SOCIAL MEDIA ANALYTICS

Wendy W. Moe, University of Maryland
Oded Netzer, Columbia University
David A. Schweidel, Emory University

Introduction

One of the most significant developments in recent years involves the proliferation of user-generated content, particularly online social media. Social media has created a power shift in the relationship between consumers and brands, providing consumers more power by allowing them to easily broadcast their views and opinions about brands to a large audience. At the same time social media has opened a window for firms into the voice of the consumer. Previously, marketers had to employ costly and time-intensive marketing research methods such as interviews, focus groups and surveys to better understand how consumers perceive their brands. Now, consumers voluntarily turn to social media and share their opinions publically for other customers as well as for the brand managers to see.

This new medium provides wealth of data from which marketing researchers can extract customer insights. However, in order to analyze and leverage social media data, we must first understand the behavior that generates the data. Thus, the first part of this chapter will discuss **online opinion behavior**, the process by which user generated content (UGC) is created, and its implications for deriving insights from such data. We then discuss **social media as a source for marketing research** and describe some of the models that have been developed for social media data mining. There are several challenges involved in converting the vast volumes of social media data to useful managerial insights. Key amongst these challenges is the fact that most social media data are unstructured and textual in nature. With a firm understanding of the consumer and the appropriate methodologies, marketers can then begin to use social media to understand and influence their customers. This leads us to a second, but equally important, function of social media, which we discuss in the third part of the chapter – **social media as a communications channel**.

Understanding the Behavior of Social Media Content Generators

Researchers have studied offline consumer word-of-mouth behavior for decades.[1] Westbrook (1987) identified three motivations that drive consumer to spread word-of-mouth: product involvement, self-involvement and altruism. Anderson (1998) found a relationship between satisfaction with the product and the likelihood of engaging in word-of-mouth behavior, where highly dissatisfied customers were more likely to share their opinions.

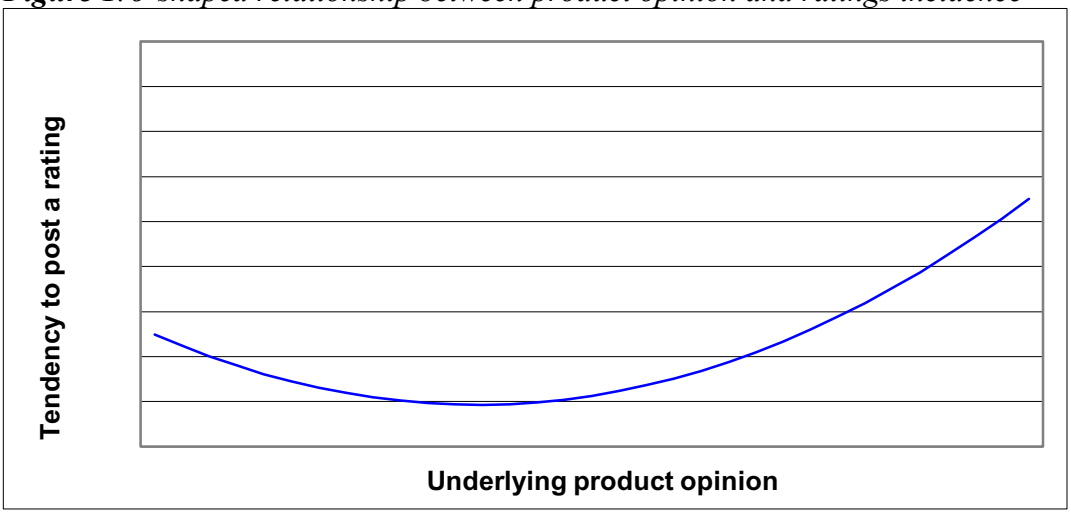
In the context of online social media environment, Hennig-Thurau et al. (2004) proposed a taxonomy of online word-of-mouth motivations. They propose that consumers are motivated to

[1] For a review of research on word-of-mouth, we refer interested readers to Berger (2014).

participate in online word-of-mouth for a variety of reasons include users seeking assistance, expressing negative feelings, helping others, self-enhancement, social benefits, economic incentives, aiding the firm and seeking advice. Berger and Milkman (2012) find that social transmission is affected in part by the arousal of content. Toubia and Stephen (2013) examined the motivations underlying consumers' posting on Twitter and differentiate between posters who derive intrinsic utility from posting in social media and those who derive image-related utility. Lovett, Peres and Shachar (2013) combine multiple sources of data to compare the underlying motivation and characteristics of online and offline word-of-mouth. They find that social and functional drivers are more prominent in online word-of-mouth, whereas emotional drivers were most important in offline word-of-mouth. Furthermore, brand differentiation plays an important role in online word-of-mouth but less so in offline word-of-mouth.

Whereas Anderson (1998) and others have found that offline word-of-mouth was predominantly negative as dissatisfied customers were more likely to engage in word-of-mouth activities, online word-of-mouth tends to be predominantly positive. This positivity bias has been documented across multiple studies (Chevalier and Mayzlin 2006; Godes and Mayzlin 2004; Resnik and Zeckhauser 2002). One of the most robust findings in product reviews form of word-of-mouth, is that of a J-shaped relationship between frequency of posts and satisfaction with the product (Moe and Schweidel 2012). Figure 1 shows the J-shape curve proposed by Moe and Schweidel (2012). The J-shape relationship suggests that while those with negative opinions are more likely to share an opinion than those with a moderate opinion (as is the case in offline word-of-mouth), those with positive opinions are even more likely to share online. While the positivity bias of online product ratings has been well documented, no explanation has yet been provided for this noticeable difference between online and offline word-of-mouth behavior.

