatch date/time, arrival date/time, etc. (Office of the Fire Marshal of Ontario, 2009). In the three calendar years immediately preceding the pandemic, there were 11,331 emergency incidents in 2017, 11,834 in 2018, and 11,313 in 2019. Compared to the 2017-2019 annual average of just under 11,493 incidents per year, the total number dropped by 12.7% to only 10,037 incidents in 2020.

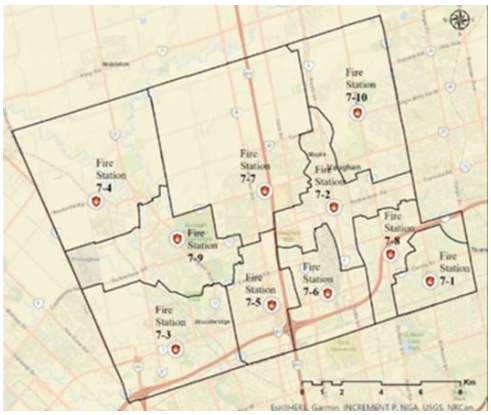


Figure 1. Fire districts and stations, City of Vaughan (as of 01 January 2019)

Table 1 summarizes the first 11 periods of the COVID-19 pandemic for the City of Vaughan. The 11 periods are based largely upon public health measures in effect in the province of Ontario and in York Region.

Table 1. First eleven COVID-19 periods applicable to the City of Vaughan

Period	# of days	Brief description of period
Period 1 (17 March - 18 May 2020)	63	State of Emergency I (lockdown began)
Period 2 (19 May - 18 June 2020)	31	Stage 1 reopening
Period 3 (19 June - 23 July 2020)	35	Stage 2 reopening
Period 4 (24 July - 18 October 2020)	87	Stage 3 reopening
Period 5 (19 October - 13 December 2020)	56	Modified Stage 2 reopening
Period 6 (14 December 2020 - 13 January 2021)	31	Lockdown
Period 7 (14 January - 21 February 2021)	39	State of Emergency II
Period 8 (22 February - 02 April 2021)	40	York Region as a 'red zone'
Period 9 (03 April - 10 June 2021)	69	'Stay-at-home' order; initially 'emergency brake' order
Period 10 (11 June - 29 June 2021)	19	Step 1 of Province of Ontario's Roadmap to Reopen
Period 11 (30 June - 15 July 2021)	16	Step 2 of Province of Ontario's Roadmap to Reopen

3. Findings

3.1. Emergency Incidents Before and During the Pandemic

Table 2 summarizes the total numbers of emergency calls received by VFRS during the 11 periods, in comparison with corresponding periods during the three years (2017-2019) immediately preceding the pandemic. With the exception of Period 9, total numbers of emergency incidents have declined during all 11 COVID-19 periods under study compared to corresponding periods in the pre-pandemic years 2017-2019.

Table 2. Emergency incidents during COVID-19 periods 1-11 vs. corresponding periods in 2017-2019

Period	# of Days	2017	2018	2019	Average 2017-2019	During COVID-19
Period 1 (17 March - 18 May 2020)	63	1,933	2,013	1,821	1,922.3	1,476
Period 2 (19 May - 18 June 2020)	31	989	1,073	974	1,012.0	873
Period 3 (19 June - 23 July 2020)	35	1,094	1,157	1,139	1,130.0	986
Period 4 (24 July - 18 October 2020)	87	2,790	2,829	2,705	2,774.7	2,466
Period 5 (19 October - 13 December 2020)	56	1,699	1,822	1,694	1,738.3	1,570
Period 6 (14 December 2020 - 13 January 2021) *	31	1,223	861	957	1,013.7	854
Period 7 (14 January - 21 February 2021)	39	1,084	1,208	1,363	1,218.3	1,044
Period 8 (22 February - 02 April 2021)	40	1,196	1,153	1,171	1,173.3	1,104
Period 9 (03 April - 10 June 2021)	69	2,122	2,292	2,099	2,171.0	2,236
Period 10 (11 June - 29 June 2021)	19	654	628	589	623.7	549
Period 11 (30 June - 15 July 2021)	16	501	517	515	511.0	484

* Numbers reported for Period 6 in 2017, 2018, and 2019 are for 14 December 2017 - 13 January 2018, 14 December 2018 - 13 January 2019, and 14 December 2019 - 13 January 2020, respectively.

Among all major incident types (e.g., property fires/explosions, medical emergencies, vehicle collisions/extrications, false fire calls), average daily occurrences of vehicle collisions and extrications during COVID-19 periods exhibited the most dramatic percentage changes (decreases) compared to the same period during pre-pandemic years (Figure 2).

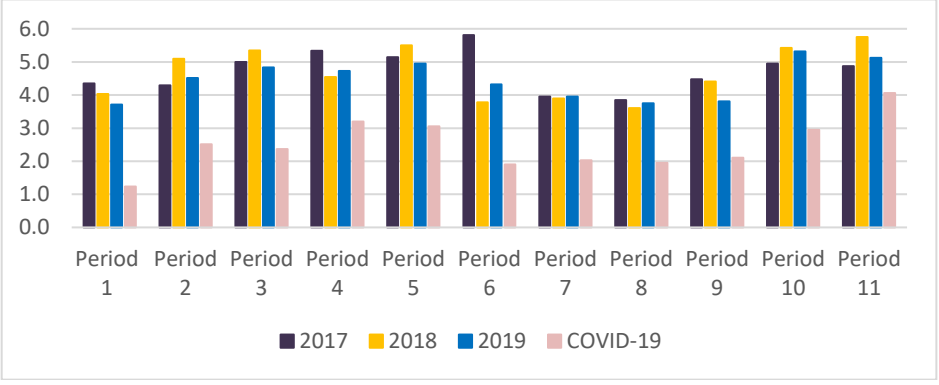


Figure 2. Vehicle collisions/extrications by period: Average per day

3.2. Density Analysis

To analyze the data on emergency incidents, we initially performed kernel density analysis using ArcGIS Pro 2.8 (ESRI, 2021) to examine changes in the spatial distribution of emergency calls (total and by major incident type) before and during various periods/stages of the pandemic. The kernel density tool calculates the density of point features around each 10 m. × 10 m. output raster cell (ESRI, 2021) based on a quartic kernel function (Silverman, 1986, eq. 4.5). Density per 1 sq. km. (1 km. × 1 km.) is evaluated.

In view of the limited space, we provide as an illustration only kernel density maps pertaining to vehicle collisions/extrications, and only for Periods 1 and 2 (Figure 3). These sample maps clearly show less colour/lighter shades in the maps for Periods 1 and 2 of the pandemic relative to the maps for the corresponding periods in the preceding year (2019, left column).

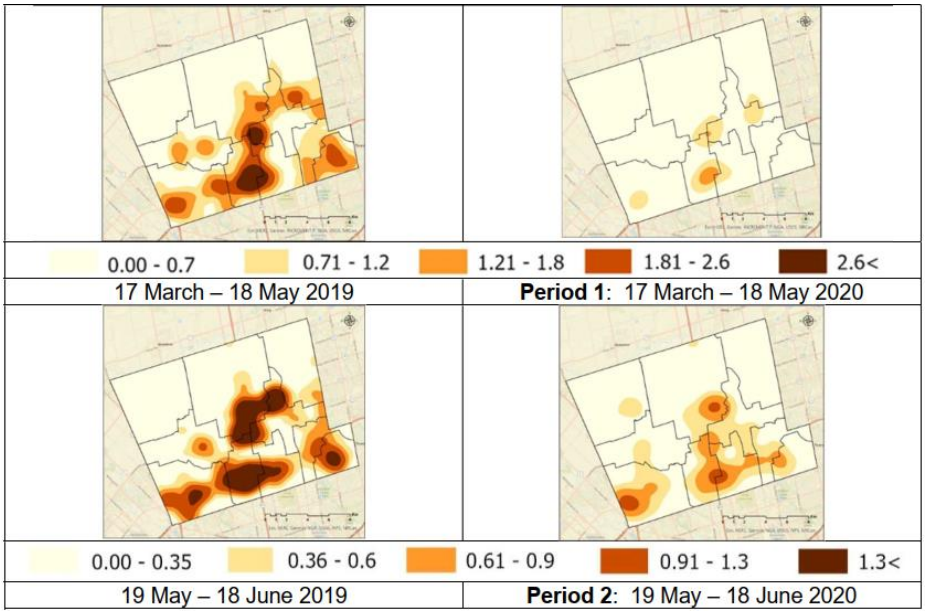


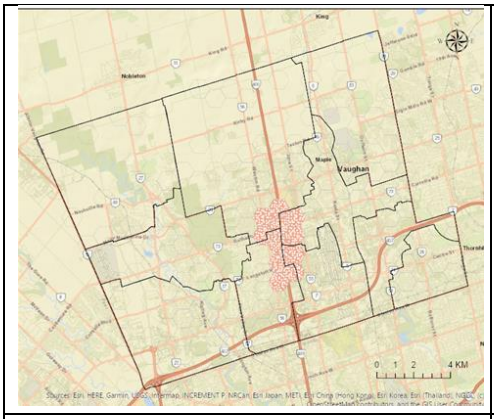
Figure 3. Kernel density maps for vehicle collisions/extrications: Periods 1 and 2

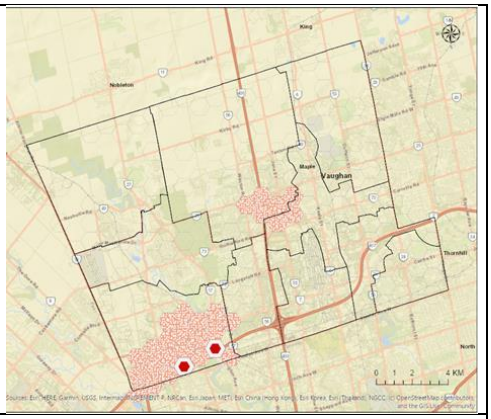
The above observation may be associated with reductions in vehicular traffic resulting from lockdowns, 'work from home', and other public health measures. With vehicle collisions occurring particularly on highways that pass through certain districts, the changes would significantly impact resource use of fire stations in districts that include the major highways.

3.3. Emerging Hotspot Analysis

We applied emerging hotspot analysis (EHA) using ArcGIS Pro 2.8 (ESRI, 2021) to understand spatiotemporal variations of emergency calls. EHA provides a summary of spatial

distribution, identifies significant clusters in the dataset, and explores patterns over time. EHA classifies the data into several patterns including: 1) 'no pattern', 2) 'new pattern', 3) 'oscillating pattern', and 4) 'sporadic pattern' (Gudes et al., 2017). Definitions/indicators of various patterns (new, consecutive, intensifying, persistent, diminishing, sporadic, oscillating, or historical hot/cold spot) are available in the ArcGIS platform (ESRI, 2021). Here we provide as an illustration a sample of EHA outputs pertaining to vehicle collisions/extrications (Figure 4).





17 March 2019 – 16 March 2020 (before pandemic) Time step Interval: 2 weeks Distance Interval: 1 km Neighborhood Distance: 2km	**17 March 2019 – 15 July 2021** (until Period 11 of pandemic) Time step Interval: 2 weeks Distance Interval: 1 km Neighborhood Distance: 2km

Figure 4. Emerging hotspot analysis before and during the pandemic for vehicle collisions/extrications

Prior to the pandemic (17 March 2019 – 16 March 2020), sporadic hotspots were concentrated mainly around the major highway (Ontario Highway 400) which cuts through the city (north-to-south). As restrictions were being lifted in Periods 10-11 and there was more movement of people on the roads, sporadic hotspots, and even a couple of new hotspots, appear to have begun to develop also in District 7-3 (closer to Highway 7).

4. Conclusion

This study has examined the geographic distributions and spatiotemporal patterns of emergency calls in the City of Vaughan during the first 11 periods of the COVID-19 pandemic and compared them with the corresponding pre-pandemic periods in 2017-2019. We believe that this is the first study to apply spatiotemporal methods in evaluating changes in the frequency and mix of emergency incidents that have been responded to by a municipal fire and rescue service over various periods of the pandemic, each period pertaining to specific public health measures/restrictions. We applied kernel density analysis and emerging

hot and cold spot analyses. The results suggest that the COVID-19 pandemic and public health measures introduced to respond to it during different periods had significant impacts on the spatiotemporal distribution of emergency incidents in the city. These may have potential implications for resource planning and allocation across fire districts/stations and provide insights on how to manage fire and rescue service operations as further stages of the pandemic unfold. Conventional data analyses can show changes to some extent, but spatiotemporal analyses enable relating such changes in space over time to further examine locational attributes that determine changing patterns in occurrences of emergency incidents. Emergency service decision-makers (in this case, those of the VFRS) can apply insights gained from the analyses in planning and management of resources – particularly the reallocation of firefighting apparatus and crews to the various fire districts/stations – in line with public health measures/restrictions associated with the latest period of the still ongoing pandemic or of similar new pandemics that may arise in the future.

Acknowledgment

This research has been conducted with financial support from the Social Sciences and Humanities Research Council of Canada (SSHRC) as part of its Partnership Engage Grants (PEG) COVID-19 Special Initiative. The Vaughan Fire and Rescue Service is the partner organization of the York University research team in this effort. The research work has been also supported by ADERSIM, funded by Ontario Research Fund (ORF).

References

Adepeju, M., Rosser, G., & Cheng, T. (2016). Novel evaluation metrics for sparse spatiotemporal point process hotspot predictions - a crime case study. *International Journal of Geographical Information Science*, 30(11), 2133-2154.

Andersen, L. M., Harden, S. R., Sugg, M. M., Runkle, J. D., & Lundquist, T. E. (2020). Analyzing the spatial determinants of local COVID-19 transmission in the United States. *Science of the Total Environment*, 754: 142396.

City of Vaughan (2022). About Vaughan. https://www.vaughan.ca/news/about_vaughan/Pages/default.aspx

de Cos, O., Castillo, V., & Cantarero, D. (2021). Scalable analysis of COVID-19 spatiotemporal patterns based on data mining tools: Using 3D bins to predict short-time focus locations. Preprint available at https://pdfs.semanticscholar.org/0788/4ca3f2b61f9fa7a874448f5416193e622c57.pdf.

ESRI (2021). ArcGIS Pro 2.8. Accessible at https://www.esri.com/en-us/home.

Gudes, O., Varhol, R., Sun, Q. C., & Meuleners, L. (2017). Investigating articulated heavy-vehicle crashes in western Australia using a spatial approach. *Accident Analysis & Prevention*, 106, 243-253.

Hart, T. C. (2021). Investigating crime pattern stability at micro-temporal intervals: implications for crime analysis and hotspot policing strategies. *Criminal Justice Review*, 46(2), 173-189.

Mollalo, A., Vahedi, B., & Rivera, K. M. (2020). GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. *Science of the Total Environment*, 728: 138884.

Mylona, E. K., Shehadeh, F., Kalligeros, M., Benitez, G., Chan, P. A., & Mylonakis, E. (2020). Real-time spatiotemporal analysis of microepidemics of influenza and COVID-19 based on hospital network data: Colocalization of neighborhood-level hotspots. *American Journal of Public Health*, 110(12), 1817-1824.

Office of the Fire Marshal of Ontario (2009). *Standard Incident Report Codes List*, issued January 2009.

Purwanto, P., Utaya, S., Handoyo, B., Bachri, S., Astuti, I. S., Utomo, K. S. B., & Aldianto, Y. E. (2021). Spatiotemporal analysis of COVID-19 spread with emerging hotspot analysis and space–time cube models in East Java, Indonesia. *ISPRS International Journal of Geo-Information*, 10(3), 133.

Rabiei-Dastjerdi, H., & McArdle, G. (2020). Identifying patterns of neighbourhood change based on spatiotemporal analysis of Airbnb data in Dublin. *2020 4th International Conference on Smart Grid and Smart Cities (ICSGSC)*. IEEE Power & Energy Society, Osaka, Japan, August 2020, 113-117.

Shafiei Sabet, M., Asgary, A., & Solis, A. O. (2019). Emergency calls during the 2013 Southern Ontario Ice Storm: Case study of Vaughan. *International Journal of Emergency Services*, 8(3), 292-314.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.

Solis, A. O., Wimaladasa, J., Asgary, A., Sabet, M. S., & Ing, M. (2022). Shifting patterns of emergency incidents during the COVID-19 pandemic in the City of Vaughan, Canada, *International Journal of Emergency Services*, 11(1), 1-37.

Špatenková, O., & Virrantaus, K. (2013). Discovering spatiotemporal relationships in the distribution of building fires. *Fire Safety Journal*, 62, 49–63.

Yao, H., Liu, Y., Wei, Y., Tang, X., & Li, Z. (2019). Learning from multiple cities: A meta-learning approach for spatial-temporal prediction. *Proceedings of WWW '19: The World Wide Web Conference*, Association for Computing Machinery, May 2019, San Francisco, CA, USA, 2181–2191.

Yao, J., & Zhang, X. (2016). Spatial-temporal dynamics of urban fire incidents: a case study of Nanjing, China. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B2, 63–69.

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15089

Implementing sentiment analysis to an open-ended questionnaire: Case study of digitalization in elderly care during COVID-19

Ida Toivanen [1], Venla Räsänen [1], Jari Lindroos [1], Tomi Oinas [1], Sakari Taipale[1 2]
[1]Faculty of Humanities and Social Sciences, University of Jyväskylä, Finland, [2] Faculty of Social Sciences, University of Ljubljana, Slovenia

Abstract

The rise of digital technology has enabled us to utilize even more integrated systems for social and health care, but these systems are often complex and time-consuming to learn for the end users without relevant training or experience. We aim to perform Named Entity Recognition based sentiment analysis using the answers of eldercare workers that have taken a survey about the effects of digitalization on their work. The collection of the panel survey data was carried out in two waves: in 2019 and 2021. For the sentiment analysis we compare these two waves to determine the effects of COVID-19 on the work of eldercare workers. The research questions we ask are the following: "Has technology affected eldercare workers' emotions in their work and how?" and "Has COVID-19 affected eldercare workers' views on digitalization in their work?". The main results suggest that criticism of modern technology persists through time – that is, before and after the pandemic the same type of negative and positive sentiments are manifested in the results. However, the familiarization with technology during COVID-19 seems to have been decreasing negative sentiments and increasing positive sentiments regarding digitalization. Due to the smallness of our data, more research should be conducted to make firmer conclusions on the matter.

Keywords: *eldercare work; digitalization; sentiment analysis; named entity recognition; BERT.*

1. Introduction

Health care and social services are usually considered to be one of the most important public services for the wellbeing of people. While assisting technologies aim to not only aid in accessing information but to form the basis for multi-tasking, the success of this process needs to be carefully examined (European Digital Agenda, 2022; Bartosiewicz et al., 2021). Although the ultimate aim for the vastly expanding digitalization is to offer better care for the patients and clients, the effects are often first experienced by the employees in social and health care services. Without proper research into the usability of the services that are intended for care professionals to operate, we may do more harm than good. Instead of decreased costs, improper incorporation of digitized tools may lead to increased costs and data security issues - but also exclusion may be prominent when trying to integrate social and health care systems in countries with low population density (Laitinen et al., 2018).

These drawbacks may be overcome more easily with carrying out user experience studies, using social and health care personnel as interviewees, that may supply more information about the difficulties of diverse types of employee groups (Ylönen et al., 2020). Especially older generations of eldercare workers find it often difficult to use information and communication technology (ICT) as a part of their work. Even more experienced and digitally skilled eldercare workers have expressed difficulties in engaging with their customers because of the growing need to report and use different technological systems. While attention has been given to this subject by several scholars (see Bartosiewicz et al., 2021; Laitinen et al., 2018; Ylönen et al., 2020; Seibert et al., 2020), the expanding need for digitalization of the social and health care sector has amplified during the COVID-19 pandemic and presents a need for continuing focus on the matter. While there are many roads to be explored, we have narrowed our approach to the elderly care sector. The research questions we ask are the following: "*Has technology affected eldercare workers' emotions in their work and how?*" and "*Has COVID-19 affected eldercare workers' views on digitalization in their work?*". Using two panel surveys, collected in two waves in 2019 and 2021, we implement a sentiment analysis to the answers to open-ended questions. Comparing the two waves we strive to make a distinction between sentiments concerning the use of technology at eldercare work before and after COVID-19.

Contemporary textual models designed for low-resource languages are still few and far between, which has prompted us to dig into the research of those languages. In this case we use Finnish language as an example to explore the difficulties (such as lack of data) of using low-resource languages in natural language processing tasks. We aim to utilize state-of-the-art language models (FinBERT) as the basis for model implementation and build up a novel use case for extended analysis.

2. Data

Our analysis is based on the first (2019) and second round (2021) of University of Jyväskylä survey on eldercare work, which is a new survey on the working conditions and digitalization of eldercare work collected by the Centre of Excellence in Research on Ageing and Care (www.jyu.fi/agecare) at the University of Jyväskylä. The aim of the survey was to collect information on the working conditions and use of ICTs (Information Communication Technologies) among eldercare workers in Finland. In this paper our focus is on textual answers for open survey questions, entailing 3971 data samples in total. We examine the answers to the open-ended question "*What kind of emotions related to the use of technology have been present in your work during the last week?*", which acts as the basis for tracking sentiments in the data. In addition to studying the feelings elicited using technology that the workers encounter in their every-day-life, it is important to conduct this type of analysis to attain information about working conditions, workload and the changes happening within their professional field.

3. Methods

In this study we refer to the analysis of emotional content by tagging sentiment vocabulary (by means of named entity recognition) as sentiment analysis. Named Entity Recognition (NER) task can be described as a sequence labeling (Virtanen et al., 2019), or token classification, task in which entities are tagged to study the way they are being used in the data. These entities can be, for example, locations, names, or dates (see Ruokolainen et al., 2020). This information extraction technique brings forth accentuated knowledge of the distinctive terminology groups there are in the data. To examine the first research question (*"Has technology affected eldercare workers' emotions in their work and how?"*), sentiment related glossary was NER tagged in the data. Example of sentiment NER tagging is visualized in Figure 1. For the second research question (*"Has COVID-19 affected eldercare workers' views on digitalization in their work?"*), we drew conclusions based on the frequencies of different NER sentiment tags and compare the frequencies between the datasets of 2019 and 2021 (Table 3). In this chapter we describe the data preprocessing, NER tagging process and model making relevant to this research.

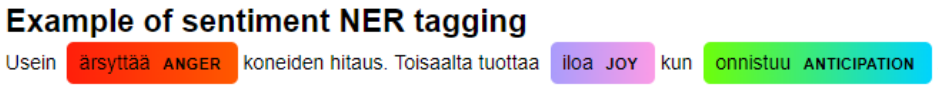


Figure 1. Example of sentiment NER tagging. The sentence translates to "Slow computers annoy (NER tag: ANGER) me often. On the other hand, it produces me joy (NER tag: JOY) when I succeed (NER tag: ANTICIPATION).".

3.1 Preprocessing and NER tagging

We preprocessed the data by removing empty and duplicate values. After preprocessing the dataset (of data from both 2019 and 2021) contains 3971 data samples. By data samples we mean one row of data that can consist of a single word, a sentence or multiple sentences. We first split the data to create a testing dataset of 10% (398 samples) of the entire dataset, to use it for the evaluation of the model. Then the remainder data was split into training (2858 samples) and validation (715 samples) datasets that were used during the model training process.

For NER tagging we used a Finnish sentiment corpus conducted by Mohammad, S. M., 2013. In the corpus there are eight sentiment classes (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust) for which we built our word lists on. From the word lists irrelevant words were excluded, and appropriate synonyms were also added to the lists. We ended up with word lists of the following sizes: 212 words for anger, 123 for anticipation, 88 for disgust, 102 for fear, 134 for joy, 98 for sadness, 72 for surprise, and 31 for trust.

To start the comparing of sentiment word lists to our data, we first lemmatized, i.e., reduced a word to its basic word form, all the data samples. Each lemma in each lemmatized data sample was compared to each semantic word in every semantic word list (angry, sad, etc.), and when the word matched with a semantic word on the semantic word list, it was NER tagged and marked to be belonging to the semantic class (e.g., angry) at hand. This is how we made a new column of NER tags to prepare the data for token classification task. Entity type dataset statistics, or the amount of NER tags of different sentiment classes that are present in different datasets, are shown in Table 1. The whole data consists of approximately 8% sentiment NER tagged words.

