2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8348

Grassroots Market Research on Grass: Predicting Cannabis Brand Performance Using Social Media Scraping

Kregor, Jennifer [a]; Gomez, Bethany [a]; Kelly, J. Steven [b]; Stevenson, Kathleen [b]
[a] Brightfield Group, USA. [b] Department of Marketing, DePaul University, USA

Abstract

Social media listening has become a useful tool to marketers in studying behavior for a wide variety of consumer applications, from political leanings and drug abuse to common product choices. Although most cannabis products are illegal at the U.S. Federal level, it is legal in 30 states for medical use and 8 states and the District of Columbia for recreational use. Despite the legal issues, cannabis is projected to reach over $31 billion in sales world-wide by 2021. The industry is both rapidly evolving and highly fragmented, making it challenging for companies operating in the space to access the insights and the data to help design communications, product development and branding strategies. The research presented here will show that the application of social media listening can be helpful for cannabis brand marketers to gauge size, scope and nuances of these markets and tailored social media mining can accurately predict a brand's future performance. Later research will show that social media scraping will help identify and segment consumers at a fraction the cost of traditional consumer research methods.

Keywords: *Social media listening, brand share, predictive analytics, cannabis industry,*

1. Introduction

Brightfield Group is a market research firm focused on the legal cannabis industry. The company holds a robust ecosystem of data on all aspects of the cannabis industry, including market sizes, brand shares, pricing and analytical reports and customized consumer research on a custom basis. Syndicated data is constructed using a multi-source methodology, including analysis of publicly available sources, expert interviews, data reported by brands and dispensaries and big data scraped from relevant industry sites.

Global marijuana sales are estimated to reach $31.4 billion by 2021 (Zhang, 2017). U.S. Cannabis consumers can acquire it in many formulations, such as edibles, concentrates, tinctures, vapes as well as the standard flower. Product subcategories include infused chocolate, savory snacks, baked goods, drinks, sugar candy, crumble, shatter, vape cartridges, resin and wax. Products have a wide variety of differentiating attributes, based on strain, dosage, cannabinoid profile (levels of THC or CBD) and quality of ingredients. Marketers in the cannabis industry are confronted with decisions for product development, packaging, and branding along with a plethora of environmental issues from federal and state regulations regarding lack of trademark protection (Schuster, 2016), banking, growing, distribution and marketing communications. With more than 1600 brands of infused products on the market in 2017 (Brightfield Group, 2018), savvy marketing strategies are crucial to a brand's success. With limited access to capital, steep competition and consumer preferences that are constantly in flux, cannabis brands need to access highly cost-effective and agile consumer research methods to drive product development, marketing and advertising strategies.

Making matters more complicated for cannabis companies, strict advertising regulations are in place limiting where and how brands can promote themselves and vary state-by-state. Since many traditional forms of advertising rely on businesses that are licensed at the federal-level (like broadcasters), few businesses have agreed to advertise cannabis-related content (IAB, 2018). This has driven cannabis businesses to focus instead on more grassroots promotions of their brands (Gunelius, 2018). Some of these methods include brand ambassadors, demo days and event sponsorships, but cause the cost of customer acquisition to increase. Social media is increasingly the preferred tool for promoting brands, leaving a tangible digital footprint to be analyzed to gain insights into brand behavior and consumer perceptions (McVey, 2017).

Reporting by the National Survey on Drug Use and Health (NSDUH) showed, through survey self-report, that in 2014, 13% or 35 million Americans over 12 years old had used marijuana in the past year (Azofeifa et al, 2018). But, for business application these reports are limited to demographics of users of cannabis and other drugs. The cannabis

products only include blunts, joints and hash and the focus is on drug abuse, not consumer purchasing patterns.

Studying social media usage has given researchers opportunity to explore the relationship of posts to other behavior. McGregor, et al. (2014) employed the monitoring of several general social media platforms (Facebook, Twitter) as well as blog sites. Their goal was to identify themes of conversations by the community of glaucoma patients. In fact, 14 different themes were identified. This kind of research demonstrates that users are openly willing to offer the language they use to discuss their issues and product usage. But there was no clear relationship to specific product usage. A study by Schwartz et al. (2013) demonstrated the correlation between language and personality. A more recent study by Antoniou (2017) demonstrated that social network posts can be related to users' cognitive profiles as measured by the Meyers Briggs, MBTI profile.