Figure 1. *J-shaped relationship between product opinion and ratings incidence*



Beyond the positivity bias in product ratings, researchers have shown that online product ratings and reviews have a distinct downward trend over time. A number of theories have been provided for this trend including product life cycle effects, where later adopters are less satisfied with their purchase decision based on the evaluations of innovators who may hold very different preferences (Li and Hitt 2008), and preference matching effects, where consumers have more

difficulty sifting through the posted ratings as the number of ratings available increases (Godes and Silva 2012).

Moe and Schweidel (2012) propose an online social dynamics account to explain the downward trend in online opinion. In the context of online product ratings, they differentiate between (1) the consumer's decision whether to post a rating and (2) the consumer's decision of what rating to post and examine how the opinions of others already posted to the same forum affects each of these two decisions. They found that individuals fall into one of two categories: activists and low-involvement consumers. The activists are frequent posters who are more likely to provide critical evaluations and make efforts to differentiate their posted opinions from those of others. The low-involvement consumers, in contrast, post less frequently and hold more positive opinions. However, the low-involvement consumers are unique in that they are less likely to post word-of-mouth when the opinions posted previously are highly varied. This is in stark contrast to the activists who thrive in these environments and are actually more likely to post when there is disagreement in the forum. These two contrasting behaviors of the activists and the low-involvement posters leads to a very interesting dynamic. As the number of opinions posted in the forum increases and disagreement among the posters emerge, the low-involvement consumers begin to withdraw from the conversation and refrain from sharing their opinions. The activists, on the other hand, relish these dissentious environments and freely contribute their critical opinions. Over time, the minority activists dominate opinion environments and their opinions are disproportionately represented. Because activists posted opinions tend to be more negative, this leads to a downward trend in posted opinions over time.

What does this mean for social media data and analytics?

These two empirical findings of J-shape relationship between product reviews and satisfaction with the product and the downward trend in ratings and reviews over time have important implication for social media analytics. The explanations provided to both of these findings suggest that there may be self-selection in the consumers' decision to participate in the online conversation. That is, the downward trend in reviews over time is not necessarily a reflection of any real decline in consumers' opinions of product quality. This is an important consideration for brand managers and social media analysts who may be considering product modifications or other changes to their overall marketing mix, as online comments may not be representative of the perceptions held by the broader customer base. Similarly, the J-shaped distribution of opinions posted may not be representative of the underlying customer population. Differences may be due to the differences between the overall customer base and those who post on social media or it may be a result of a systematic shift in the opinion posted due to the social dynamics described above. Either way, in analyzing user generated content, the researcher needs to consider such biases. That being said, because user generated content affects purchase decision of a much larger population than those who generated it (Ludwig et al. 2013), these, possibly self-selected, opinions may affect actual purchases of a much larger and more representative population.

Offline researchers have also identified an *audience effect* that can affect the opinions one share with others. For example, Fleming et al. (1990) showed that an individual's evaluation changed depending on who he/she believed would be the audience of the evaluation. In the online environment, Schlosser (2005) showed that product evaluations changed depending on whether

or not the individual providing the opinion believed it would be made public. The existing research is clear in that audience effects can alter individual posting behavior. In aggregate, this leads to a bias in the social media data that is dependent on the nature of the audience that a social media user expects to face. For example, a user who is posting to a community of close friends and family (e.g., on Facebook) is likely to express him or herself differently than a user who is posting to a professional network of colleagues (e.g., on LinkedIn). Thus, because various social media venues tend to attract different audiences, we should expect systematic differences across various social media venues.

Schweidel and Moe (2014) directly examine the differences in social media posting behavior across various social media venues such as, forums, blog and micro blogs. They found that opinions shared on different venues systematically vary due to the unique audience that each venue attracts and the social dynamics that each venue encourages. For example, discussion forums allow for the most social interaction. As a result, opinions posted to these forums tend to be more negative and experienced the most severe downward trend over time (consistent with the findings of Moe and Schweidel 2012). Blogs facilitated the least social interactions while allowing the poster to provide more depth in their comments. As a result, opinions expressed in blogs were more moderate and did not experience any downward trends over time.

The discussion above suggests that aggregation of social media data over time and/or venue may mask or even bias the results of the analysis performed on that data. Furthermore, as consumer self-select themselves to participate in the discussion the data available for analytics and modeling may not be necessarily representative of underlying opinions of the entire customer base. In analyzing social media data we encourage researchers to develop models that explicitly acknowledge the individual level behaviors and the factors that influence them. In the next section, we review a number of models with applications for marketing research, with particular focus on models that de-bias social media data for social dynamics and venue effects for the purposes of brand tracking (e.g., Schweidel and Moe 2014) and methods to analyze the text that dominates so much of social media data (e.g., Netzer et al. 2012).

Using UGC for Marketing Research

From numerical ratings to textual information

One of the main difficulties in utilizing consumer-generated content for quantitative analysis is that the data are primarily qualitative in nature. In discussing future directions for social interaction research, Godes at al. (2005) note that one of the difficulties in tapping into UGC is the ability to analyze the content. Similarly, Liu (2006) analyzed messages posted on Yahoo Movies message board and reported “an extremely tedious task” in mechanically analyzing over 12,000 movie review messages using human reviewers.