Table 1. Entity type dataset statistics.

Entity	Train	Valid	Test	Total
Anger	690	213	90	993
Anticipation	597	118	63	778
Disgust	98	20	13	131
Fear	163	41	17	221
Joy	371	80	56	507
Sadness	189	47	22	258
Surprise	332	77	40	449
Trust	167	27	21	215
Total	2607	623	322	3552

3.2 Model

For building a language model that is supported by a low-resource language such as Finnish, we used the currently openly available state-of-the-art model, the FinBERT base (Virtanen et al., 2019) as the backbone and build a NER model upon that. We also used a ConvBERT (Jiang et al., 2020) model variation ConvBERT base Finnish, that is pretrained on a large Finnish corpus, as the backbone. Modest hyperparameter optimization was conducted while building the models. The models were trained for 10 epochs with a batch size of 16, learning rate of 5e-5 using a linear scheduler with two warmup steps, optimizer AdamW and weight decay set as 0, and maximum sequence length of 250.

4. Results

Results for NER task were obtained with several evaluation metrics: precision, recall, F1 score and accuracy. All values obtained with the evaluation metrics are shown in Table 2. The results suggest that ConvBERT model generally produces better results than the FinBERT model. This might be because replaced token detection (RTD) objective was used in the pretraining of the ConvBERT model (Jiang et al., 2020), while BERT's pretraining is based on the masked language modeling (MLM) objective (Virtanen et al., 2019).

Table 2. Results for testing dataset that is used to evaluate the model after the model training process is concluded.

Approach	Entity	Precision	Recall	F1	Accuracy
FinBERT	**Overall**	0.7892	**0.8424**	0.8150	0.9787
ConvBERT		**0.8208**	0.8392	**0.8299**	**0.9788**
	Anger	0.8068	0.8353	0.8208	
		0.8295	**0.8588**	**0.8439**	
	Anticipation	0.7727	**0.8500**	0.8095	
		0.8065	0.8333	**0.8197**	
	Disgust	1	0.7500	0.8571	
		1	0.7500	0.8571	
	Fear	0.7222	0.7647	0.7429	
		0.8125	0.7647	**0.7879**	
	Joy	0.9273	0.9107	0.9189	
		0.9811	**0.9286**	**0.9541**	
	Sadness	**0.8000**	**0.9091**	**0.8511**	
		0.6923	0.8182	0.7500	
	Surprise	0.6400	**0.8421**	**0.7273**	
		0.6522	0.7895	0.7143	
	Trust	0.7143	0.7143	0.7143	
		0.8889	**0.7619**	**0.8205**	

Additionally, we drew a subset of data from people who filled out the survey both in 2019 and 2021. We used this subset (n=1388) to compare how sentiment contents changed over the two-year period. The frequencies between 2019 and 2021 data proved to be so small that no comprehensive conclusions can be made from the results, shown in Table 3, alone. The biggest frequency decreases were in the sadness and anger entities, where the amount of NER tags were decreased by 22.22% and 18.57%, respectively. The amount of NER tags for trust and joy were conversely increased by 75.76% and 14.95%, respectively. This could implicate that the use of technology is more common among the eldercare workers after COVID-19, resulting in lesser use of sad and anger sentiments, that may be considered as negative sentiments, while reinforcing trust and joy sentiments, that can be seen as positive sentiments.

Table 3. Frequencies of words belonging to sentiment classes in a subset drawn from data from both 2019 and 2021 data.

Entity	2019	2021
Overall	713	709
Anger	210	171
Anticipation	142	157
Disgust	34	30
Fear	40	40
Joy	107	123
Sadness	54	42
Surprise	93	88
Trust	33	58

5. Discussion

The study at hand was carried out for two reasons: training language models to understand Finnish language and finding out the possible applications to different research settings. Prospects for the implementation of NER based sentiment analysis within the research field of social science (and others) are still largely to be discovered, especially because of the lack of trained tools for low-resource languages. Although the size of our dataset is relatively small, it still provides a basis to build domain-specific sentiment analysis. Identifying and describing difficulties that healthcare professionals face while using new technology at work is crucial to improve processes of software development and application in the future. Applying our approach could be used to rate the success of existing deployments – that is, to determine how accessible and usable the systems under evaluation really are.

Our preliminary results suggest ConvBERT model performs generally better than FinBERT model for our data. Additionally, we can make observations of the sentiment NER tag frequencies, that have changed from 2019 to 2021, that would suggest that COVID-19 has

made eldercare workers familiarize themselves more with technology. More research with bigger data should be done to draw more precise conclusions. As our study is a retrospective analysis of these views, reflecting on them may provide useful results to demonstrate how more rigorous approaches to usability already ingrained in the research process are needed. For example, Jokela and Polvi (2010) present a way to conduct a test study by refining the usability requirements with as little ambiguity as possible. This was done by setting up a target level, a confidence value of 95 % that at least 75 % of the users will succeed in a task, that was supposed to be fulfilled or else they would have fallen short on the requirements. Jokela and Polvi (2010) make the conclusion that instead of asking about cosmetic details of the interface from test users, putting them under a test that is based on empiricism is needed for real user-friendliness. Defining requirements this way might make the process heavier to conduct but reduce the need for re-making the same systems in the future. Future research into these matters is required to form a comprehensive view on defining usability and how to serve eldercare workers with as little prejudice as possible.

The study includes also some limitations that should be taken into account. When designing a study to perform classification tasks, bias may be a fundamental part of the study when defining the classes. Especially in sentiment analysis, the way sentiments or single words in answers to open-ended questions are being divided into any of the eight classes already begs the question of the validity and reliability of our disposition to define the classes. As the finetuning of the tools is still in process with the more recent data, our latter part of the research is under thorough revision to meet up with scientific standards.

References

Bartosiewicz, A., Burzyńska, J., & Januszewicz, P. (2021). Polish Nurses' Attitude to e-Health Solutions and Self-Assessment of Their IT Competence. *Journal of clinical medicine*, 10(20), 4799.

European Digital Agenda (2022). *Digital Agenda for Europe*. Available online: https://www.europarl.europa.eu/ftu/pdf/en/FTU_2.4.3.pdf (accessed 21.3.2022).

Jiang, Z. H., Yu, W., Zhou, D., Chen, Y., Feng, J., & Yan, S. (2020). Convbert: Improving bert with span-based dynamic convolution. *Advances in Neural Information Processing Systems*, *33*, 12837-12848.

Jokela, T., & Polvi, J. (2010). Miten vaatia käytettävyyttä terveydenhuollon tietojärjestelmien tarjouspyynnöissä? Tapaus Oulun omahoitopalvelu. *Finnish Journal of eHealth and eWelfare*, 2(3), 129–135.

Laitinen, M., Hantunen, T., Heino, T., Hilama, P., Huttunen, A., Janhunen, P., . . . ammattikorkeakoulu, K. (2018). "Digi vie, sote vikisee": Kokemuksia sote-alan digitalisaatiosta DigiSote-hankkeessa Etelä-Savossa. Kaakkois-Suomen ammattikorkeakoulu Oy.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word – emotion association lexicon. *Computational intelligence*, 29(3), 436-465.

Ruokolainen, T., Kauppinen, P., Silfverberg, M., & Lindén, K. (2020). A Finnish news corpus for named entity recognition. *Language Resources and Evaluation*, 54(1), 247-272.

Seibert, K., Domhoff, D., Huter, K., Krick, T., Rothgang, H., & Wolf-Ostermann, K. (2020). Application of digital technologies in nursing practice: results of a mixed methods study on nurses' experiences, needs and perspectives. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, 158, 94-106.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., ... & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv*:1912.07076.

Ylönen, K., Salovaara, S., Kaipio, J., Tyllinen, M., Tynkkynen, E., Hautala, S., & Lääveri, T. (2020). Sosiaalialan asiakastietojärjestelmissä paljon parannettavaa: käyttäjäkokemukset 2019. *Finnish Journal of eHealth and eWelfare,* 12(1), 30–43.

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15090

Changes in corporate websites and business activity: automatic classification of corporate webpages

Joan Manuel Valenzuela Rubilar, Josep Domenech, Ana Pont
Universitat Politècnica de València, Spain.

Abstract

Every time a firm or institution performs an activity on the Web, this is registered, leaving a "digital footprint". Part this digital footprint is reflected on their websites as these officially represent them on the Web. We plan to automatically monitor the changes that periodically occur in a website to relate them with the business activity. The aim of this paper is to propose a theoretical classification of corporate webpages to associate changes that occur on them with the regular activity of the firms, and to evaluate the possibility of an automatic categorization using classification models. To generate the classification of corporate webpages, a significant number of today corporate webpages were analyzed and observed, distinguishing four theoretical types of corporate webpages. To evaluate the automatic categorization of corporate webpages, a dataset of 1005 today corporate pages was generated by manually labeling them and evaluating their automatic categorization using classification models.

Keywords: *Corporate websites; Webpages classification; Websites changes.*

1. Introduction

Nowadays, the daily activities of entities and individuals on WWW (the Web) generate tons of fresh and digitized data. These data are commonly referred to as "Big Data". Properly analyzing these data could help to reveal trends and monitor economic, industrial and social behavior or magnitude (Blazquez and Domenech, 2018). Every time a firm or institution performs an activity on the Web, this is registered, leaving a "digital footprint". Part this digital footprint is reflected on their websites as these officially represent them on the Web.

Corporate websites help to obtain direct information about the strategic variables that can define firms and allow to describe the corporate profile of the firms (Llopis et al., 2019) and institutions as they are regularly updated by the firms and institutions themselves. Thus, firm data are available on corporate websites almost in real time (Crosato et al., 2021) and can therefore be considered up-to-date and reliable sources of information on the activities of firms and institutions that also reflect, to a large extent, their health and behavior.

Today's websites are composed of several webpages. An important characteristic of web content in general is the constancy of its change (Han et al., 2019). The webmasters regularly add new products, modify textual content, include new photos and insert and remove links and webpages, etc. (Llopis et al., 2010). But in addition, web content can change because of the spontaneous interactions of the users (Calzarossa and Tessera, 2018). Besides, webpages can appear and disappear on the Web all the time (Calzarossa and Tessera, 2018). Thus, changes in corporate webpages represent changes in the firms.

These changes on webpages of corporate websites are of varying significance at the level of business activity. For example, a structural change in the organization chart page does not have the same significance as the day-to-day addition of new product pages. To detect the types of changes on corporate webpages and to evaluate their business activity significance, it is important to generate a classification of today corporate webpages.

To monitor the changes in corporate websites, the aim of this paper is to propose a theoretical classification of firm webpages to associate changes on them with the regular activity of the firms, and to evaluate the possibility of an automatic categorization using classification models.

2. Literature review

This section reviews research papers on the dynamics of web content changes and classifications for web content.

Today's corporate websites are unstructured and non-traditional sources of social and economic Big Data. They are also dynamic over time. That is, during its finite lifespan —

an average of 2 years and 7 months, according to Crestodina (2017) —, they undergo structural, content, technological and other changes.

To cope with the highly dynamic behavior of corporate websites, it is necessary to predict how often and to what extent their content changes (Calzarossa and Tessera, 2018). That is, it is necessary to identify and classify the changes occurring on the corporate webpages to convert them into knowledge about the health and behavior of firms and institutions and ultimately to make economic forecasts about the firms. In this way, Calzarossa and Tessera (2018) presents a methodological framework — based on time series analysis — for modeling and predicting the dynamics of the web content changes.

Focused in understanding temporal dynamics and evolution of topics on the Web, Santos et al., (2016) developed a methodology to monitor webpages that belong to a same topic. Their results show, among other things, that distinct topics have different change patterns.

This is in line with Radinsky and Bennett (2013), who state that webpages are dynamic channels whose contents change with time, in their work on predicting content change on the Web.

Regarding the classification of the content of the web content of webpages, Zhou and Sun (2014) distinguished that the web content often comes in two camps: Evergreen content frequently don't change with time, whereas ephemeral content easily become dated.

From a more general view of corporate sites, based on users' expectations and the direct purpose of the websites, Cebi (2013) classifies the websites into (i) commercial websites, to make money selling products or services; (ii) service websites, which present various services free of charge; and (iii) mixed-type websites which present two or more purposes at the same time.

Finally, in relation to how to rescue and treat information extracted from non-traditional, non-structured Big Data sources, Blasquez and Domenech (2018) propose a flexible Big Data architecture for nowcasting and forecasting social and economic changes, applicable to data sources such as corporate websites.

Based on this review, it was decided to observe and analyze a significant number of today corporate websites and generate a theoretical classification for today corporate webpages.

3. Proposed classification of corporate webpages

As websites have become more massive and evolved, stereotypical structures have emerged from their use, i.e., stable and recognizable content configurations have emerged for visitors. Although there is a wide diversity of webpages within a single website, it is possible to identify certain typologies that are repeated in many websites. Thus, based on

the observation of a significant number of today corporate websites at different points in their lifespan, a theoretical classification for corporate webpages has been generated.

Corporate: They present content related to the firm itself. These webpages have a long-life expectancy, i.e., they are likely to remain enabled throughout the lifespan of the website. The content of these pages undergoes few changes during their lifespan; e.g., corporate pages or data protection policy pages. Normally, changes in the content of this type of page will be related to deep changes, for example, corporate pages will change due to restructuring in the firm/institution.

Post: They present current content for firm's stakeholders, especially for the general public, e.g., a news page or a job offer. These webpages have a long-life expectancy, but their content is likely to grow through spontaneous comments from visitors, e.g., forum posts. The constant appearance of this type of page, added to the constant growth of its content due to user interactions, indicates that there is active communication between the firm and its stakeholders. If no such pages appear during the lifespan of the website, it will be a bad sign of the firm's digital evolution.

Service: They show the products and/or services offered by the firms. These webpages have a short-life expectancy, that is, there is a short time between their creation and their disappearance. In addition, the content of these web pages normally does not change during their lifespan. The constant appearance and disappearance of this type of pages, during the lifespan of a website, indicates that the firm has a regular business activity. An example of this type of pages are Amazon's product pages.

Catalogue: They conglomerate content linked to the Service and Post type pages. These webpages have a long-life expectancy but their content change constantly as new products or posts are added to the website, e.g., Amazon's homepage (https://www.amazon.es). The constant updating of content of this pages indicates that the firm behind the website has a regular business activity. If a page of this type does not undergo changes in content during its lifespan, it will be a bad sign of the firm's evolution.

4. Data and methods

4.1. Dataset description

To carry out the experiments, a dataset was defined and generated four times, once for each defined target class, as follows: A sample of 999 Spanish firms was extracted from the Bureau Van Dijk's SABI (Sistema de Análisis de Balances Ibéricos) database using stratified sampling with a uniform fixation to three strata: large firms, medium-sized firms and small firms according to the EU definition of SMEs. Then, a subset of 100 corporate websites was randomically extracted from the sample of Spanish firms. After that, 1005

webpages from the subset of corporate websites were visited and, after analyzing them structurally, assigned each one a label according to the classification defined above making sure that the dataset classes were balanced. Thus, 250 pages of Corporate type, 259 of Post type, 248 of Service type and 248 of Catalogue type were labeled. Then the Scrapy framework was used to automatically extract from the 1005 labeled webpages a series of relevant features, common at the source code level, that allowed to classify the webpages. Finally, the features extracted from each page were processed and converted them into categorical values.

Table 1. Summary of types of webpages defined, the changes that typically occur on them and the link of these changes to the firm's regular activity.

Type	Changes that occur	Link to business activity
Corporate	Specific changes in content or sudden appearance/disappearance	Structural changes in the firm or institution
Post	Content grows due to user interactions	There is active communication between the firm and its stakeholders
Service	Constant appearance and disappearance	The firm or institution has a regular economic activity
Catalogue	Constant changes in content	

4.2. Description of variables

As a result of the process described above, the dataset is made up of seven quantitative variables (HTML_size, Text_size, Tags_number, Links_number, Words_number and Images_number and Modified), and three sets of categorical variables (Files_extensions, HTML_tags and NL_words) that together form more than 37,000 independent variables that allowed to classify the webpages according to the typologies defined. The sets of variables Files_extensions, HTML_tags and NL_words contain the frequency of occurrence of the file extensions, HTML tags and Natural Language (NL) words of the scraped pages. It should also be noted that, the values of the HTML_size, Text_size, Tags_number, Links_number, Words_number and Images_number quantitative variables are transformed into quantiles (categorical variables) using the Pandas library.

4.3. Classification models used

From the literature review and after performing several preliminary tests with the classification models offered by Scikit-learn, for this study four of the most widely used classification models in the literature were used: Logistic Regression (aka logit, MaxEnt), Linear Support Vector Classification (Linear SVM), Neural network model Multi-layer

Perceptron classifier and AdaBoost classifier. The models were used with the default Scikit-learn configuration.

Table 2. Description of quantitative sets of categorical variables that make up the dataset.

Variable	**Description of quantitative variable**
Type	Target variable of the classification (Corporate, Service, Catalogue or Post).
HTML_size	Size in bytes of the body of the scraped pages.
Text_size	Size in bytes of the text content within the body of the scraped pages.
Tags_number	Number of HTML tags within the source code of the scraped pages.
Links_number	Number of links in HREF attribute of Anchor tags of the scraped pages.
Words_number	Number of natural language words within of the scraped documents.
Images_number	Number of images in SRC attribute of IMG tags of the scraped pages.
Modified	Last modification date of the of the scraped pages.
Variable	**Description of set of categorical variables**
Files_extensions	All file extensions present within the source code of the scraped pages.
HTML_tags	All HTML tags present within the source code of the scraped pages.
NL_words	All NL words present within the source code of the scraped pages.

5. Results and discussion

The performance of the classifiers was evaluated on a class-by-class basis, so the results are presented below organized according to the target classes.

Figure 1 shows AUC values of 5-fold Stratified cross-validation obtained for the Logistic Regression, Linear SVM, Multi-layer Perceptron and Ada_Boost classifiers for Post target class (which obtained the highest AUC values).

The highest AUC values were obtained when classifying the Post class using the Logistic Regression and Linear SVM classification models, both with values of 0.96 AUC and ± 0.01 of Standard Deviation. Logistic Regression was the model that presented the best performance in classifying each of the 4 target classes with respect to the other models used, also obtaining the highest AUC values when classifying the Corporate (0.90 ± 0.02), Service (0.90 ± 0.01) and Post Catalogue (0.84 ± 0.03) classes. The lowest AUC values were obtained when classifying the Catalogue class using the Multi-layer Perceptron (0.79

± 0.04) and AdaBoost (0.78 ± 0.04) models. Overall, the Catalogue class appears to be a difficult class for the models to classify.

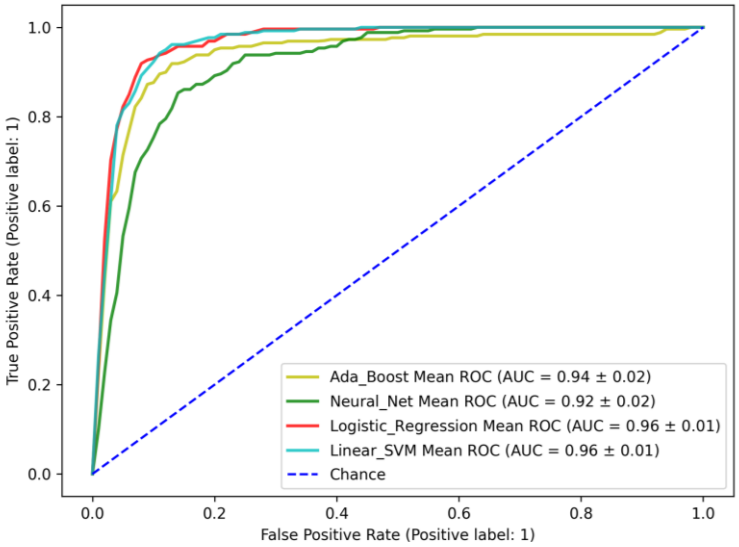


Figure 1. AUC values for all classification models for Post target class.

Table 3. AUC and Standard Deviation for all models, sorted by target class.

Target class	Logistic Regression	Linear SVM	ML Perceptron	AdaBoost
Catalogue	0.84 ± 0.03	0.83 ± 0.03	0.79 ± 0.04	0.78 ± 0.04
Corporate	0.90 ± 0.02	0.89 ± 0.02	0.89 ± 0.02	0.84 ± 0.03
Post	0.96 ± 0.01	0.96 ± 0.01	0.94 ± 0.02	0.92 ± 0.02
Service	0.90 ± 0.01	0.89 ± 0.01	0.89 ± 0.01	0.84 ± 0.03

5. Conclusions

This study, on the one hand, set out a theoretical classification of corporate webpages to associate their changes with the regular activity of firms. Through the observation of a significant number of today corporate websites at different points in their lifespan, four types of corporate webpages were defined: Catalogue, Corporate, Post and Service. On the other hand, the study also evaluated the possibility of an automatic categorization of corporate webpages using classification models. Logistic Regression presented the best performance in classifying each of the 4 target classes followed by Linear SVM. The best classified class was the Post class. Finally, it is necessary to emphasize that an automatic

categorization of today corporate webpages is important to firms to associate changes on their websites with their business activity and to evaluate its significance of these changes to the business. This could make a positive difference for a particular firm or organization.

Acknowledgments

This work was partially supported by grants PID2019-107765RB-I00 and funded by MCIN/AEI/10.13039/501100011033.

References

Blazquez, D., and Domenech, J. (2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, *130*, 99-113.

Calzarossa, M. C. and Tessera, D. (2018). Analysis and forecasting of web content dynamics. In 32nd *International Conference on Advanced In-formation Networking and Applications Workshops (WAINA)*, pages 12–17.

Cebi, S. (2013). Determining importance degrees of website design parameters based on interactions and types of websites. *Decision Support Systems*, 54(2), 1030 – 1043. 4.

Crestodina, A. (2017). What is the average website lifespan? 10 factors in website life expectancy.

Crosato, L., Domenech, J., and Liberati, C. (2021). Predicting sme's default: Are their websites informative? *Economics Letters*, 204, 109888

Han, S., Brodowsky, B., Gajda, P., Novikov, S., Bendersky, M., Najork, M., Dua, R., and Popescul, A. (2019). Predictive crawling for commercial web content. In *Proceedings of the 2019 World Wide Web Conference*, pages 627–637.

Llopis, J., Gonzalez, R., and Gasco, J. (2010). Web pages as a tool for a strategic description of the spanish largest firms. *Inf. Process. Manage*.,46, 320–330.

Llopis, J., Gonzalez, R., and Gasco, J. (2019). The evolution of web pages for a strategic description of large firms. *Economic Research-Ekonomska Is-traˇzivanja*,0(0), 1–21.

Santos, A., Pasini, B., and Freire, J. (2016). A first study on temporal dynamics of topics on the web. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, page 849–854.

Radinsky, K. and Bennett, P. N. (2013). Predicting content change on the web. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, page 415–424, New York, NY, USA. Association for Computing Machinery.

Zhou, E. and Sun, L. (2014). Evergreen or ephemeral: Predicting webpage longevity through relevancy features. 5.

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15091

Understanding the effects of Covid-19 on P2P hospitality: Comparative classification analysis for Airbnb-Barcelona

Juan Pablo Argente del Castillo Martínez, Isabel P. Albaladejo
Universidad de Murcia, Spain

Abstract

The Covid-19 pandemic has produced significant changes in tourism markets around the world. The large amount of data available on the Airbnb platform, one of the world's largest hosting services, makes it an ideal prospecting place to try to find out what the aftermath of this event has been.

This paper explores the entire Airbnb housing stock in the city of Barcelona with the aim of identifying the key characteristics of the homes that have remained operational during the 2019-2021 period. We carried out this analysis by using two classification methods, the random forest and logistic regression with elastic net. The objective is to classify the houses that have remained on the platform against those that have not. Finally, we analyze the results obtained and compare both the general performance of the models and the individual information of each variable through partial dependence plots (PDP).

We found a better performance of Random Forest over logistic regression, but not significant differences in the relevant variables chosen by each method. It is worth noting the importance of the geographical location, the number of amenities in the home or the price in the survival of the homes.