Research focusing on product usage, Culotta and Cutler (2016) established that they could monitor Twitter posts to study consumer perceptions of 200 brands along three perceptual attributes. Their social media monitoring showed high correlation with more expensive survey techniques. As for using social media to predict behavior, Lievens and Van Iddekinge (2016) used social media scraping in the staffing area where employer keywords or signals were compared to social media conversations to predict who might be good potential employees.

Other important research also has demonstrated the predictive ability of social media monitoring analytics. St Louis and Zorlu (2012) demonstrated the relationship between Twitter posts and the spread of flu. Sul et al. (2016) found that monitoring emotional sentiment about a company from Twitter conversations demonstrated impact on same-day and longer-term stock prices for those same companies. Yaden et al. (2017) demonstrated correlation between word usage on social media posts and the religion of the media discussant. Social media scraping has been used to study drug abuse such as when Sarker et al. (2016) showed a clear relationship between Twitter posts and drug abuse.

Chen et al. (2015) conducted smoking research in their analysis of users of some blogs about vape devices and Reddit posts to discover experience of users of electronic cigarettes. However, this work was about tobacco use and did not explore the specifics of the product brands. Research in the cannabis field was carried out by Nguyen et al. (2016). They studied Twitter posts as related to marijuana usage. The customer profiling was minimal, correlating to type of phone used, times of day, etc. The research did not focus on or indicate user perceptions of cannabis product types or brands.

There seems to be no academic research done to relate cannabis users to the brands available in the various marketplaces. What will be presented here is a study of how social

media scraping and the results of the analysis therein relates to cannabis product usage by brands and brandshare.

2. Methodology

A total of 3,050,725 words and phrases from 38,014 twitter messages, 2,319 forum messages, and 1,695 professional articles were collected for 86 of the leading cannabis brands spanning a period from January 2016-September 2017. Web crawlers were developed using python, Twitter's API (tweepy), Reddit's API, and Beautiful Soup (Richardson 2007), a search technology system that uses the html structure of websites to more easily extract iterable information. Researchers qualitatively compiled a list of 427 hashtags or search functions that uniquely identified brands (e.g. #kivaconfections for Kiva Confections). The web crawlers used this list to collect messages for each respective hashtag or phrase. Approximately 50,000 posts were scraped that comprise the dataset. Table 1 provides an excerpt of the dataset and structure.

Following the formation of the dataset, the sentiment and topics of these messages were analyzed using python packages which leverage differential language analysis techniques. The sentiment of the Twitter and Forum post language was obtained using VADER (Valence Aware Dictionary and sEntiment Reasoner), a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. It is fully open-sourced under the MIT License (Hutto & Gilbert, 2014; Rebeiro et al., 2016).

Table 1. Full dataset structure

Field	Description	Example
brand	The brand name	Kiva Confections
source	Where this is coming from	Leafly, Reddit, etc.
search	The search function used	Kiva%20Confections
timestamp	Timestamp	27-Dec-16
quarter	Quarter	1
text	Text from Twitter Posts, Reddit Posts	I am feeling awesome after eating this edible from Kiva Confections
User_type	Whether or not a twitter username is a dispensary, brand, or individual	Dispensary handle
id	Reddit has a unique Identifier for each post	342
composite	Sentiment Composite Score for Text	0.5
Topic	A topic id that signifies a topic	20111 = days of the week

3. Results

Researchers then took the full dataset and aggregated by brand, compiling the total number of twitter posts, professional articles, and forum posts, number of followers, as well as the mean overall sentiment across each of the seven quarters. Social media performance was then aggregated at the monthly basis and compared with monthly brand performance.

Monthly brand shares come from Brightfield Group's proprietary database (Brightfield Group, 2017). Brand shares are calculated based on a combination of sales data provided directly from brands and retailers as well as monthly menu audit scrapes identifying distribution and number of SKUs carried for each brand across each state. Baseline brand share calculation algorithms use distribution of SKUs as a proxy for sales, with algorithms weighted based on sales data provided by dispensaries and brands and validated by qualitative and primary research. An example of the social volume tracking for one brand can be seen in Figure 1.