Because of the difficulties associated with analyzing text, many researchers have resorted to characteristics of the consumer-generated data to provide quantitative summaries of the content, such as product ratings to represent the content of the consumer’s opinion. Three common measures of product ratings have been primarily used in the literature: volume, valence and variance. These measures have been used by numerous researchers when investigating the relationship between ratings and sales (e.g., Chevalier and Mayzlin 2006; Dellarocas, Zhang and

Awad 2007; Moe and Trusov 2011) as well as the dynamics in the social media discussion (e.g., Godes and Mayzlin 2004; Moe and Schweidel 2012).

In a study investigating how online conversations affect television ratings, Godes and Mayzlin (2004) operationalized variance (or dispersion) using a measure of entropy to reflect the extent to which conversations occur in different online newsgroups or discussion forums:

$$Entropy_{it} = \begin{cases} -\sum_{n=1}^{N} \frac{POST_{it}^{n}}{POST_{it}} \log\left(\frac{POST_{it}^{n}}{POST_{it}}\right) & \text{if } POST_{it} \geq 0 \\ 0 & \text{if } POST_{it} = 0 \end{cases}$$

where $POST_{it}^{n}$ is the total number of posts on newsgroup n about program i between the airing times of episodes t and t+1. $POST_{it}$ is the total number of posts about program i between the airing of episodes t and t+1 airing, aggregated across all newsgroups. Entropy is maximized when the number of posts is equally distributed across the N newsgroups. As Godes and Mayzlin (2004) note, in contrast to variance, entropy does not depend on the total volume of posts as it instead relies on the share of posts appearing in each newsgroup. The authors find that that increased entropy is associated higher television ratings, which may reflect reaching a broader audience.

Dellarocas and Narayan (2006) provide a volumetric measure of product reviews that combines social media data with sales data. They propose a measure of *density,* operationalized as the ratio of the number of people posting online comments during a fixed period of time to the number of people purchasing the product.

In many contexts, users directly provide a measure of valence in the form of an ordinal rating, such as 1-5 star ratings in Amazon.com product reviews. However, many social media environments, such as blogs, forums and social networks, do not include quantitative summary of the consumers' evaluations. Instead, posts to these venues comprise solely of text. In these contexts, valence measures are often constructed using sentiment analysis (e.g., Pang and Lee 2008) where words associated with positive, neutral or negative emotions are extracted to capture the overall sentiment expressed in the message.

Richer content analyses beyond just valence measures have also been employed. However, as the field of text mining is still relatively new, the degree of expertise needed to develop and employ a solid sentiment analysis tool is quite high, and the accuracy of such automated processes is still evolving. Accordingly, many researchers still employ manual coding. For example, Schweidel and Moe (2014) have used a commercial firm to manually code the sentiment and product attribute mentions of over 7,500 messages across three social media venues. Nevertheless, as social media data is often voluminous, it is clear that a more automatic and multifaceted view of the online text can help generate meaningful insights.

Text mining

Text mining (sometimes called knowledge discovery in text) refers to the process of extracting useful information from unstructured text (Fellbaum 1998; Feldman et al. 1998; Feldman and Sanger 2006). For example, Swanson and colleagues found relationships between magnesium and migraine (Swanson 1988) and between biological viruses and weapons (Swanson and Smalheiser 2001) by text mining disjoint literatures and uncovering words common to both literature bases.

Text mining has become particularly popular and successful in fields in which meaningful information must be extracted from "mountains" of data in a relatively short period of time. Such fields include security and intelligence organizations looking for signs of irregular activity in the stream of public and less public written media (Fan et al. 2006) or doctors searching for biomedical information in the superabundance of medical information (Rzhetsky et al. 2004). Academics have used text mining in order to automatically meta-analyze the knowledge base on a particular topic (Börner, Chen and Boyack 2003). With the increasing availability of voluminous digitized data sources, the business world started to take notice of the opportunities offered by text-mining tools to automatically analyze the infinite stream of financial report data, to open a window to consumer online discussions and to collect competitive intelligence information. Many companies are offering text-mining services to businesses to help them with these tasks.

Computer scientists and information system researchers have made the greatest leap in developing advanced text-mining apparatuses. Collaborations between computer scientists or information systems researchers and business researchers helped facilitate the dissemination (albeit limited) of these tools to business research (e.g., Das and Chen 2007; Feldman et al. 2008; Ghose et al. 2012; Lee and Bradlow 2011; Netzer et al. 2012). In marketing, the first attempts to text-mine UGC used manual text-mining involving humans reading the messages and judging their content (e.g., Godes and Mayzlin 2004, Liu 2006). This inefficient and inaccurate methodology was described by the authors as a "tedious task" and a "costly and noisy process." Computer scientists such as Dave, Lawrence and Pennock (2003), Hu and Liu (2004), Liu, Hu, and Cheng (2005) and Feldman et al. (2007) offered a solution to the tedium by building apparatuses that could automatically summarize and quantify consumer reviews. These advances have facilitated a fast and continuing diffusion of text mining applications to research in marketing (e.g., Decker and Trusov 2010; Ghose et al 2012; Lee and Bradlow 2011; Netzer el al. 2012) we further describe these and other applications of text mining in marketing later on in this chapter.