Keywords: *Airbnb; Covid-19; Survivability; Random-Forest; Logistic-Regression .*

1. Introduction

The Covid-19 pandemic has caused all kinds of effects in the different affected markets and economies, especially in tourism. Spain, and more specifically the city of Barcelona, has notably suffered the consequences of the different policies that have been carried out to control the devastating effects of the virus. According to the Statistics of tourist movements at the border of the Spanish Statistics Institute (INE), the number of international tourists arriving in Cataluña has dropped from 19.4 millions in 2019 to 5.7 millions in 2021. This fall in arrival of turists has had non-zero effects in hosting services.

Currently, one of the most popular accommodation services in the city of Barcelona is the Airbnb platform. It brings together both professionals from the world of hospitality and non-professionals, offering a range of different solutions far superior to conventional media. This fact, together with the greater competitiveness in prices, has made the platform become the reference for the vast majority of tourists who visit the city (Gutierrez J. et al, 2017).

In this paper, we analyze the characteristics that best define the group of dwellings that has remained available during 2019-2021. This is a matter of importance for the owner in terms of achieving regular income in the long term (Lladòs-Masllorens et al., 2020). It is also a key factor for local managers in tourism-dependent economies (Wachsmuth, D., & Weisler, A. 2018), one of the most external-shock dependent, like the city of Barcelona. Proof of this are the scenarios that both covid-19 and 2008 crises left behind. We use datasets from Airbnb which contain the different characteristics of the dwellings of Barcelona and correspond to the month of November for both 2019 and 2021.We decided to use these timestamps because in November 2019 the pandemic had not yet started and in November 2021 the process of suspending anti-covid measures in Spain had already begun, along with the end of travel bans to other countries and the end of the second round of vaccinations for the majority of the population. In addition the reason for choosing the month of November is to avoid the seasonality effects of tourism data, since Barcelona, along with most tourist destinations in the world, suffers an increase in tourists in the summer months.

To find the characteristics that best define the "dwelling-survivavility" we have carried out two classification methods with the same set of variables: random-forest and logistic regression with elastic-net regularization. We have found evidence that the most determining variables have been the geographical location, the price, the experience of the host and the level of equipment of the dwelling.

2. Data Handling

Both datasets from November 2019 and November 2021 were collected from Inside Airbnb, an online service that provides these datasets for different cities around the world. It has been used in similar studies (Gibbs et al. 2018) due to the large amount of data it provides and the ease with which it can be obtained.

After a minimal cleaning of the datasets, with the aim of maintaining as much of the sample as possible, the datasets decreased to 12,337 dwellings in 2019 and 9,540 in 2021. Furthermore different operations were carried out in order to operationalize the dataset. We did 'one hot encoding' of categorical variables, character count of variables that are strings, or count of items in lists, wich is the case of the amenities.

Once the datasets were cleaned separately, we created the dummy variable that will be used as the dependent variable, old_homes. We obtained this variable by comparing the unique identifiers of the different ads, so if we happened to find the same identifier in 2019 and 2021, we will add a 1 to the variable and 0 in the event that this assignment does not occur. This variable allowed us to find that only 4,366 ads were shared between both moments of time, so 8,001 homes were lost since 2019. Only 5,174 have been recovered as of November 2021. We can observe the spatial distribution of the 2021 dataset in Figure 1. The red dots indicate the three main tourist points of the city, from top to bottom: Sagrada Familia, Ramblas and Puerto. In yellow we see the dwellings that were maintained throughout the period and in blue the new ones, slightly more grouped in these three neighbourhoods.

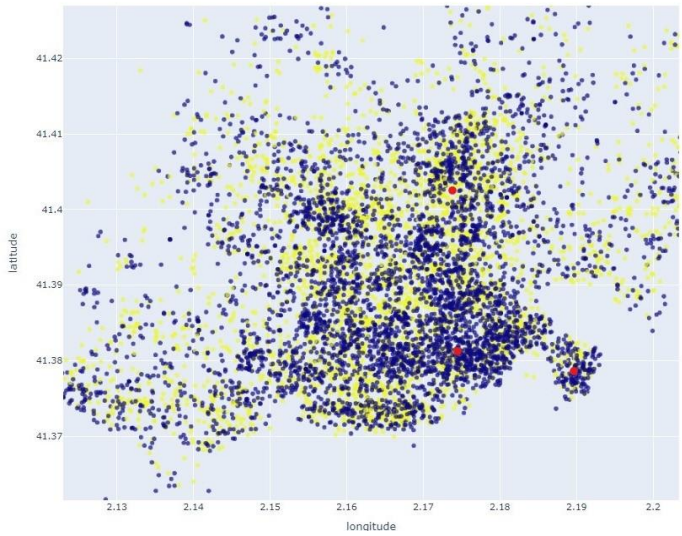


Figure 1. Spatial Distribution of 2021 Airbnb Dataset

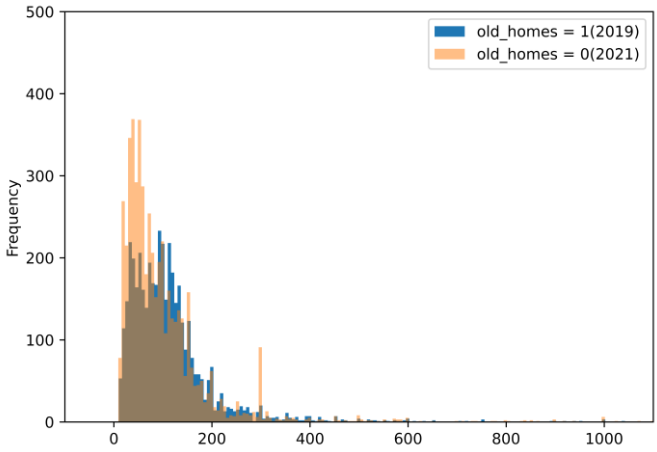


Figure 2. Price frecuency distribution of 2021 Airbnb Dataset.

Figure 2 shows the frequency distributuion of home's prices. We can also find certain differences, the most important being the predominance of the cheapest prices in the newer houses, and a more uniform distribution in the houses that remained during the entire period. One last aspect to highlight is the professionalism of the hosts. This quality is usually measured in the literature (Lladós-Masllorens et al, 2020) by the total number of ads that the host has on the platform. This variable is known as calculated_host_listings_count and has had significant variations from 2019 to 2021, increasing the average level in the period, as we can see in figure 3.

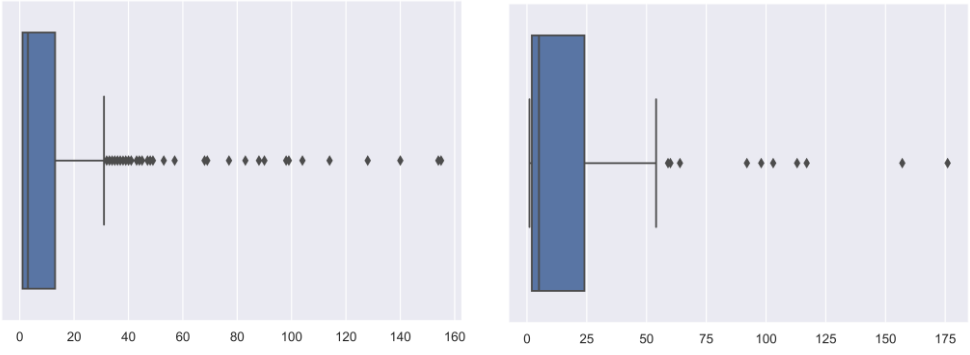


Figure 3. Box plots of calculated_host_listings_count. (left 2019 and right 2021).

All the variables that were provided with the original datasets were used, with the exception of those who implied somehow temporality (f.e. first_review, last review, …) or were descriptions that had little to no relationship with the analysis (f.e. host_neighbourhood, …). In total, 72 variables were used.

3. Modeling the survivability

We have used two different aproaches to understand the differences between those dwellings that were kept on the platform for the whole period, wich we refer as old houses, and those who did not, wich we refer as new houses: Random Forest Classifier and Logistic Regression with Elastic Net. To assure we do not overfit the samples, two different datasets were established within the sample, a training set with 80% of the values and a test set with the remaining 20%.

3.1 Random Forest Classifier

The hyperparameters for the Random Forest (Brieman, 2001), wich were optimized by Grid Search, are the following: 760 trees, the square root of the total available variables as the maximum number of variables to consider in each tree, and an execution without Bootstrap. It should be noted that entropy has been used as a criterion to measure the quality of the sample divisions and not the Gini impurity.

3.2 Logistic Regression with elastic net classifier

Also for the Logistic Regresssion a Grid Search was carried out to obtain the hyperparameters, these being: elastic net penalty function with an l1/l2 ratio of 50%, 1000 maximum iterations and the SAGA algorithm (Defacio, Bach, Lacoste- Julien, 2014) for the optimization of the problem.

4. Results

Table 1 shows the classification output of both methods and Table 4 the confussion matrix. As can be seen in both, random forest presents a precision 15% higher than the logistic regression with elastic net. According to random forest the five most important variables to define survivability are latitude, longitude, price, the amount of amenities and the professionalism of the host. It is important to notice that these variables are relevant in both methods but only the random forest allows us to see the non-linearities between them.

To analize the effects of these variables we use the partial dependence plots (Friedman, 2001). Each point of these plots indicate how many of the trees (in average) that random forest builds have been classified for the class "old_homes" across all observations, given a fixed level of the variable we're looking at. Figure 6 indicates these partial dependence for the obtained key variables.

Table 1. Random Forest Classifier and Logistic Regression Output

	Variable	Precision	Recall	F1-Score	Support
RF	0	0.78	0.69	0.73	658
	1	0.78	0.85	0.82	872
Log_e-n	0	0.65	0.31	0.42	658
	1	0.63	0.87	0.73	872

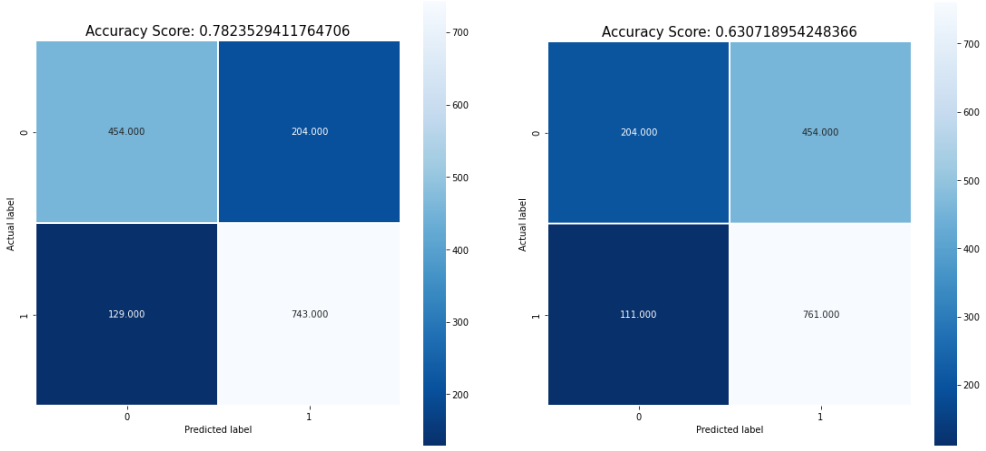


Figure 4. Confusion matrix of Random Forest Classifier(left) and Logistic Regression-elastic net(right).

In the first place, the partial dependencies of latitude and longitude indicate a greater tendency to predict a "new" house in the geographical points (41.38,2.175), wich coincides with Las Ramblas point. This shows that the houses in the pool with the best geographical location, in terms of restrictions for the covid-19 pandemic, have managed to remain active during this time. Secondly, we can relate higher prices to houses that have remained during the pandemic. The bulk of the new houses are located in the cheapest levels, which normally correspond to private rooms in shared houses or collaborative housing solutions, which were more likely to close in the evaluated period. In relation to the amenities of the properties, we can infer that fewer amenities are associated with lower probabilities, that is, with "new" homes, while the best-equipped homes are those that have achieved endure. Finally, the professionalism of the owner of the home, measured by the total number of homes owned by the host (calculated_host_listings_count), provides information that is only valid in the lower ranges of the variable, since that is where the majority of the sample is found. In this case, a decrease in probability is observed as professionalism increases,

indicating that owners with a greater number of homes have withdrawn their homes from the market, unlike the majority of small owners, who have been more open to staying.

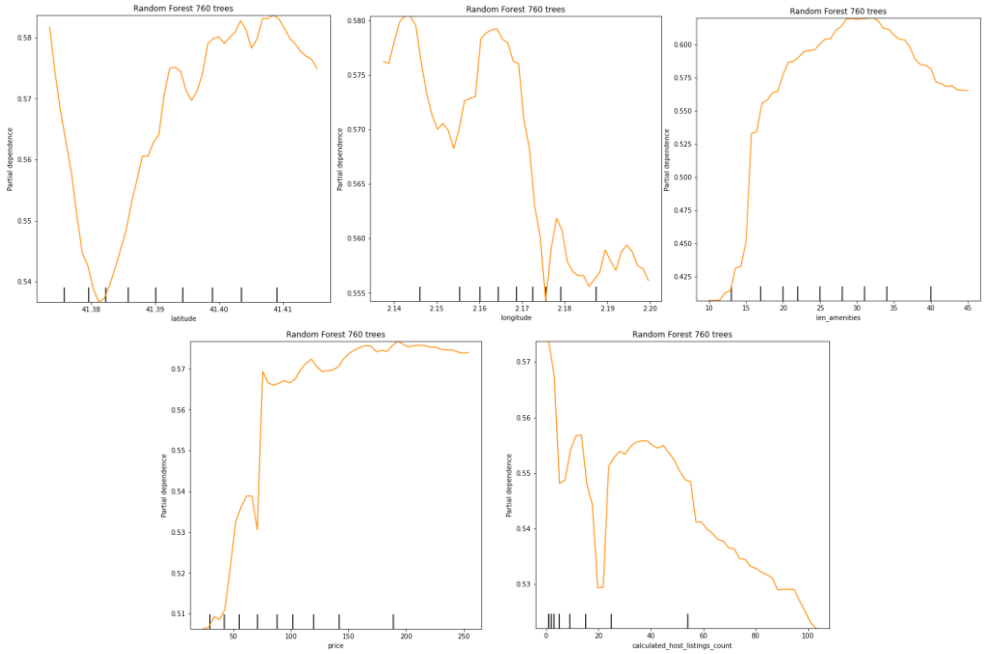


Figure5. Partial Dependence Plots of most important features according to Ranfom-Forest.

4. Conclusion and Prospects

The classification proposed in this study of the Airbnb accommodations of Barcelona allows us to conclude that the main variables affecting their survaivability are those that also define quite well the prices on hedonic models that are generally used to determine the factors defining the prices of this kind of accommodations. (Casamatta et al. 2022). In particular, the most touristic neighborhoods and the most densely populated with housing are going to be what suffer the most during non-normal stages. We can also infer that professionalism plays a key role in survival, and that although it normally implies better decision-making, in the face of this type of event it also implies a certain level of vulnerability.

The results of the study allow us to intuit that the robustness of the market, understanding it as the survivability rate of the dwellings on such difficult periods, could be improved by regulating property in order to avoid professional proprietaries. This regulation would also help to avoid gentrification that most tourist cities suffer, as we have seen in Barcelona in

Las Ramblas neighborhood. These densely populated areas are more prone to develop big clusters of accommodations that are not based in shared economy but in hospitality firms.

In future works we will explore more specific effects of these attributes, mediated by macroeconomic variables, as well as also analyze the effect on robustness of a pure shared economy model. New techniques that allow deeper analysis, such as neural networks and the use of natural language processing, are also interesting to complete the analyses.

References

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

Casamatta, G., Giannoni, S., Brunstein, D., & Jouve, J. (2022). Host type and pricing on Airbnb: Seasonality and perceived market power. *Tourism Management*, 88, 104433.

Defazio, A., Bach, F., & Lacoste-Julien, S. (2014). SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Gibbs, C., Guttentag, D., Gretzel, U., Morton, J., & Goodwill, A. (2018). Pricing in the sharing economy: A hedonic pricing model applied to Airbnb listings. *Journal of Travel & Tourism Marketing*, *35*(1), 46-56.

Gutiérrez, J., García-Palomares, J. C., Romanillos, G., & Salas-Olmedo, M. H. (2017). The eruption of Airbnb in tourist cities: Comparing spatial patterns of hotels and peer-to-peer accommodation in Barcelona. *Tourism management*, 62, 278-291.

Lladós-Masllorens, J., Meseguer-Artola, A., & Rodríguez-Ardura, I. (2020). Understanding peer-to-peer, two-sided digital marketplaces: pricing lessons from Airbnb in Barcelona. *Sustainability*, 12(13), 5229.

Wachsmuth, D., & Weisler, A. (2018). Airbnb and the rent gap: Gentrification through the sharing economy. *Environment and Planning A: Economy and Space*, 50(6), 1147-1170.

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15093

Simulating the inconsistencies of Google Trends data

Eduardo Cebrián, Josep Domenech
Department of Economics and Social Sciences, Universitat Politècnica de València, Spain.

Abstract

Google Trends (GT) allows users to obtain reports on the evolution of the popularity of searches made through the Google Search engine. Its main output is the Search Volume Index (SVI), a relative measure of the popularity of a term, which is computed using a sample of the searches. Due to the sampling error, the reports are not completely consistent, as the same query produces different time series that can widely change from day to day. This paper simulates the process of generating the SVI time series in the same way as GT does. By doing this, it has been shown that the sampling error could be an important issue if the popularity of the term under study is relatively low. Averaging multiple extractions from GT can only partially alleviate this.

Keywords: *Google Trends, Consistency, Measurement Error, Online Data*

1. Introduction

Google Trends (GT) is a freely available tool developed by Google that allows users to obtain reports of the evolution of the popularity of searchers made through the Google Search engine. In the last decade, GT has become popular in the scientific literature because its reports can be used to measure the population's interest on any topic. Moreover, this data can be easily accessed and is constantly updated.

The main output of GT reports are time series data representing the Search Volume Index (SVI), a relative measure of the popularity of a term. To compute the SVI, Google does not consider the whole set of searches they received in a given time period, but a sample with unknown characteristics. Due to the sampling error, the reports are not completely consistent, as the same query can produce different time series which change from day to day (Choi and Varian, 2012).The importance of these inconsistencies is often minimized (Choi and Varian, 2012; Dilmaghani, 2019) although Cebrián and Domenech (2022) report that variations in GT data may be significant enough to hinder the interpretability and reproducibility of the models estimated with them.

To understand the inconsistencies of GT data, this paper proposes a simulation model of the GT data generating process. This model is then used to analyze how a typical time series with seasonality is distorted due to the sampling process and how the averaging of extractions can mitigate the error, but only partially.

2. Related Work

The inconsistencies of GT time series have long been described, although most researchers do not consider them relevant enough to affect their results (Choi and Varian, 2012; Preis et al., 2013). Other research works identify these inconsistencies as an important source of error and average multiple GT requests of the same time series on different days, or using some tricks to force a new sample. This way, the time series are smoothed, thus reducing the sampling error.

However, the number of extractions which are averaged widely vary across the literature. On the one hand, D'Amuri and Marcucci (2017) take 24 different extractions for the search term "jobs" and report cross-correlations of at least 0.99 between extractions. On the other hand, Cebrián and Domenech (2022) extract queries related to Austrian cities on 6 different occasions and find correlations between 0.79 and 0.94, while Carrière-Swallow and Labbé (2013) use the average standard deviation to measure the sampling error and report values above 15% for the term "Chevrolet" after 50 extractions.

To the best of our knowledge, there is no method for determining how many extractions should be averaged to alleviate the sampling error, or which factors may affect it. To

understand the intricacies of how the SVIs are produced and the effects of averaging multiple extractions of the same GT time series, this paper proposes a simulation model to generate the SVIs (and its sampling error) in the same way as Google Trends does.

3. Google Trends sampling

The process to compute SVIs is illustrated in Figure 1. It starts with the whole set of searches that Google has received from 2004. From this set (Total Searches), GT draws a random sample that is replaced over time. This introduces an unknown sampling error because the parameters of this sampling, such as the coverage or how often the sample is replaced, are not disclosed by Google. When a user requests the GT report for a given term and time period, the sample is filtered to keep only those rows matching the request so that frequencies by time period can be computed (Raw Popularity). Finally, the time series are normalized by setting the SVI in the period with highest frequency to 100 and scaling the frequencies in other periods proportionally (and rounding them to integers).

The sampling error introduced in GT reports is also illustrated in Figure 1. In the example provided, the term a is reported with an SVI of (0, 100, 33), but if it were computed with the Total Searches set, the result should have been (67, 100, 67).

4. Simulation and Results

This section provides some simulations of the GT process described in Section 3 to check how the SVI time series change depending on the popularity of the search term and what the effect of averaging multiple extractions is.

For illustrative reasons, it has been assumed that the total number of searches of the term *y* follows a function with linear trend and a seasonal component, modeled with a sinusoidal function with a period of 12 time units, as defined in Equation 1.

$$Y_t = \beta_1 t + \beta_2 \sin\left(\frac{2\pi}{12} t\right) \tag{1}$$

where Y_t is the number of searches of the term *y* at time *t*, β_1 is the parameter defining the strength of the linear trend, and β_2 defines the strength of the seasonal component.

For each time period t, the presence of the term *y* in the sample (y_t) follows a binomial distribution with parameters *n* equals the sample size, and *p* equals the proportion of searches of term *y* among all the searches received by Google at that time period.

Therefore, the expected number of occurrences in the sample of term *y* at time t is:

$$E[Y_t] = n * p_t \tag{2}$$

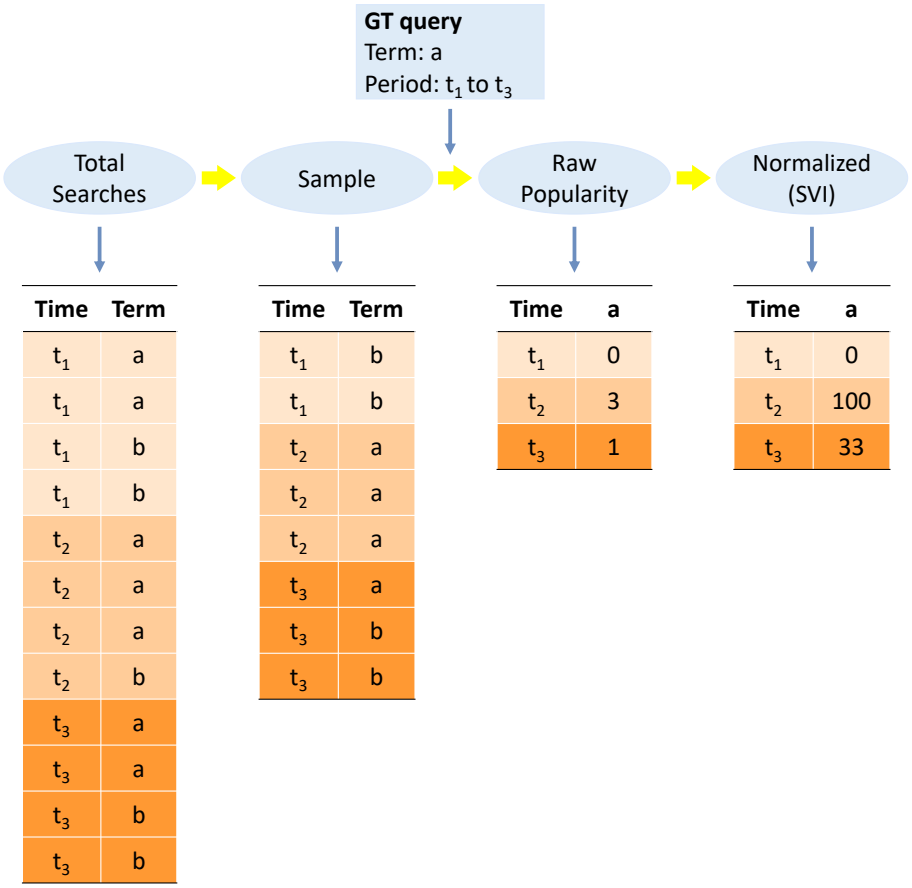


Figure 1: Process GT follows to compute an SVI time series. It is illustrated with an example with three time periods (t_1, t_2 and t_3) and two terms (a and b). Only the GT report for term a is requested.

