Applied Marketing Analytics 5(2) 2019

PREPRINT: See https://www.ingentaconnect.com/content/hsp/ama/2019/00000005/00000002/art00007

Using weak supervision to scale the development of machine learning models for social media-based marketing research

Jennifer Cutler and Aron Culotta

Abstract

Marketers have expressed substantial enthusiasm about the potential of social media data to enhance marketing research, and the computer science literature provides many examples of using the text and network connections of social media users to infer measurements of broad interest to marketers. Yet, adoption of such machine learning approaches has been surprisingly limited in marketing practice. We believe that, for many prediction and classification tasks, the hurdle of procuring the labeled training data that is generally necessary to build such models may be limiting widespread adoption. Such training data can be very expensive to generate, and for many tasks, may not be feasible at all. Further, the rapidly changing nature of social platforms leads to an often limited lifespan for models once trained, further decreasing the value proposition of investment. We propose that the organic structure of social media itself can be leveraged to circumvent the need for such curated training data for a variety of marketing-relevant prediction and classification tasks, making such models much more accessible and useful to marketers. We describe two emerging methodological themes of weak supervision (training on exemplars and training on groups) that are broadly promising towards this goal, and discuss examples of how they have been applied towards a variety of marketing tasks—in all cases, without requiring any manually labeled training data, and in some cases, requiring nothing more than a single keyword as input. Our hope in presenting these methodological themes and implementation examples is that it will inspire and facilitate the development of more flexible, scalable, and cost-effective models for marketing applications, and stimulate additional research in this area.

1

Over the past decade, the effective use of social media as a marketing resource has emerged as a top priority among global CMOs and marketing managers[1], and by 2020 it is expected that a full quarter of firms' marketing budgets will be allocated towards social media[2]. The growing excitement is not just about disseminating marketing messages to customers, but also about the potential use of secondary social media data for generating insights about consumers and markets to supplement or replace traditional marketing research methods involving primary data collection (such as surveys) which can be costly and subject to many limitations [3, 4].

Yet, the vast majority of marketing practitioners are reporting unpreparedness for the digital data explosion [1], and as of the end of 2015, only 6.2% were even attempting to mine unstructured text and network data on social media for consumer insights [2]. There has been limited guidance from the academic marketing literature, where articles on machine learning

and artificial intelligence methods for mining social media data for marketing insights have been surprisingly sparse.

At the same time, the computer science literature has exploded with thousands of papers on machine learning-based social media analytics, many with relevance to marketing. These papers have shown, among other things, that a wide range of measures about consumers that are of interest to marketing research can be reliably inferred from the text and network data trails left by consumers on social media [5, 6, 7]—and that these inferred measures can be of practical value to marketers, for example, by improving the performance of social media ad campains [8]. Many of these findings have received a great deal of attention not only within the computer science field, but also among much broader audiences. For example, studies showing that detailed personality and characteristic profiles of Facebook users can be inferred from what they "like" in their Facebook profiles[9, 10] were the subject of a broadly disseminated TED talk receiving over two million views[11].

Given the focus on marketing-relevant social media problems in the machine learning literature and the publicity around this work, why, then, has adoption in marketing research and practice been so limited? One reason may lie in communication barriers between marketing and data science [12], and importantly, interdisciplinary efforts are being established at many universities and conferences to help bridge these gaps. Another reason may relate to data access and privacy concerns. In the wake of high-profile scandals involving the sharing and use of personal data[13, 14], new regulations such as the California Consumer Privacy Act (CCPA) and the European General Data Protection Regulation (GDPR) have been enacted to attempt to protect privacy and provide consumers with more control over their information, and many social media platforms have begun revising and tightening their data access policies[15, 16]. Marketers may be unsure how to navigate this changing environment, and how to legally and ethically acquire and extract optimal value from personal data. Importantly, ongoing research is helping to clarify trade-offs and systematize and formalize risks in various methods for processing sensitive private data[17]. However, even as policies and standards evolve for accessing and sharing private data, there is still a wealth of public social media data that is not considered protected by modern regulations and can be freely accessed and used for marketing purposes, with fewer ethical quandaries to navigate. Thus, we believe there is another key reason for the limited adoption of social media data mining in marketing practice: the limitations of the extant machine learning approaches themselves, in terms of the practicality and cost-effectiveness of applying them in practice towards customized marketing research tasks. In this paper, we discuss a emerging thematic approach for overcoming this obstacle, making the development of such models more scalable and cost-effective for marketers.