Text mining approaches

It is beyond the scope of this chapter to provide a detailed and exhaustive description of different text-mining tools and the methodology involved. We refer the interested reader to books that specialize in text mining (e.g., Feldman and Sanger 2006). Instead, our objective is to describe, at a high level, the most commonly used types of text-mining analyses in marketing and some of the considerations one should be aware of in applying such tools in marketing contexts.

At the most basic level text mining has been used in marketing to extract individual entities such as brands, product attributes, emotions and adjectives used to describe products. Numerous commercial companies are offering buzz monitoring services, tracking how frequently a brand is

being mentioned across alternative social media. Similarly, academic researchers have looked at how often brands are mentioned in social media venues, which emotions are being mentioned (e.g., Berger and Milkman 2012; Ludwig et al. 2013), or which attributes are being mentioned in a review related to a particular product (e.g., Lee and Bradlow 2011). Netzer et al. (2012) note that the task of accurately identifying brands (e.g., Audi or Volvo) is easier than identifying product models (e.g., Audi A4 or Volvo S6). In the context of cars they report F1 accuracy levels[2] of 98.1% for car brands and 91.6% for car models. Extracting more difficult entities such as adverse drug reactions (ADRs) led to lower F1 accuracy levels of 81.6%, all within acceptable ranges in the text mining literature. Dictionaries such as WordNet (Fellbaum 1998) and the linguistic inquiry and word count (LIWC; Pennebaker, Francis and Booth 2001) have been used as easy tools to conduct such basic text mining analysis. However, as we discuss later for most text mining application more advanced tools such as natural language processing (NLP) are needed.

A slightly more advanced set of tools is needed if one is interested in capturing the sentiment of a particular textual unit such as a product review (See Pang and Lee 2008 for a review of methods). Most of the sentiment analysis tools (sometimes called opinion mining) rely on NLP, statistics tools, or machine learning.[3] Often a combination of approaches is used together with a sentiment dictionary (e.g., Sentiwordnet - sentiwordnet.isti.cnr.it or Sentistrength - sentistrength.wlv.ac.uk) to obtain more accurate sentiment analysis. For example, Ghose et al. (2012) used part-of-speech tagging combined with crowd sourced Amazon mechanical Turks scoring of adjectives to derive their sentiment tool. Netzer et al. (2012) used a machine-learning approach, combined with a sentiment dictionary and human coded rules to identify common problems reported in various car models. Das and Chen (2007) used a statistical approach involving classifiers to capture sentiment for stocks from message boards. Tirunillai and Tellis (2012) used a combination of statistical classifier and a support vector machine approach to capture the valence of UGC about products. The accuracy level of sentiment analysis methods is still limited and the sentiment tool often needs to be tailored to the specific domain of analysis. For example, the word "high" would be considered as positive sentiment in the context of stock prices but negative sentiment in the context of blood pressure. Thus, marketing researchers are advised to use domain-specific tools or tools that can be adapted to a specific domain. Furthermore, it is recommended to manually examine the level of accuracy of the tool using human coders on a sample of textual units.

At the next level of complexity of analysis lies the process of relation extraction. Relation extraction refers to the process of identifying textual relationships among extracted entities. For example, in the context of pharmaceutical drugs, Netzer et al. (2012) and Feldman et al. (2015) identified the textual relationships between drugs and ADRs that imply that drug X causes ADR Y. Such textual relationship extraction often requires linguistic analysis, usually involving NLP, to allow the text-mining algorithm to understand the textual context of the sentence. Netzer et al. (2012) reported F1 accuracy of 73.6% in identifying the relationship between drugs and ADRs.

[2] F1 is measured as the harmonic mean of the levels of recall and precision, where recall is the proportion of instances that were identified, and precision is the proportion of correctly identified instances of the set of identified instances.

[3] For more information about NLP, we refer interested readers to Manning and Schütze (1999).

Applications of such relation extractions are still few in marketing, primarily due to the text mining complexity involved in accurately making such relational inferences from unstructured data. However, we believe this area is one of the most promising directions for future work. Marketing researchers are often more interested in extracting the relationships between products, attributes and the sentiments or context of the relationship between them, than simply measuring the volume of mentions of their brands or even the overall sentiment about their brands.

Once the relationships between entities have been extracted one needs to summarize the co-occurrences of entities in the textual corpora. One may be tempted to simply report co-occurrences of how often any pair of entities appears together in the text. For example, how many times did the brand BMW appear with the term "sporty?" The problem with reporting simple co-occurrence is that if an entity appears very frequently in the text it will also co-occur with more entities than an entity that occurs less frequently. Thus, one should normalize the measure of co-occurrence for how often each entity appears in the text independently, to capture how often two entities occurred in the text over and beyond chance. Various such "normalization" approaches have been proposed.

