3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11560

Model degradation in web derived text-based models

Piet J.H. Daas[1,2], Jelmer Jansen[3]
[1]Division of Corporate services, IT and Methodology, Statistics Netherlands, the Netherlands, [2]Department of Mathematics and Computer Science, Eindhoven University of Technology, The Netherlands, [3]Faculty of Social Science, Radboud University Nijmegen, the Netherlands.

Abstract

Getting an overview of the innovative companies in a country is a challenging task. Traditionally, this is done by sending a questionnaire to a sample of large companies. For this an alternative approach has been developed: determining if a company is innovative by studying the text on the main page of its website. The text-based model created is able to reproduce the results from the survey and is also able to detect small innovative companies, such as startups. However, model stability was found to be a serious problem. It suffered from model degradation which resulted in a gradual decrease in the detection of innovative companies. The accuracy of the model dropped from 93% to 63% over a period of one year. In this paper this phenomenon is described and the data underlying it is studied in great detail. It was found that the combination of the inactivity of a subset of websites and changes in the composition of the words on company websites over time produced this effect. A solution for dealing with this phenomenon is presented and future research is discussed.

Keywords: *Innovation; Text analysis; Webscraping; Big data.*

1. Introduction

In our modern world, more and more data are generated on the web and produced by sensors in the ever growing number of electronic devices surrounding us. These data have very interesting potential applications such as provide novel insights on the activities of companies (Gökk *et al.*, 2015), to inform policymakers (Höchtl *et al.,* 2015) and also for official statistics (Florescu *et al.,* 2014), especially when performed at large scale. However, extracting relevant and reliable information from Big Data sources in a reproducible way is not an easy task (Kitchin, 2015, Daas *et al.*, 2015).

In this paper we describe the results of our research on the development and application of text-based models for official statistics. The focus is on the development of a model that is able to detect innovative companies based on the text displayed on their website (Daas & van der Doef, 2020). Here, the occurrence of particular words and their co-occurrences with other words are used to determine if a company is innovative or not. The findings of the Community Innovation Survey (CIS) in the Netherlands are used for model development. The initial logistic regression model created had an accuracy of 93% and an F1-score of 93% on the test set (Daas & van der Doef, 2020). Kinne and Lenz (2019) demonstrated that the text on the websites of German companies could also be used to detect innovative companies. Their Deep Learning based model had an F1-score of 80%.

1.1. Goal of this study

During the work on Dutch company websites it was discovered that the performance of the initial model gradually decreased over time (more on that below). The model could no longer properly detect positive examples of innovation when exposed to more recently crawled data. The work described in this paper focusses on this issue, which very much resembles a phenomenon known as ‘concept drift’ in the Machine Learning world (Lu *et al.,* 2018). In this paper, we particularly focus on the underlying causes of the fact that the model produced less and less accurate results, to better understand what is exactly happening. Ultimate goal of the work is to assure the production of high quality output derived from new data sources collected at different points in time.

2. Concept drift

Certainly for the analysis of texts, models need to be used to measure the phenomenon the researcher is interested in. This indicates that the concept cannot be directly measured and thus one has to rely on the statistical properties of features in the text (Jo, 2019). When these properties change over time, the model-derived results are not stable (Zhang *et al.*, 2017). This phenomenon is known as concept drift (Lu *et al.,* 2018) which has attracted much attention in the Machine Learning community, especially by those that study socalled data streams. Concept drift describes unforeseeable changes in the underlying distribution of streaming data over time. In general, four types of drift are discerned which are: i) Sudden drift, ii) Gradual drift, iii) Incremental drift and iv) Reoccurring concepts (Lu *et al.*, 2018, Fig 4). In the context of text-based models, reduction of the long-term stability has, for instance, been observed in news topic classification (Kim & Hovy, 2006) and event detection and tracing (Atefeh & Khreich, 2015).

2.1. Detecting innovation

The initial innovation detection model was developed on a set of 2,529 innovative and 2,236 non-innovative websites that all contained 10 or more words after processing (Daas & van der Doef, 2020). The processing steps were implemented in Python and were as follows. First, all script and style sections were removed from the scraped and parsed web pages, followed by language detection of the visible text. Since the majority of the pages were either written in Dutch or English, only those languages were discerned; i.e. any non-Dutch page was considered written in English. Subsequently, all words were converted to lower case and all punctuation marks, numbers, and all words with less than 3 characters were removed. Next, depending on the language detected, any words included in the stop words list for that language were removed. This was followed by stemming with the SnowballStemmer library. The words and language were subsequently used as features in model development using the well-known representation of frequency-annotated bag-of-words (Aggarwal, 2016). As already indicated above, ten words per website was the minimum amount of words considered for classification. For more details the reader is referred to Daas and van der Doef (2020).