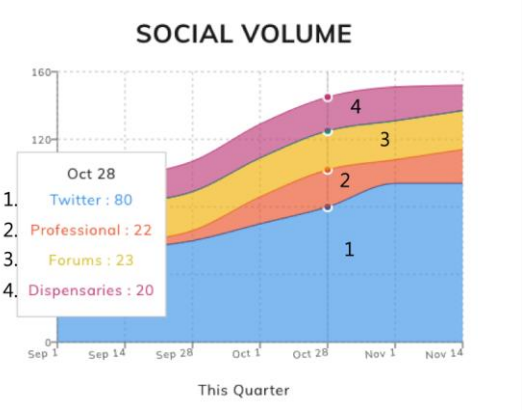

Figure 1. Social volume tracked over time for sample brand

Multiple regression analysis was used to test if the social volume metrics significantly predicted brand SKU's. The results of the regression indicated the six predictors explained 95% of the variance (R^2 = .95, F(120,33) = 4.88, p < .000). Table 2 explains the coefficients and Figure 2 displays the regression analysis.

Table 2. Analysis of coefficients

Variable	B	SEB	t	P-Value
quarter	30.45	55.55	0.55	.04
twitter	4382.39	1112.32	3.94	.00
sentiment	364.88	1331.5	0.27	.08
forums	56.77	19.70	2.88	.00
professional	26.87	11.19	2.40	.02
followers	.0000375	.00000141	2.66	.01

Notes R2 = .9466 (p <.001)

Researchers then used this model to predict 1 quarter in advance after information for the eighth quarter was obtained. The predicted brandshares used by this model accurately explained a significant proportion of the actual brandshares collected in the following quarter, R2 = .95, F(1,152) = 2695.6, p < .000.

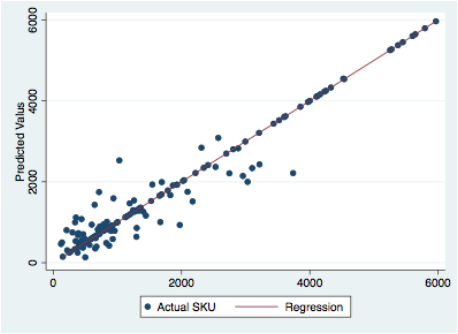

Figure 2. Regression line actual SKU vs. predicted SKU

4. Discussion and Next Steps

The approach presented leverages social listening: tracking both social volume and analysis of linguistic features within Twitter and other online platforms. Social volume metrics predicted outcomes like brandshares and SKUs for 86 cannabis brands. Brands can use social volume metrics to track and predict future brand performance.

The availability and prevalence of social media data now makes it possible to automatically derive characteristics from language use. This research is an example of a highly cost-effective and efficient way to obtain deep insights into an opaque and rapidly changing industry. By mining and analyzing data present on relevant social media channels and key publications, analysts were able to gain insights into the brands, performance and consumers of the legal cannabis industry. This dataset can be used for a variety of applications, including predicting future brand performance, identifying the ROI from social media presence for brands in the space and gaining deeper insights into modern consumer base of this highly sensitive and dynamic industry. An extension of this analysis can be conducted by collecting posts and following trends from individuals that post or follow a particular brand, enabling consumer segmentation into individual personas of each brand (e.g. millennial moms, techie bros) to emerge, which can cut the cost of consumer research down to a fraction of its original cost. Learning more about language patterns and following tendencies may help brands more effectively message to and reach receptive audiences. In a space as grassroots as cannabis advertising, analyses like ours may lead to illuminating insights about a budding industry.

This technique can extend to other industries and consumer research. Companies new to market with low budgets or quickly changing industries can use methods like these to derive automatic cost-effective insights into their consumers. The ability to not only track the messaging around the product, but the aspects of consumers (e.g. collecting posts and

liking trends of those messaging or following your brand) allows for thorough capturing of both the active vocal consumers as well as their silent followers.

References

Antoniou, A. (2017). Social network profiling for cultural heritage: combining data from direct and indirect approaches. *Social Network Analysis and Mining*, *7*(1), 39. https://doi.org/10.1007/s13278-017-0458-x

Auerbach, B. (2018, March 15). Looking At The Year Ahead In Cannabis, Technology, Microdosing, Home Brew and Blockchain. Retrieved March 18, 2018, from https://www.forbes.com/sites/bradauerbach/2018/03/15/looking-at-the-year-ahead-in-cannabis-technology-microdosing-home-brew-and-blockchain/

Azofeifa, A., Sherman, L. J., Mattson, M. E., & Pacula, R. L. (2018). Marijuana buyers in the United States, 2010–2014. *Drug & Alcohol Dependence*, *183*, 34-42.