One of the most commonly used measures in the text-mining literature is the term frequency–inverse document frequency (tf-idf) weighting. tf-idf is used to weigh the occurrence of each term by its role in the document. The term frequency for term *j* in document *m* is defined by $tf_{jm} = X_{jm} / N_m$, where X_{jm} is the number of times term *j* appeared in document *m*, and N_m is the number of terms in document *m*. $idf_j = \log(D / M_j)$, where D is the total number of documents and M_j is the total number of documents where term *j* appeared. tf-idf is given by $tf - idf_{jm} = tf_{jm} \times idf_j$. The term frequency term captures how prominent a particular term is in the document. For example, if the word "sporty" appeared three times in a car review that include 10 words the prominence of sporty is higher than a similar but much longer review that includes three mentions of the word sporty in a review that includes 100 words. However, if all reviews include the words sporty three times, the appearance of three mentions of sporty in a particular review is less informative. Accordingly, the inverse document frequency "normalizes" the term frequency measure to how likely we are to see the term in a typical document. Thus, multiplying term frequency by inverse document frequency gives us an overall measure of prominence of a term in a particular document, after controlling for the likelihood of the term to appear in the entire textual corpora.

Another classic measure of normalized co-occurrence commonly appearing in the co-word analysis literature as well as the market basket literature is the measure of lift. Lift is the ratio of the actual co-occurrence of two terms to the frequency with which we would expect to see them together. The lift between terms A and B can be calculated as

$$Lift(A,B) = \frac{P(A,B)}{P(A) \times P(B)},$$

where *P(A)* is the probability of occurrence of term *A* in a given textual unit, and *P(A,B)* is the probability that both *A* and *B* appear in a given textual unit. A lift ratio of less than (more than) 1 suggests that the two terms appear together less than (more than) one would expect by the mere

occurrence of each of the two terms in the text separately. Other frequently used measures of co-occurrence are the Salton's cosine similarity and the Jaccard index (e.g., Toubia and Netzer 2016).

The process of text mining often involves five steps. In the context of identifying products and product attributes and the textual relationships among them the process can be defined as:

1. *Downloading*: The textual information is downloaded (often in an html format). This process can be either done manually using a pre-specified set of URLs or using a Web scraper that searches the Web for instances of a particular topic or product.
2. *Cleaning*: html tags and non-textual information such as images and commercials are cleaned from the downloaded files.
3. *Information Extraction*. Entities such as products and product attributes are extracted from the messages. The researcher may use *n*-gram approach to extract entities that include more than one word. In the *n*-gram approach the text-mining algorithm extracts all possible sequences of up to *n* words found in the text. Additionally, the researcher may wish to use stemming algorithms to reduce and combine words into their word stem or root (e.g., using stemming, the words, "run," "ran," "running," would all be captured by the entity "run").
4. *Chunking*: The textual parts are divided into informative units such as threads, messages, and sentences.
5. *Semantic relationships*. The linguistic algorithm identifies the co-occurrence of entities in the same textual unit (e.g., two cars that are mentioned together in the same forum message). At a deeper level of textual relationship, the researcher may want to not only identify that the two entities (e.g. a drug and an ADR) were mentioned together, but also what is the nature of the textual relationship between the two entities (e.g., the drug *causes* the ADR).[4]

When text-mining UGC the researcher is often faced with a problem of high dimensionality. That is, the text-mining process often results in thousands of possible unique words that appeared in the mined corpora. This dimensionality problems leads to both statistical and interpretation difficulties. From a statistical point of view if one wishes to use the terms extracted as independent variables or predictors to predict some market outcome, estimating a model with hundreds or thousands of predictors is difficult. From an interpretation point of view, deriving meaningful insights from such large space of variables is challenging. Stemming the words to their stem or root helps to reduce the dimensionality of the entity space as multiple words are combined to their common stem. However, the number of derived stems is still often highly unwieldy. The simplest approach to reduce the dimensionality of the problem is to trim the list of entities to only entities that appeared at least a certain number of times. However, determining the threshold from which to remove entities is often ad-hoc and this process may still leave the researcher with a large number of entities.

[4] If the researcher is interested in sentiment analysis or other types of output rather than relationship extraction, Step 5 could be replaced with the eventual goal of the text-mining task. For example, if the researcher is interested in understanding which topics were mention in a review, step 5 may be replaced with a topic modeling approach.

A more statistically driven approach to reduce the dimensionality and derive insights from textual information is to use the latent Dirichlet allocation (LDA), often called topic modeling (Blei, Ng and Jordan 2003). The idea behind LDA is that each document (e.g., product review) contains a mixture of topics. The distribution of topics is assumed to have a Dirichlet prior. Each topic is then (probabilistically) associated with a set of words. The advantage of the LDA approach is that one can automatically infer which topics were most likely to be mentioned in each document. The set of topics could be either learned in a fully unsupervised manner (i.e., purely informed from the data) or in a supervised or semi-supervised manner (i.e., informed fully or partially by the researcher). Tirunillai and Tellis (2014) have demonstrated the potential of using LDA in marketing by using unsupervised LDA to capture the latent dimensions underlying product quality in product reviews. For example, for mobile phones the authors report that the top six dimensions of quality (topics) in order of importance are "ease of use," "secondary features," "performance," "visually appealing," "reliability" and "customer service." For footwear, on the other hand, the most important dimension was "physical support" and the second most important "visually appealing." Some of the limitations of the standard LDA approach is that it does not consider the order in which words appear in a document or sentence structure. More generalized approaches have been developed that take into account sentence structure (e.g., Buschken and Allenby 2015) and the identity of the author (e.g., Rosen-Zvi et al. 2004).

