JOURNAL OF INTERACTIVE ADVERTISING
https://doi.org/10.1080/15252019.2019.1700851

Computationally Analyzing Social Media Text for Topics: A Primer for Advertising Researchers

Joseph T. Yun[a,b], Brittany R. L. Duff[b], Patrick T. Vargas[b], Hari Sundaram[c], and Itai Himelboim[d]

[a]Department of Accountancy, Gies College of Business, University of Illinois at Urbana–Champaign, Champaign, Illinois, USA; [b]Charles H. Sandage Department of Advertising, College of Media, University of Illinois at Urbana–Champaign, Champaign, Illinois, USA; [c]Department of Computer Science, Grainger College of Engineering, University of Illinois at Urbana–Champaign, Champaign, Illinois, USA; [d]Department of Advertising and Public Relations, Grady College of Journalism and Mass Communication, University of Georgia, Athens, Georgia, USA

ABSTRACT
Advertising researchers and practitioners are increasingly using social media analytics (SMA), but focused overviews that explain how to use various SMA techniques are scarce. We focus on how researchers and practitioners can computationally analyze topics of conversation in social media posts, compare each to a human-coded topic analysis of a brand's Twitter feed, and provide recommendations on how to assess and choose which computational methods to use. The computational methodologies that we survey in this article are text preprocessed summarization, phrase mining, topic modeling, supervised machine learning for text classification, and semantic topic tagging.

KEYWORDS
Social media analytics; topic discovery; topic modeling; computational advertising; machine learning

Social media analytics (SMA) is the process of using computational methods and tools to extract insights from social media data (Fan and Gordon 2013) as well as measure the performance of social media campaigns (Murdough 2009). SMA is widely used by companies and academia to measure and analyze consumers' reactions to advertising phenomena such as brand promotions and new trends (e.g., Daniel, Crawford Jackson, and Westerman 2018; Fulgoni 2015; Kwon 2019; Moe, Netzer, and Schweidel 2017; Rosenkrans and Myers 2018; Yun and Duff 2017). The growth of SMA parallels the growth of social media use, as a recent Pew Research Center survey showed that 75% of adults in the United States use social media, and this percentage grows to 94% for 18- to 24-year-olds (Smith and Anderson 2018).

While public social media posting creates a large set of data for advertising researchers and practitioners, the use of these data has been largely constrained to those with more technical knowledge in computation and algorithms or those with access to professional tools that may be cost-prohibitive. While there have been several recent academic papers that help researchers discern how to go about analyzing computational data for marketing and advertising research, most of these papers function as high-level overviews (e.g., Humphreys and Wang 2018; Liu, Singh, and Srinivasan 2016; Malthouse and Li 2017; Murdough 2009; Rathore, Kar, and Ilavarasan 2017) without giving details on how to actually go about conducting analyses. More focused "how-to" articles on specific computational methods for research are limited, especially how to detect topics of conversation in social media text for the purposes of advertising. As far as we know, there are no articles detailing what to consider when one wants to detect topics of

CONTACT Joseph T. Yun jtyun@illinois.edu Department of Accountancy, Gies College of Business, University of Illinois at Urbana–Champaign, 2046 Business Instructional Facility, 515 East Gregory Drive, Champaign, IL 61820, USA.
Joseph T. Yun (PhD, University of Illinois at Urbana–Champaign) is a research assistant professor of accountancy and director of the Data Science Research Service in the Gies College of Business, University of Illinois at Urbana–Champaign.
Brittany R. L. Duff (PhD, University of Minnesota) is an associate professor of advertising in the Charles H. Sandage Department of Advertising, College of Media, University of Illinois at Urbana–Champaign.
Patrick T. Vargas (PhD, Ohio State University) is a professor of advertising in the Charles H. Sandage Department of Advertising, College of Media, University of Illinois at Urbana–Champaign.
Hari Sundaram (PhD, Columbia University) is an associate professor of computer science and advertising in the Charles H. Sandage Department of Advertising, College of Media, University of Illinois at Urbana–Champaign.
Itai Himelboim (PhD, University of Minnesota) is the Thomas C. Dowden Professor of Media Analytics, director of the Social Media Engagement and Evaluation Suite, and an associate professor of advertising in the Grady College of Journalism and Mass Communication, University of Georgia.

© 2019 American Academy of Advertising

conversation from social media text for advertising. This is an important area of investigation because computational advertising is the primary revenue stream for companies like Google (Schomer 2019), and one of the core analytics methods used in computational advertising is computational text analysis of topics (Soriano, Au, and Banks 2013).

The focus of this article is to provide consideration to advertising researchers when computationally detecting topics of conversation from social media text. First, we discuss and compare several ways to computationally detect topics of conversation in social media text: text preprocessed summarization, phrase mining, topic modeling, supervised machine-learned text classification, and semantic topic tagging. Second, we present the results of comparing the various methods against a human-coded Twitter data set in which coders identified topics of conversation via a human content analysis approach. Finally, we provide a reference matrix that outlines a summary of each method, when to consider using that method, and tools to execute that method.[1]

Topic Discovery in Mass Communication Research, Consumer Research, and Advertising Research

There has been previous research regarding topic discovery in the areas of mass communication research, consumer research, and advertising research. Humphreys and Wang (2018) recently outlined research steps for incorporating automated text analysis into consumer research. In their breakdown of approaches, they discuss topic discovery but focus only on unsupervised learning (topic modeling). We broaden the possibilities of topic discovery within our article to include text summarization (counting) as well as supervised learning. In their overview of journalism-focused computational research Boumans and Trilling (2016) suggested that to see more computational research produced, researchers should build custom-made tools and share not only their results but also their code. Our focus digs more deeply into how to detect topics of conversation within social media text. In addition, we provide the links to our code for those researchers who have more advanced computational skills, and we went a step further and built the methods we describe into an open-source tool called the Social Media Macroscope (SMM), which is available to all researchers, so that those with less of a technical background can use these methods for their research.