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

Brightfield Group (2017. July). *Understanding Cannabidiol: CBD*. Retrieved from Brightfield Group database.

Chen, A. T., Zhu, S. H., & Conway, M. (2015). What online communities can tell us about electronic cigarettes and hookah use: a study using text mining and visualization techniques. *Journal of medical Internet research*, *17*(9). doi: 10.2196/jmir.4517. http://www.jmir.org/2015/9/e220/

Culotta, A., & Cutler, J. (2016). Mining brand perceptions from Twitter social networks. *Marketing science*, *35*(3), 343-362. https://doi.org/10.1287/mksc.2015.0968

Gunelius, S. (2018, January 24). 5 Proven Marijuana Marketing and Advertising Ideas that Work. Retrieved March 24, 2018, from https://cannabiz.media/5-proven-marijuana-marketing-and-advertising-ideas-that-work/

Interactive Advertising Bureau. (2018). *Marijuana Legalization and Advertising Restrictions in the United States*. Retrieved from https://www.iab.com/marijuana-legalization-advertising-restrictions-united-states/

Lievens, F., & Van Iddekinge, C. H. (2016). Reducing the Noise From Scraping Social Media Content: Some Evidence-Based Recommendations. *Industrial and Organizational Psychology*, *9*(3), 660-666. https://doi.org/10.1017/iop.2016.67

McGregor, F., Somner, J. E., Bourne, R. R., Munn-Giddings, C., Shah, P., & Cross, V. (2014). Social media use by patients with glaucoma: what can we learn?. *Ophthalmic and Physiological Optics*, *34*(1), 46-52. https://doi.org/10.1111/opo.12093

McVey, E. (2017, November 13). Chart: Most effective forms of advertising for cannabis businesses. Retrieved March 24, 2018, from https://mjbizdaily.com/chart-effective-forms-marketingadvertising-marijuana-businesses/

Nguyen, A., Hoang, Q., Nguyen, H., Nguyen, D., & Tran, T. (2017, January). Evaluating marijuana-related tweets on Twitter. In *Computing and Communication Workshop and*

Conference (CCWC), 2017 IEEE 7th Annual (pp. 1-7). IEEE. doi: 10.1109/CCWC.2017.7868364

Richardson, L. (2007). Beautiful soup documentation. Retrieved from https://media.readthedocs.org/pdf/beautiful-soup-4/latest/beautiful-soup-4.pdf

Sarker, A., O'Connor, K., Ginn, R., Scotch, M., Smith, K., Malone, D., & Gonzalez, G. (2016). Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from Twitter. *Drug safety*, *39*(3), 231-240. https://doi.org/10.1007/s40264-015-0379-4

Schuster, W. M., & Wroldsen, J. (2018). Entrepreneurship and Legal Uncertainty: Unexpected Federal Trademark Registrations for Marijuana Derivatives. *American Business Law Journal*, *55*(1), 117-166. https://doi.org/10.1111/ablj.12118

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, *8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791

St Louis, C., & Zorlu, G. (2012). Can Twitter predict disease outbreaks?. *BMJ: British Medical Journal (Online)*, *344*. doi: 10.1136/bmj.e2353

Sul, H. K., Dennis, A. R., & Yuan, L. I. (2017). Trading on twitter: Using social media sentiment to predict stock returns. *Decision Sciences*, *48*(3), 454-488. https://doi.org/10.1111/deci.12229

Yaden, D. B., Eichstaedt, J. C., Kern, M. L., Smith, L. K., Buffone, A., Stillwell, D. J., ... & Schwartz, H. A. (2017). The Language of Religious Affiliation: Social, Emotional, and Cognitive Differences. *Social Psychological and Personality Science*, 1948550617711228. https://doi.org/10.1177/1948550617711228

Zhang, M. (2017, November 7). The Global Marijuana Market Will Soon Hit $31.4 Billion But Investors Should Be Cautious. Retrieved March 20, 2018, from https://www.forbes.com/sites/monazhang/2017/11/07/global-marijuana-market-31-billion-investors-cautious/#2c79945c7297