As the term "machine learning" is quite broad, for this paper, we limit our scope to discussing social media tasks involving the prediction or classification of market research measures (such as the perceived reliability of a brand, or the amount of consumer interest in a particular topic), historically obtained through surveys or human observation. Prevalent such applications in the academic literature include tracking brand sentiment [5, 6] and inferring user personality [7, 9, 10]. Such prediction and classification tasks have great use to marketers, and are typically performed through supervised machine learning. The researchers begin with a large set of labeled training data, from which they determine what (observable) social media features are predictive of the (unobservable) measure(s) of interest. Unfortunately, such data can be difficult and costly to obtain (which makes sense—if the labels were easy to acquire, then the need for a

predictive model would not be very high!). For example, in order to predict users' personalities from the text of their Facebook statuses, one team of researchers relied on over 70,000 Facebook users self-reporting their personality traits via an app, and also providing researchers access to their status updates[18]. To predict the political alignment of Twitter users, another research team relied on two researchers manually reading through the profiles and activities of 1,000 Twitter users to subjectively assess their individual political leanings[19]—a highly tedious and time-consuming process, and also (for many talks) error-prone.

Because of this need for extensive labeled training data, building such predictive models comes with many of the same limitations as primary data collection (along with the often additional costs of manual tuning by skilled data scientists). Furthermore, even if one does invest in building such a model, it may bear a high risk of going out of date quickly in the rapidly evolving context of social media (for example, many features found in prior studies to be predictive of key personality traits [10] no longer exist on the platform). The high up-front costs and limited lifespans of these supervised machine learning models means that it may often not be feasible or cost effective to invest in the (often very substantial) resources needed to build and maintain them for the particular measures and context of interest to a given marketing researcher.

However, a fortunate feature of social media is that end users are continuously and organically providing valuable structure to the text and social network data therein. Importantly, this organic structure can be exploited to circumvent the need for curated training data for a wide variety of prediction and classification tasks useful to marketers— thus opening the door for substantially more feasible, cost-effective, and potentially effective implementation in practice.

2 Leveraging Structure from Users

A major contributor to the revolution in Internet search in the late 1990's and early 2000's was the incorporation of the idea that web users were organically providing trackable information about what web pages they considered important, through the creation of backlinks— and that the ranking of search results could be improved by effectively interpreting this user-provided structure[20, 21]. While the ranking algorithm now used by Google comprises a wide range of complex and proprietary algorithms, a core component of the innovation was the shift away from static predictive models and towards "a machine to capture living time and living labour and to transform the common intellect into network value"[22]. Social media did not come to prominence among consumers (and then marketers) until well after Internet search did—but it similarly contains organic, evolving structure that can help scale algorithms for marketing measurement if leveraged appropriately. Most analogously to Google's use of backlinks as a source of information about a webpage's importance, many have observed, for example, that "follow" and "like" relationships (in addition to hyperlinks) now provide easily accessible evidence of user judgments and affinity, and metrics such as a brand's follower count and a post's engagement rate are commonly used in marketing practice to help gauge popularity[23]. However, in marketing applications that seek to provide more nuanced and detailed information about consumers and brands, such relationships are by and large treated as features from which another measure of interest (such as user personality) can be predicted, generally through the

aforementioned process of curating labeled data and training a supervised machine learning model[9, 10].

However, the inherent and organic structure of social media make it particularly amenable to the emerging method of *weak supervision* [24], i.e., the use of noisier but easier-to-acquire training data. The goal of this paper is to discuss two promising methodological themes of weak supervision that have been emerging towards the goal of enabling accurate social media prediction and classification models using minimal human input: training on exemplars, and training on groups. We draw on recent and in-progress examples to conceptually illustrate how these approaches can be applied towards three common marketing tasks: classifying text by topic, measuring dimensions of brand image, and identifying user characteristics, without the need for any individually labeled training data. It is our hope that these examples will convey the flexibility of the broad methodological themes, and stimulate additional marketing research applications.

3 Learning from Exemplars

As discussed earlier, a key obstacle in the development of supervised machine learning models for social media is the onerous task of manually labeling enough individual data items to effectively train a high quality model. However, because organic social media data are actually quite organized, it is not always necessary to manually label individual items; greater efficiency can be obtained by leveraging this structure. Social media data such as user texts and relationship-based endorsements are not distributed randomly over a platform; rather, such data are naturally organized and correlated by accounts. The first insight in the process of exemplar-based training is that, given this structure, in many cases we can label and train by accounts, rather than by individual data items (such as individual posts or individual "likes").