The initial logistic regression model contained 180 words had an accuracy of 93% on the test set. But on freshly scraped data -from the same list of websites- 6 months after the original model was developed, the accuracy of the model was found to be 76% and after one year it was as low as 63% (Figure 1). As is clear from Figure 1, the accuracy slowly decreased over time, suggesting a gradual degradation of the concept measured. The big question here is ‘What exactly caused this decline?’. For this we will compare the first (t = 0) and last (t = 12) data sets in more detail.

2.2. Data sets compared

The study started with an initial set of 3,338 innovative and 2,876 non-innovative companies for which a website was found. As is clear from the numbers of websites finally used to create the model (see above), not all websites could be scraped and not all resulted in 10 or more words after processing. When comparing the companies for which a website was scraped and used in model building, it was found that these numbers were not identical

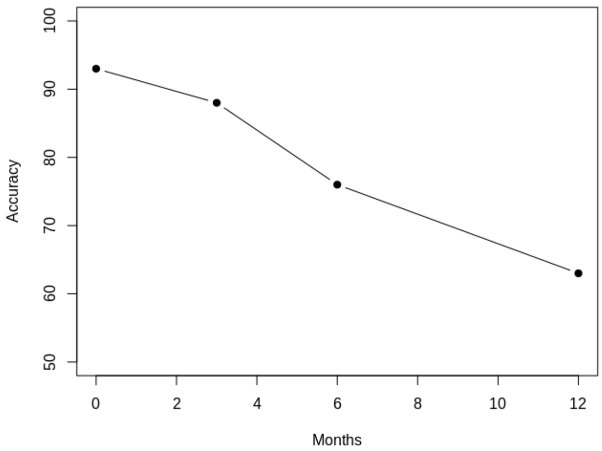

Figure 1. Accuracy of the innovation detection model over time on websites of companies included in the CIS survey scraped at different points in time. The model was developed at month 0. The average accuracy of the model, after 10-fold cross-validation, on the scraped webpages is shown.

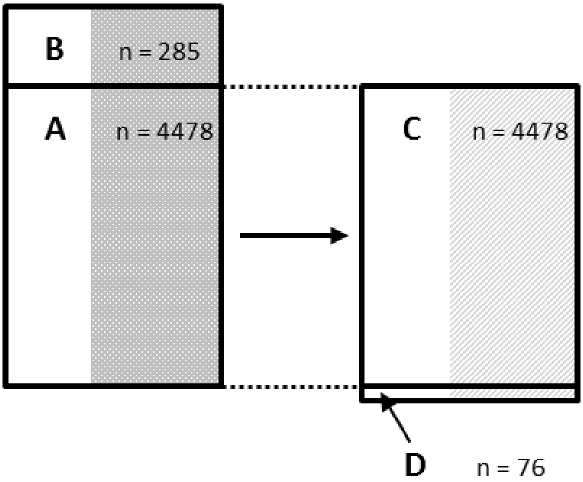

Figure 2. Graphical representation of the difference in the composition of the first (t = 0) and last (t = 12) data sets scraped. The data sets have 4,478 company web pages in common. For these the text extracted may differ which is illustrated by the different shades of gray.

in both data sets. This is illustrated in Figure 2. From this figure it is clear that the websites of 4,478 companies are included in both data sets. A total of 285 companies only occur in the first data set and, to our surprise, 76 only occur in the last data set. Comparison of the texts obtained after processing of the 4,478 websites included in both data sets revealed that these also differed. The similarity of the texts, expressed as the number of words in common in both versions of each website divided by the total number of unique words in