Notice that p_t varies in time, being this variation the change in popularity of the term.

Since n and p_t are unknown (as they are not disclosed by Google), we have studied the SVIs of two terms with different popularity. Term H has an average frequency in the sample of 200 times, while term L is less popular and has an average frequency in the sample of 20 times through all the considered periods. Simulations are conducted considering GT requests for 60 periods. The random process of generating the SVI for each term has been repeated 20 times, each one representing one extraction from GT.

Figure 2 shows the simulation results of 1 (left), 10 (center), and 20 (right) SVI extractions. Light blue lines represent each individual extraction, while dark blue lines illustrate the

average of all the extractions in each plot. Plots in the top row refer to the most popular term (H), while plots in the lower row refer to the less popular term (L). Each plot includes the Pearson's correlation coefficient (r) of the time series in dark blue with the actual popularity of the term (defined by Equation 1).

Plots in the left part of Figure 2 evidence that a single extraction has significant noise. This noise, which is introduced by the sampling process, can be alleviated by averaging multiple extractions. After averaging 10 extractions, the curve for the term with high popularity (H) is smoother, as evidenced by the $r = 0.997$ value as well. However, the less popular term (L) requires more extractions to obtain a good approximation to the actual trend. Indeed, after averaging 20 extractions of the less popular term, the correlation coefficient r is still below the one of 10 extractions of the high popular term.

These simulation results highlight the relationship between the sharpness of the SVI and the absolute popularity of the term and, therefore, the need for averaging more GT extractions when studying less popular terms. However, as one can observe in Figure 2, there is a side effect related to the construction of a time series as the average of a number of extractions: the range of the SVI is reduced. In the case of the less popular term (L), the SVI takes values from 18 to 100 in a single extraction (bottom-left plot). When the average of 20 extractions is considered, SVI ranges from 27.1 to 81.3 (bottom-right plot). This implies that the value of a single extraction (for instance, when using GT for nowcasting purposes) cannot be directly compared to the series obtained after averaging multiple extractions.

5. Conclusions

Although Google Trends has become a very popular data source among researchers, its sampling error has not been intensively studied. This paper has replicated the process of generating the SVI time series in the same way as GT does. By doing this, it has been shown that the sampling error could be an important issue if the popularity of the term under study is relatively low, as the quantity of noise it introduces in the series is noticeable.

The technique of extracting GT data multiple times and using the average of the series has also been studied. Our results showed that it certainly alleviates some of the variability introduced in the random sampling, but the number of repetitions needed to smooth the curve heavily depends on the absolute popularity of the term. Moreover, this procedure changes the range and scale of the SVI time series, thus increasing the complexity of using GT data for nowcasting and forecasting, as additional transformations should be considered.

Acknowledgments

This work was partially supported by grants PID2019-107765RB-I00 and funded by MCIN/AEI/10.13039/501100011033.

References

Barreira, N., Godinho, P., & Melo, P. (2013). Nowcasting unemployment rate and new car sales in south-western Europe with Google Trends. *NETNOMICS: Economic Research and Electronic Networking*, 14(3), 129–165.

Borup, D., Christian, E., and Schütte, M. (2022). In search of a job: Forecasting employment growth using Google Trends. *Journal of Business & Economic Statistics*, 40(1), 186–200.

Carrière-Swallow, Y., & Labbé, F. (2013). Nowcasting with google trends in an emerging market. *Journal of Forecasting*, 32(4), 289–298.

Cebrián, E., & Domenech, J. (2022). Is google trends a quality data source? *Applied Economics Letters*, pages 1–5.

Choi, H., & Varian, H. (2009). Predicting initial claims for unemployment benefits. *Google Inc*, 1, 1–5.

Choi, H., & Varian, H. (2012). Predicting the present with google trends. *Economic Record*, 88(s1), 2–9.

Dilmaghani, M. (2019). Workopolis or The Pirate Bay: what does Google Trends say about the unemployment rate? *Journal of Economic Studies*.

D'Amuri, F., & Marcucci, J. (2017). The predictive power of google searches in forecasting US unemployment. *International Journal of Forecasting*, 33(4), 801–816.

Preis, T., Moat, H. S., & Stanley, H. E. (2019). Quantifying trading behavior in financial markets using google trends. *Scientific Reports*, 3(1), 1684.

Saxa, B. (2015). Forecasting mortgages: internet search data as a proxy for mortgage credit demand. *National Bank of the Republic of Macedonia*, 107.

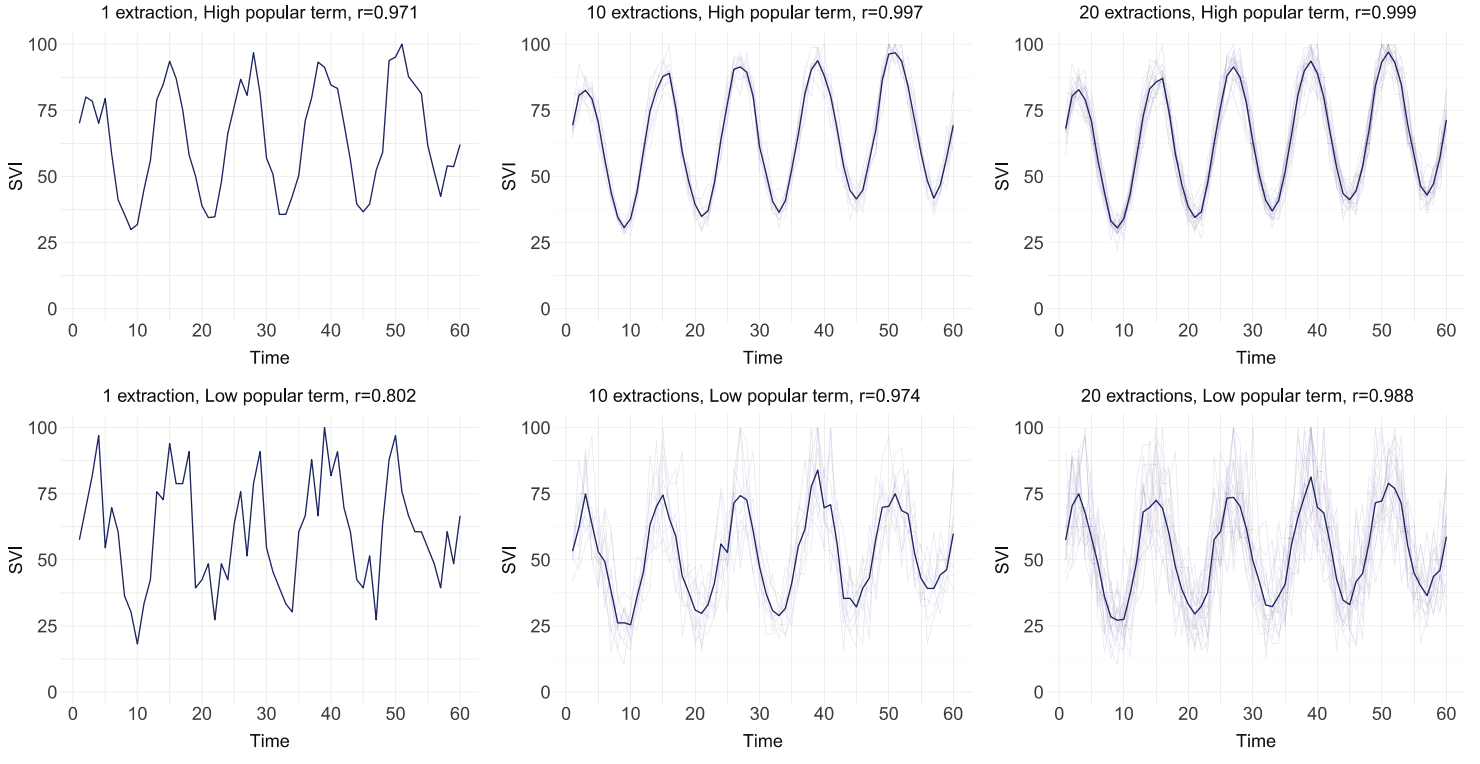


Figure 2: Simulation of GT data generation process of a search with a 12-period seasonality. Each individual extraction is shown in light blue. The dark blue line represents the average of all extractions in each plot.

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15098

Covid 19 and lodging places

Estefania Ruiz-Martinez, Francisco Porras-Bernardez, Georg Gartner
Technischen Universität Wien, Vienna, Austria.

Abstract

Tourism is a very important source of income for national economies all over the world. Before Covid-19, this sector contributed with 10.4% of the global GDP. Innovative tools for tourism study and promotion are very necessary for a future recovery of the industry. Thus, we have explored Airbnb data as a source of information about the lodging sector, very relevant within the tourism industry. We have analyzed these data to explore the experience of tourists before and after the pandemic. Our aims included identifying and visualizing opinion changes through semantics extracted from semi-structured data generated by the Airbnb customers. We used Natural Language Processing and techniques such as sentiment analysis combined with spatial analysis with KDE in order to characterize and spatially visualize user opinion. Results did not show significant differences in user opinion before and after the outbreak of Covid, however spatial patterns related to sentiments were made visible. Moreover, a large dataset covering 3.6M Airbnb lodging spots from 108 cities was compiled and will be made available in the future. This paper can be useful for the lodging industry, tourism organizations as well as social media researchers by providing an alternative approach that involves the role of location in the study of customer behaviour.

Keywords: *Airbnb; Sentiment Analysis; Covid-19; Kernel Density Estimation.*

1. Introduction

The current situation generated by Covid-19 has affected most of the economic sectors in the world. Tourism is one of the most impacted areas due to a fundamental dependence in mobility and safety. The travel and tourism sector lost near to 3.8€ billion in 2020, whereas its contribution to global GDP sunk by a huge 49.1% compared to the previous years. Thus, innovative tools for tourism study and promotion are even more necessary for a future recovery of the industry.

One very relevant part of the tourism sector is the lodging industry. People need a physical location to stay when visiting a new region. These locations and their surroundings are places. For the study of places and human perception and behaviour in them, traditional methods include questionnaires or travel diaries among other tools. However, traditional data collection methods are often limited by the number of participants that can be involved in collection campaigns. These tools can now be complemented or replaced by the use of big spatiotemporal data. User-Generated Content (UGC) (Wyrwoll, 2014) sources offer huge amounts of data usable for the analysis of spatially related phenomena. In particular, in the lodging industry, online reviews are widely used to explore the experience of customers (Xiang et al., 2015). According to (Li et al., 2018), only in tourism research, UGC accounted for almost half of the literature during the last decade.

We aim to study the influence of Covid-19 on the lodging industry by analyzing the experiences of tourists when being hosted in lodging places before and after the pandemic. For this purpose, we analyze Airbnb data including the users' reviews associated with places listed in the platform. We used Natural Language Processing (NLP) techniques such as sentiment analysis to classify online reviews as positive, neutral and negative and then characterize and visualize the spatial distribution of listings accordingly by using Kernel Density Estimation (KDE).

This work can contribute to a better understanding of the impact of Covid-19 and other phenomena on the lodging industry and on the perception of places very relevant for tourism. This paper can be useful for the tourism industry, property owners or the public administration. Moreover, it can provide a better insight into available Airbnb datasets and analysis methods for worldwide researchers.

2. Method

We collected the data and pre-processed it in two steps. The first step involved the collection, preparation and storage of the raw data from listings existing since 2015 until February 2021. The second step had to do with the pre-processing of the online reviews to

implement the semantic analysis using sentiment analysis then visualize the spatial distribution of listings according to the average sentiment.

2.1. Data preparation

The data used in this work was collected from the website Inside Airbnb, which uses public information compiled from the Airbnb website referred to the hosting places (a.k.a. listings). The location information of these listings is anonymized by Airbnb by introducing a geolocation error of 0-150 meters.

The preparation process involved the collection and further processing of the data. The first step consisted in the collection of all the datasets available until February 2021 at Inside Airbnb. Data accounts for 242 GB and includes the listings existing in 108 cities worldwide since 2015. In a second step, a Python script was developed for the data processing and a PostgreSQL spatial database was created for the storage. The raw files contained monthly snapshots of the listings. The temporal coverage for each city varies between 72 and only a few months in some cases. The processing parsed the monthly files selecting 110 attributes and generating a point element for each listing. In order to build our geodataset, we stored all the listings that have existed at any time in each city during the whole collected period. The final number of unique listings reached more than 3.6 Million.

2.2. Text pre-processing

In this step, reviews were prepared for the implementation of sentiment analysis, which required the removal of reviews written in a language different from english. To do so, a script was developed to use the available libraries for text pre-processing in Python. First of all, automated postings (e.g., “This is an automated posting”), non-English, duplicated, empty reviews and reviews consisting in only two characters, numbers or NaN were discarded. Non-English reviews were identified using the python library Fasttext (Joulin et al., n.d., 2016), which employs a pre-trained model to predict the language of a sentence. This model was trained using data from Wikipedia, Tatoeba and SETimes and can identify 176 languages. When used in python, it returns a tuple with an ISO code of the language recognized and a confidence value indicating the probability of the sentence belonging to that language.

2.3. Semantic Analysis

2.3.1. Sentiment Analysis

We performed sentiment analysis on the reviews in order to estimate the polarity of the texts. The sentiment analysis determines sentiment orientation and classifies the reviews into classes of polarity: positive, neutral or negative. For the analysis, we used the VADER

model (Hutto & Gilbert, 2014). This model follows an approach based on valence and considers the sentiment as well as its intensity. VADER is a lexicon and rule-based sentiment analysis tool. Its effectiveness has been compared against eleven state-of-the-art sentiment benchmarks with more favourable results in different contexts and even offering better performance than human raters.

The model relies on a list of lexical features that are labelled as positive or negative depending on their semantic orientation, i.e a sentiment lexicon. VADER also quantifies how positive or negative sentiment is. A text is analysed and its constituent words are searched in the lexicon: positive words have higher ratings whereas negative words lower. All lexicon ratings are combined in a compound score formed by the summarization of all of them and standardized between -1 and 1. A score of -1 represents a fully negative sentiment, a value of 0 denotes a neutral text, and a score of 1 represents a fully positive sentiment.

2.4. Spatial Visualization

2.4.1. Kernel Density Estimation (KDE)

Listings were categorized into positive, neutral and negative polarity, according to the average compound score of all the reviews one year before the outbreak of covid and one year after it. KDE was used to generate density surfaces and visualize the distribution of listings according to sentiment polarity and time period. A bandwidth was determined for the density surfaces of each polarity but was the same for both time periods. It was selected based on an iterative process that finished when the density surface was not either over-smoothed or under-smoothed. The values of pixels from density surfaces were normalized between 0 and 1 using *Map Algebra*. The new values indicate the density of listings in proportion to the maximum density obtained.

3. Results

As a case study, up to now, we have focused on two cities from two continents, i.e. Rio de Janeiro (Brazil) and New York (U.S). Both cities have a different tourism orientation and at the same time, are in two of the most severely affected countries by Covid-19.

To analyse the experience of users after the outbreak of covid, we took as reference the experience before the outbreak as well. To do so, only listings with reviews one year before and after the outbreak were included in the analysis. As a result a total of 26,262 reviews from 3,522 property listings in Rio de Janeiro and 486,438 reviews from 18,751 properties in New York were processed.

3.1. Sentiment Analysis

The proportion of positive, neutral, and negative reviews from both Rio de Janeiro and New York did not significantly change after the outbreak of Covid-19 (see Figure 1). Positive reviews from Rio de Janeiro decreased 1%, neutral reviews remained the same amount and negative reviews increased 1%. Similarly, positive reviews from New York decreased 2% and negative and neutral reviews increased 1%.

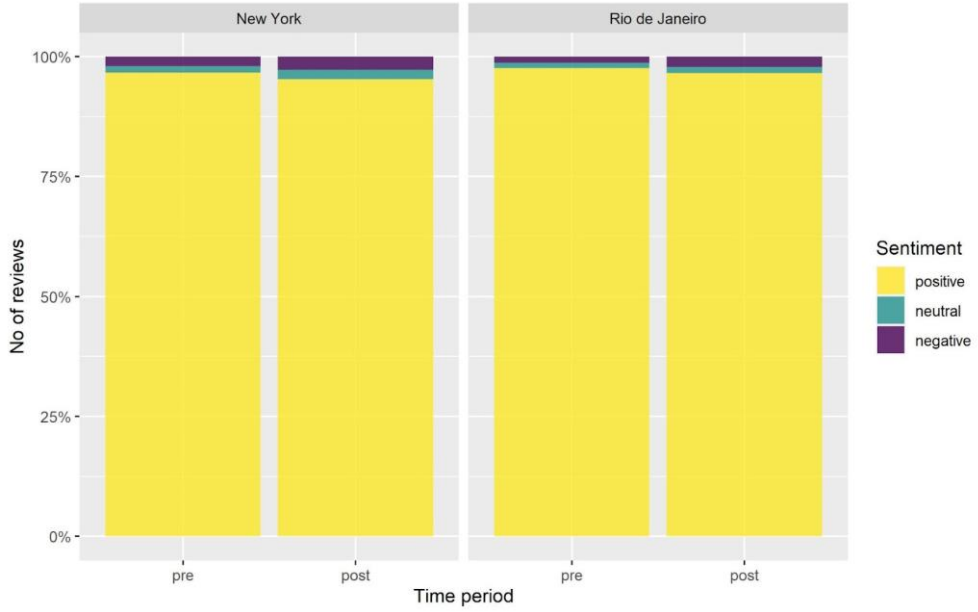


Figure 1. Percentage of positive, neutral and negative reviews from listings in NY and RJ before and after the outbreak of covid.

3.2. Spatial Visualization

Figure 2 and Figure 3 show contrasting results. While in Rio de Janeiro hotspots of overall positive, neutral, and negative listings remained basically in the same area after the pandemic, in New York, a relevant amount of listings located in Brooklyn, received mostly negative and neutral opinions which created a new hotspot on this area.

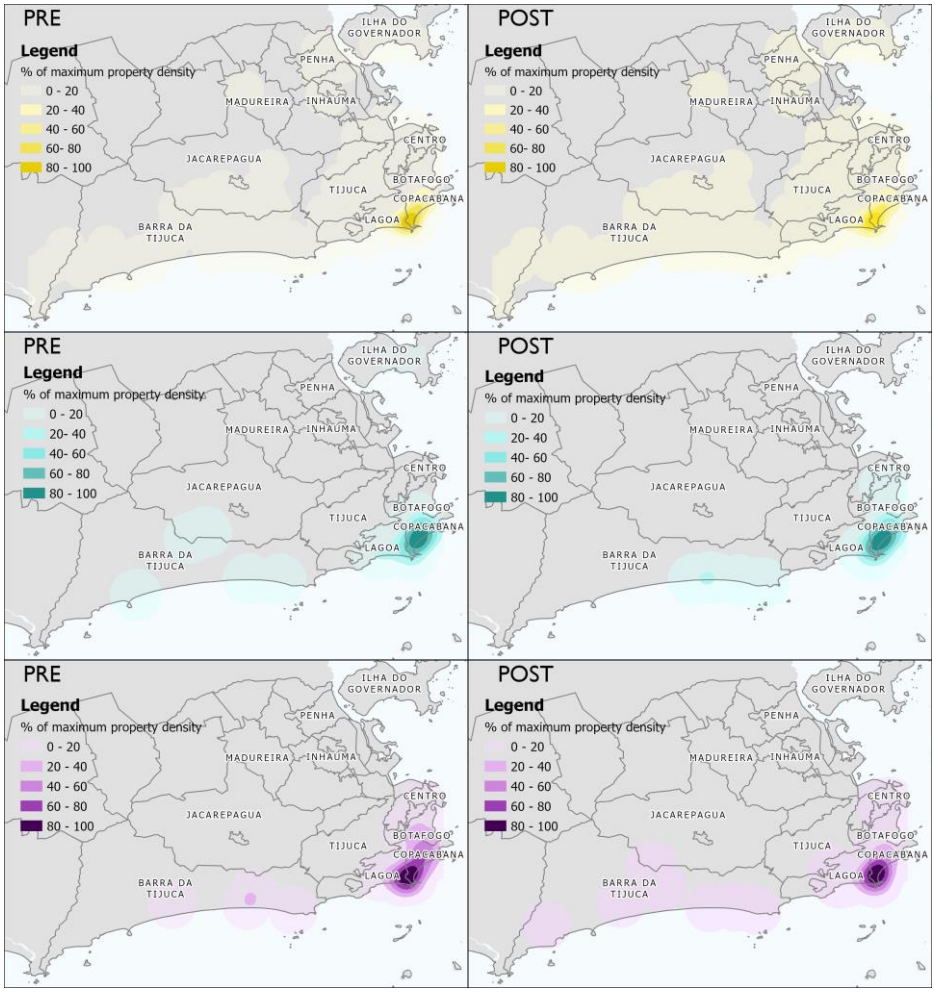


Figure 2. Spatial distribution of listings in Rio de Janeiro with overall positive (yellow), neutral (aquamarine), and negative (purple) polarity before and after the outbreak of Covid-19.

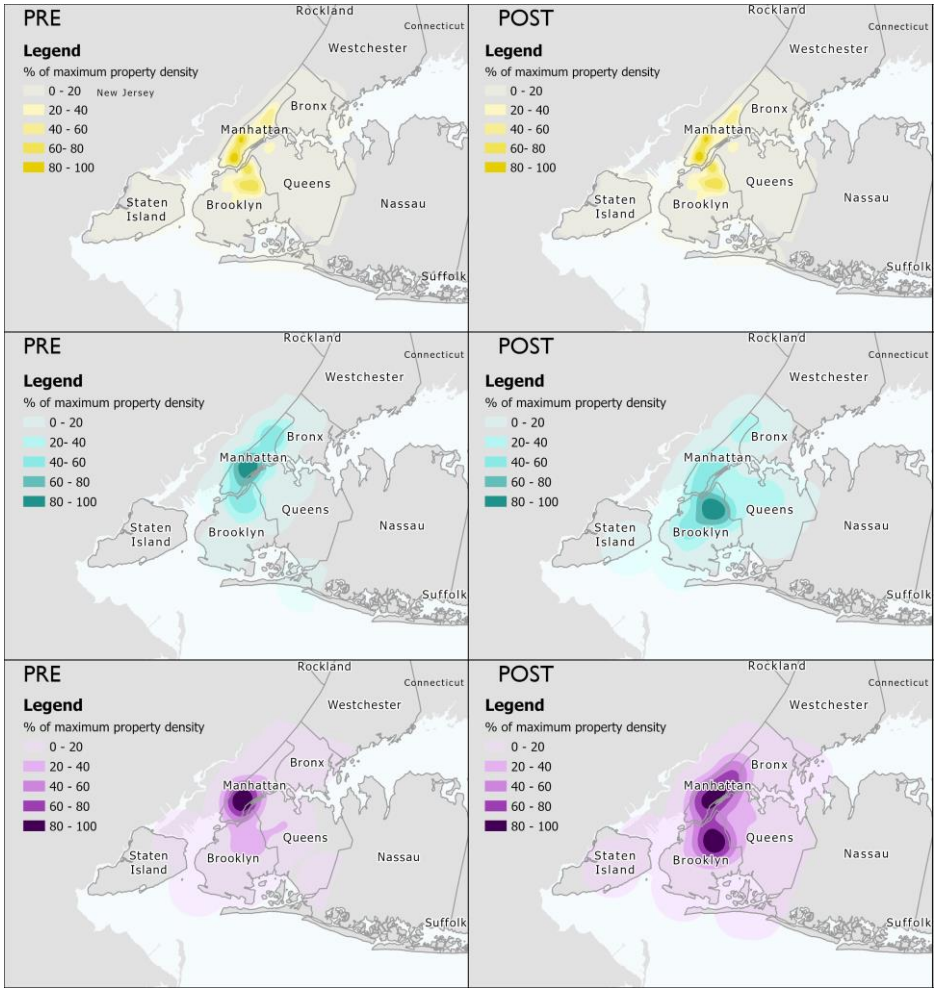


Figure 3. Spatial distribution of listings in New York with overall positive (yellow), neutral (aquamarine), and negative (purple) polarity before and after the outbreak of Covid-19.