For example, if we can label a set of accounts that we know are dedicated to sustainability (such as the Twitter account @GreenPeace), we can infer that the textual contents of such exemplar accounts' collective statuses will reference the topic of sustainability with greater frequency than that of similar, but non-environmentally oriented accounts (such as @RedCross); and that their followers are, on aggregate, more likely to value sustainability. Of course, not every post written by GreenPeace will be about the environment, and not every one of their followers will value the environment. But, the relative proportions of such will be predictably higher for a set of well-chosen exemplars, compared to non-exemplars.

Labeling accounts as exemplars of a quality of interest provides two important practical advantages compared to labeling individual data items. First, it dramatically reduces the amount of data that needs to be labeled. Second, it allows the resulting models to easily stay up-todate even if linguistic and popularity trends change rapidly. Because accounts are live, rather than static data items, the account-level labels can be used to re-train a model with the most up-to-date features (e.g., newly invented hashtags) at any time.

The next insight in the process of exemplar-based training is that it's not only features that are organized into accounts—accounts are also often organically organized by users into meaningful groups. For example, on the Twitter platform, there are over three million user-generated *Lists* of accounts, which are curated by users to create thematic news feeds. In many cases, such structure can be exploited to automatically identify a broadly representative set of

correlated exemplar accounts— for example, by entering in a single keyword query to automatically search for relevant lists and extract suitable exemplars.

In the following sections, we describe in more concrete terms how exemplar-based training has been used to fully automate two common marketing tasks: classifying social media posts by topics, and estimating the strength of brand image dimensions.

3.1 Example Marketing Application: Tracking Social Media Posts by Topic

A common goal in social media marketing research is to identify, quantify, and track user-generated (UGC) or marketer-generated (MGC) content about a particular topic. For example, brand managers may be interested in tracking how their competitors are positioning themselves through MGC. Marketing researchers may be interested in tracking the dynamics between MGC and consumer attitudes. Marketers deciding which platforms to advertise on may wish to measure the volume of conversations about different topics on different platforms. Product researchers may wish to track trends in user conversations over time.

Common approaches for building text classifiers include using lexiconbased tools (such as LIWC [25]), which rely on a set of expert-contributed keywords that are believed to indicate relevance to a topic of interest; and supervised machine learning approaches, which (typically) require multiple judges to hand label a large set of posts according the their relevance to the topic at hand. To date, the lexicon approach has been utilized with far greater frequency in the marketing literature (LIWC itself has been used in hundreds of social science studies), possibly due to the relative ease of applying it. However, pre-existing lexicons exist only for a limited number of topics, and the extant lexicons may not be consistently reliable for social media applications, where posts can be short and linguistic cues (such as hashtags and emoticons) evolve rapidly[26]. Furthermore, new lexicons can be difficult to build for topics with many ambiguously relevant keywords (e.g., is the word “green” a reliable indicator of sustainability-relevance, when it can also refer simply to the color or other non-environmental meanings?). While traditional supervised machine learning methods may promise customizing to context, in practice, manually labeling enough posts to enable accurate training can in many cases be prohibitively expensive or difficult, in particular for sparsely discussed topics, which are often the most useful to track. Furthermore, the short length of posts may limit such models’ abilities to adapt as new hashtags, slang, and cultural references evolve. For some topics, exemplar-based training can provide a low-cost and highly accurate alternative approach to text classification of social media posts.

One context that is particularly well-suited for this approach is the task of classifying cause-related MGC on social media—for example, to track competitive positioning, monitor authenticity, or study its relationship to brand image. This method has been applied to build a classifier to track sustainability-related marketing generated text on Twitter and Facebook[27]. The process begins with using CharityNavigator.org to semi-automatically identify forty environmental nonprofits (the exemplars) and forty non-environmental non-profits (the controls). Twitter accounts are identified for each non-profit, and Twitter’s API is used to access the most recent tweets from each. The tweets are cleaned and normalized according to standard processes, and tokenized into hashtags and bigrams. Tokens not appearing in tweets from at least three distinct accounts are removed to minimize organization-specific terms. Feature selection is used to identify tokens most predictive of being in the exemplar set, and an

"environmental" classifier is built using standard methods from these predictive terms. When tested against hand-labeled brand tweets, the precision obtained from such methods was estimated to be as high as 99%. Though this work is still in progress, we expect that this approach would work well for other social causes as well, such as health and social justice (for which there are also non-profit organizations that can be downloaded from CharityNavigator.org).