Numerous advertising and marketing studies have incorporated automated topic detection in their research. Liu, Burns, and Hou (2017) applied latent Dirichlet application (LDA) topic modeling to tweets to understand what topics brands were discussing overall. They separated the brand-generated tweets into positive and negative, and applied LDA topic modeling to the positive tweets and negative tweets to understand if different topics were being discussed. Liu (2019) applied text preprocessing and LDA topic modeling in her efforts to discover a relationship between tweets and financial stock prices using machine learning. Vermeer et al. (2019) analyzed Facebook brand pages and used supervised learning to detect whether consumer posts fell into the following categories: rejection, complaint, comment, question, suggestion, acknowledgment, and/or compliment. Although these categories are a bit different than detecting the topic a consumer may be discussing, the process they used to categorize the posts (supervised learning) is the same process through which myriad topics could be detected. Okazaki et al. (2015) conducted a similar study using supervised learning to understand how consumers were discussing IKEA (a home furnishing brand) on Twitter and divided consumer conversations into the categories of sharing, information, opinion, question, reply, and exclude. Liu, Singh, and Srinivasan (2016) used principal component analysis (PCA) to summarize what topics were being discussed on Twitter by consumers of television programming. PCA is like LDA topic modeling in the sense that it provides groups of words that statistically belong in the same groupings as each other, but PCA ignores class labels as compared to LDA. We focus on LDA topic modeling in this article due to its popularity over its PCA counterpart, but interested readers should see Martínez and Kak (2001) for more information about the pros and cons of LDA versus PCA.

Recently, Malthouse and Li (2017) suggested that big data analyses in advertising research could uncover "exploratory insights to create better messages and motivate new theories … [and] can provide insights on what consumers think and feel about a brand" (p. 230). Our article directly builds from Malthouse and Li (2017), as we provide instructions on how researchers can conduct analyses of topics of conversation within social media text to understand what consumers think and feel about a brand. Taking this even further, topics of conversation can be analyzed to understand the profile of an audience more accurately, and this profile could be used to better create advertising messages (Kietzmann, Paschen, and Treen 2018).

Detecting Topics of Conversation in Social Media Text

Here we briefly discuss what constitutes a topic of conversation computationally. For the purposes of this article, we adhere to the definition of a conversational topic as "short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks such as classification, novelty detection, summarization, and similarity and relevance judgments" (Blei, Ng, and Jordan 2003). In addition, the realm of social media is vast, making "social media data" a very broad term. Therefore, for the purposes of this article, we limit our scope of social media data to short form/microblogged text, such as status updates (e.g., the text of Twitter, Facebook, Instagram, and Reddit posts).

Our list of methodologies is not meant to be exhaustive, but they should give advertising researchers who are newer to social analytics a satisfactory starting place for engaging in detecting topics. We provide a high-level summary of five methods of detecting topics: text preprocessed summarization, phrase mining, topic modeling, supervised machine-learned text classification, and semantic topic tagging. Some of these methods are subsets of another (e.g., semantic topic tagging is a subset of supervised machine-learned text classification), but we have pulled each of them out individually due to special considerations that we outline in each respective section.

Text Preprocessed Summarization

Text preprocessing can be defined as bringing "your text into a form that is predictable and analyzable for your task" (Ganesan 2019). It is almost always the first step when preparing text for advanced analytics techniques such as machine learning, but it can also be used on its own to understand what is being discussed within a corpus of text. Using text preprocessing to summarize text could be considered a naive approach to topic detection in social media text, but it is almost always a proper starting place to understand what is being talked about within a grouping of social media posts. Our previously stated definition of "topic" suggests that short summarized descriptions of a text to help understand the larger whole qualifies as topic detection. Text preprocessed summarization uses words that occur within social media posts as proxies for topics.

In their book on text mining, Miner et al. (2012) outline the general steps of preprocessing text: choose the scope of text to be processed, tokenize, remove stop words, stem, normalize spelling, detect sentence boundaries, and normalize casing. Depending on the context of what they are trying to extract from text, researchers can include/exclude some of these general steps according to their research objective. The steps that we followed to analyze the social media posts for this article are as follows:

1. *Collected social media posts.* Although a specific how-to on collecting and aggregating social media posts is beyond the scope of this article, researchers who have a limited background in this area could consider some existing tools that do not require extensive programming to collect social media data (Smith et al. 2009; Yun et al. 2019). We pulled the tweets for our analyses using the Tweepy Python package (Roesslein 2009), but some alternative considerations would be the twitteR R package (Gentry 2015) or using SMM. SMM leverages the Tweepy Python package.
2. *Tokenized the text from the social media posts.* Tokenization is the process of reducing text into smaller chunks (tokens). We suggest using white-space tokenization to separate social media posts into separate words. For example, a tweet may read "I love Fridays!" White-space tokenization after our previous preprocessing step would reduce the tweet to "i" and "love" and "fridays." Other more nuanced forms of tokenization exist (e.g., Hochreiter and Schmidhuber 1997; Trim 2013), but we focus on the most common technique for this article. We used Python's NLTK (Natural Language Toolkit) package (Loper and Bird 2002) for this step, as well as for steps 3 through 5.
3. *Normalized the text from social media posts.* Normalization involves tasks such as converting all text from uppercase to lowercase, removing numbers (unless numbers mean something within the context of the topical analyses), and removing punctuation.
4. *Removed stop words from social media posts.* Stop words are the most common words that appear in a text but are not actually important to understanding the topical content of what is being discussed (e.g., *the*, *with*, *to*, *a*, *and*).

5. *Stemmed all the remaining words within social media posts.* Stemming is the process of reducing words to their word stems. For example, the words *walking*, *walked*, and *walk* would all stem to the word *walk*.
6. *Sorted the remaining words in descending order of frequency of the words.* Sorting in this way will identify the most important words (topics) within the body of posts.

We built all of these steps into the SMM's natural language processing function within SMM's SMILE (Social Media Intelligence and Learning Environment) tool. SMILE's natural language processing function is a graphical interface presentation of our NLTK Python code, which was created for researchers who are not experts at coding.[2]

We ran this text preprocessed summarization on Fitbit's Twitter timeline snapshot taken in October 2017 using the SMM's SMILE tool (Yun et al. 2019). We chose Fitbit's Twitter timeline because Yun (2018) provided human-coded topic analyses of six brands' Twitter timelines (Fitbit, Wyndham Hotels, Royal Caribbean, American Heart Association, National Rifle Association, and World Wildlife Foundation), as well as collecting all of the brands' postings using the Twitter public application programming interface (API) from September 6 through 20, 2017. For the sake of consistency and ease of comparison across so many different methods, we chose to focus on only one of the brand's Twitter timelines: Fitbit. Within Yun's (2018) Fitbit data set, Fitbit's Twitter timeline data included 3,200 Fitbit tweets. Twitter's public API limits timeline downloads to the most recent 3,200 tweets. An example of the top topics for Fitbit's social media posts using text preprocessed summarization is found in Table 1.