4. Discussion and conclusions.

This research offers an alternative approach to the study of customer experience in tourism and hospitality literature. The methodology contributes by illustrating how spatial analysis can be combined with NLP techniques to visualize the role of location in customer experience during health crises. The findings show that after the pandemic, in New York, a new area of the city became a hotspot of neutral and negative reviews, which raises the question of why after the pandemic, in this area a relevant amount of listings experienced an increase in negative and neutral reviews and whether the characteristics of this area

could have negatively influenced the perception of customers. Thus, reviews from listings located in this area deserve further analysis, as they might reveal useful insights that lead to a better understanding of customer needs and expectations during health crises.

Results were contrasting, but also suggest that different areas of a city might play a new role in customer experience during health crises. However, at this stage, these findings can not be generalized, and therefore the need to extend this analysis to other cities including those that are not popular touristic destinations or that were not severely affected by the pandemic.

References.

Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. 10.

Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. Tourism Management, 68, 301–323. https://doi.org/10.1016/j.tourman.2018.03.009

Wang, S., & Chen, J. S. (2015). The influence of place identity on perceived tourism impacts. Annals of Tourism Research, 52, 16–28. https://doi.org/10/f7dgqj

Wyrwoll, C. (2014). User-Generated Content. In C. Wyrwoll (Ed.), Social Media: Fundamentals, Models, and Ranking of User-Generated Content (pp. 11–45). Springer Fachmedien. https://doi.org/10.1007/978-3-658-06984-1_2

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15100

Cracking the Code of Geo-Identifiers: Harnessing Data-Based Decision-Making for the Public Good

Patricia Snell Herzog
Indiana University Lilly Family School of Philanthropy; Department of Human-Centered Computing, School of Informatics & Computing; Department of Sociology, IUPUI, USA

Abstract

The accessibility of official statistics to non-expert users could be aided by employing natural language processing and deep learning models to dataset lexicons. Specifically, the semantic structure of FIPS codes would offer a relatively standardized data dictionary of column names and string variable structure to identify: two-digits for states, followed by three-digits for counties. The technical, methodological contribution of this paper is a bibliometric analysis of scientific publications based on FIPS code analysis indicated that between 27,954 and 1,970,000 publications attend to this geo-identifier. Within a single dataset reporting national representative and longitudinal survey data, 141 publications utilize FIPS data. The high incidence shows the research impact. Yet, the low proportion of only 2.0 percent of all publications utilizing this dataset also shows a gap even among expert users. A data use case drawn from public health data implies that cracking the code of geo-identifiers could advance access by helping everyday users formulate data inquiries within intuitive language.

Keywords: *Geospatial data; Big data; Official statistics; Bibliometrics.*

1. Introduction

Many official statistics provide publicly available datasets of social patterns that could be harnessed in making informed decisions. The funding to collect and share these data is often supported through public entities due to the potential for data to have broad applications that benefit the public good. However, the inaccessibility of the data structure is a barrier to broader use. While data can readily be downloaded, users need to understand the data lexicon, including the meta-data, data dictionary, and most importantly the meaning that can be extracted from data variable names and labels.

A fundamental problem that prevents broader accessibility of publicly available data is the expert-level vocabulary embedded within the syntax of complex datasets. Everyday people and knowledge workers are required to decode this syntax in order to understand the information the data offer. Datasets that have existed for a long time carry a layered legacy of complex codes and dictionary structures that are difficult to make sense of and disentangle. This data syntax is crucial for understanding the meaning of the available variables, and ultimately the kinds of answers that a dataset can provide. However, the complexity of the data syntax obscures the meaning of the data for non-expert users. Explicating this syntax can lessen the expert-level barriers inhibiting broader data usage.

This paper focuses on a common attribute of datasets that is crucial for extracting actionable insights: geospatial data. The geographic location of data are often coded within a relatively controlled vocabulary of geo-identifiers. GeoIDs are frequently coded within a fairly standardized and finite set of codes, and thus the lexicon of geospatial data is a prime syntax to detect in automation procedures. Machine learning can be utilized to detect semantics of GeoIDs by developing data dictionaries with common location attributes.

2. Geospatial Data

Geospatial data can connect across otherwise disparate facets the data pipeline, from data acquisition to analysis and visualization (Breunig et al., 2020). Moreover, geospatial data aid replicability of information and analysis techniques across distinct datasets, questions, researchers, and stakeholders: from academia to urban planning (Lee and Kang, 2015). Yet, despite shared semantic foundations in geospatial data ontology, heterogeneity in data lexicons remains a barrier to broader sharing and accessibility (Sun et al., 2019).

2.1. Controlled Vocabularies

Rieder (2020) compares column names to the contracts or promises that software products make with users, with the caveat that published data tables provide a service to users that is more ambiguous. Data consumers are offered information but without clear parameters. Within this context, column names provide an informal contract with the data user

regarding what information can be harnessed from which variable. As contracts benefit from a relatively standard set of vocabulary to express the promises that users can expect, Rieder asserts that engaging a controlled vocabulary within column names can serve as a latent contract between data producers and consumers, with accessibility, integration, and transferability promised within a recognizable lexicon. For example, ID can be used to indicate a uniquely identified entity in the dataset, and N can be used for sample counts.

2.2. Geographic Identifiers (GeoIDs)

Geo-identifiers indicate to which geographic units the data can be aggregated, and geo-identifiers are nearly ubiquitous within publicly available datasets. Many public entities are geopolitically structured at the level of country, state, city, county, and thus associated data are also imbued with these geographic units. The United States Census Bureau (2021) states that: "GEOIDs are numeric codes that uniquely identify all administrative/legal and statistical geographic areas for which the Census Bureau tabulates data. From Alaska, the largest state, to the smallest census block in New York City, every geographic area has a unique GEOID." In official statistics, there are several different GeoID coding systems, such as the American National Standards Institute (ANSI), Geographic Names Information System (GNIS), and Federal Information Processing Series (FIPS) codes.

2.3. Federal Information Processing Series Codes (FIPS)

There are two primary appeals of focusing on the lexicon for Federal Information Processing Series (FIPS) codes. First, FIPS codes have a high degree of standardization and nested structure in the data dictionary (US Census Bureau 2020; see Figure 1). Second, there is broad utilization of FIPS codes across a range of analyses and within subsidiary datasets (see for example: Mullen and Bratt, 2018; Brown et al., 2020; Boland et al., 2017; Roberts et al., 2014). The next section quantifies the scientific research impact of FIPS.

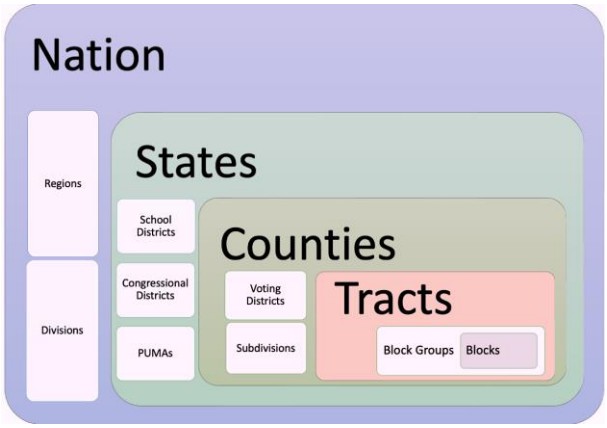


Figure 1. Nested Geographic Entities. Source: Author creation based on US Census (2020).

3. Bibliometric Data

Bibliometrics is a methodological approach that focuses on scientific literature as the subject of analysis (Ball, 2017). Many of these techniques focus on citation analysis, including content analysis of titles, keywords, abstracts, and full text of published journal articles, books, conference proceedings, dissertations, and reports (Zhao and Strotmann, 2015). Applying scientific techniques to citation analysis facilitates a statistical evaluation and measurement of influence within the scientific community (Iftikhar et al., 2019). Citation analyses have been utilized in studying research impact from social work (Holden et al., 2012) to the humanities (Ochsner et al., 2016). This paper presents a bibliometric analysis of FIPS code impact within scientific literatures and official statistical datasets.

3.1. Census Data

Searching census data within a popular scholarly bibliometric database Google Scholar returns 4,230,000 results, and 1,970,000 of these entries were published since 2010. Moreover, 6,400 publications cite FIPS codes within census data analysis. Computing the same set of analyses within Scopus respectively returns: 47,761; 27,954; and 27. Combined, these results indicate the high degree of research impact that FIPS codes have.

3.2. PSID Data

Additionally, FIPS codes have been utilized within subsidiary datasets. For example, the Panel Study of Income Dynamics (PSID) is a longitudinal survey of a nationally representative sample of U.S. based more than 18,000 individuals within 5,000 families. Data have been collected in over 40 waves of data spanning multiple generations of descendants from the original respondents. In the PSID bibliographic database of citations (PSID, 2022), there are a total of 7,033 publications in this database, of which 4,892 are journal articles, 785 book chapters, 92 books, 1,180 dissertations, and 84 reports. The PSID uses FIPS codes, and Table 1 displays the bibliometric data respective to each geography.

Table 1. Table captions should appear *above* tables.

Geographic Entity	Count	Percent
Tract	31	0.4
County	03	0.0
Metropolitan Area	22	0.3
Region	20	0.3
Urban / Rural	65	0.9
TOTAL	141	2.0

Source: Author creation based upon the PSID (2022).

4. Deep Learning

In order to harness the power of geo-identifiers to unlock the information in the thousands of scientific publications summarized in the previous section, it is necessary to correctly detect the semantic structure. Though the geo-identifier lexicon is fairly complex and typically embedded in dirty data, it is also finite and more standardized than unstructured text. The controlled vocabulary of FIPS codes offers an opportunity to apply a multi-input deep neural network for detecting semantic types, such as Sherlock (Hulsebos et al., 2019).

4.1. GeoID Structure

As displayed in Figure 1, the GeoID structure of FIPS codes is nested. Specifically, the structure begins with a 2-digit state code, such as 18 for Indiana. This is followed by a 3-digit county code, such as 097 for Marion County, which includes the city of Indianapolis. The nested structure to FIPS codes is such that these identifiers can be combined into a 5-digit code = 2 for state + 3 for county: 18097. The standardization of the digit format in FIPS codes lends itself to a detectable lexicon, as the data dictionary can be trained to identify recurring combinations of 2-digit, 3-digit, and 5-digit string data as a geography. Moreover, the deep learning process can be improved by harnessing column names, which would be a finite variety of for instance: state, STATE, sta, county, COUNTY, cty.

Continuing to smaller geographies, the Census approximation of a neighborhood is a tract, which has a 6-digit code. For example, one tract within Marion County is 310104, and a neighboring tract is 310105. These are sometimes designated with a period after the first four digits, as: 3101.04 and 3101.05. This indicates that a broader set of neighborhoods can be grouped together within the 3101 identifier. However, some datasets would omit the period and others would not. This presents a complication to automatically detecting the semantic lexicon, yet again the dictionary in column names is limited: tract, trc, tra, ct. Moreover, an 11-digit string variable (state+county+tract) is readily identifiable.

4.2. Public Health Data

Applied to a specific data use case, the Indiana Department of Health provides a data hub of publicly available data regarding health issues (IN.MPH 2022a). Currently, this data hub has a prevalent array of datasets reporting Covid-19 rates: tests, cases, deaths, trends over time. One available dataset is for Covid-19 county statistics (IN.MPH 2022b). The meta-data include this additional information: Spatial/Geographic Coverage – State of Indiana; Granularity – Aggregate, County-Level. The data dictionary reports four fields: County (Reported county where patient resides), Case_Count (Number of reported Covid-19 cases), Death_Cases (Number of reported Covid-19 deaths), and Lab_Tests (Number of reported individuals with a resulted Covid-19 test).

Yet, the dataset actually contains six fields, with the addition of _id and Location_ID. Plus, County is not labeled as described in the data dictionary but is rather labeled: County_Name. Thus, to train a machine learning model well, it is necessary to identify the concactanated column name of county to match it to the data dictionary. More importantly, in terms of harnessing the power of FIPS codes, it is crucial to identify the string pattern of Location_ID. In this example, all the values in this field begin with 18, which is the state FIPS for Indiana. This is readily visually detectable to a human, at least one with the necessary expertise to recognize the state digits. Another human-detectable clue is that there are a total of 92 counties in the state of Indiana, and the dataset has 92 entries.

If a deep learning model is trained to recognize Location_ID as a FIPS code, then the power of the dataset can be harnessed through identification of its semantic structure. Only then could a non-expert user be automatically provided with a natural language description of the information: Covid-19 tests, cases, and deaths by Indiana county. Building upon the intuitive human curiosity structure of questions, this data could generate answers to the question: Which counties in Indiana have the highest Covid-19 rates? Even more complex, the dataset could also offer answers to an inquiry regarding the preventive contributions of testing through responding to the question: Do counties in Indiana that have a lower deaths to cases ratio have a higher test rate?

5. Conclusion

In conclusion, the accessibility of official statistics to non-expert users could be aided by employing natural language processing and deep learning models to dataset lexicons. Specifically, the semantic structure of FIPS codes would offer a relatively standardized data dictionary of column names and string variable structure to identify. The bibliometric analysis indicated that decoding this structure would enable access to the insights included in thousands of scientific publications. The datasets embedding FIPS codes span from macro-level geopolitical units in census data, to public health data aggregated to counties, to individual and family survey data aggregated to tracts, counties, stages, and regions. Thus, cracking the code of geo-identifiers could advance access by everyday people by helping users to formulate data-based inquiries within their intuitive language.

Acknowledgments

The author is grateful to Davide Bolchini, Rama Sai Arun Varma Pensmatsa, Anshuman Dixit, and Rahul Yadav for collaborations on the question-generation project. Additionally, the author is grateful to Laurie Paarlberg for her collaborations in conceiving of data as a public and philanthropic resource; Una Okonkwo Osili and Mark Ottoni-Wilhelm for their

contributions to the Panel Study of Income Dynamics; and the National Science Foundation for funding a human-technology frontier workshop that informed this project (1934942).

References

Ball, R. (2017). *An Introduction to Bibliometrics: New Development and Trends*. Chandos Publishing.

Boland, M. R., Parhi, P., Gentine, P., & Tatonetti, N. P. (2017). Climate Classification is an Important Factor in Assessing Quality-of-Care Across Hospitals. *Scientific Reports*, *7*(1), 4948. https://doi.org/10.1038/s41598-017-04708-3

Breunig, M., Bradley, P. E., Jahn, M., Kuper, P., Mazroob, N., Rösch, N., Al-Doori, M., Stefanakis, E., & Jadidi, M. (2020). Geospatial Data Management Research: Progress and Future Directions. *ISPRS International Journal of Geo-Information*, *9*(2), 95. https://doi.org/10.3390/ijgi9020095

Brown, C. C., Moore, J. E., Felix, H. C., Stewart, M. K., & Tilford, J. M. (2020). County-Level Variation in Low Birthweight and Preterm Birth: An Evaluation of State Medicaid Expansion under the Affordable Care Act. *Medical Care*, *58*(6), 497–503. https://doi.org/10.1097/MLR.0000000000001313

Holden, G., Rosenberg, G., & Barker, K. (Eds.). (2012). *Bibliometrics in social work*. Routledge.

Hulsebos, M., Hu, K., Bakker, M., Zgraggen, E., Satyanarayan, A., Kraska, T., Demiralp, Ç., & Hidalgo, C. (2019). Sherlock: A Deep Learning Approach to Semantic Data Type Detection. *ArXiv:1905.10688*. http://arxiv.org/abs/1905.10688

Iftikhar, P. M., Ali, F., Faisaluddin, M., Khayyat, A., De Gouvia De Sa, M., & Rao, T. (2019). A Bibliometric Analysis of the Top 30 Most-cited Articles in Gestational Diabetes Mellitus Literature (1946-2019). *Cureus*, *11*(2), e4131. https://doi.org/10.7759/cureus.4131

Lee, J.-G., & Kang, M. (2015). Geospatial Big Data: Challenges and Opportunities. *Big Data Research*, *2*(2), 74–81. https://doi.org/10.1016/j.bdr.2015.01.003

Mullen, L. A., & Bratt, J. (2018). USAboundaries: Historical and Contemporary Boundaries of the United States of America. *Journal of Open Source Software*, *3*(23), 314. https://doi.org/10.21105/joss.00314

Ochsner, M., Hug, S. E., & Daniel, H.-D. (Eds.). (2016). *Research Assessment in the Humanities*. Springer International Publishing. https://doi.org/10.1007/978-3-319-29016-4

PSID. (2022). *Panel Study of Income Dynamics Bibliography Search*. https://psidonline.isr.umich.edu/Publications/Bibliography/search.aspx

Riederer, E. (2020, September 9). *Column Names as Contracts: Using controlled dictionaries for low-touch documentation, validation, and usability of tabular data*. GitHub. https://github.com/emilyriederer/website/issues/9

Roberts, J. D., Voss, J. D., & Knight, B. (2014). The Association of Ambient Air Pollution and Physical Inactivity in the United States. *PLOS ONE*, *9*(3), e90143. https://doi.org/10.1371/journal.pone.0090143

Sun, K., Zhu, Y., Pan, P., Hou, Z., Wang, D., Li, W., & Song, J. (2019). Geospatial data ontology: The semantic foundation of geospatial data integration and sharing. *Big Earth Data*, *3*(3), 269–296. https://doi.org/10.1080/20964471.2019.1661662

US Census Bureau. (2020, November). *Standard Hierarchy of Census Geographic Entities*. https://www2.census.gov/geo/pdfs/reference/geodiagram.pdf

US Census Bureau. (2021, October 8). *Understanding Geographic Identifiers (GEOIDs)*. https://www.census.gov/programs-surveys/geography/guidance/geo-identifiers.html

Zhao, D., & Strotmann, A. (2015). *Analysis and visualization of citation networks*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00624ED1V01Y201501ICR039

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15103

Non-conventional data and default prediction: the challenge of companies' websites

Lisa Crosato[1], Josep Domenech[2], Caterina Liberati[3]
[1]Department of Economics, Ca' Foscari University of Venice, Italy, [2]Department of Economics, Universitat Politècnica de València, Spain, [3]Department of Economics Management and Statistics, University of Milano-Bicocca, Italy.

Abstract

Small and Medium Enterprises (SMEs) contribution to the European Union economy has always been relevant, for both value added and the creation of jobs. That is why the prediction of their survival is considered one of the economic pillars UE keeps under observation. Default prediction models, accounting for SMEs idiosyncratic traits, are based on several types of data, mainly accounting indicators. Balance sheet data, indeed, are considered the standard predictors for classification models in this field, although they do not allow to completely overcome the information opacity that is one of the main barriers preventing these firms from accessing credit. In our work, we explore the possibility of complementing accounting information with data scraped from the firms' websites. We modeled the data using a nonlinear discriminant analysis and we benchmarked the results with the Logistic Regression. The evidence of our study is promising although the combination of online and offline data shows better results in case of survival firms than for defaulted companies.

Keywords: *Website Data; SMEs; Default Prediction; Kernel Discriminant.*

1. Introduction

Economic studies on businesses present a wide literature on forecasting SMEs default. The great interest that scholars and practitioners have been showing toward this particular topic is given by the combination of two aspects: first, SMEs are 99% of all enterprises in the European Union, covering the largest part of the European value added and jobs (56.4% and 66.6% respectively, European Commission, 2019); second, the access to credit for these companies is difficult, especially in their early stages, due to information opacity. This makes the assessment of the creditworthiness of a SME a relevant issue, particularly in a policy maker and credit lender perspective (Cultrera, 2020; Belghitar et al, 2021).

To get low misclassifications' rates between survival and defaulted companies, researchers has focused mainly on the derivation of credit-scores based on Machine Learning (ML) algorithms (eg. Random Forest, Support Vector Machines), because they have shown the best performances with respect to the standard linear classifiers (Baesens et al, 2003; Lessmann et al, 2015). Input variables of such models are, generally, accounting indicators derived from balance sheets (Fantazzini & Figini, 2009; Succurro & Mannarino, 2014; Ciampi, 2015) provided by databases as Bureau van Dijk. Indeed, the quality of these data are really high - thanks to well established data-collection procedures- but on the same time, they suffer of a large delay between their availability and their reference period. Unfortunately, this drawback prevents a prompt prediction of default, diminishing the value of the results in a forecasting perspective.

In our paper we propose the use of websites as an additional source of information for detection of SMEs default (Crosato et al, 2021). The analysis of corporate websites to get new proxies of the standard business indicators has been already investigated in previous works: for capturing different corporate culture dimensions (Overbeeke & Snizek, 2005), or for measuring firm performances (Merono-Cerdan & Soto-Acosta, 2007) and the level of innovation (Axenbeck & Breithaupt, 2021).

The employment of these unconventional data requires an articulated architecture of data pre-processing, including specific procedures for data retrieval, cleaning and dimension reduction. Naturally, the knowledge extraction process is not straightforward: it is directly correlated with the complexity of the analyzed data. On the other hand, the gain in terms of additional information, which is available and free, largest coverage of the firms' population as well as the recency of the obtaining data rewards the analytical efforts.

The web-based indicators could be generated reviewing the corporate websites one by one using a manual evaluation of the features (Blazquez & Domenech, 2018), although it is very time consuming and not recommended when monitoring a high number of companies. In this study we apply instead an automatic process that photographs the websites at a given time and also tracks their changes. Using the Wayback machine, a digital archive of the World

Wide Web, we were able to see archived versions of web pages across time and compare their evolutions.

We analyze a sample of 700 SMEs sampled from the SABI - Sistema de Análisis de Balances Ibéricos (Bureau van Dijk). The database we built merges the accounting (offline) indicators and the web-based (online) indicators. We aim to verify whether website features help to predict corporate bankruptcy and in particular which indicators, among the online ones, can be selected to discriminate between surviving and defaulted firms. We also study if the joint use of online and offline information aids to reduce misclassifications error rates both using a nonlinear discriminant model and a Logistic regression.

2. Websites Data

To understand the process to create online indicators, it is necessary to illustrate how websites are built and organized. A website is a set of documents stored in a web server. The Hyper Text Markup Language (HTML) is the language used for setting the formatting and the layout of the hypertextual documents, including all the specifications of the webpages structure. A corporate website can be studied using two approaches: mining the web structure or the web content. The two approaches focus on different characteristics of the sites: the first one detects the linkages among the web-pages, the second one concentrates on understanding the semantics and the meaning of the contents. In our paper we rely on the latter, because it best describes the business activity.

In practice, we carry out the study from a twofold perspective:

- **Textual analysis** - It is useful to retrieve all the relevant information shared on the website by the companies: the economic sector in which they operate, market orientation (e.g., national or international, final consumer or other businesses) the locations and the activities done. Additionally, any editing/updating in the text can be reviewed as a sign of investment in the online communication and consequently of active behavior. The textual analytics, which encompasses a large variety of data manipulation processes (eg. cleaning and words stemming), helps to identify meaningful terms with the highest occurrences. This way, the unstructured text is converted into a set of dummies variables that can be later used by the classifiers.

- **HTML code**- The analysis of the HTML code provides important information about the tag of a webpage. The tags carry knowledge about the interaction (e.g., defining hyperlinks or forms), appearance (e.g., bold or italics), and the structure (e.g., defining lists or different blocks) of a web page. They evolve according to the progress of HTML language, so they indirectly represent the complexity and the level of technology of a

website. For instance, EMBED is generally employed to include Flash technology, which is currently being abandoned; FORM is usually employed to interact with the company/site. The hyperlinks connections are listed as tag A, they include all the internal/external connections present in the website (href), the extensions of the files shared (pdf, excel, word) and the underlying technology (php/asp/htm). The images are listed as tag IMG with their correspondent extension.

3. Online features

Before dealing with any classification model, we start by simply reducing the number of available features to the ones showing significant differences between the groups of survived and the group of defaulted SMEs. To this purpose we calculate, within each group, the proportion of firms on whose website the considered feature was present and then apply a standard test for the difference in proportions to each of the features. Significant differences were found for three main groups of features: 7 in the category "Hrefwords", 20 in the category "Stems" and 18 in the category "Words", plus 4 in a residual miscellaneous category. The majority of features making the difference appear more often in the survived group, particularly in the "Hrefwords" category.