For classification of other topics (besides social causes), for which high-quality exemplars may not be as easily identifiable, Twitter Lists provides a compelling alternative. Because users are continually curating news feeds of accounts that tweet with frequency about specific topics, searching these lists by a keyword of interest can return a list of relevant lists, each of which is a list of (potentially) relevant accounts. Subject to some automated quality controls (e.g., accounts must be active and should appear on more than one list), these accounts can become meaningful exemplars (and exemplars identified through a similar process, but for a different keyword, can be used as controls, after removing any overlapping accounts). Researchers have had success using this approach to build tweet classifiers for a range of topics (e.g., "books", "art", "technology"), requiring nothing more than a single word of input[28]. Because exemplars identified through lists (or through identification sites such as CharityNavigator) tend to be organizations, or professional influencers with large followings, their linguistic patterns are often more similar to that of MGC than UGC. However, in some cases, domain adaptation may be applied to achieve accuracy in classifying UGC (competitive with a fully supervised approach) without sacrificing the automation[28].

3.2 Example Marketing Application: Tracking Attribute-Specific Brand Perceptions

Another common goal in marketing research is to measure consumer brand perceptions[29], often with the goal of creating perceptual maps. A perceptual map is a foundational tool in marketing research that plots a set of brands by how strongly they are perceived along two different dimensions (e.g., taste vs. nutrition) and is used for tasks such as identifying market opportunities and tracking brand image changes[30]. For decades, they have been generated by surveying consumers directly [31], but this reliance on primary data has been frequently lamented by marketing researchers as costly, limited in scale, and subject to a myriad of self-selection and self-report biases[3, 32, 33]. These limitations are becoming increasingly urgent in modern times, as survey response rates have fallen precipitously in recent years[34].

Although some have attempted to investigate brand image through mining user-generated text on social media[4], much of this work has focused on the discovery of new associations through clustering approaches[35, 36], rather than on the quantification of image along of pre-determined dimensions of interest. Sentiment analysis has been applied many times towards the goal of measuring overall brand sentiment[37, 38, 39]; however, sentiment is a topic-independent feature that can be measured for any brand mention. For many attributes of interest (e.g., environmental friendliness or luxuriousness), the sparsity of UGC mentions of the brand and that attribute may limit the reliability of UGC-based analyses. However, although most Twitter uses don't actively post text[40, 41] (and fewer still post about a given brand in concurrence with a perceptual attribute of interest to marketers), the silent majority of Twitter users are still providing valuable information about their values and perceptions through their "mere virtual presence"[42]—that is, through their location in the social network.

Researchers have used exemplar-based training to automatically generate perceptual maps for brand image attributes of interest by assessing the similarity between a brand's followers and the followers of exemplar accounts of that attribute[43]. Specifically, a keyword representing an attribute of interest (e.g., "nutrition", or "luxury") is used to search Twitter lists for lists users have made about that attribute. As in the prior example with text classification, a large set of exemplars is identified by selecting active accounts that appear on multiple lists (and, if applicable, excluding any accounts that are for the brands whose image is to be estimated). Once this set is established, the IDs of the followers of both the exemplars and the brands to be mapped can be extracted from Twitter's API, and compared. Because the followers of the exemplar accounts are, on average, expected to value the attribute at hand at a higher than typical rate, having a high degree of followership overlap with these exemplars provides a signal that the brand is valued for that attribute. In this paper, the authors quantified brand image as the brand's average Jaccard similarity with the exemplars, inversely weighted by the exemplar's number of followers—though showed that the general approach is robust to a variety of network similarity measures. Using a sample of over two hundred brands across three sectors, for three example attributes, the fully automated exemplar-trained network similarity ratings correlated strongly with directly elicited survey ratings for the attributes at hand.

Exemplar-based training has the advantage of being fully automated; because both lists and accounts are being actively updated by users, this approach allows training on the most up-to-date text and network material at the time of running, often requiring no more than a keyword of input. As such, it offers feasibility and scalability of development relative to more traditional supervised training approaches. It is our hope that the two examples discussed—classifying text and quantifying brand perceptions—demonstrate the flexible potential of this general approach towards increasing the accessibility of data mining tasks to marketers. While these methods are not yet mainstream, it is our hope that these examples will stimulate extensions and refinements of these methods to additional marketing applications.