Table 1. Top 20 most common words of Fitbit's Twitter posts.

Topic of Conversation	Number of Occurrences
Happystep	462
Step	418
fitbit	347
goal	221
fitbitfriend	209
hear	207
great	201
fit	193
job	169
awesom	160
congrat	159
make	148
tip	137
day	136
tracker	123
good	119
workout	119
share	115
work	112
love	108

Text preprocessed summarization is usually a great place to understand topics of conversation within social media posts because the computational complexity is low and the computing resources required are low. With this said, one of the most apparent limitations to text preprocessed summarization in discovering topics of conversation within social media is that phrases will sometimes be unnaturally separated by the tokenization process (e.g., "united" and "states" would be separated). Therefore, we turn to the next method of discovering topics within social media posts: phrase mining.

Phrase Mining

The primary limitation of text preprocessed summarization is that it relies on a bag-of-words assumption, which is simply that each individual word is handled as isolated from the other words, and multiword phrases (also referred to as *n*-grams) are separated. For example, the phrase "United States" would be separated out as "united" and "states" in our previous approach. One way to address this issue is to predefine the size of the tokens prior to tokenization, thus choosing to make all the tokens 2-grams (two-word tokens) or 3-grams (three-word tokens), for example, but this would clearly cause problems because many words are supposed to be single words. Another problem with the bag-of-words approach is that phrases that include stop words and punctuation may be incredibly important for understanding the overall conversation.

Brown et al. (1992) introduced a concept they called "sticky pairs," in which they used a statistical formula to calculate the probability of whether two words are most likely supposed to be paired together as a phrase or handled separately as two distinct words. In one analysis, they used a 59,537,595-word sample of text from the Canadian Parliament and found that the words "humpty" and "dumpty" occur together as "humpty dumpty" 6,000,000 times more frequently than one would expect from the individual frequencies of "humpty" and "dumpty." Although this was a major step forward, this type of statistical word comparison is quite computationally expensive, as it requires parsing through all 59,537,595 words and calculating the probabilities of each word occurring next to another word, and it is also limited by focusing solely on 2-grams. This also does not address phrases that include stop words and punctuation that are important for understanding the conversation within text.

More advanced statistical techniques in detecting phrases have since been published that handle varying lengths of *n*-grams (e.g., Deane 2005) and even include stop words (e.g., Parameswaran, Garcia-Molina, and Rajaraman 2010), but, as mentioned by Liu et al. (2015), many of these purely statistical techniques still suffer from other limitations to detecting quality phrases. For example, what if the word *just* was part of the larger phrase "just do it"? Many phrase-mining techniques would struggle with differentiating these types of situations. Recently, Shang et al. (2018) built on previous phrase-mining work focusing on the popularity, informativeness, and independence of phrases. Their goal was to create an automated phrase-mining method that was "domain-independent, with minimal human effort or reliance on linguistic analyzers" (Shang et al. 2018, p. 1825). The result of their work is a framework called AutoPhrase (with associated code found at https://github.com/shangjingbo1226/AutoPhrase). At the core of AutoPhrase is the use of existing databases of human-created quality phrases, such as Wikipedia, to refine their phrase detection. For example, they might detect two distinct phrases within a body of text: "Barack Obama" and "this is." Wikipedia has an entry for "Barack Obama" but not for "this is." Thus, the phrase "Barack Obama" would be weighted higher in their results as compared to "this is." AutoPhrase also incorporates parts of speech (POS) tagging (i.e., separating text into various parts of speech, such as nouns and verbs) to contribute to the weighting of results. An example of the top topics for Fitbit's social media posts using phrase mining is found in Table 2.

Phrase mining provides a more complete picture of topics being discussed within social media posts as compared to text preprocessed summarization, but many of the phrase-mining algorithms require more advanced levels of computer programming knowledge. We recently incorporated AutoPhrase into SMM, thus allowing for programming-free access to the framework.[3] As far as we know, phrase mining has not been applied to advertising studies or practice as of yet, but we believe this is a promising new technique to consider due to the probabilistic determination of phrases and the importance of understanding phrases, such as taglines, for advertising research.

Table 2. Top 20 AutoPhrase phrase-mining topics of Fitbit's Twitter posts.

Topic of Conversation	AutoPhrase Confidence Score
Red carpet	0.9731376
Enrollment program	0.96542927
Personal trainer	0.96466617
Peanut butter	0.96342927
[Unsupported characters][a]	0.96067927
Ultramarathon man	0.95777018
Case number	0.95597927
Stay tuned	0.9545134
Water resistant	0.9537634
Losing weight	0.95368565
Woody scal	0.95216303
Bay area	0.95116938
Corporate wellness	0.94877018
Slow cooker	0.9487499
Weight gain	0.94727732
Continuous heart rate	0.94268006
Weight loss	0.94174546
Wireless headphones	0.94141542
Alta hr	0.94097494
Yoga poses	0.94007653

[a] AutoPhrase parses English words only; thus, characters such as emojis and other nonsupported Unicode characters do not report correctly.

Phrase mining is an elegant and powerful way to analyze topics within social media posts, but one of its major limitations is that while it can collect phrases, it cannot assess larger topics of conversation and themes that are not directly written in the posts. For example, if the phrases "united states," "congress," and "supreme court" were found within social media posts, the higher-level topic that is being discussed is most likely "U.S. government." This is a scenario in which topic modeling is more appropriate to address.

Topic Modeling

Topic modeling is a method to summarize a large corpus of documents to provide a quick summary of what the documents are about (Blei, Ng, and Jordan 2003). The most common form of topic modeling is LDA topic modeling, which assumes that there are a set number of latent topics across a collection of documents, and a probabilistic equation is used to allocate words to each latent topic. This is one of the major differences between topic modeling versus text preprocessed summarization and phrase mining; topic modeling does not actually output topics but rather outputs words from the documents that are related to one another (e.g., instead of outputting a topic such as "Burger King," a topic modeling exercise would output a grouping of words such as "whopper, flame-broiled, nuggets"). Topic modeling simply groups words together that probabilistically belong together, and the researcher needs to review the word clusters to determine what the underlying topic is. This is essentially a coding exercise of determining topics, which is not a trivial task; and proper coding of topics should be based on theoretical knowledge and context-specific expertise (Humphreys and Wang 2018;

Table 3. LDA topic modeling topics of Fitbit's Twitter posts.