Software packages such as the *tm* package in *R* (Feinerer and Hornik 2015) and *NLTK* in *Python* (Bird, Loper and Klein 2009) have made text mining more accessible. That being said, in employing text mining techniques, researchers should exercise care to ensure that the tools is appropriate for the problem at hand and the underlying data generating mechanism, and that the tools is properly executed for the idiosyncrasies of the specific problem it is used for.

Applications of Text Mining in Marketing

One can, broadly speaking, divide the applications of text mining UGC data in marketing into two main groups based on the goals of the analysis: 1) using UGC to *describe* and monitor markets, and 2) using UGC to *predict* relevant market outcomes.

1. Text mining UGC to describe markets – by text mining UGC companies can listen to and monitor consumer discussions and opinions about their own products as well as the competition. Furthermore, because the UGC data stream keeps updating in real time, one can monitor the changes in consumer perceptions over time. In that sense one can think of text mining UGC data as leveraging Web 2.0 as a marketing research playground or as an almost infinite size, and re-occurring, focus group. Accordingly, researchers have explored the type of insights that can be generated from mining UGC data as a descriptive listening tool. For example, Schweidel and Moe (2014) analyzed approximately 7,500 user-posted text across multiple social media venues. The authors measured sentiments both at the overall brand level and at product specific and attribute specific levels. They define a single measure of brand health as well as sentiment measures of specific aspects of their product.

   Because UGC provides firms with a window into the discussion about their own and their competitive products, one can use UGC to assess the competitive market structure. For

example, Netzer et al. (2012) looked at the co-occurrence between pairs of cars in the same message in nearly 900,000 messages from the sedan cars forum Edmunds.com to create a competitive market structure map of the car industry, simultaneously analyzing nearly 170 different car models. Leveraging that longitudinal nature of the data the authors show that, following a marketing campaign, Cadillac re-positioned itself away from the group of American brands and towards the luxury import brands. Similarly, Lee and Bradlow (2011) used product reviews to extract the attributes and attribute levels that were mentioned with each brand of digital camera to create market structure maps based on the similarity in attribute mentions across digital camera brands.

Tirunillai and Tellis (2014) used LDA to describe the dimensions of product quality mentioned in product reviews across five product categories. By tracking the quality dimensions over time the authors explore the competitive brand positions on the quality dimensions. For example, in the context of computers, the authors found that Dell's perception of "ease of use" were highly volatile over time, whereas those of Hewlett Packard were relatively steady. Similarly, Zhang, Kim and Xing (2015), used dynamic topic modeling to monitor the evolution of the competitive landscape.

These studies highlight the value text-mining tools provide in converting a largely qualitative source of data such as UGC, to a quantitative data and useful descriptive and perspective information and insights for business decision making.

2. Text mining UGC to predict market outcomes – in addition to leveraging text mining of UGC data to listen to consumers' discussions, researchers have also explored relating the text-mined information to market outcomes such as consumer choices, aggregate sales or stock prices.

   Several studies have shown that textual analysis of UGC can predict stock prices. For example, Tirunillai and Tellis (2012) have linked the volume and valence (as measured by sentiment analysis) to the firm stock performance. The authors find positive relationship between the volume of the chatter about the brand and the brand's stock return. The authors report that negative UGC can lead to negative stock returns but positive UGC had little effect on the stock prices. Yu, Duan and Cao (2013) used sentiment analysis across multiple social media sources such as blogs, forums and Twitter to demonstrate relationship between social media sentiment and stock prices. Similarly, Bollen, Mao and Zeng (2011), showed that the inferred mood from Twitter posts can predict overall stock market performance.

   Another important outcome measure is firm sales. Archak, Ghose and Ipeirotis (2011) demonstrate that the textual information in product reviews can help extract consumer preferences for product attributes, which in turn adds predictive power in predicting sales over and beyond the reviews' numerical ratings. Similarly, Trusov and Decker (2010), mine product reviews to estimate consumer preferences and predict the overall evaluations of products. Ludwig et al. (2013) find that positive affective content in book reviews affects conversion rates for books in Amazon.com. At the individual level, Ghose et al. (2012) used a text mining analysis of both product reviews and hotel

descriptions, together with image recognition, and crowd sourcing to predict consumers' choices among hotels and design a ranking algorithm for hotels.

Mining UGC has been also shown to predict other business relevant outcomes such as diffusion of information and success of ideas and movies. Berger and Milkman (2012) used a combination of automated sentiment analysis, the Linguist Inquiry of Word Count (LIWC) dictionary, and manual coding to assess the drivers and predict the sharing of New York Times articles. Toubia and Netzer (2016) show that automatically mining the text of individual ideas generated by consumers, can help flagging promising ideas and recommend, in real time, words for consumer to improve their ideas. Text mining non-UGC type of data, Eliashberg, Hui, and Zhang (2007), demonstrate that text mining movie scripts using NLP and statistical learning tools can help predict the success of the movies.

Looking at social welfare outcomes, Feldman et al. (2015), have demonstrated that by mining UGC medical forums one could predict drug label changes as early as 10 years prior to the label change. Similarly, Culotta (2010) showed that one can detect influenza epidemics by mining Twitter messages. Twitter messages have been also shown to predict overall political opinion. O'Connor et al. (2010), found correlations of as high as 80% between the political sentiment of Twitter messages and the political public opinions measured in surveys.

Taken together, we see that the rich textual information available in UGC data can help predict consumer preferences and evaluations, aggregate firm sales, stock prices, major events such as drug label change and success of ideas and information goods.