4 Learning from Groups

Another promising approach towards lower-cost prediction and classification involves the use of learning from customized groupings of social media users. While individually labeled data can be difficult, costly, or impossible to acquire for many variables of interest, aggregate data labels are often much more readily available (e.g., 14% of individuals in this zip code have depression vs. this specific individual has depression), especially as regulations and concerns regarding the protection of individual-level data strengthen. Inasmuch as prediction and classification models can be trained on label proportions for groups, rather than on individually labeled data, it opens the door to far more applications at far lower cost. Over the past decade, computer scientists have been developing and refining general methods for learning from label proportions (LLP)—enabling the utilization of existing secondary distant labels, rather than manually-curated individual labels, for some classification and prediction tasks [44, 45, 46]. However, little of this literature has focused of social media-based applications. We believe that social media is particularly amenable to this approach, as users can be grouped meaningfully a number of ways (by location, brand followership, interest in a topic, etc.), which provides substantial flexibility in creating groups that map to useful sources of aggregated labels. As such, we suspect that using

LLP methods with different groupings of social media users can open the door to much simpler and more scalable social media data mining for marketers. In the following section, we illustrate through the example problem of inferring user characteristics how LLP can be used applied towards market research tasks on social media.

4.1 Example Marketing Application: Estimating User Characteristics

Marketers are often interested in knowing the demographic or characteristic profile of groups of consumers. Being able to measure this on social media can help marketers understand questions such as how their brand communities differ from their competitors', how their brand communities are changing over time, how to characterize consumers interested in a particular emerging topic, and key differences are between consumers who post positively vs. negatively about their brand. Going further, being able to classify individual users according to demographics and characteristics opens the door to highly personalized marketing, more nuanced models of advertising response, as well at the ability to control for demographics in online research studies.

Because there are so many ways to meaningfully group users on social media, there is considerable opportunity for creating groupings that match meaningfully to extant sources of aggregated label proportions. For example, the overall demographic profiles of visitors to a large variety of brand websites is freely available on websites such as Quantcast.com and Alexa.com. While such information is not directly available for all brands, and may not be available for brands of interest to a given marketing team, it is available for thousands of brands that can be used for training a more general model. Researchers have, for example, matched proportion labels for demographics of brands' website visitors (including categories for age, ethnicity, education, political views, and parental status) to those brands' follower sets on Twitter[1].

The text and network features of the brands' followers were used as features to train a model to estimate the percent of followers who belonged to each demographic category. Using hold-out samples from the website visitor profiles for validation, high accuracy rates were achieved, and the linguistic features and followed accounts most indicative of membership in any given demographic category were identified[28]. In a similar manner, other researchers have mapped county-level racial profiles reported in U.S. Census to tweets geotagged from each county, and trained a classifier to predict the racial composition of a county based on the text of the tweets originating from that county[47].

These are two examples of sources of public data for which groupings on social media can be easily created to map to the aggregated labels; we expect, given the flexibly of groupings on social media, that there are potentially many more. A potential advance of this method is that, once a predictive model is built using Twitter user groups that map to extant aggregated data labels (in the above cases, brand visitors or county residents), then users can be grouped in different ways, and the model can be applied to predict detailed demographic profiles of groupings relevant to a wide range of marketing questions— for example, users who tweet about #BigData; users who complain frequently online about a brand; or users who are most (and least) likely to engage with brand posts. Encouragingly, both of these example studies proceeded to adapt their

[1] with the assumption that, if brand Twitter followers differed systematically from brand website visitors, these differences would be consistent across brands

resulting predictive models into models that classify individual users by membership into a demographic category. Using individually labeled data for a subset of demographic categories that could be reliably inferred from profile observation, these classification models, which were trained entirely on distant secondary label proportions, were competitive with traditionally supervised approaches requiring manually labeled individual-level training data.

It is our hope that these examples will motivate other researchers to take advantage of the flexible groupings on social media and seek additional applications for learning from label proportions from cost-effective, up-to-date aggregately labeled data, as an additional approach towards circumventing the need for individually labeled training data.