Topic of Conversation	Words in LDA Cluster
Marketing Fitbit	happystep, step, fitbitfriend, know, fitbit, thank, good, awesome, happi, come
Exercise	goal, great, happystep, step, goalday, crush, challeng, work, like, hear
Exercise encouragement	congrat, awesome, share, congratul, workout, love, fitbit, cook, free, butt
Healthy lifestyle	fit, tip, healthi, time, stay, health, help, weight, fitbit, sleep
Fitbit customer service	fitbit, hear, sorri, help, hope, thank, tracker, email, phhrmhlmt, check

Saldaña 2009). Some recent research using topic modeling in the context of marketing and business analytics are Liu, Burns, and Hou (2017) and Liu (2019).

Topic modeling assumes there are many documents within a corpus and that each document has many words. The problem with applying topic modeling to social media posts is the question of what to consider a document. If you decide to consider each individual post a document, then data within each post will likely be insufficient to adequately do what topic modeling was originally conceptualized for. More recent topic modeling methods created specifically for Twitter have attempted to work around this limitation through various strategies, such as considering each conversation string between Twitter users as separate documents (Alvarez-Melis and Saveski 2016), considering all the tweets that mention the same hashtag a document (Mehrotra et al. 2013), or considering all the tweets from the same author a document (Hong and Davison 2010).

An example of the word clusters and projected topics for Fitbit using LDA topic modeling is found in Table 3. For LDA topic modeling, the number of latent topics must be predefined before running the clustering algorithm, because LDA topic modeling cannot estimate how many latent topics are within a series of documents on its own. For the topics in Table 3, we ran LDA topic modeling multiple times with different numbers of topic clusters on Fitbit's social media posts, considering all tweets from the same author as a document (Hong and Davison 2010), and found the most intuitive results with the value of five total latent topic clusters. We conducted this topic modeling using the genism Python library (Řehůřek and Sojka 2011), and we also built our code into SMM's topic modeling function for researchers who would like to use that tool.[4]

One of the weaknesses of topic modeling is that much of the work is ultimately done by humans in determining what topics are indicated by various word clusters. This human-decision part of the topic modeling process is not typically documented and saved in a format that can be assessed or used by future researchers. This is where supervised machine-learned text classification can help.

Supervised Machine-Learned Text Classification

Supervised machine learning is "the search for algorithms that reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances" (Kotsiantis 2007, p. 249). One of the most common tasks for machine learning is the task of classification, which is the process of applying meaningful labels to data to help make sense of that data. Supervised machine learning for classification can be applied to text, especially when the goal is to classify (or categorize) text according to a predefined goal. At a basic level, the steps involved are (1) identification of required data and preprocessing the data (using techniques mentioned in the text preprocessed summarization section); (2) determining which part of the data will be used as the training set from which the computer learns patterns (each data point must be labeled or categorized by either a human coder or an automated labeling process); (3) selecting which machine-learning algorithm will be used for the computer to learn patterns within the labeled data training set; and finally (4) allowing the algorithm(s) to predict a test set of data that has also been labeled to evaluate how well the algorithm(s) can guess the prelabeled test set (Kotsiantis 2007).

This same process can be applied to social media posts in which humans label posts into topic categories based on theoretical knowledge and context-specific expertise (Saldaña 2009). Thus, supervised machine-learned text classification brings together the best of both worlds in leveraging human labeling of topics and pairing this with a computer's ability to find the statistical relationships between words in social media posts and topics. We do not provide an example of topics using supervised machine-learned text classification on the Fitbit social media posts in this section because this classification is dependent on an initial phase of human-labeled topic training.[5] For additional reading, some recent studies that have used supervised machine learning within the realm of advertising and marketing are Vermeer et al. (2019) and Okazaki et al. (2015).

Supervised machine-learned text classification is excellent for creating topic categories that can be trained as granularly as a researcher may want to train them, but the downside is that the process of hand-coding a set of social media posts large enough to adequately train a machine-learned model can be extremely time-consuming. Thus, we turn to our final method of detecting topics within social media posts: semantic topic tagging. This method leverages a combination of techniques borrowed from many of the previously mentioned topic detection methods.

Semantic Topic Tagging

Semantic topic tagging is "the extraction and disambiguation of entities and topics mentioned in or related to a given text" (Jovanovic et al. 2014, p. 39). At the heart of semantic topic tagging is the recognition that humans have already prelabeled large numbers of topics (called concepts within semantic tagging) in public Web repositories, such as Wikipedia. This is similar to phrase mining, which has leveraged Wikipedia entry titles to better rank the importance of existing phrases within social media posts (Liu et al. 2015), but semantic topic tagging extends the usage of Wikipedia-like repositories by also using data such as whole description words within Wikipedia entries themselves. The typical semantic topic tagging process for social media posts involves text preprocessing of the posts (e.g., "Try an Impossible Whopper!" to "try" "impossible" "whopper"; much like text preprocessed summarization), semantic similarity probabilistic calculations of the posts to determine *n*-grams and phrases using repositories like Wikipedia (much like phrase mining), and machine-learning techniques using the content of Wikipedia to help predict topics being discussed (as explained in supervised machine-learned text classification). Various frameworks exist for applying semantic topic tagging to text (for a comprehensive list, see Jovanovic et al. 2014), but two frameworks that have been proven to work well with microblog social media posts are TAGME (Ferragina and Scaiella 2010) and the work of Meij, Weerkamp, and De Rijke (2012). TAGME is currently presented as a pretrained topic tagger that can be accessed via API (Parameswaran, Garcia-Molina, and Rajaraman 2010).[6] Examples of the topics tagged for Fitbit using semantic topic tagging are provided in Table 4.

Table 4. Top 20 TagMe semantic topic tagging of Fitbit's Twitter posts.