Moving forward, we expect that advances in text mining tools will allow researchers and marketers to go beyond capturing volume, valence, or particular words that can describe markets and predict outcomes, and towards understanding the textual relationships and deeper information content expressed in UGC. Combined with content analysis applied to other aspects of UGC, such as hyperlinks (e.g., Liu 2007) and tags (e.g., Nam and Kannan 2014), such analyses can aid in, for example, detecting problems with products based on consumer discussions in consumer forums or understanding the comparative language consumers use to describe competitive products.

Social Media as a New Marketing Channel

In addition to being a new source of marketing data that can be used to generate consumer insights, social media provide an important channel through which consumers can communicate with one another and with organizations, and organizations can communicate back with consumers. Indeed several researchers have demonstrated that consumer discussions in social media as a standalone channel, have an effect on sales (e.g., Chevalier and Mayzlin 2006; Godes and Mayzlin 2009; Moe and Trusov 2011). Furthermore, researchers have also shown how social media can be leveraged by firms as part of a broader marketing strategy (e.g., Stephen and Galak 2012; Srinivasan, Rutz and Pauwels 2015).

Social Media's Effect on Sales

Social media has a created a major shift in power between consumers and firms. It provided to consumers a vehicle with a wide reach to easily and publicly express their liking and disliking for products. Additionally, consumers frequently consult social media and UGC venues such as forums, blogs and product reviews before making a purchase. Accordingly it is likely that the content of social media would causally affect sales. While we have mentioned previously that several studies have demonstrated the ability of UGC data (e.g., product review ratings) to predict sales (e.g., Godes and Mayzlin 2004; Moe and Trusov 2011) arguing for a causal relationship is much more difficult. Vector autoregressive (VAR) models (e.g., Stephen and Galak 2012; Srinivasan et al. 2015) econometric approaches (e.g., Anderson and Magruder 2012; Mayzlin, Dover and Chevalier 2014), and field experiments (Godes and Mayzlin 2009) were proposed to isolate and identify the impact of social media activity on performance.[5]

The seminal study in establishing the relationship between product ratings and product sales was conducted by Chevalier and Mayzlin (2006). In their research, they examined how star ratings, provided by customers, impacted book sales at both Amazon.com and Barnesandnoble.com. Chevalier and Mayzlin used the fact that the inherent quality of the books on both websites is likely to be the same, thus differences in sales variation over time between Amazon.com and Barnesandnoble.com for any single book can likely be attributed to changes in the online ratings that are unique to each site. This study was the first to establish the role that user-provided product ratings have on product sales. Interestingly, their study also found that the impact of 1-star reviews exceeds that of 5-star reviews, suggesting that negative user-generated content may be more impactful than positive user-generated content.

Anderson and Magruder (2012), used a clever approach of regression discontinuity approach to establish causal relationship between Yelp ratings and restaurant reservations. The authors found that increase in star rating causes increase in restaurant reservations, which results in a higher likelihood of the restaurant being fully booked.

Moe and Trusov (2011) further investigated the effects of product ratings on product sales by considering both direct and indirect effects. Specifically, they measure the effects that previous ratings have on the arrival of subsequent ratings as well as on product sales. In other words, posted ratings can directly affect product sales; they can also affect subsequent ratings thereby indirectly affecting future product sales. Methodologically, the authors develop an exponential hazard model for the arrival of each ratings level for a given product and model the effects of lagged measures of valence (the average ratings for the product), variance (the variance in ratings for the product), and volume (the total number of ratings for the product).

[5] For a review of the impact of online WOM on sales, we refer readers to the meta analyses conducted by Babic et al. (2015) and You, Gautham and Joshi (2015).

The authors deconstruct the impact of product ratings on sales into components attributable to baseline effects, social dynamics and idiosyncratic error. This decomposition is found to provide superior model fit compared to benchmarks that rely directly on summary measures of social media (i.e., the volume, valence and variance).

The impact of UGC on sales is likely to differ based on the volume, valence and variance of the UGC, as well as across product categories and the review platform. For, example, Sun (2012) combines a theoretical model and empirical evidence to show that high variance of product reviews is better for products that were rated low, as the high variance signals a niche product.

The findings presented thus far suggest that user-generated content can positively affect key metrics such as sales. Thus, from a managerial perspective, it would be very tempting to try and manipulate the ratings environment to improve sales. Dellarocas (2006) investigated the outcome of such strategic manipulation by the firm and proposed a game theoretic model in which profit-maximizing firms seek to manipulate ratings by contributing fake anonymous product ratings. Dellarocas (2006) describes an outcome in which firms expend resources to artificially inflate their perceived quality through online ratings, but such behavior is expected by consumers. In this game theoretic outcome, consumers will assume that manipulation of online forums occurs and thus discount the quality signal they obtain from online forums. Given this eventual outcome, firms will thus choose to conduct a minimum level of manipulation since there is no long-term benefit of such behavior. Mayzlin, Dover and Chevalier (2014) argue and demonstrate that independent single property hotels are more likely to participate in generating deceptive hotel reviews than branded chain hotels.