5 Moving Forward

We have described two methodological themes (learning from exemplars and learning from groups) that are emerging as promising approaches towards making machine learning for social media-based marketing research more flexible, scalable, and cost-effective. Although the presented methods of weak supervision provide many advantages over the more prominent supervised machine learning methods, they, of course, have their own limitations and cannot be applied in all contexts. Exemplar-based training can only be performed when exemplars of the trait to be predicted can be identified, and when the trait can be predicted from observable account features such as language use or network connections. Group-based training can only be performed when groupings can be made on social media that map to sources of aggregated data for variables of interest. However, we believe that the nature of social media—with meaningful organic user networks— makes it particularly amenable to a wide range of such applications. While these approaches do not preclude the need for training and validation, they offer a means for doing so in a more efficient and scalable manner than is required for fully supervised methods.

We have discussed initial implementations of these approaches towards three very different marketing tasks: classifying posts by topic, quantifying dimensions of brand image, and inferring user characteristics. Results thus far have been promising, and there is much room to grow in both refining and extending the methods and applications. It is our hope that the ideas and examples presented here will stimulate additional research in this emerging area, leading to an increase in the variety and automaticity of reliable social media-based marketing research algorithms, and more widespread adoption in marketing practice.

6 Acknowledgments

The research goals described in this paper have been partially funded by the National Science Foundation Grant #1618244.

References

[1] IBM. From Stretched to Strengthened: Insights from the Global Chief Marketing Officer Study; 2011. IBM Institute for Business Value: CMO C-Suite Studies.

[2] Moorman C. The CMO Survey Report: Highlights and Insights; 2015. Sponsored by Duke University, McKinsey&Company, and the American Marketing Association.

[3] Aaker DA. Measuring brand equity across products and markets. California management review. 1996;38(3):102–120.

[4] Fader PS, Winer RS. Introduction to the special issue on the emergence and impact of user-generated content. Marketing Science. 2012;31(3):369–371.

[5] Chamlertwat W, Bhattarakosol P, Rungkasiri T, Haruechaiyasak C. Discovering Consumer Insight from Twitter via Sentiment Analysis. J UCS. 2012;18(8):973–992.

[6] Mostafa MM. More than words: Social networks' text mining for consumer brand sentiments. Expert Systems with Applications. 2013;40(10):4241–4251.

[7] Quercia D, Kosinski M, Stillwell D, Crowcroft J. Our Twitter profiles, our selves: Predicting personality with twitter. In:IEEE third international conference on social computing; 2011. p. 180–185.

[8] Matz SC, Kosinski M, Nave G, Stillwell DJ. Psychological targeting as an effective approach to digital mass persuasion. Proceedings of the National Academy of Sciences. 2017;114(48):12714– 12719.

[9] Golbeck J, Robles C, Turner K. Predicting personality with social media. In: CHI'11 extended abstracts on human factors in computing systems. ACM; 2011. p. 253–262.

[10] Kosinski M, Stillwell D, Graepel T. Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences. 2013;110(15):5802– 5805.

[11] Golbeck J. Your social media likes expose more than you think. In: TEDxMidAtlantic. TED; 2013. .

[12] Frank MR, Wang D, Cebrian M, Rahwan I. The Evolution of Citation Graphs in Artificial Intelligence Research. Nature Machine Intelligence. 2019;1:79–85.

[13] Cadwalladr C, Graham-Harrison E. Revealed: 50 million facebook profiles harvested for Cambridge Analytica in major data breach. The Guardian. 2018 Mar;.

[14] McGoogan C. NHS illegally handed Google firm 1.6m patient records, UK data watchdog finds. The Telegraph. 2017 July;.

[15] Dwoskin E. Facebook's Mark Zuckerberg says he'll reorient the company toward encryption and privacy. Washington Post. 2019 March;.

[16] Hautala L. Twitter's new privacy policy: Here's what we do with your data. CNet. 2018 April;.

[17] Rocher L, Hendrick JM, deMontjoye YA. Estimating the success of re-identifications in incomplete datasets using generative models. Nature Communications. 2019;10(3069).

[18] Park G, Schwartz HA, Eichstaedt JC, Kern ML, Kosinski M, Stillwell DJ, et al. Automatic personality assessment through social media language. Journal of personality and social psychology. 2015;108(6):934.

[19] Conover MD, Gon¸calves B, Ratkiewicz J, Flammini A, Menczer F. Predicting the political alignment of Twitter users. In: 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing. IEEE; 2011. p. 192–199.

[20] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems. 1998;30(17):107–117.