Topic of Conversation	Number of Occurrences
Fitbit	562
Physical fitness	105
Physical exercise	90
E-mail	64
Health	57
"Woohoo" (Christina Aguilera song)	55
Motivation	43
Fun	43
U.S. dollar	39
Billboard 200	35
Steps (group)	34
"With You" (Chris Brown song)	34
The Who	33
Deutsche Mark	33
Calorie	32
Heart rate	31
Today (U.S. TV program)	31
Recipe	30
Thanks for sharing	26
For good	26

To the best of our knowledge, semantic topic tagging has not yet been incorporated into advertising and marketing studies, but Jovanovic et al. (2014) discussed how semantic topic tagging is currently being used in contextual advertising "to enable better positioning of advertisements on webpages based on the semantics of the main content of the page" (p. 39). Semantic topic tagging should almost always be preferred for detecting topics of conversation in social media posts because it leverages many strengths of other topic discovery methodologies, but there are still important weaknesses to take into consideration. One of the biggest weaknesses of semantic topic tagging is its heavy reliance on the public repositories that were used to train the topics. For example, with TAGME, the topic tagger was trained with data from a Wikipedia snapshot taken on November 6, 2009. Wikipedia has clearly evolved and changed from 2009 to current day, and many of the topics that are on Wikipedia now may not have existed in concept in 2009. For example, Kendall Jenner, a currently popular social media influencer, did not have a Wikipedia entry until 2010.

Computational Topic Methods versus Human-Coded Topics

Our next step was to compare each topic detection method against a human-coded baseline, as human-coded content analyses of social media data has been used successfully in previous studies (e.g., Chen et al. 2015; Kwon and Sung 2011; Lin and Peña 2011). For this study, we chose an available data set with topics already coded from a study that looked at Twitter timelines from brands and charities (Yun 2018). Yun (2018) used traditional content analysis techniques (e.g., Harwood and Garry 2003; Krippendorff 2013; Skalski, Neuendorf, and Cajigas 2017) to analyze Twitter timelines of the brands and charities of Fitbit, Wyndham Hotels, Royal Caribbean, American Heart Association, National Rifle Association, and World Wildlife Foundation. He recruited two full-time employees from a large midwestern university to conduct the content analyses.

The first content analysis phase was an effort to decide which topics should be included in the set of topics that they would look for when reviewing the Twitter timelines for the brands and the causes. Yun (2018) adopted a provisional coding approach (Saldaña 2009), which begins the coding process with a "start list" of potential codes to use as a base prior to coding. His start list was taken from survey responses where participants were asked to list what topics they believed each brand and cause would talk about. After reviewing the suggested topics across all the brands and causes, it was determined that there was consistency in responses for the following topics: business travel, charitable giving, climate change, cruises, customer service, diet, environmentalism, exercise, fashion, firearms, gun politics, gun violence, health, heart, hotels, hunting, marketing, medicine, nature, oceans, religion, sports, technology, vacations, weather, and wildlife. He then presented each coder with the Twitter timelines from the three brands (Fitbit, Royal Caribbean, and Wyndham Hotels) and the three charities (American Heart Association, National Rifle Association, and World Wildlife Foundation) separately as HTML web pages. Upon discussion, the coders suggested that 100 to 150 tweets were as much as they could process visually and cognitively when attempting to look for topics being discussed, so 150 tweets from each brand and each cause were chosen via a random selection script and pre-sented as separate HTML pages. The coders reviewed each Twitter timeline and indicated on a separate spreadsheet the topics they believed were being discussed from the previously constructed start list of topics. They also were given the opportunity to write in suggestions of other topics being discussed that were not covered by the provided start list of topics. Upon completion of this coding task, Yun (2018) ran a reliability analysis and found substantial agreement (κ =.70) between the two coders according to Viera and Garrett (2005). In addition, after discussion of potential topics that were not part of the original start list, the coders came to a joint conclusion that the start list was comprehensive enough without any necessary topics missing. However, they did indicate that two topics from the start list (business travel and religion) did not seem to be discussed across the three brands and the three causes. Table 5 contains an example of the topics identified by the results from the two human coders. Table 6 presents the results from each method (excluding the general supervised

Table 5. Human content analysis of Fitbit's Twitter posts.

Content of Posts
Charitable giving
Customer service
Diet
Exercise
Fashion
Health
Marketing
Technology

Table 6. Results comparison for one brand's (FitBit) topics as determined by each method.

Human Coded	Text Preprocessed Summarization	Phrase Mining	Topic Modeling	Semantic Topic Tagging
Charitable giving	happystep	red carpet	Marketing Fitbit	Fitbit
Customer service	step	enrollment program	Exercise	Physical fitness
Diet	fitbit	personal trainer	Exercise encouragement	Physical exercise
Exercise	goal	peanut butter	Healthy lifestyle	Email
Fashion	fitbitfriend	[Unsupported characters]	Fitbit customer service	Health
Health	hear	ultramarathon man		Woohoo (Christina Aguilera song)
Marketing	great	case number		Motivation
Technology	fit	stay tuned		Fun
	job	water resistant		United States dollar
	awesom	losing weight		Billboard 200
	congrat	woody scal		Steps (group)
	make	bay area		With You (Chris Brown song)
	tip	corporate wellness		The Who
	day	slow cooker		Deutsche Mark
	tracker	weight gain		Calorie
	good	continuous heart rate		Heart rate
	workout	weight loss		Today (U.S. TV program)
	share	wireless headphones		Recipe
	work	alta hr		Thanks for Sharing
	love	yoga poses		For Good

Table 7. Social media topic detection decision matrix.