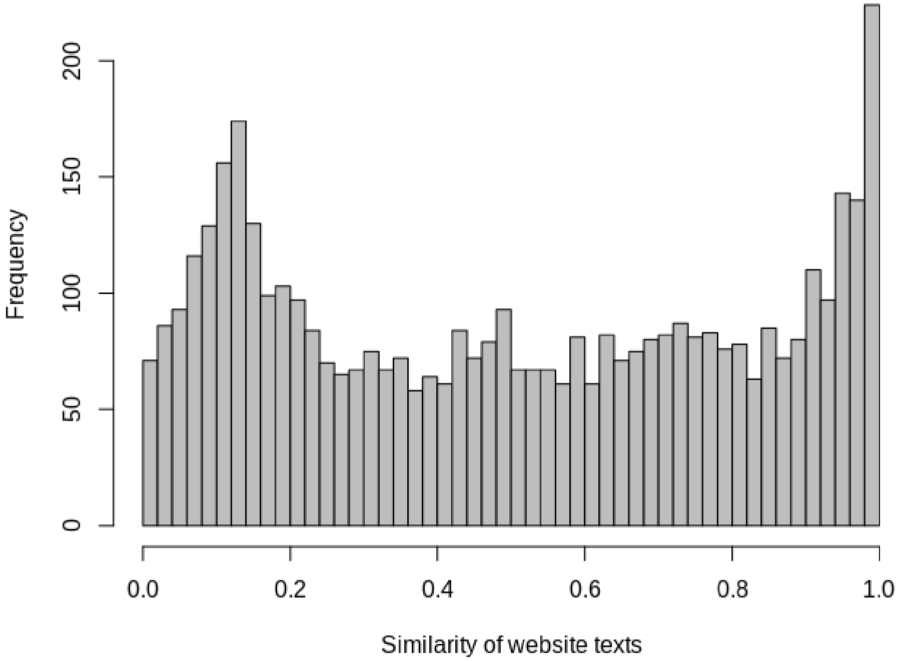

Figure 3. Distribution of the similarity measures of the words extracted from the 4,478 company websites included in the first and last data set. A 1 indicates that all the words on both versions of the website are identical; a 0 indicates no words are in common. The distribution indicates considerable textual changes for many websites.

the combination of both, revealed an average value of 0.50. The distribution of those values indicates a whole range of changes (Figure 3). From this it is obvious that the content of many websites changed considerably over time. To better understand the effect of these changes, the accuracy of the models built on various subsets of the first and last data set and their combinations were determined. The subsets are identified as indicated in Figure 2. By comparing these results, we wanted to know which part(s) of the data contributed to the high accuracy of the first data set. The findings are summarized in Table 1. In this table only the findings of informative combinations are shown.

The results in Table 1 reveal the importance of the data of the units in subset B and the text in subset A for the creation of a highly accurate innovation detection model. In subset B, the website texts of 156 innovative and 129 non-innovative companies are included. Using the data solely in B already results in a model with an accuracy of 76%. Combined with the data in subset A, it is clear that this produces a highly accurate model; an accuracy of 91% is obtained. Subset C on its own has an accuracy of 60%. Adding B to subset C has no additional effect, indicating that the data in C must be of low quality. This becomes obvious when the text in C is replaced by the text of the same units in A (this essentially recreates subset A). Because of this, the accuracy increases from 60 to 85%. Adding B to the latter increases the accuracy even more; to 89%. Clearly the texts of the first websites are much more informative than the texts in the last dataset. This indicates that -over time- vital information on the innovative, and probably also the non-innovative, character of a company is lost. The effect of B indicates that, in addition, the websites of a set of highly informative units is lost as these websites were no longer active online. The latter suggest that these companies are no longer economically active.

Table 1. Accuracy of the model developed on various combinations of subsets of the first and last scraped data sets to detect technological innovation.

Data set composition	Accuracy (%)	Description
A+B	91	First data set
A	86	Common units, text of first data set
B	76	Unique units in first data set
C+D	60	Last data set
C	60	Common units, text of last data set
D*	-	Unique units in last data set
C+B	60	Common units, text of last data set + unique units in first
C(A)**	85	Common units, text of last data set replaced by text of first
C(A)+B	89	Common units, text of last data set replaced by text of first + unique units in first

*Subset D contained not enough units for model development and did not influence the results of other subsets combinations. **C(A) indicates that the text in subset C is replaced with the text included in A for identical units.

3. Discussion and future work

3.1. Model degradation

From the results described in this paper its obvious that the degradation of the accuracy of the innovation detection model is the result of the combined effect of i) the loss of a part of the company websites included in the CIS survey and ii) changes in the text on the websites that remained active. Both findings are indicative for a dynamically changing population, which is not surprising for the kind of topic studied. It also reveals that innovation is a challenging topic to capture. Because of the high accuracy of the model based on the first data set, the websites of the companies included in the CIS survey must be scraped at a very appropriate point in time; a time in which the innovative and non-innovative character of the company was clearly reflected in the text on their websites. Future research will focus on comparing the period covered by both survey and web data and will include extracting and studying websites from a web archive. The latter may enable us to determine if and when the optimal point in time occurs between the ending of the CIS survey period and the moment websites should be scraped. What is also apparent from the results is the fact that the findings reveal that it is not the concept that is drifting (changing) but that it is gradually fading away (degrading). Hence, the title of the paper. One of the explanations for this could be a gradual change in the innovative character of the companies studied. Because of this, one even starts to wonder how long the innovation classification data provided by the CIS survey actually holds over time. These are interesting questions that emerge because of the high frequency at which the data can be collected and analyzed.