[21] Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab; 1999.

[22] Pasquinelli M. Google's PageRank algorithm: A diagram of cognitive capitalism and the rentier of the common intellect. Deep search: The politics of search beyond Google. 2009;p. 152–162.

[23] Larson J, Draper S. Internet Marketing Essentials. Stukent, Inc.; 2017.

[24] Bach SH, Rodriguez D, Liu Y, Luo C, Shao H, Xia C, et al. Snorkel drybell: A case study in deploying weak supervision at industrial scale. In: Proceedings of the 2019 International Conference on Management of Data. ACM; 2019. p. 362–375.

[25] Pennebaker JW, Chung CK, Ireland M, Gonzales A, Booth RJ. The development and psychometric properties of LIWC2007. Austin, TX; 2007.

[26] Crystal D. Language and the Internet. New York: Cambridge University Press; 2001.

[27] Cutler J, Culotta A. Green Marketing in Social Media. Working Paper. 2018;.

[28] Culotta A. Training a text classifier with a single word using Twitter Lists and domain adaptation. Social Network Analysis and Mining. 2016;6(1):1–15.

[29] Lehmann DR, Keller KL, Farley JU. The structure of surveybased brand metrics. Journal of International Marketing.
2008;16(4):29–56.

[30] Shocker AD, Srinivasan V. Multiattribute approaches for product concept evaluation and generation: A critical review. Journal of Marketing Research. 1979;16(2):159–180.

[31] Steenkamp JB, Van Trijp H. Attribute elicitation in marketing research: a comparison of three procedures. Marketing Letters. 1997;8(2):153–165.

[32] Day GS. The threats to marketing research. Journal of Marketing Research. 1975;12(4):462–467.

[33] McDaniel SW, Verille P, Madden CS. The threats to marketing research: An empirical reappraisal. Journal of Marketing Research. 1985;22(1):74–80.

[34] Kohut A. Assessing the representativeness of public opinion surveys; 2012. Pew Research Center.

[35] Archak N, Ghose A, Ipeirotis PG. Deriving the pricing power of product features by mining consumer reviews. Management Science. 2011;57(8):1485–1509.

[36] Netzer O, Feldman R, Goldenberg J, Fresko M. Mine Your Own Business: Market-Structure Surveillance Through Text Mining. Marketing Science. 2012;31(3):521–543.

[37] Ludwig S, de Ruyter K, Friedman M, Bru¨ggen EC, Wetzels M, Pfann G. More Than Words: The Influence of Affective Content and Linguistic Style Matches in Online Reviews on Conversion Rates. Journal of Marketing. 2013;77(1):87–103.

[38] Sonnier GP, McAlister L, Rutz OJ. A dynamic model of the effect of online communications on firm sales. Marketing Science. 2011;30(4):702–716.

[39] Tirunillai S, Tellis GJ. Does chatter really matter? Dynamics of user-generated content and stock performance. Marketing Science. 2012;31(2):198–215.

[40] Toubia O, Stephen AT. Intrinsic vs. Image-Related Utility in Social Media: Why Do People Contribute Content to Twitter? Marketing Science. 2013;32(3):368–392.

[41] Wu S, Hofman JM, Mason WA, Watts DJ. Who says what to whom on Twitter. In: Proceedings of the 20th international conference on World Wide Web. ACM; 2011. p. 705–714.

[42] Naylor RW, Lamberton CP, West PM. Beyond the "like" button: The impact of mere virtual presence on brand evaluations and purchase intentions in social media settings. Journal of Marketing. 2012;76(6):105–120.

[43] Culotta A, Cutler J. Mining Brand Perceptions from Twitter Social Networks. Marketing Science. 2016;35(3):343–362.

[44] Mann GS, McCallum A. Generalized expectation criteria for semisupervised learning with weakly labeled data. Journal of machine learning research. 2010;11(Feb):955–984.

[45] Quadrianto N, Smola AJ, Caetano TS, Le QV. Estimating labels from label proportions. Journal of Machine Learning Research. 2009;10(Oct):2349–2374.

[46] Yu FX, Choromanski K, Kumar S, Jebara T, Chang SF. On learning from label proportions. arXiv preprint arXiv:14025902. 2014;.

[47] Mohammady E, Culotta A. Using county demographics to infer attributes of twitter users. In: Proceedings of the joint workshop on social dynamics and personal attributes in social media; 2014. p. 7–16.