Method	How Does This Method Work?	When Is This Method Appropriate?	How Can I Execute This Method?	Advantages	Disadvantages
Text preprocessed summarization	This method breaks down each social media post to stemmed words, removes stop words, and ranks by word frequency across all posts.	It is almost always a proper starting point to quickly view the text, but not appropriate to be used as the sole means for detecting topics.	Various Python and R packages/code are available (e.g., Benoit et al. 2018; Loper and Bird 2002); SMM incorporates this method without the need for coding.	It is very simple to run via available packages and for the most part does not take many computational resources.	Due to numerous limitations such as mishandling of phrases, this method alone cannot be trusted as a final determinant of topics.
Phrase mining	This method algorithmically discerns which words should remain together as phrases and ranks by phrase certainty across all posts.	It is appropriate when researchers believe that topics are accurately represented in social media posts as explicitly mentioned phrases.	AutoPhrase's code can be found via GitHub (Shang et al. 2018); also can be used via SMM.	It is a major upgrade to text preprocessed summarization that still allows for fully automated topic detection.	It is natively a bit complex to run via original codebase, although it is easier to run via SMM.
Topic modeling	This method algorithmically aggregates words from social media posts that are related to one another, depending on the researchers to then qualitatively define topics from the word clusters.	It is appropriate when topics are not explicitly mentioned in the posts, and computational help to aggregate topically related words is desired before qualitatively assigning topics by human coding.	Various Python and R packages/code are available (e.g., Mabey 2015; Řehůřek and Sojka 2011); also can be used via SMM.	It is a nice balance between human content analysis and automated help in reducing words to analyze when coding topics.	The quality of topics is dependent on the skill level of the human coder(s).
Supervised machine learned text classification	This method takes social media posts that have been human coded into topics and uses them to train a mathematical model in which a computer can then classify future posts.	It is appropriate when domain-specific topics have been predefined and coded by humans and researchers subsequently want to detect those predefined topics from social media posts.	The most recognized package for machine learning is Python's sklearn (Pedregosa et al. 2012); also can be used via SMM.	It allows for the highest level of fine-tuning of topics.	It is extremely time- and resource-consuming to train topics correctly.
Semantic topic tagging	This method uses repositories like Wikipedia as a large-scale, human-coded training set of topics to algorithmically label social media posts with topics.	It is appropriate when repositories like Wikipedia contain the appropriate topics and are updated enough to be considered a proper computational training set for discovering topics in social media posts.	TAGME is currently the most appropriate tool for short texts such as social media and has an API that is functioning (Jovanovic et al. 2014).	It draws on the vast crowdsourced knowledge base of Wikipedia, which allows for very up-to-date topic possibilities.	Wikipedia can be edited and updated by anyone; therefore, there is the possibility for erroneous categorization of topics.

machine-learned text classification) side by side for comparison.

As seen in Table 6, the only methods that included any overlap in topic phrases from the human coding were topic modeling and semantic topic tagging. Each method of topic detection has unique nuances that stem from how it determines topics from social media text. This emphasizes the importance for researchers and practitioners to understand how computational algorithms work, as using them with a misunderstanding of how they determine the output can dramatically affect research results. Thus, rather than a one-size-fits-all approach to analyzing social media text, researchers should consider their goals and the strengths and weaknesses of the various methods considering those goals.

When to Use Which Computational Methods

For quick reference, a brief comparison of method considerations is presented in Table 7. One of the first things researchers and practitioners should consider when analyzing topics of social media posts computationally is whether they believe the actual topics of conversation are explicitly mentioned in the posts themselves. If this is the case, then phrase mining is an appropriate choice. However, running a text preprocessed summarization analysis is a good first step before running a more computationally expensive phrase-mining exercise. Phrase mining on large bodies of social media posts can take substantial computational resources to run, whereas text preprocessed summarization can usually be achieved with minimal computational resources. Unfortunately, text preprocessed summarization suffers too greatly from the issue of segmented phrases to be useful as the only means of detecting topics of conversation within social media posts. However, when comparing text preprocessed summarization and phrase mining to the human-coded topics in Table 6, it is clear that the way actual people may conceive of topics is different from what may explicitly be mentioned in the social media posts.

Topic modeling forms somewhat of a hybrid between text preprocessed summarization and human coding of topics. As can be seen in Table 3, LDA topic modeling helps aggregate stemmed words that belong together, but it is up to the researcher or practitioner to determine the underlying latent topic. We assigned the LDA word clusters to the topics of marketing Fitbit, exercise, exercise encouragement, healthy lifestyle, and Fitbit customer service, but this is clearly a subjective assignment process. Therefore, having a sound qualitative coding process is important for the human-coding step of topic modeling. Our recommendation with topic modeling is that it is more useful as a preparatory step for supervised machine-learned text classification. Topic modeling helps provide a sense of what chunks of words could form certain topics that are emerging from the posts. After a proper coding scheme is developed, researchers can then code a larger number of social media posts for a hyperspecific supervised machine-learned topic classifier for the exact topics they are seeking. Semantic topic tagging then becomes just one example of a detection algorithm that is built on shared principles of supervised machine-learned text classification (in fact, some semantic topic taggers using supervised machine learning; e.g., Meij, Weerkamp, and De Rijke 2012).

Thus, it would benefit researchers and practitioners to consider running through most, if not all, computational topic detection methods to gain a greater understanding of what is being discussed within their aggregated social media posts. The purpose of the text analysis and the project's research questions should then inform the best approach to finally choose. A major consideration for researchers should be the realization that there is quite a bit of subjectivity with these computational methods, as there is a tendency to believe that computational methods are objective in nature. In addition, an important future research direction is the creation of tools and environments that can reduce the barrier of entry to these kinds of computational methods. Just as web-authoring technology went through an evolution of direct HTML-based coding to software like Dreamweaver to most recently websites like Squarespace and Wix, computational data science methods need to be made more accessible via efforts such as SMM to enable researchers to focus more on their research questions, as opposed to needing to focus on becoming an expert in any given algorithmic computational method.

Concluding Remarks

In this article, we discussed various ways of computationally detecting topics of conversation in social media text, namely, text preprocessed summarization, phrase mining, topic modeling, supervised machine-learned text classification, and semantic topic tagging. We also compared those methods against human coding of topics on social media brand posts and presented a matrix that researchers and practitioners can use to make decisions when they want to analyze social media posts for topics of conversation. While all methods are currently used for analyzing what is being discussed on social media, the results of each show that there are significant differences in what a researcher might conclude is being discussed based on the method used. Rather than leading to standardized or objective outcomes, we can see the role of subjectivity and the strengths and weaknesses of the various methods. Thus, researchers should start with their research question. We also referenced advertising and marketing papers that have used the various methods so that researchers can get a more in-depth illustration of each. Finally, we included the code used for each analysis reported in this article so that researchers can assess and use that code themselves. In addition, we built that code into an existing open-source, free environment for conducting social media data analysis (http://socialmediamacroscope.org) to allow computationally nonexpert advertising researchers to try the various methods out directly for their own research questions. Although this article was not meant to be an exhaustive list of topic detection methods and considerations, it should function as a strong starting point for advertising researchers and practitioners interested in detecting topics in social media posts.