3.2. Model resurrection

A solution was also developed to deal with model degradation. We were inspired by the literature on dealing with concept drift (Lu *et al.*, 2017) here. The solution most often applied when dealing with this phenomenon is simply retraining the model on new data. In our case this would be the texts derived from ‘freshly’ scraped websites. From the findings described in this paper its obvious that doing this would never solve the model stability problem; the second data set simply does not contain the information needed for the development of a highly accurate model (see Table 1). We therefore followed the second most often mentioned suggestion in the Machine Learning literature, which is to add newly classified data to the original data set and retrain the model (Janardan, 2017, Gama *et al.*, 2014). We found that this worked well when the classified websites of innovative startups, a total of 855, and those of a large number of websites, i.e. 20,000, for companies in the Business Register were jointly added to the training and test set (Daas & van der Doef, 2020). This increased the training and testset from 4,763 to 25,618 records. Comparing the effect of adding varying amounts of classified Business Register data (samples from 1,000 to 40,000 were tested) on the accuracy of the model, revealed that adding more then 20,000 records did not further improve the model and development took much longer to complete. An accuracy of 88% on the test set was obtained for the new logistic regression model. In our opinion this approach worked well since adding more examples allowed the classifier to find more words which were either positively or negatively related to innovation; i.e. more synonyms were included. This is apparent from the number of words included in the new model; a total of 584. A detailed study of both models revealed that nearly all the words in the old model, 177 to be exact, are included in the new model as well. Future studies will focus on the new model’s composition including its stability over time.

References

Aggarwal, C.C. (2016). Mining Text Data. In Aggarwal, C.C. (Ed.) *Data Mining: the Textbook* (pp. 429-455). New York: Springer.

Atefeh, F. & Khreich, W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence,* 31(1), 132-164. doi.org/10.1111/coin.12017.

Daas, P.J.H., Puts, M.J.H. Buelens, B., & van den Hurk, P.A.M. (2015). Big Data and Official Statistics. *Journal of Official Statistics,* 31(2), 249-262. doi.org/10.1515/jos-2015-0016.

Daas, P.J.H. & van der Doef, S. (2020). Detecting Innovative Companies via their Website. *Journal of the IAOS*, accepted for publication.

Florescu, D., Karlberg, M., Reis, F., Rey Del Castillo, P., Skaliotis, M. & Wirthmann, A. (2014). Will ‘big data’ transform official statistics? Paper for the Quality in Official Statistics Conference, June 2-5, 2014. Vienna. Retrieved from http://www.q2014.at/fileadmin/user_upload/ESTAT-Q2014-BigDataOS-v1a.pdf. (Accessed June 2019).

Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M. & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys,* 46(4), 1-37. doi.org/10.1145/2523813.

Gökk, A., Waterworth, A. & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics* 102(1),653-671. doi.org/10.1007/s11192-014-1434-0.

Höchtl, J., Parycek, P. & Schöllhammer, R. (2015). Big Data in the Policy Cycle: Policy Decision Making in the Digital Era. *Journal of Organizational Computing and Electronic Commerce,* 26(1-2), 147-169. doi.org/10.1080/10919392.2015.1125187.

Janardan, S.M. (2017). Concept drift in Streaming Data Classification: Algorithms, Platforms and Issues. *Procedia Computer Science,* 122, 804-811. doi.org/10.1016/j.procs.2017.11.440.

Jo, T. (2019). *Text mining Concepts, Implementation, and Big Data Challenge*. New York: Springer.

Kim, S-M. & Hovy, E. (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. Proceedings of the Workshop on Sentiment and Subjectivity in Text, pp. 1-8. Association for Computational Linguistics. Retrieved from https://www.isi.edu/natural-language/people/hovy/papers/06ACL-WS-opin-topic-holder.pdf.

Kinne, J. & Lenz, D. (2019). Predicting Innovative Firms using Web Mining and Deep Learning. ZEW Discussion paper no 19-001, Mannheim, Germany. doi.org/10.13140/RG.2.2.22526.84809.

Kitchin, R. (2015). The opportunities, challenges and risks of big data for official statistics. *Statistical Journal of the IAOS,* 31(3), 471- 481. doi.org/10.3233/SJI-150906.

Lu, J., Liu, A., Dong, F., Gu, F. Gama, J. & Zhang, G. (2018). Learning under Concept Drift: A Review. *IEEE Transactions on Knowledge and Data Engineering,* 31(12), 2346-2363. doi.org/10.1109/TKDE.2018.2876857.

Zhang, Y., Chu, G., Li, P., Hu, X. & Wu, X. (2017). Three-layer Concept Drifting Detection in Text Data Streams. *Neurocomputing,* 260, 393-403. doi.org/10.1016/j.neucom.2017.04.047.