4. Methods

The online indicators we are working with are now binary and this could restrict the application of certain models. Therefore, we transform them into numerical orthogonal factors via Multiple Correspondence Analysis (Greenacre, 1984). As about accounting indicators, the literature suggests to take care of possible associated non-linear patterns, and we expect that including the online features in the analysis will add complexity to the within-variable relationships. Another aspect compounding the classification task is the overbalance between survived and defaulted companies. For assessing both issues, we refer to the wide range of statistical techniques in the relevant literature on SMEs bankruptcy prediction.

Machine-learning methods have been generally found to outperform the linear ones. Non-linear models such as Deep Learning (Mai et al., 2019), Boosting (Kou et al., 2021) and Neural Networks (Baesens et al., 2003) have been successfully employed, maintaining the z-score (Altman, 1968) or the logistic regression (Hosmer and Lemeshow, 2000) as a benchmark. Here we use Kernel Discriminant Analysis (KDA, Mika et al., 1999) due to very good performances of kernel-based algorithms in screening SMEs (Gordini, 2014; Zhang, 2015).

The goal of KDA is to provide a decision function *f(x)* from the combination of features that best separates two classes of objects. Given a training set $I_{XY} = \{(x_1, y_1), \dots, (x_n, y_n)\}$

describing n units, with data $x_i \in R_p$ and labels $y_i \in \{0,1\}$, due to the recurrent non-linear separability of training data in the input space R_p, the training data are usually associated to some feature space F via a non-linear mapping:

$$\varphi: x_i \rightarrow \varphi(x_i) \qquad (1)$$

F can be referred to as a Reproducing Kernel Hilbert Space (RKHS) when the Mercer's theorem is satisfied (Mercer, 1909).

The ratio of the Between and Within covariance matrices (computed in the Feature Space) is then minimized to obtain a separating hyperplane in F, as in the Fisher Discriminant Analysis (Fisher, 1936):

$$g(x) = w^T \varphi(x) + b \qquad (2)$$

where w is the weight vector in RKHS, and $b \in R$ is the bias term.

Getting an optimal generalization of kernel-based methods still requires choosing a suitable kernel map. Among the many proposed in the literature we have selected the Radial Basis Function (RBF), Laplace, Cauchy and Multiquadric kernels due to their remarkable performances.

Acknowledgments

This work was partially supported by grants PID2019-107765RB-I00 and funded by MCIN/AEI/10.13039/501100011033.

References

Altman, E.I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance 23(4), 589–609

Axenbeck, J. & Breithaupt, P. (2021). Innovation indicators based on firm websites - which website characteristics predict firm-level innovation activity? PloS one 16(4), e0249, 583

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society 54(6), 627–635

Belghitar, Y., Moro, A. & Radić, N. (2021). When the rainy day is the worst hurricane ever: the effects of governmental policies on smes during covid-19. Small Business Economics pp. 1–19

Blazquez, D. & Domenech, J. (2018). Web data mining for monitoring business export orientation. Technological and Economic Development of Economy 24(2), 406–428

Ciampi, F. (2015) Corporate governance characteristics and default prediction modeling for small enterprises. an empirical analysis of Italian firms. Journal of Business Research 68(5), 1012–1025

Crosato, L., Domenech, J. & Liberati, C. (2021) Predicting SMEs default: Are their websites informative? Economics Letters 204, 109,888

Cultrera, L. (2020). Evaluation of bankruptcy prevention tools: evidences from COSME programme. Econ. Bull. 40(2), 978–988

European Commission (2019). Annual Report on European SMEs 2018/2019. Tech. rep.

Fantazzini, D. & Figini, S. (2009) Default forecasting for Small-Medium Enterprises: Does heterogeneity matter? International Journal of Risk Assessment and Management 11(1-2), 138–163

Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems. Ann. Eugen. 7(2), 179–188

Gordini, N. (2014)A genetic algorithm approach for SMEs bankruptcy prediction: Empirical evidence from Italy. Expert Systems with Applications 41(14), 6433–6445

Greenacre, M.J. (1984) Theory and applications of correspondence analysis

Hosmer, D., Lemeshow, S. (2000) Applied Logistic Regression. Wiley

Kou, G., Xu, Y., Peng, Y., Shen, F., Chen, Y., Chang, K., Kou, S. (2021) Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection. Decision Support Systems 140, 113,429

Llopis, J., Gonzalez, R., Gasco, J. (2010) Web pages as a tool for a strategic description of the spanish largest firms. Information processing & management 46(3), 320– 330

Mai, F., Tian, S., Lee, C., Ma, L. (2019) Deep learning models for bankruptcy prediction using textual disclosures. European Journal of Operational Research 274(2), 743–758.

Mercer, J. (1909) Functions of positive and negative type, and their connection the theory of integral equations. Philosophical transactions of the royal society of London, A 209, 415–446.

Mika, S., R$\ddot{a}$tsch, G., Weston, J., Sch$\ddot{o}$lkopf, B., M$\ddot{u}$ller, K.R. (1999) Fisher discriminant analysis with kernels. In: Neural networks for signal processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop, pp. 41–48

Overbeeke, M. & Snizek, W.E. (2005) Web sites and corporate culture: A research note. Business & Society 44(3), 346–356

Succurro, M. & Mannarino, L. (2014). The Impact of Financial Structure on Firms' Probability of Bankruptcy: A Comparison across Western Europe Convergence Regions. Regional and Sectoral Economic Studies, Euro-American Association of Economic Development, 14(1), 81–94

Zhang, L., Hu, H., Zhang, D. (2015) A credit risk assessment model based on SVM for small and medium enterprises in supply chain finance. Financial Innovation 1(1), 1–21

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15104

Automated Information Retrieval from the Bibliographic Metadata: A Way to Facilitate the Systematic Literature Review

Marie Vítová Dušková[1], Martin Víta[2]
[1]Department of Marketing, Faculty of Business Administration, Prague University of Economics and Business, Czech Republic, [2]Department of Mathematics, Faculty of Informatics and Statistics, Prague University of Economics and Business, Czech Republic.

Abstract
The aim of this paper is to demonstrate the possible enrichment of the traditional procedure of bibliographic literature review using Natural Language Processing (NLP) methods – automated information retrieval. Our task was to conduct a systematic review of academic literature focused on the classical music audience research in the context of arts management and arts marketing. As a core base, we used bibliographic metadata, extracted from the Scopus database. The limits of the most commonly used methods of bibliographic analysis of the literature, which are co-citation analysis and bibliographic coupling, are well known. Therefore, we also used one of the NLP methods for metadata analysis, which allows automated processing of large numbers of texts to overcome these known problems. Thanks to this, we managed to obtain a higher granularity of the researched topics, to reveal emerging topics and to identify gaps in research. To the best of our knowledge, such an approach to the systematic literature review in the field of social sciences has not yet been applied.

Keywords: *Bibliographic Analysis; Natural Language Processing; Bibliographic Metadata; Art Audience Research.*

1. Introduction

Literature reviews play a crucial role in academic research in gathering existing knowledge and examining the state of the art. Among the many types of reviews that exist (from critical to post-publication reviews), systematic reviews of the literature are the most informative and scientific, however, only if they are consistently implemented and well justified (Paul et al., 2021).

It is common for scholars from the field of marketing and management to justify the search for a research question only on the basis of a cursory and narrative review of the literature (Linnenluecke et al., 2020). Unlike narrative literature review methods, meta-analysis, which is used in systematic literature reviews, allows us to statistically integrate and synthesize previous marketing and management research to prevent inconsistencies in the selection of documents included in the review and to create accumulated knowledge in given area.

As part of our long-term research focused on the classical music audiences, we analyzed published works during the period when, on the one hand, academic research in the field of marketing and art management developed (Colbert et al., 2014; Rentschler et al., 2006; Walmsley, 2019) and at the same time there was also a development and changes in the behavior and preferences of the audience (Prieur et al., 2013) (i.e., the development of the researched topic). In view of these facts, the goal of the systematic literature review was to 1) explore the scope of classical music audience research in the context of marketing and art management over time, 2) identify the most influential articles that were (or still are) the starting point of the research, 3) reveal current trends and perspectives in music audience research, and 4) identify unresolved issues and research subareas.

To achieve these goals, we performed two automated analyzes – bibliographic analysis and document affinity analysis.

2. Methods and Data

A systematic literature search requires a replicable, scientific and transparent process of evaluating existing knowledge (published in the literature) in order to minimize biases resulting from the random inclusion or exclusion of specific studies in the literature search process (Linnenluecke et al., 2020).

2.1. Data Collection

The basic precondition for a systematic review is the creation of a comprehensive or at least representative data set, which includes data on available research (Tranfield et al., 2003). Relevant documents for our research were first searched in the Scopus database using specific keywords and search strings. This search has been carefully documented. The result of this

search is a core base containing 188 documents. Subsequently, this set of documents was enriched with other documents found using Google Scholar based on reference analysis. The resulting file contains 257 documents. The dataset contains citation information, bibliographical information, the text of the abstract, keywords and references.

2.2. Bibliographic Analysis

Co-citation analysis is a bibliometric technique proposed by Small (1973), which aims to map the structure of the research field by analyzing groups of documents that are commonly cited together. The main disadvantage of co-citation analysis is that it is seen as an approach to the "past" of the research field, as it is more likely to capture older contributions and well-established researchers, rather than the current state of research. The papers in each cluster tend to share some common themes and are considered the basic knowledge base of the research area: the key concepts and methods on which the researchers build.

Bibliographic coupling can be interpreted as the opposite process to co-citation: two publications are marked as bibliographically paired if there is a third publication that is cited by both publications. Bibliographic coupling assumes that when two articles show similar bibliographies, they are likely to represent the same or at least related research topics. Because the citing documents are more recent than the cited documents, this method is suitable for examining newer contributions.

Both bibliometric analyzes were performed using the bibliometrix package programmed in `R` (Aria et al., 2017).

2.2. NLP Application -- Document Affinity Analysis

After performing data preprocessing, the corpus is represented as a standard document-term matrix – i.e., as a table whose rows correspond to documents (in our case papers), the columns then correspond to words (terms).

If a word does not appear in the document, then the value at the intersection of the corresponding row and column is set to zero, otherwise, a positive real number is used -- it expresses both the frequency of occurrence of the word in the document (“more frequent is more important'”) and frequency across the corpus (“words that are in a large number of documents are not so important'”) – more precisely, tf-idf weighting is used. Therefore, on each line of the matrix we find a vector that represents the given document (the components of the vector correspond to the dictionary we have available, which originated from the whole corpus). This way of representing entities belongs to vector representations.

Having a vector representation available for each document, we are able to measure the mutual distance of these documents, i.e., their vector representations. For this purpose, we use a standard cosine similarity – this expresses the affinity between each pair of documents.

This process implicitly leads to an undirected graph with weighted edges (vertices correspond to individual papers, the width of the edge/line between them expresses the degree of their similarity). For illustration we present a graph showing the similarity of authors – we do not present a graph showing similarity of documents for spatial reasons (Fig.1).

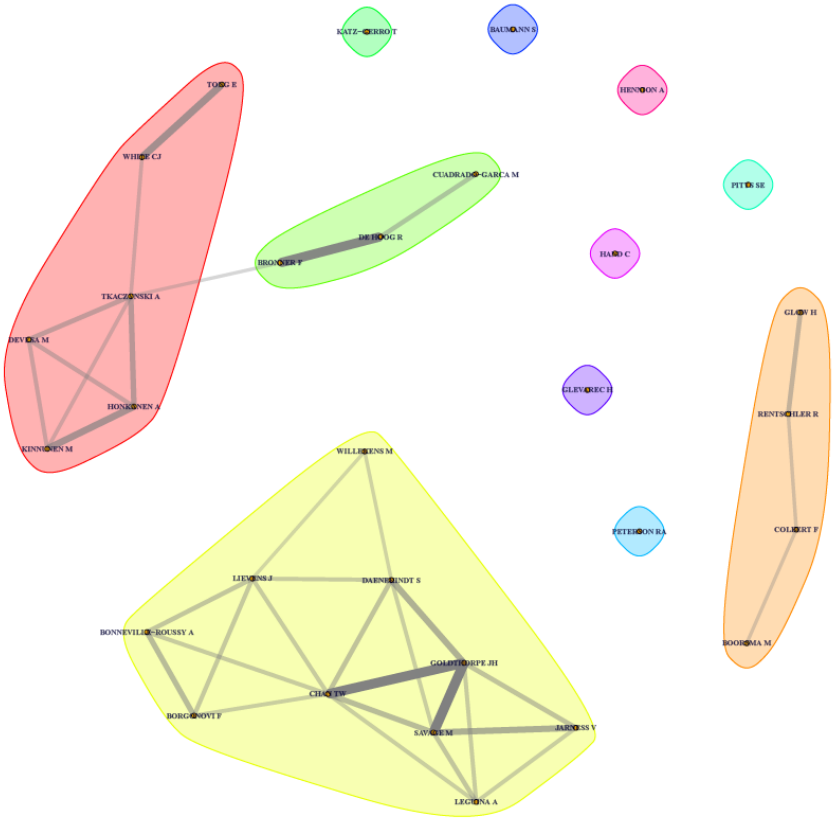


Figure 1 - Similarity of Authors

The WalkTrap algorithm implemented in `igraph` package (Sousa et al., 2014) was used to obtain clusters, i.e., communities of documents in our affinity/similarity graphs. As shown in Trigo et al. (2014), WalkTrap algorithm provides stable and useful results comparing to other approaches. The result can be represented in the form of a graph with highlighted sets of vertices. Formally, the output is a "key-value" table, in our case the title of the article (when analyzing the similarity of documents) or the name of the author (when analyzing the similarity of authors). Clusters can generally consist of a different number of elements (papers or authors), we will be particularly interested in those with the largest number of elements. Communities of authors are in fact induced by clustering (communities) of lists of

authors' papers. Hence the resulting author communities may contain authors with no common papers and no common citations etc.: they are grouped because they are related in the sense of working on the same or similar topic. This approach naturally extends approaches elaborated in Donthu et al. (2021).

3. Results

Each of the performed analyzes resulted in a number of clusters. In the case of co-citation analysis and bibliographic coupling, five clusters were created. In the case of document affinity/similarity analysis, more than twenty clusters were created. Upon closer analysis of the resulting clusters and after manual processing, their number was reduced to thirteen, because some clusters were very close thematically. Typically, four to six key references were published for each cluster; here, for space reasons, we only mention some of them (most cited in the bibliographic analysis, the most relevant in terms of the topic in the similarity analysis of documents).

3.1. Co-citation Analysis

The result of the co-citation analysis was five clusters, which were named according to the main topic of the documents included in the cluster. To save space, here are the cluster names and the key references for each cluster. More detailed papers' information, see the references in this paper.

- Cultural Capital
- Omnivorousness in Cultural Consumption
- Social Boundaries of Cultural Consumption
- Marketing Challenges in Audience Research
- Consumer Behavior

Among the key references for each cluster, we find very well-known texts and authors, e.g. Bourdieu (1984), Peterson (1992), Peterson et al. (1996), Lamont et al. (1992) or Van Eijck (2001).

3.2. Bibliographic Coupling

The bibliographic coupling resulted in five clusters. Their names correspond to the common topic of the documents included in the clusters.

- Cultural Stratification and Omnivorousness
- Musical Tastes and Musical Preferences
- Cultural Consumption Determinants, Arts Participation Boundaries
- Audience Research
- Festival Audiences and Event Marketing

Among the key references for these clusters, we alsofind very well-known texts and scholars, e.g. Peterson et al. (1996), Borgonovi (2004), Pitts (2005) or Bonneville-Roussy et al. (2013).

3.3. Document Affinity Analysis

The result of the document affinity analysis was thirteen clusters:

- Audience Development
- Audience Experience and Engagement
- Consumer Behavior
- Audience Segmentation
- Arts Marketing - Audience Research Theory
- Festival Audience
- Symphony Orchestra Audience
- Opera Audience
- Cultural Consumption and Social Stratification
- Arts Participation
- Tastes and Preferences
- Age and Musical Preferences
- Other Factors Influencing Arts Participation

At first glance, we see higher specialization of individual topics. In addition to older work, clusters include the latest work in the field, e.g. Soares-Quadros Junior et al. (2019), Daenekindt (2019), Vries et al. (2021), Kinnunen et al. (2019) and many others.

4. Discussion and Further Work

As we can see, the topics of the clusters from the individual analyzes overlap to some extent. Because the analyzes follow each other to a certain extent chronologically, we can see the development of the researched topics over time: in the citation analysis the topics are more general, in bibliographic coupling they are more specific in relation to our main research topic (i.e. classical music audience research) and document affinity analysis shows very specific research topics and directions. It is not without interest that, for example, the cluster "Other Factors Influencing..." includes only papers from 2014-2019, that means relatively new works.

The document affinity analysis divides all documents from the dataset into clusters, thanks to which we can then see the latest literature in the context of the older one. We have already mentioned that the main disadvantage of the bibliometric analyzes is that they provide a retrospective view of the researched field and draw attention to the most cited papers. To a certain extent, this can also lead to some distortions in future research – especially in research fields that are not widely exposed, there are certain communities of researchers in which there

is almost an obligation to cite some scholars, although their work may not be so crucial for specific research. Therefore, we see as a great advantage of document affinity analysis that it eliminates this influence – document affinity analysis only classifies documents into clusters based on their similarity and not importance within the scientific community.

Since affinity analysis seems to be a promising way to enrich "classical approaches"' to create systematic literature reviews, one of the directions of further research is to improve the computation of affinity by replacing tf-idf vector representations by state-of-the-art text embeddings arising from deep learning approaches (i.e., BERT).

Another direction is the development of a powerful visualization that incorporates both affinity results and co-citation/coupling results. Next step is then a development of a web based application, i.e., a web interface to our scripts in order to allow the user to create such reports and results interactively (without using our raw scripts exploited in this paper).

5. Conclusion

The aim of this paper is to demonstrate a possible approach to conducting a systematic literature search. Commonly used approaches – narrative literature processing on the one hand and bibliographic analysis on the other – may not always produce the desired results, especially if used in isolation. As we worked with a lot of data extracted from the web, we used tools that allowed us automated processing of such data. When processing a literature review, the researcher never avoids a certain amount of time-consuming manual work and careful study of found literature, yet our approach to the review makes it easier to discover research topics in the field (and to some extent quantify and visualize them in the context of other research topics), find research gaps or new research trends, thanks to the application of automated data processing.

References

Aria, M. & Cuccurullo, C (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of informetrics* 11.4, 959–975.

Bonneville-Roussy, A., Rentfrow, P.J., Xu M.K., & Potter J. (2013). Music through the ages: Trends in musical engagement and preferences from adolescence through middle adulthood. *Journal of personality and social psychology* 105.4, 703.

Borgonovi, F. (2004). Performing arts attendance: an economic approach. *Applied Economics* 36.17, 1871–1885.

Bourdieu, P (1984). Distinction: A Social Critique of the Judgement of Taste. *Cambridge, MA: Harvard UP.*

Colbert, F., & St-James Y. (2014). Research in arts marketing: Evolution and future directions. *Psychology & Marketing* 31.8, 566–575.

Daenekindt, S. (2019). Out of tune. How people understand social exclusion at concerts. *Poetics* 74, 101341.

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133, 285-296.

Kinnunen, M., Luonila, M., & Honkanen, A. (2019). Segmentation of music festival attendees *Scandinavian Journal of Hospitality and Tourism* 19.3, 278–299.

Lamont, M. et al. (1992). Money, morals, and manners: The culture of the French and the American upper-middle class. *University of Chicago Press.*

Linnenluecke, M. K., Marrone, M., & Singh, A. K. (2020). Conducting systematic literature reviews and bibliometric analyses. *Australian Journal of Management* 45.2, 175–194.

Paul, J., Lim, W. M., O'Cass, A., Wei Hao, A. & Bresciani, S. (2021). Scientific procedures and rationales for systematic literature reviews (SPAR-4-SLR). *International Journal of Consumer Studies.*

Peterson, R. A. (1992). Understanding audience segmentation: From elite and mass to omnivore and univore. *Poetics* 21.4, 243–258.

Peterson, R. A. & Kern, R. M. (1996). Changing highbrow taste: From snob to omnivore. *American sociological review,* 900–907.

Pitts, S. E. (2005). What makes an audience? Investigating the roles and experiences of listeners at a chamber music festival. *Music and letters* 86.2, 257–269.

Prieur, A. & Savage, M. (2013). Emerging forms of cultural capital. *European Societies* 15.2, 246–267.

Rentschler, R., Radbourne, J., Carr, R., & Rickard, J. (2006). Relationship marketing, audience retention and performing arts organisation viability. *International journal of nonprofit and voluntary sector marketing* 7.2, 118–130.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science* 24.4, 265–269.

Soares-Quadros J., Fortunato, J., Lorenzo, O., Herrera, L., & Santos, N. S. A. (2019). Gender and religion as factors of individual differences in musical preference. *Musicae Scientiae* 23.4, 525–539.

Sousa, F. B., de, & Zhao, L. (2014). Evaluating and comparing the igraph community detection algorithms. *2014 Brazilian Conference on Intelligent Systems. IEEE*, 408–413.

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British journal of management* 14.3, 207–222.

Trigo, L., & Brazdil, P. (2014). Affinity analysis between researchers using text mining and differential analysis of graphs. *ECML/PKDD 2014 PhD session Proceedings*, 169-176.

Van Eijck, K. (2001). Social differentiation in musical taste patterns. *Social forces* 79.3, 1163–1185.

Vries, R. de & Reeves, A. (2021). What does it mean to be a cultural omnivore? Conflicting visions of omnivorousness in empirical research. *Sociological Research Online*, 13607804211006109.

Walmsley, B. (2019). Understanding Audiences: A Critical Review of Audience Research. *Audience Engagement in the Performing Arts*, 25

4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15077

Refusing to be safe.
The Social Network Communication of deniers

Rosario D'Agata, Simona Gozzo
University of Catania, Italy.

Abstract

This essay aims at showing the results of analysis concerning communication on social networks by focusing on the collection of comments related to the pandemic. The analysis describes the structure of the communication, showing the presence of parallel communities and the different configurations of relational dynamics, selected contents, flows of communication, category of users, and language. Complex network structures are identified branching from keywords like no-mask, covid-19, and greenpass. Further attention is paid to the connection between online communication and the triggering of protests.

Keywords: *Social Media, Pandemic, Network Analysis, hubs.*

1. Introduction

This paper shows the results of an analysis conducted through the continuous monitoring of pandemic related posts on Twitter for the period 2020-2022. The specific objective is to analyse the structure of comments among those subjects criticizing governments' choices about how to face the risk of contagion. For this purpose, we extracted tweets containing 3 reference keywords. The extracted keywords have changed twice, in line with the changes in the public debate. Given the high number of views in the reference period, the first lemma was "No-mask" (from November 2020 to February 2021). Subsequently, the reduction of the communication led to a new selection: "covid-19", monitored until August 2021. The last phase of monitoring concerned the "Greenpass" lemma and lasted until December 2021. A new wave of protest emerged in this phase, whose media visibility is also evident, which echoes in the increase in communication with the "Greenpass" hashtag. This work is the first phase of the whole project, which also includes a semantic in-depth study of the emerging comments and thematic clusters. In this paper, however, we only present results concerning the underlying network structure of online communication.

2. Data and methods

The work refers to networks of communication among users. In particular, samples of tweets were extracted every two weeks from November 2020 until December 2021. The extraction was carried out via NodeXL Pro Twitter data importers (Smith et *al*., 2009), monitoring the communication every two weeks (Tab. 1).

Table 1. Number of tweets for each extraction.

Hashtag	**Data of Extraction**						
no mask	***30 Nov 2020***	***14 Dec 2020***	***30 Dec 2020***	***15 Jan 2021***	***05 Feb 2021***		
	1352	1749	2923	901	606		
covid 19	***23 Apr 2021***	***13 May 2021***	***03 Jun 2021***	***24 Jun 2021***	***15 Jul 2021***	***05 Aug 2021***	
	16802	17979	9767	12045	8825	18000	
greenpass	***05 Aug 2021***	***26 Aug 2021***	***16 Sep 2021***	***07 Oct 2021***	***28 Oct 2021***	***18 Nov 2021***	***09 Dec 2021***
	9141	9269	9407	7666	9004	8868	9450

The communication was analysed by building graphs and applying Social Network Analysis tools (Borgatti and Halgin, 2011), where users are defined as nodes and the comments constitute the links among them (Hansen *et al*., 2010; 2012). The proposed method permits the rapid extraction of information on network structure, shared meanings,

and main user categories given a large amount of data and comments on social networks. This can be useful for various reasons such as, for example, the evaluation of political choices, the spread of fake news, or marketing analysis.