Notes

1. While we provide details and code for those who would like to use it, we wanted to ensure that researchers who do not have backgrounds in computer science or coding would be able to follow along or conduct their own research without needing to gain significant additional technological expertise. Therefore, we reference a tool—Social Media Macroscope (SMM)—numerous times throughout this article. SMM is a science gateway that allows researchers without computer science backgrounds to execute open-source data science analytic methods without the need to code, and use of this gateway is free for academic and nonprofit use (Yun et al. 2019). The methods detailed in this article do require some knowledge of coding; however, SMM can be used as an alternative option for those who are not comfortable with coding. Therefore, we built our code into the SMM project, as well as providing links to our direct code throughout the article for researchers who want to apply the code apart from SMM. All code for SMM can be found at https://opensource.ncsa.illinois.edu/bitbucket/projects/SMM.
2. For researchers who would like to run the Python code themselves, all associated code for this method can be found at https://opensource.ncsa.illinois.edu/bitbucket/projects/SMM/repos/smm-analytics/browse/lambda/lambda_preprocessing_dev/preprocessing.py.
3. Researchers desiring to run our code/scripts themselves can access the Dockerized script that we used to run AutoPhrase at https://opensource.ncsa.illinois.edu/bitbucket/projects/SMM/repos/smm-analytics/browse/batch/smile_autophrase/dockerfile.
4. For researchers interested in the Python code, it can be found at https://opensource.ncsa.illinois.edu/bitbucket/projects/SMM/repos/smm-analytics/browse/batch/batch_topic_modeling/gensim_topic_modeling.py.
5. Machine-learned text classification is often conducted using Python's sklearn package (Pedregosa et al. 2012), but we have also built in text classification to SMM. Our code can be found at https://opensource.ncsa.illinois.edu/bitbucket/projects/SMM/repos/smm-analytics/browse/batch.
6. We used the TAGME API to conduct our semantic topic tagging. All documentation on how to call the TAGME API can be found at https://sobigdata.d4science.org/web/tagme/tagme-help.

ORCID

Joseph T. Yun http://orcid.org/0000-0001-6875-4456
Brittany R. L. Duff http://orcid.org/0000-0002-3206-0353

References

Alvarez-Melis, D., and M. Saveski (2016), "Topic Modeling in Twitter: Aggregating Tweets by Conversations," presented at the 10th International AAAI Conference on Web and Social Media, Cologne, Germany, May.

Benoit, K., K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller and A. Matsuo (2018), "quanteda: An R Package for the Quantitative Analysis of Textual Data," *Journal of Open Source Software*, 3 (30), 774.

Blei, D.M., A.Y. Ng, and M.I. Jordan (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3 (4–5), 993–1022.

Boumans, J.W., and D. Trilling (2016), "Taking Stock of the Toolkit: An Overview of Relevant Automated Content Analysis Approaches and Techniques for Digital Journalism Scholars," *Digital Journalism*, 4 (1), 8–23.

Brown, P.F., P.V. deSouza, R.L. Mercer, V.J. Della Pietra, and J.C. Lai (1992), "Class-Based *n*-Gram Models of Natural Language," *Computational Linguistics*, 18 (4), 467–79.

Chen, K.-J., J.-S. Lin, J.H. Choi, and J.M. Hahm (2015), "Would You Be My Friend? An Examination of Global Marketers' Brand Personification Strategies in Social Media," *Journal of Interactive Advertising*, 15 (2), 97–110.

Daniel, E.S., Jr., E.C. Crawford Jackson, and D.K. Westerman (2018), "The Influence of Social Media Influencers: Understanding Online Vaping Communities and Parasocial Interaction through the Lens of Taylor's Six-Segment Strategy Wheel," *Journal of Interactive Advertising*, 18 (2), 96–109.

Deane, P. (2005), "A Nonparametric Method for Extraction of Candidate Phrasal Terms," presented at the 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, Michigan, June.

Fan, W., and M.D. Gordon (2013), "The Power of Social Media Analytics," *Communications of the ACM*, 57 (6), 74–81.

Ferragina, P., and U. Scaiella (2010), "TAGME: On-the-Fly Annotation of Short Text Fragments (by Wikipedia Entities)," presented at the 19th ACM International Conference on Information and Knowledge Management, Toronto, Canada, Octobe.

Fulgoni, G.M. (2015), "How Brands Using Social Media Ignite Marketing and Drive Growth: Measurement of Paid Social Media Appears Solid, But Are the Metrics for Organic Social Overstated?," *Journal of Advertising Research*, 55 (3), 232–36.

Ganesan, K. (2019), "All You Need to Know about Text Preprocessing for NLP and Machine Learning," *KDnuggets*, April, https://www.kdnuggets.com/2019/04/text-preprocessing-nlp-machine-learning.html.

Gentry, J. (2015), "twitteR," https://www.rdocumentation.org/packages/twitteR/versions/1.1.9.

Harwood, T.G., and T. Garry (2003), "An Overview of Content Analysis," *The Marketing Review*, 3 (4), 479–98.

Hochreiter, S., and J. Schmidhuber (1997), "Long Short-Term Memory," *Neural Computation*, 9 (8), 1735–80.

Hong, L., and B. Davison (2010), "Empirical Study of Topic Modeling in Twitter," *Proceedings of the First Workshop on Social Media Analytics*, New York: ACM, 80–88.

Humphreys, A., and R.J.-H. Wang (2018), "Automated Text Analysis for Consumer Research," *Journal of Consumer Research*, 44 (6), 1274–1306.

Jovanovic, J., E. Bagheri, J. Cuzzola, D. Gasevic, Z. Jeremic, and R. Bashash (2014), "Automated Semantic Tagging of Textual Content," *IT Professional*, 16 (6), 38–46.

Kietzmann, J., J. Paschen, and E. Treen (2018), "Artificial Intelligence in Advertising: How Marketers Can Leverage Artificial Intelligence along the Consumer Journey," *Journal of Advertising Research*, 58 (3), 263–67.

Kotsiantis, S.B. (2007), "Supervised Machine Learning: A Review of Classification Techniques," *Informatica*, 31 (2007), 249–68.

Krippendorff, K. (2013), *Content Analysis: An Introduction to Its Methodology*, 3rd ed., Thousand Oaks, CA: Sage.

Kwon, E.S., and Y. Sung (2011), "Follow Me! Global Marketers' Twitter Use," *Journal of Interactive Advertising*, 12 (1), 4–16.

Kwon, K.H. (2019), "Public Referral, Viral Campaign, and Celebrity Participation: A Social Network Analysis of the Ice Bucket Challenge on YouTube," *Journal of Interactive Advertising*, 19 (2), 87–99.

Liu, J., J. Shang, C. Wang, X. Ren, and J. Han (2015), "Mining Quality Phrases from Massive Text Corpora," presented at the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Australia, June.

Lin, J.-S., and J. Peña (2011), "Are You Following Me? A Content Analysis of TV Networks' Brand Communication on Twitter," *Journal of Interactive Advertising*, 12 (1), 17–29.

Liu, X. (2019), "Analyzing the Impact of User-Generated Content on B2B Firms' Stock Performance: Big Data Analysis with Machine Learning Methods," *Industrial Marketing Management*, published electronically March 8,

———, A.C. Burns, and Y. Hou (2017), "An Investigation of Brand-Related User-Generated Content on Twitter," *Journal of Advertising*, 46 (2), 236–47.

———, P.V. Singh, and K. Srinivasan (2016), "A Structured Analysis of Unstructured Big Data by Leveraging Cloud Computing," *Marketing Science*, 35 (3), 363–88.

Loper, E., and S. Bird (2002), "NLTK: The Natural Language Toolkit," in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teach Natural Language Processing and Computational Linguistics*, Stroudsburg, PA: Association for Computational Linguistics, 63–70.

Mabey, B. (2015), "pyLDAvis." https://github.com/bmabey/pyLDAvis.

Malthouse, E.C., and H. Li (2017), "Opportunities for and Pitfalls of Using Big Data in Advertising Research," *Journal of Advertising*, 46 (2), 227–35.

Martínez, A.M., and A.C. Kak (2001), "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23 (2), 228–33.

Mehrotra, R, S. Sanner, W. Buntine, and L. Xie (2013), "Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York: ACM, 889–92.

Meij, E., W. Weerkamp, and M. De Rijke (2012), "Adding Semantics to Microblog Posts," presented at the Fifth ACM International Conference on Web Search and Data Mining, Seattle, Washington, February.

Miner, G., J. Elder, IV, A. Fast, T. Hill, R. Nisbet, and D. Delen (2012), *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications* Waltham, MA: Academic Press.

Moe, W.W., O. Netzer, and D.A. Schweidel (2017), "Social Media Analytics," in *Handbook of Marketing Decision Models*, 2nd ed., B. Wierenga and R. Van der Lans, eds., Cham, Switzerland: Springer, 483–504.

Murdough, C. (2009), "Social Media Measurement," *Journal of Interactive Advertising*, 10 (1), 94–99.

Okazaki, S., A.M. Díaz-Martín, M. Rozano, and H.D. Menéndez-Benito (2015), "Using Twitter to Engage with Customers: A Data Mining Approach," *Internet Research*, 25 (3), 416–34.

Parameswaran, A., H. Garcia-Molina, and A. Rajaraman (2010), "Towards the Web of Concepts: Extracting Concepts from Large Datasets," *Proceedings of the VLDB Endowment*, 3 (1–2), 566–77.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, and M. Brucher, M. Perrot, and É. Duchesnay (2012), "Scikit-Learn: Machine Learning in Python," *Journal of Machine Learning Research*, 12, 2825–30.

Rathore, A.K., A.K. Kar, and P.V. Ilavarasan (2017), "Social Media Analytics: Literature Review and Directions for Future Research," *Decision Analysis*, 14, 229–49.

Řehůřek, R., and P. Sojka (2011), "Gensim—Statistical Semantics in Python," statistical semantics; gensim; Python; LDA; SVD.

Roesslein, J. (2009), "Tweepy," https://tweepy.readthedocs.io/en/latest/.

Rosenkrans, G., and K. Myers (2018), "Optimizing Location-Based Mobile Advertising Using Predictive Analytics," *Journal of Interactive Advertising*, 18 (1), 43–54.

Saldaña, J. (2009), *The Coding Manual for Qualitative Researchers*, London: Sage.

Schomer, A. (2019), "Google Ad Revenue Growth Is Slowing As Amazon Continues Eating into Its Share," *Business Insider*, May 1, https://www.businessinsider.com/google-ad-revenue-growth-slows-amazon-taking-share-2019-5.

Shang, J., J. Liu, M. Jiang, X. Ren, C.R. Voss, and J. Han (2018), "Automated Phrase Mining from Massive Text Corpora," *IEEE Transactions on Knowledge and Data Engineering*, 30 (10), 1825–37.

Skalski, P.D., K.A. Neuendorf, and J.A. Cajigas (2017), "Content Analysis in the Interactive Media Age," in *The Content Analysis Guidebook*, K.A. Neuendorf, ed., Thousand Oaks: Sage Publications, 2 ed., 201–42.

Smith, A, and M. Anderson (2018), "Social Media Use in 2018. Pew Research Center," https://www.pewresearch.org/internet/2018/03/01/social-media-use-in-2018/.

Smith, M.A., B. Shneiderman, N. Milic-Frayling, E. Mendes Rodrigues, V. Barash, C. Dunne, T. Capone, A. Perer, and E. Gleave (2009), "Analyzing (Social Media) Networks with NodeXL," in *Proceedings of the Fourth International Conference on Communities and Technologies*, New York: ACM, 255–64.

Soriano, J., T. Au, and D. Banks (2013), "Text Mining in Computational Advertising," *Statistical Analysis and Data Mining*, 6 (4), 273–85.

Trim, C. (2013), "The Art of Tokenization," *IBM Community*, January 23, https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en.

Vermeer, S.A.M., T. Araujo, S.F. Bernritter, and G. van Noort (2019), "Seeing the Wood for the Trees: How Machine Learning Can Help Firms in Identifying Relevant Electronic Word-of-Mouth in Social Media," *International Journal of Research in Marketing*, 36 (3), 492–508.

Viera, A.J., and J.M. Garrett (2005), "Understanding Interobserver Agreement: The Kappa Statistic," *Family Medicine*, 37 (5), 360–63.

Yun, J.T. (2018), "Analyzing the Boundaries of Balance Theory in Evaluating Cause-Related Marketing Compatibility," doctoral dissertation, University of Illinois at Urbana–Champaign, https://www.ideals.illinois.edu/handle/2142/101522.

———, and B.R.L. Duff (2017), "Consumers As Data Profiles: What Companies See in the Social You," in *Social Media: A Reference Handbook*, K.S. Burns, ed., Denver, CO: ABC-CLIO, 155–61.

———, N. Vance, C. Wang, L. Marini, J. Troy, C. Donelson, C.L. Chin, and M.D. Henderson (2019), "The Social Media Macroscope: A Science Gateway for Research Using Social Media Data," *Future Generation Computer Systems*, published electronically, doi:10.1016/j.future.2019.10.029.