As a first step, we selected the top-10-tweets for each extraction, so that it was possible to carry out an in-depth investigation of them. In this way, the hubs of the network are identified (i.e. those nodes on which the entire network structure depends) and the main information of the networks is reconstructed. As a second step, we applied the group analysis function and measured centrality, betweenness, and closeness (Junlong and Yu, 2017) to obtain further information.

The preliminary and in-depth analysis of the main comments refers to the hubs of the communication networks, selected through an automatic but controlled procedure. The analysis of these comments allows us to understand who are the main users, messages, groups, languages, etc. Then a quantitative study, referred to all comments, was carried out (Borgatti and Halgin, 2011). At this point, we selected - for each extraction - the entire communications structure and the main groups (as sub-networks or components) obtained by extracting clusters connected, with greater internal homogeneity and external heterogeneity of links.

3. The network hubs

The first phase of our analysis requires the identification of the main information beneath the whole network of communication so that we can understand the main content implied in a great number of comments just reading a limited number of these. We notice that the major part of the comments about “No-mask” is in English, Italian and French while those about “Covid-19” are in Asian and English languages (tweets from India, Japan, Pakistan, Thailand, United States, and Canada) and the most of “Greenpass” comments are in Italian.

Further information we obtained through this procedure concerns categories of users and topics. Overall, the detected communication is mainly private, but the “No-mask” one is almost exclusively private, while the most important tweets about “Covid-19” have a higher proportion of comments related to parties or institutions. This implies the presence of more reliable information compared to the one contained in no-mask and no-greenpass networks. The comments extracted using the lemma “Covid-19”, on the other hand, are more generic and referred to different topics. Furthermore, the major part of communication about “Covid-19” is produced both by individuals and institutions. The “covid-19” tweets, addressed both to other users and the institutions, came mainly from worried people protesting against the increase of contagions, the opening of shops, the lack of social distancing, and the lack of personal protection measures (use of masks, vaccinations, etc.).

Further comments are about the official information regarding the number of infections, availability of oxygen, and beds in the hospitals.

The communication that focuses on "Greenpass", mainly from private citizens, is among users with a high sense of political effectiveness and self-direction, largely oriented towards reaching and "influencing" political institutions and decision-makers. In particular, a strong Italian protest against the restriction policies ("greenpass") emerges during August-December 2021, while in November2020-February2021 Italians were mostly against the No-mask movement (Miller, 2020)

This first phase of our analysis, mainly descriptive, is useful when you want to identify information about languages, kinds of comments, "sentiments" and evaluations. We can identify the typical characteristics of three phases/keywords (Fig. 1). Otherwise, looking at the content of the communication, we have noticed that No-Mask nets form many groups and popular rumors (except during the Christmas period when a single large movement against restrictions and limitations appears). These comments are divided into two clear categories: derisive or ironic purposes (mostly in Italian), and the opposite: against any form of limitation of freedom (Fig. 2). The dynamics guiding the communication change when the "Covid-19" tweets are extracted. These comments are distributed among the different voices identified, despite the prevalence of neutral positions (which are usually marginal).

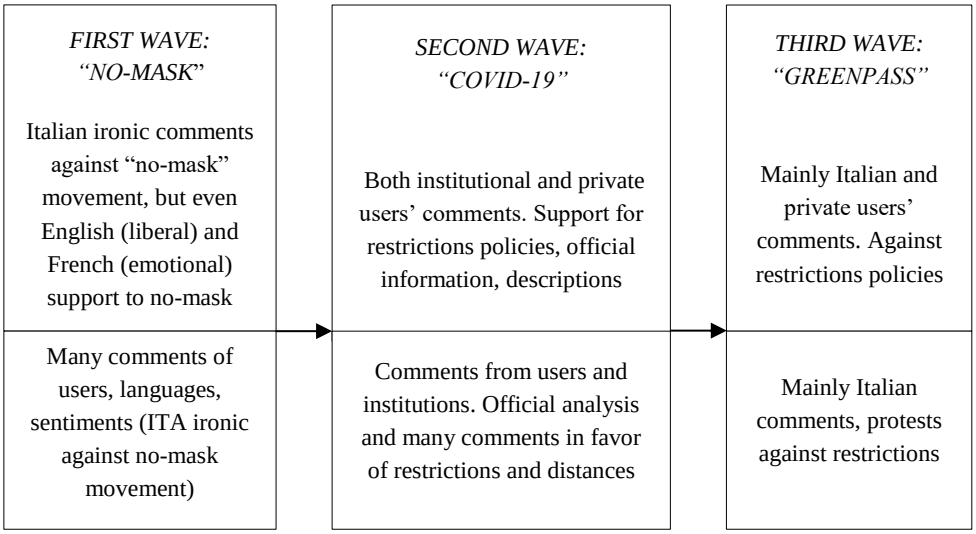


Figure 1. The three phases of communication on Twitter

As "No-mask" comments, also "Greenpass" comments come mainly from single users and are characterized by the positions against restrictions/institutions, showing the presence of a (local, mainly Italian) politically oriented movement. This model of communication, mainly private and self-referential, implies the diffusion of disinformation among users (Bessi and Quattrociocchi, 2015; Del Vicario et al., 2016).

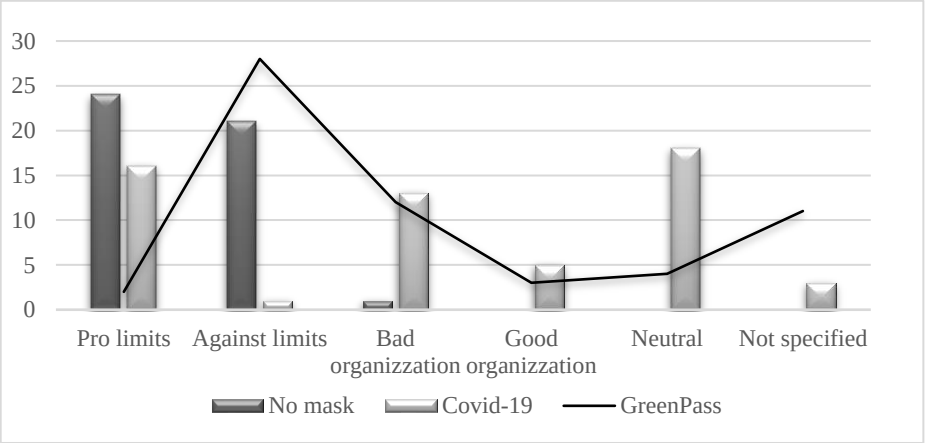


Figure 2. Number of Hashtag for evaluation about politics against the spread of infections

The keywords also affect the structure of communication (Fig. 3). The "No-mask" communication shows the presence of parallel communities with only one or two big main components, while the "Covid-19" communication is subsetted into a huge number of small groups.

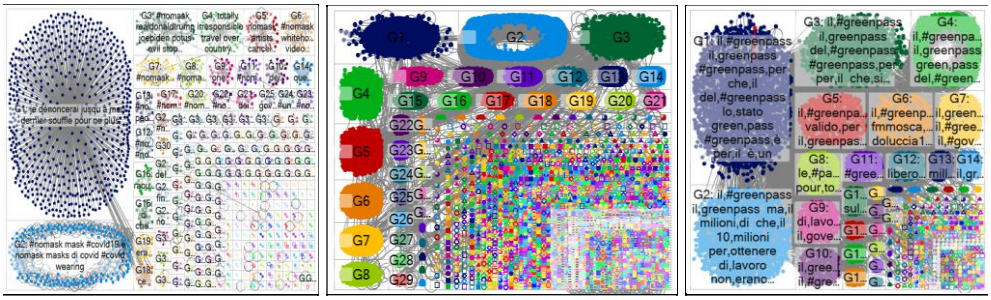


Figure 3. The representative graphs about "No-Mask", "Covid-19" and "Greenpass"

Only in June and in the first part of July there is a larger component that identifies a greater tendency of users to focus on common themes while in August the communication is divided into many groups with few users. Finally, the "Greenpass" communication has a structure that is intermediate between the others, with a fairly high number (7-12) of numerous groups.

4. Longitudinal Trend Description

In an attempt to understand the diachronic dynamics underlying the structure of the groups and their communication, the major network measurements were applied to the *tweet* analysis (Priyanta et al., *2019*). The first of these measurements is *closeness* centrality – calculated as the sum of reciprocals of the smallest distance between each node, in formula:

$$C_c(n_i) = \left[\sum_{j=1}^{g} d(n_i, n_j)\right]^{-1} \quad (1)$$

where $d(n_i, n_j)$ is the distance between the *i*-th node and all the other *g-th* nodes. In other words, the closeness average value returns information about the presence of peculiarly themed networks. The higher the closeness, the higher the presence of compact communities sharing a specific subject. From the analysis run (Fig.4), *closeness* values appear to be constant over the considered period. On the other hand, low levels of closeness suggest communication with no specific focus and composed of small groups, which implies that the various nodes characterizing the network are rarely intertwined.

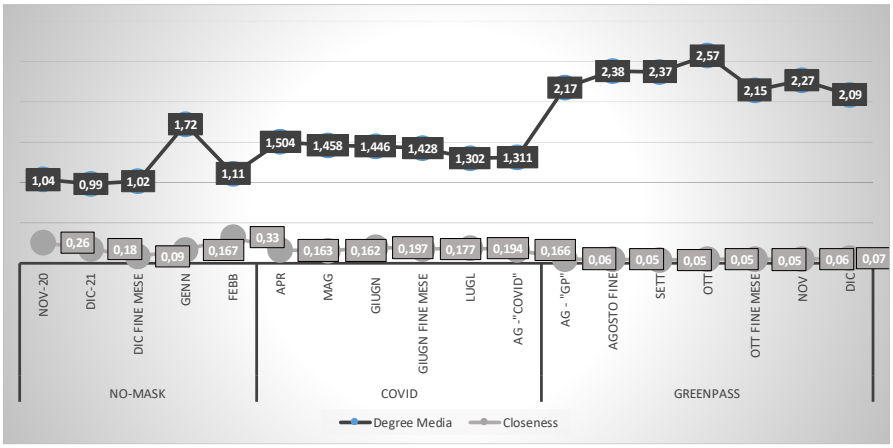


Fig. 4 – Network measurements: Degree e Closeness centrality

While closeness stays constant, it seems important to stress how centrality degree (C'_d) levels significatively increase throughout the study. (Standardized) Centrality degree is obtained considering the number of linkages each node has - $d(n_i)$ on the total amount of possible linkages underneath the network (g-1), in formula:

$$C'_d = \frac{d(n_i)}{g-1} \quad (2)$$

In the specific case, centrality degree refers to the number of relations among nodes, detected by looking at the reactions linked to the tweet, i.e. visualization, retweet, likes, etc.

(Bild *et al*., 2015). As it is possible to notice from figure 1, the degree shows a trend of growth over time. Such an increase seems to be characterized by the specific topic observed and it reaches its acme when dealing with communication on greenpass.

Unlike what has been observed monitoring the hashtag "No Mask" (scattered communication, not centered in any special node, and outcome of private users only), the analysis of the hashtag "Covid 19" has shown a growing communicative dynamicity. If it is true that being it a general topic the interaction among nodes grows, it is also true that in this case, compared with the previous one, public-derived nodes emerge: association, politicians, and healthcare-related users. On the other side, the private nature of the communication related to the "No Mask" theme generates a huger number of reactions that, due to specific characteristics (being them professional, political, or institutional), turns into a major interest even in terms or triggering the research of further information, which makes the communicative flow growing (Miller, 2020). This flow, however, reaches its acme when observing the network structures related to the third focus of our analysis: the hashtag "greenpass". In the last case, the centrality degree, which reaches its peak in October (2,57), highlights the existence of a wider communicative dynamic. In such a case, communication does not only imply reactions – which are still present – but it becomes more direct, creating an actual debate between private and public users who "communicate" with one another.

This peculiar aspect of the last analysed communication seems to be confirmed by the third measurement calculated on the network: betweenness centrality, obtained through the sum of all of the partial betweenness calculated for each couple of nodes, in formula:

$$C_b(n_i) = \sum_{j<k} g_{jk}(n_i)/g_{jk} \quad (3)$$

where $g_{jk}(n_i)$ is the number of geodesics that connect two nodes containing an *i*-th node.

This means that betweenness centrality highlights the presence of users that play an intermediate role between either users or groups of users.

Nodes are not only constituted by reactions in this case but, rather, they involve many occasions of sharing tweets, contributing to its diffusion. This appears more evident looking at the betweenness centrality in the first period, characterized by the hashtag "No Mask". In this case, the communicative structure seems sparser, being it a sort of pseudo-dialogue among people sharing the same thoughts (the maximum value is 1200). The presence of intermediaries emerges instead when looking at the "Covid" centered communications (the maximum value is 26987) and with "greenpass" centered discussions, configuring complex relational structures, though emerging from social networks (the maximum value is 25874). This created occasions and conditions for protest movements in Italian (and beyond) squares to rise.

5. Conclusions and further developments

What's immediately noticeable from the obtained results is how the structure of social communication has turned into a protest movement destined to become widespread. No-mask-centered communication has two peculiar traits: structurally, it's closed and formed by many small and self-referential groups. The communication, spread locally, nationally, and internationally, reaches a peak in December 2020. Then, it gradually reduces its extent and significance. The network structure thus shows the presence of a fluid online movement that gradually fades away around February 2021. Another hashtag progressively gained attention between April and June 2021: Covid-19. In this case, the communicative structure shows less weak than the previous one, with more links and nodes and fewer scattered groups. Moreover, Covid-centered communication appears to be more heterogenous and spread: it's not a movement anymore; rather it shows as the center of many discussions. The last hashtag extracted ('greenpass') is mainly an Italian topic, selected because of the huge number of comments. This last focus seems to lay the *statu nascenti* of the movement against the Greenpass in Italy. Compared with the No-Mask, the green-pass communication is more connected and compact.

The data presented are not always suitable to describe the behaviour of citizens of different nationalities at different stages of the pandemic situation. The work does not aim at this but identifies trends. Certainly, the online comments allow comparisons and identify salient themes also on a trans-national level and this will be the subject of further developments. Here the main focus is, however, the analysis of the network structure underlying the communication (albeit including a qualitative investigation of the comments of the network hubs). It will be necessary, as a further step, to link this analysis of the structure to a more comprehensive and in-depth investigation of the content.

References

Bessi, A. & Quattrociocchi, W. (2015). Disintermediation: Digital Wildfires in the Age of Misinformation. *AQ: Australian Quarterly*, 86(4), 34-40.

Bild, D.R., Liu, Y., Dick, R.P., Mao, Z.M., DS Wallach, D.S. (2015). Aggregate characterization of user behavior in Twitter and analysis of the retweet graph. *ACM Transactions on Internet Technology (TOIT)*, 15(1), 1-24.

Borgatti, S.P. & Halgin, D.S. (2011). On Network Theory. *Organization Science*, 22(5) 1168-1181.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E. & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences of the United States of America*, 113(3), 554-559.

Hansen, D.L., Shneiderman, B. & Smith, M.A. (2010). *Analyzing Social Media Networks With NodeXL: Insights From a Connected World*. Amsterdam, Elsevier Science.

Hansen, D. L., Rotman, D., Bonsignore, E., Milic-Frayling, N., Rodrigues, E. M., Smith M. & Shneiderman, B. (2012). Do You Know the Way to SNA?: A Process Model for Analyzing and Visualizing Social Media Network Data. *International Conference on Social Informatics*, 304-313.

Junlong, Z. & Yu, L. (2017). Degree Centrality, Betweenness Centrality, and Closeness Centrality in Social Network. *Advances in Intelligent Systems Research*, 132, 300-303.

Miller J. 2020. Do COVID-19 Conspiracy Theory Beliefs Form a Monological Belief System?, *Canadian Journal of Political Science*, 53, 2, pp. 319-326.

Priyanta, S. & Prayana Trisna, I.N. (2019). Social Network Analysis of Twitter to Identify Issuer of Topic using PageRank. *International Journal of Advanced Computer Science and Applications*, 10 (1), 107-111.

Smith M., Shneiderman B., Milic-Frayling N., Rodrigues E.M., Barash V., Dunne C., Capone T., Perer A., Gleave E. (2009). Analyzing (Social Media) Networks with NodeXL. in C&T '09: *Proc. Fourth International Conference on Communities and Technologies*, pp. 255-264.

Leveraging mobile network data to understand pandemic-era population human mobility

Aidan Condron
Central Statistics Office, Ireland.

Abstract

This paper discusses how datasets based on mobile phone usage were used to understand population mobility at aggregate levels during 2020-2021, the time of the global COVID-19 pandemic. The paper introduces the data, discussing security and privacy considerations, before presenting a web based dashboard style app, built with open source software, used to interact with and explore the data. A case study of how stadium events impacted on local and regional mobility is used to showcase the data, before concluding with reflections on opportunities and capabilities the data and software present for innovative national statistics.

Big Data and Official Statistics: Challenges and Applications at Statistics Netherlands

Piet J.H. Daas[1,2]
[1]Department of Data Services, Research and Innovation, Statistics Netherlands, the Netherlands, [2]Department of Mathematics and Computer Science, Eindhoven University of Technology, the Netherlands.

Abstract

The use and application of Big Data in official statistics has made considerable progress at Statistics Netherlands. The major contributors are the increased attention for Big Data in the methodological research program, in the creation of experimental statistics and in its use for regular statistics production. To stimulate this the Center for Big Data Statistics was setup in 2016. The most important research challenges identified are:

1. Concept: What (derived) concept is measured in Big Data?
2. Population: What part of the target population is included in Big Data?
3. Methods: What new methods (or new ways of thinking) are needed?
4. Infra: What infrastructural requirements are needed?

The infrastructural (IT) challenge is ignored here. The fact that there is a steady increase in the application of Big Data at the office indicates the need (and progress made) in the study of the research challenges identified. The statistics that make use of Big Data and are either in production or for which an implementation process has started at Statistics Netherlands are:

1. Using scanner data and scraped prices for the Consumer Price Index
2. Using road sensor data for Traffic Intensity statistics
3. Using website texts for Online Platform Economy statistics
4. Using social media for the Social Tension indicator
5. Using texts of online job advertisement for Vacancy statistics
6. Using solar panel output and weather data for Solar Energy production

The presentation will discuss the research challenges and how this has affected the use of Big Data for official statistics at the office.

Keywords: *Key topics; Applications; Challenges; Methodology.*

Quality Guidelines for the Acquisition and Usage of Big Data with additional Insights on Web Data

Alexander Kowarik, Magdalena Six
Statistik Austria, Vienna, Austria.

Abstract

The increasing knowledge and experience within the European Statistical System (ESS) in the acquisition, processing and use of new data sources provides now a clearer picture on quality demands. These guidelines use the quality based experiences in the ESSnet Big Data II to formulate guidelines for NSIs who already use and/or plan to use new data sources for the production of Official Statistics. Looking at the production process of statistics, the usage of new data sources mostly affects quality aspects of processes related to the input and the throughput phase. Taking this into account the guidelines concentrate on the input and the throughput phase of the statistical production process.

With new data sources, the access to as well as the processing of input data makes it necessary to consider new and very source- and data-specific sub-processes. The variety of sub-processes is much broader compared to the use of traditional data sources. What is relevant for one data class and one data access might be of no interest for others.

The Web Intelligence Network (WIN) builds on the work of the previous ESSNets Big Data I and Big Data II and adapts and expands the focus on the more specialized usage of web data.

In this paper we describe the outline of the formulated Quality Guidelines as well as the challenges to structure the wild field of new (sub-)processes and aspects when new data sources are used in the production process of Official Statistics. We give examples of specific quality guidelines and further, we risk an outlook how quality guidelines will develop when the usage of new data sources has become a normal part of the production process of Official Statistics.

Suggested Framework for Big Data Analysis of Enterprise Websites. A Case Study for Web Intelligence Network

Jacek Maślankowski, Dominika Nowak
Statistics Poland, Poland.

Abstract

Big Data gives an opportunity for the researchers and scholars to make surveys in various domains. In this paper we will concentrate on websites as a data source which can be used to provide lots of valuable information for enterprise statistics. In this field, Big Data allows to get various information, including the type of the enterprise (e-commerce etc.), whether the enterprise is present in social media, the frequency of updating the website etc. The main goal of the paper is to present what Big Data methods are the most efficient in acquiring and processing the information from websites. The discussion shows different variants of conducting the work, based on the case studies conducted as experimental statistics at European Union level over the last 6 years.

This paper is based on the experience in processing the data from websites in ESSnet grants on Big Data I (2016-2018), Big Data II (2018-2020) and Web Intelligence Network (2021-2025).

The process of getting enterprise data from websites can be divided into the following steps: (1) Defining the population of enterprise websites; (2) Web scraping; (3) Data processing (extracting); (4) Data validation (de-duplication, quality indicators); (5) Data analysis; (6) Data dissemination.

Each of the steps needs additional validation, especially the first step in this process have an impact on the final results that may not be comparable to the official statistical data. The essential part is also the way the data will be extracted to find the interesting data. In this sense, we need to choose between text mining methods, e.g. machine learning and regular expressions, that gives different results according to the information which should be provided. The paper shows how the use of appropriate methods can increase the overall value of the analysis.

The Web Intelligence Hub – A tool for integrating web data in Official Statistics

Fernando Reis
Eurostat, European Commission, Luxembourg.

Abstract

The use of the World Wide Web as a source of big data has become quite common. However, the use of web data in regular statistical production following the quality standards expected in official statistics has particular requirements and is not easy. The Web Intelligence Hub (WIH) is a European Statistical System (ESS) statistical infrastructure built to address those requirements.

The WIH is being developed in the context of the Trusted Smart Statistics, an ESS initiative launched to address the issues raised by the use of new types of data and sources. The initiative is organised in hubs, with each hub specialising on data sources with similar characteristics and processing similar types of data. The WIH is the pillar of the TSS that provides the fundamental building blocks for harvesting information from the Web to be used in statistics.

The WIH implements the principles adopted for the trusted smart statistics. It follows a modular architecture in order to be easily evolvable to the technological changes occurring on the Web. The processes running in the hub and the corresponding outputs are open, transparent and auditable. The WIH will provide several services to the ESS, such as commonly agreed partnership models with web portals, commonly agreed data gathering, processing and statistical methodologies, algorithms and ready to run scripts. The data and scripts running in the WIH will also be available to the partners as a service.

Exploration and experience with new web data sources. A Case Study for innovative tourism statistics

Galya Stateva[1], Marek Cierpiał-Wolan[2]
Bulgarian National Statistical Institute, Bulgaria, [2] Statistics Poland, Poland.

Abstract

The aim of the first part of presentation is to tap into the potential of new web data sources, which will have the potential to be integrated in the Web Intelligence Hub, developed by Eurostat. Parallel to the exploration of the data sources, we aspire to produce experimental statistics, using these new web data sources, given that they meet the quality criteria.

The presentation will delve deeper into Work Package 3, part of the European Statistical System Collaborative Network (ESSnet) Web Intelligence Network (WIN) project, dedicated to the exploration of non-traditional data sources for official statistical production.

Work package 3's activities are divided into six use cases, each having distinct characteristics and specific goals:

• Use Case 1 aims to explore new data sources and monitor the real estate market.

• Use Case 2 aims to derive early estimates of construction activities, pertaining to both already built and planned buildings, based on real estate web portals.

• Use Case 3 aims to collect data about online prices of household appliances and audio visual, photographic and information processing equipment by web scraping of online shops and at a later stage compare the data with scanner data for the shop's sales.

• Use Case 4 aims to develop new indices for tourism statistics, using the data from booking portals, air traffic portals, travel agencies portals and portals related to quality of life.

• Use Case 5 is concentrated on mass web scraping, primarily for the enhancement of the quality of the business register via linking URLs of enterprises and predicting main economic activity codes (NACE)

• Use Case 6 aims to explore the use of publicly available traffic camera data in order to produce new indicators. In this use case a peculiar data source is used – pictures from traffic cameras and induction loops.

Use cases 1-4 share similar characteristics in terms of data sources and expected experimental indicators and adhere to pre-defined process steps in compliance with Big data life cycle, which include "New data sources exploration", "Programming, production of software", "Data acquisition and recording", "Data processing", "Modelling and interpretation" and "Dissemination of the experimental statistics and results". Use cases 5 and 6 take a slightly different approach due to their extraordinary data sources and do not adhere to the aforementioned process steps.

During the first project's year, the Work package 3 achieved meaningful results, such as a Checklist used as a tool for assessment and justification of web data sources, defined a set of mandatory and optional variables to be extracted from the data sources, sets of minimal indicators, based on the mandatory variables, successfully set up and tested their working environment and software solutions for the upcoming data collection, literature review focused on URL finding methodology and tools and the use of business websites to predict economic activity of enterprises, preparation of training and tests sets and accompanying methodology for URL finding, preparation of the upcoming NACE prediction and classification, exploration of the available assessment of the model results, implementation of Machine-learning pipeline for publicly accessible traffic camera data.

We are also scheduled to begin testing of Eurostat's Web Intelligence Hub for specific use cases from our Work package, which volunteered in the endeavor.

While we have successfully implemented our initial planned activities for the first project year we continue our work, constantly monitoring the available resources, arising issues and quality of the data, which is to be collected and processed during the second project year.

The different use cases have already encountered potential and expected issues like the possible changes in the source of web data structure and web site changes, checks for legal and copyright constraints, non-standard variables, mechanisms blocking extraction of data (e.g. javascript, captchas, etc), viability of training and test sets for both URL finding and NACE prediction, difficulties when comparing results with other partners, since NACE code classification is knowledge-intensive and language-specific

sources have to be used, regular update of the data source. Due to the peculiar data sources for some use cases we have also encountered unsolvable issues like weather variation (e.g. snow,rain, darkness). Some of the issues have been solved, while others still remain.

A Case Study for innovative tourism statistics aims to show the achievements of two projects: ESSnet Big Data II and ESSnet WIN concerning the use of unstructured data sources in the field of tourism.

The work in the Big Data II project started with an inventory of data sources related to tourism statistics, which can be used for research of tourist accommodation establishments as well as for estimating tourist traffic and related expenditures. The VisNet tool was developed to visualise the links between the identified sources.

The gathering of data from digital sources required the preparation of a scalable solution for data retrieval using web scraping techniques. The developed author's method allowed for continuous and non-invasive extraction of data from selected accommodation booking portals.

The process of integrating statistical databases with data derived from web scraping required the development of a fully automated innovative tool, which unified the structure of identification data and assigned them geographical coordinates. The preparation of appropriate structures allowed the implementation of methods of combining data from different sources.

The project also developed a methodology for estimating the volume of tourist traffic and tourist expenditures using spatial-temporal disaggregation methods or the method of flash estimates of accommodation establishments.

As a result of the work carried out, a prototype of the Tourism Integration and Monitoring System (TIMS) was prepared, together with dedicated micro services, which will support statistical production in the area of tourism statistics and assist in monitoring changes in the tourism sector.

The continuation of the work initiated in ESSnet Big Data II is the ESSnet WIN project, in which new methods for assessing the quality of external data sources have been introduced and web scraping has been expanded to other types of portals related to tourism. The main objective of the project is to develop new indicators, which will be an integral part of the developed prototype.

An interpretable machine learning workflow with an application to economic forecasting

Marcus Buckmann, Andreas Joseph
Bank of England, United Kingdom.

Abstract

We propose a generic workflow for the use of machine learning models to inform decision making and to communicate modelling results with stakeholders. It involves three steps: (1) a comparative model evaluation, (2) a feature importance analysis and (3) statistical inference based on Shapley value decompositions. We discuss the different steps of the workflow in detail and demonstrate each by forecasting changes in US unemployment one year ahead using the well-established FRED-MD dataset. We find that universal function approximators from the machine learning literature, including gradient boosting and artificial neural networks, outperform more conventional linear models. This better performance is associated with greater flexibility, allowing the machine learning models to account for time-varying and nonlinear relationships in the data generating process. The Shapley value decomposition identifies economically meaningful nonlinearities learned by the models. Shapley regressions for statistical inference on machine learning models enable us to assess and communicate variable importance akin to conventional econometric approaches. While we also explore high-dimensional models, our findings suggest that the best trade-off between interpretability and performance of the models is achieved when a small set of variables is selected by domain experts.

Keywords: *machine learning, model interpretability, forecasting, unemployment, Shapley values.*

Textual analysis of a Twitter corpus during the COVID-19 pandemics

Valerio Astuti, Marta Crispino, Marco Langiulli, Juri Marcucci
Bank of Italy, Directorate General for Economics, Statistics and Research, Italy.

Abstract

Text data gathered from social media are extremely up-to-date and have a great potential value for economic research. At the same time, they pose some challenges, as they require different statistical methods from the ones used for traditional data. The aim of this paper is to give a critical overview of three of the most common techniques used to extract information from text data: topic modelling, word embedding and sentiment analysis. We apply these methodologies to data collected from Twitter during the COVID-19 pandemic to investigate the influence the pandemic had on the Italian Twitter community and to discover the topics most actively discussed on the platform.

Using these techniques of automated textual analysis, we are able to make inferences about the most important subjects covered over time and build real-time daily indicators of the sentiment expressed on this platform.

Keywords: *Text as data, Twitter, Big data, Sentiment, COVID-19, Topic analysis, Word Embedding.*

Opportunities and risks in the residential sector during a green transition: House prices, energy renovations and rising energy prices

Alessandro T. Martinello, Niels F. Møller
Danmarks Nationalbank, Denmark.

Abstract

Transitioning to a low carbon economy implies both risks and opportunities in the Danish housing sector, which accounts for one fourth of Denmark's CO2 emissions. We study the heterogeneous impacts on house prices of energy prices and energy refurbishments by combining micro-level data on sales and housing characteristics with geolocation data and data from the official mandatory energy rating and housing condition reports.

While energy refurbishments are generally convenient in the long run due to large future flows of savings, households who do not plan to stay in the same housing unit for long do not have an incentive to renovate unless the refurbishment costs are reflected in the sale prices. Yet the extent to which refurbishment costs are reflected on sale prices might vary by housing, market, and refurbishment characteristics. We exploit machine learning tools both to preprocess geolocation data and to identify sources of effect heterogeneity.

We find that most refurbishments will not increase sales prices enough to cover the costs. Those refurbishments whose price effect will cover are typically located in and around smaller towns and mid-sized cities, and other areas with higher population density and well-developed road networks connected to towns and cities. They are also cheap and have lower-than-average impact on CO2 emissions.

Our results imply that private incentives may not be sufficient to facilitate climate change mitigation. We show that if home owners financed the most profitable refurbishments before selling, CO2 emissions of these houses would have decreased by only 13,000 tonnes per year, or less than 0.02 per cent of total Danish greenhouse emissions. Hence, there may be a scope for tax deductions and the allocation of subsidies for energy renovation among private households.

We conclude that while opportunities for profitable energy renovations are concentrated in these areas, transitional risks are instead associated with peripheral rural areas, where both the exposure to rising energy prices and the risk of financing renovations is highest.

Assessing the green transition priorities of SMEs: A large-scale web mining approach

Josep Domenech[1], Maria Rosalia Vicente[2], Hector Martinez-Cabanes[1], Pablo de Pedraza[3]
[1]Department of Economics and Social Sciences, Universitat Politècnica de València, Spain, [2]Applied Economics, Universidad de Oviedo, Spain, [3]Joint Research Center, European Commission, Ispra, Italy.

Abstract
Company websites are a rich source of data that exhibit the activities, intentions, and strategies of the respective companies. Aggregating that information at a sector, company size, and country level has the potential to reveal the underlying behavior of the different units. This paper presents a pilot study in which a sample of more than 32,000 companies from Germany, France, Italy, and Spain has been analyzed to assess their evolution in the sustainability transition over the last 15 years.

Keywords: *Web data mining, sustainability transition, SMEs, web scraping.*

Unsupervised Learning for the Analysis and Detection of Fraud in the Insurance Industry

José A. Álvarez-Jareño, José Manuel Pavía
Applied Economic Department, University of Valencia, Spain.

Abstract

Analysis and detection of fraud in the insurance sector has traditionally been carried out through supervised learning. The main problem is that the data presents a strong imbalance and techniques are used to balance the variable. Unsupervised learning is an alternative to consider, especially anomaly detection methodologies. If the fraud variable has a significant imbalance, then it can be treated as an anomaly. That is, the behavior of the fraudsters must be different from the rest of the insured.

The main methodologies used are Isolation Forest, Attribute wise learning for scoring outliers (ALSO), Trimmed K-means, Autoencoders (neural networks) and Principal Component Analysis. The objective is through dimensionality reduction techniques to obtain a model with which to make predictions. The instances that present greater differences between the real values and the values estimated with this methodology will be considered anomalies and analyzed as if they were fraud.

The results obtained show that these methodologies can be used as a complement to supervised learning. The assembly of models will allow the integration of both methodologies and improve the detection of fraud in the insurance sector.

Keywords: *Insurance; Fraud, Unsupervised Learning, Isolation Forest, ALSO, Autoencoders.*

Policy indicators from private online platforms

José Luis Cervera[1], Yolanda Gómez[1], José Vila[1,2]
[1]DevStat, Spain, [2]Department of Economic Analysis, Universitat de València, Spain

Abstract

The information collected by private online platforms is very relevant for policy design and evaluation. Big data technologies and applications can unlock the potential of these increasing data volumes and analysis requirements for decision-makers in industry and policy and make them usable. However, the use of big data to inform public policy decision-making is still scarce. To contribute to fill this gap, this paper proposes and discusses some relevant examples of policy indicators that could be obtained from selected and reliable online private gamified and non-gamified platforms. The proposed indicators are SMART indicators that are relevant for policy-making, in particular construction soustenability and territorial policies.Proposed indicators can be computed using one of a combination of the following strategies:

• Point-process estimation, to be obtained just by aggregating the value of a variable.

• Distance-based estimation: the value of the indicator is obtained as the aggregation of a pre-defined distance measure: geodesic distance, shortest driving/walking/public transportation distance, etc

• Area estimation. Supervised machine learning algorithms can be used to identify and measure the percentage of an area with a given relevant feature.

• Neighbourhood structure estimation: Graph theory can be applied to the definition of connection indicators of geographical units.

• Gamification of configuration or recommendation private platforms. Information downloaded from gamified private environments cane be used as an alternative to more resource demanding economic experiments in order to define behavioural pocily indicators.

Keywords: *policy indicators; big data; gamification; private online platforms.*

Investigating mechanisms for compensating for an inability to touch products

Lili Zheng[1], Michel Plaisent[2], Prosper Bernard[2]
[1]Excelia Business School, France, [2]University of Quebec in Montreal, Canada.

Abstract

A key challenge that online retail faces is associated with the lack of physical contact (touch) with products, which has an impact on customers' product evaluations and shopping experience and, as a result, influences consumer decision making. It is important to investigate mechanisms for compensating for an inability to touch products. We conducted two experiments to test hypotheses. Two experiments were conducted to test the hypotheses. Study I was designed to measure the role of instrumental NFT (vs. autotelic NFT) in the effects of brand on consumers' purchase intention by manipulating the brand conditions (leading brand vs. non-leading brand). As expected, the results showed that instrumental NFT (rather than autotelic NFT) moderated the relation between brands (leading vs. non-leading) and consumers' intention to purchase. Furthermore, it was found that a leading brand increased consumers' intention to purchase for people who are high in instrumental NFT but not for those who are low in instrumental NFT. Specifically, a leading brand (rather than a non-leading brand) increased intention to purchase for participants who are high in instrumental NFT, whereas leading brand/non-leading brand had no effect on participants who are low in instrumental NFT. The results provide insights for developing effective strategies to address the challenges of online retail channels. To overcome shoppers' inability to touch products online, it appears that building highly recognizable brands is a key to the success of online retail. Strong brands can compensate for the intangibility of e-commerce.

Study II further examined the effects of brand on purchase intention in conditions of high and low SI by varying the shopping tasks in two situations. It was found that when SI was included in the analysis, the results obtained in Study I regarding the effect of low instrumental NFT on the relation between brand and intention to purchase became more complex. Specifically, in the leading brand condition, people who are low in instrumental NFT showed willingness to purchase under low versus high SI with the product. However, people who are high in instrumental NFT were not affected by the involvement condition. This research contributes to the understanding of mechanisms for compensating for an inability to touch products in the online environment under conditions of different levels of SI. Furthermore, the findings represent an important contribution to the understanding of the influence of search- and experience-attribute information on purchase intention.

Report on Amazon´s Project - statistical evaluation on socioeconomic variables across Germany

Sebastian de la Serna
Universität Bayreuth-Geographisches Institut, Germany.

Abstract

In this report we define a proxy that can explain Germany´s precariousness at the district level by relating socio-economic variables to the distribution of parcel centers for the years 2011 and 2019. This precariousness indicator is an aggregated indicator which is composed by 5 socio-economic variables and its consequent normalization processes. These 5 socio-economic variables, which are mainly related to unemployment, form the normalized indicator "precariousness" on a scale from 0 (less) to 8 (most), with equal weighting. The challenge in this project is to webscrape all relevant logistics centres of different competitors in the courier industry and map them on a district level in order to later, when matching the socioeconomic indicators with this dataset, highlight the regions where Amazon operates. Therefore, we could raise the question whether Amazon operates systematically in regions with relative precariousness or not. To answer this question, we address the chosen socio-economic variables by running descriptive statistics by looking at how the standard deviations perform. Consequently, we examine the collinearity of the 5 variables by means of a correlation matrix and PCA. Finally, we compare the two resulting maps for 2011 and 2019 and assess their precariousness.
Keywords: *webscraping, normalization, georeferenced, statistics.*

Research on the Construction of Agroecological Park Under the Background of Smart Agriculture – Take Zengcheng Chuangxian Smart Agriculture Model Park as an Example

Chen Jian[1], Cai Yingying [2], Gaoge Yunhan, Lv Haiyan[3]
[1]Belt and Road School, Beijing Normal University, International Business Faculty, Beijing Normal University, Zhuhai, China, [2]Tianjin Foreign Studies University, China. [3]Guangzhou City Construction College, China.

Abstract
In the process of agricultural modernization in China, smart. agriculture is the new direction of rural development, and scientific development of agroecological parks is an important starting point to promote agricultural transformation and upgrading, as well as an effective way to promote rural revitalization. At present, the application of smart agriculture in agricultural ecological parks in China is still in the experimental stage, and the pilots have been carried out in various regions, but most of them are almost stagnant at the level of agricultural production, failing to maximize the use of smart agricultural resources to build agroecological parks. This paper analyzes the case of Zengcheng Chuangxian Smart Agriculture Model Park on the basis of describing its current development and the existing problems after the introduction of smart agriculture, and puts forward relevant suggestions.

Keywords: *smart agriculture; agroecological park; agricultural modernization.*

Analyzing the Natural Language Processing technology field using Tech mining

Gaizka Garechana, Rosa Río-Belver, Izaskun Álvarez-Meaza, Enara Zarrabeitia
Business Management Department, University of the Basque Country (UPV/EHU), Spain.

Abstract

The Natural Language Processing (NLP) field is the branch of computational science devoted to the automated interpretation of human language, having several technical applications in areas such as speech recognition and information retrieval/summarization, among others. In this paper we analyze NLP patent data corresponding to the yearly interval 2006-2020 in order to characterize the main agents, purposes and analytical tools behind this field. With this goal in mind, we use text mining software to extract the relevant information from patent abstracts and identify the specialization of the main players in the area. In addition to this, we detect the dominance of artificial intelligence applications of NLP and the versatility and acceptation of deep learning algorithms in this field. These concepts are at the same time the dominant ones and show the highest growth rate, being present in roughly 15% of the patents forming our dataset. Two clear conclusions are extracted when analyzing the conceptual maps and cluster analysis of the data: voice/speech recognition systems and the automated medical diagnosis systems are well-consolidated specialties in NLP patenting activity.

Keywords: *Tech mining; NLP; Patent.*

Bibliometric analysis of Experimental Economics and progress of the research field

Myriam González-Limón[1], Asunción Rodríguez-Ramos[2], Cristina Maldonado[3]
[1]Department of Economic Analysis and Political Economy, University of Seville, Spain, [2]Department of Economics and Economic History, University of Seville, Spain, [3]University of Seville, Spain.

Abstract

Bibliometric studies are mainly based on the quantitative analysis of publications belonging to a specific research field. The main objective of this paper is to analyse and observe the progress and future lines of research in the field of Experimental Economics (EE) from a bibliometric perspective. To do so, the scientific production in this field of research is identified from the Web of Science Core Collection (Clarivate). As it is one of the most complete in existence, guarantees the reliability of the bibliometric analysis. The period of analysis covers from 1990 to the end of 2021. The bibliometric analysis was carried out on the basis of the information obtained on the journals in which these articles were published, the authors and their institutional affiliation, the years of publication and the number of citations received. The analysis was carried out with the support of the VOSviewer software, version 1.6.18, which allowed us to identify groups of countries, researchers and thematic areas. There is a clear upward trend in scientific production in this field of study. The theme most prolific is behavior. The most popular keywords in recent years are: behavior, preferences, cooperation and fairness.

Keywords: *Experimental Economics; bibliometric analysis; research trends; scientific production; WoS databases; VOSviewer*

Evaluating E-Learning systems success to understand student's performance during Covid Pandemic

Eliseo Bustamante[1], Mónica Martínez-Gómez[1], César Berna-Escriche[1, 2]
[1] Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, València, Spain. [2] Instituto de Ingeniería Energética, Universitat Politècnica de València, València, Spain.

Abstract
The present work arises as a consequence of the current situation that is being experienced worldwide due to the COVID 19 and the associated repercussions on education. E- Learning has demonstrated to be the only resource capable of replacing traditional in-person learning methods in the present global gridlock due to the COVID-19 pandemic. Academic institutions around the world have heavily invested in E-Learning and it is necessary to ensure the success of E-Learning initiatives in order to make a learning model with the same guarantees of traditional models. The objective of this study is to propose and validate a model to measure the E-learning success based on different dimensions and it is an extended of the Technology Acceptance Model (TAM) and Information Systems Success (ISS), which has been empirically tested. For this purpose, an e-learning questionnaire was used in order to test academical performance with students from secondary education in València during the pandemic. Results show a significant relationship between Intention To Use for Sustainability and Students Satisfacction, a significant relationship between Students Satisfacction and Student Performance and finally a significant relationship between Student Performance and Learning Achievements.

Keywords*: E-learning, Academic Performance, Students Satisfaction (SS), Structural Equation Modelling (SEM), Partial Least Squares (PLS).*

A comparative study of Bitcoin's Price fluctuations by Twitter sentiments

Saida Bruce
University of Malaga, Spain.

Abstract

Cryptocurrencies are digital currencies that utilize blockchain technology, a radical, decentralized, and cryptographic technology that allows for the digitalization of trust. In the case of cryptocurrencies, blockchain technology theoretically eliminates the function of governments as currency providers and the role of intermediary (third-party) parties in transaction verification. During the last few years, the scrutiny of bitcoin and other cryptocurrencies as legally regulated components of financial systems has been increasing insignificantly. Bitcoin is one of the biggest cryptocurrency in terms of capital market share and trading volumes. This study is going to determine whether the sentiments in Twitter about Bitcoin have an influence on overall market and pricing of the Bitcoin. The volatility of Bitcoin's price can be related to the Twitter sentiments and that's what this study is going to reveal and how it's correlated with each other. The tweets of Bitcoins have been taken by using Twitter's API from June 2021 till September 2021. For this study there are different predictive and descriptive models have been applied that are important for data analysis. The sentiments are being categorized into three parts, positive, negative and neutral. These are done by using data scrapper using python scripts. VADER sentiment analysis will be used to analyze the tweets and it provides several benefits including the fact that it doesn't only classifies text as positive, negative or neutral but also measures the polarity and intensity of the words used. Another benefit of using VADER sentiment analysis is that it doesn't require the data to be simplified and remove punctuations or emojis'. It will process the data and check intensity and emotions of the text with punctuations. As a result, this study gives an understanding of Bitcoin's price fluctuation related to twitter sentiments but it has limitations as this data has been processed manually and there will be a chance that it does has a deeper correlation other than only trading price. By utilizing this study it gives an idea that users can make better informed purchase and selling decisions based on twitters current sentiments. The results will significantly prove that twitter sentiments do impact bitcoin's price and trading volumes.

Keywords: *Blockchain, Bitcoin, Twitter, VADER, Sentiment Analysis, Social Media, Cryptocurrencies.*

Analysis of factors involved in the teaching-learning system in a state of emergency

Alba Lira Pérez Avellaneda, Diana Cueva, Mónica Martínez-Gómez
Universitat Politècnica de València, Spain.

Abstract

The present work arises as a consequence of the current situation that is being experienced worldwide due to COVID 19 and the repercussions it has had on households, especially in the adaptation to new forms of work and teaching, affecting social, economic and cultural aspects. The pandemic has hit the poor and vulnerable hardest, and could push millions more people into poverty, especially in Latin American countries. In this context, in Ecuador, due to the COVID-19 pandemic, extreme poverty is likely to increase dramatically.

The study focuses on the application of different multivariate techniques, such as, Discriminant Analysis, Cluster Analysis and ROC curve,to develop a multidimensional poverty index based on non-monetary household resources such as equipment and basic services.

With the combined use of the proposed techniques, an indicator was obtained that can serve as a tool for measuring the factors that can affect extreme poverty in Ecuador and it can be used to introduce continuous improvement actions in the resource assigments.

Keywords: *Poverty Index. Digital Divide, Multivarte Techniques, Latin America*

Topic modeling in court rulings

Juan Diego Cuenca Camacho
PhD candidate, Department of Economics, Universitat Politècnica de València, Spain.

Abstract

Judges usually have to deal with the valuation of a company's shares in the event of bankruptcy, merger and acquisitions and other disputes (DiGabriele, 2006). This paper aims to take a first step in answering the question of what is the treatment of company valuation in Spanish jurisprudence. We will construct a classifier of court rulings that allows us to discriminate rulings that effectively deal with company valuation from those that do not. Three unsupervised models are proposed: Latent Dirichlet Allocation (LDA), Latent Semantic Indexing (LSI) and Nonnegative Matrix Factorization (NMF). LDA identifies two topics that have a interpretation aligned with the classification that a human being would give. Topics from LSI are not easy to interpret,whereas results from NMF are closer to the LDA ones. To estimate the goodness-of-fit of the models and compare them we use the coherence measure. According to it, LDA gets the highest score. However, for all models the number of topics that maximize coherence are greater than two, what highlights that, despite the usefulness of objective measures of topic modeling evaluation, human judgment may be more appropriate for topic modeling and model comparison.

Keywords: *Company valuation; Court ruling; Topic modeling.*

Automated real estate valuation disruption in the Smart Cities context

Andrea San José Cabrero[1], Gema María Ramírez Pacheco[2]
[1]Universidad Politécnica de Madrid, Spain, [2]Department of Construction and Architectural Technology, Universitad Politécnica de Madrid, Spain

Abstract

The technological revolution has given rise to a new scenario for action, the Smart Cities. This growing demand for personalization, differentiation, authenticity, sustainability and responsibility, maximization of ease of use, on-demand access and experience are determining factors in all fields and, specifically in the field of real estate valuation. Its influence in this area is transferred to the Automated Valuation Models (AVM) as tools that allow us to deal with the volume, complexity and speed of growth of massive data. Traditional models and methodologies become obsolete in the face of the intangible relationships between the new attributes of each property: its intrinsic characteristics and their urban, political, economic and social context. Among the multiple consequences that are looming in property management, three of them are worth highlighting: the valuation consultant, predictive analysis and real-time valuation. This paper aims to provice an updated framework for the future of property valuation, defining the possible consequences of managing real estate in a new sustainable way.

Keywords: *Property valuation; Real estate management; Big Data; Automated Valuation Model; Artificial Intelligence; Sustainability.*