This is not to say that all firm-generated word-of-mouth is intended to deceive customers. Godes and Mayzlin (2009) conduct a field test to understand the impact of firm-generated word-of-mouth on sales. The authors worked with a marketing agency focused on creating word-of-mouth communications for its clients by providing consumers with small incentives to spread word-of-mouth about its clients. For each incident of word-of-mouth that is spread for the client, the consumers spreading word-of-mouth is asked report their relationship to the recipient. Godes and Mayzlin (2009) showed that exogenously generated word-of-mouth positively affects week-to-week sales. The authors' results also suggest that firms may be better suited to recruit less loyal customers to spread word-of-mouth. Doing so may enable the firm to reach consumers who are unaware of the firm's offerings and whose opinions may be more malleable compared to those consumers reached by more loyal customers.

Social Media as Part of the Broader Marketing Strategy

While the above discussion establishes the relationship between social media word-of-mouth and firm performance, social media is not a channel that operates in isolation of other elements of the marketing mix. In addition to user-generated social media, marketers also manage earned media and paid advertising activities.

Stephen and Galak (2012) investigate the effects of traditional earned media (e.g., media mentions in traditional outlets of newspapers, magazines, television and radio) relative to social earned media (e.g., blogs, forum posts, and new members registering for the forum). Using a zero-inflated autoregressive double Poisson model for the marginal distributions, the authors employ a multivariate normal copula to correlate the marginal distributions for sales and media variables. They find evidence to suggest that both traditional and social earned media positively impact sales. Their results also indicate that, in addition to the direct impact of earned social media on sales, Social earned media may indirectly affect sales by increasing the amount of traditional earned media.

Gopinath, Thomas and Krishnamurthi (2014) go beyond looking at the volume of online word of mouth to investigate how the content of word of mouth, along with advertising, impacts sales. In addition to considering the volume of online activity, the authors account for the valence specific to comments that focus on recommendations, attribute features and emotional attachment. They find that it is the topic-specific valence measures and advertising that directly impact sales, while advertising also exhibits an indirect effect by impacting the topic-specific valence measures and the volume of online word of mouth.

Srinivasan et al. (2015) extend the work of Stephen and Galak (2012) by considering the impact of marketing mix activity on consumers' online activity and, ultimately, sales. Using a vector autoregressive (VAR) model, the authors examine the impact of price and distribution channels, television advertising, paid search clicks, website visits, and Facebook activity on sales. While the elasticities of price and distribution on sales have the highest magnitudes, consumers' online activities including paid search, website traffic and Facebook likes have larger elasticities compared to television advertising. Importantly, the authors illustrate that these activities are inter-related. For example, paid search is affected by television advertising and Facebook likes, while it impacts distribution, site visits, Facebook likes and sales.

Taken together, this stream of research highlights the inter-connected nature of UGC in organizations' marketing efforts. In evaluating the impact of marketing efforts on performance measures, we must be cognizant to account for both the direct impact of marketing efforts on performance, as well as the indirect effect of marketing efforts through the production of UGC. Viewed in this light, UGC may provide an early indication of marketing effectiveness.

Conclusion

User-generated content (UGC) has been investigated from different perspectives by marketing scholars. Some have investigated the process by which it is produced and diffused to better understand the phenomenon. Others, looking at it as the digital manifestation of the voice of the consumer, have employed it as a rich and economical means of conducting marketing research, from tracking brand health to inferring brand associations and the competitive landscape. We have also seen it interpreted and employed as a new tool for marketing to consumers. As these streams of research have largely evolved independently of each other, one of the goals of this chapter is to provide an integrated perspective on the evolution and future of UGC in marketing.

New methodology, like deep learning, is being developed in related disciplines and gradually being adopted by marketing researchers in academia and practice. As these methodologies evolve, coupled with techniques developed by marketing academics, our ability to characterize social media and incorporate it into subsequent analysis improves. For example, while familiar metrics like valence and volume still have a place in marketing models, our understanding of consumers is much richer thanks to the content of the UGC and topics identified through text analytic methods.

As we discussed earlier, several papers have investigated the integrity of UGC data and the risk of deceptive UGC (Anderson and Simester 2014; Dellarocas 2006; Mayzlin et al. 2014). However, possibly due to the difficulty in credibly identifying deceptive reviews, we do not have a firm assessment of the degree of this phenomenon and its effect on reliably using UGC as a marketing research or a predictive tool. We encourage future researchers to tackle this important problem.

We encourage future research to investigate targeting individual consumers based on UGC and its propagation. For example, can social media posts reveal consumers' interests? What messages resonate most with them? Who is influential within their social networks? Such analyses may prove beneficial for marketers, as marketing messages can be delivered to those who will be most responsive to them. In doing so, researchers may examine the extent to which insights gleaned from social media analytics complement what can be learned about consumers using other data sources. For example, to what extent does the incorporation of social media activity affect perceptions based on analyzing transactional activity? Beyond informing the likelihood of future transactions (e.g., Schweidel et al. 2014), merging social media data with CRM data could provide information regarding the likely categories in which customers will be most likely to purchase in the future.

From the perspective of brand management, UGC can provide insights into how the brand is perceived in the market place. But, these perceptions are likely to vary considerably across consumers and markets. Future work may build on the text analytic work (e.g., Netzer et al. 2012; Tirunillai and Tellis 2014) to examine other forms of unstructured data in which brands appear and understand brand associations using such data. In particular, research may consider the development of predictive models for the trajectory of conversations.

As research into social media continues to develop, academic researchers should be mindful of the potential application of their research. Popular social media monitoring platforms often lack advanced analytic tools, in part because of the magnitude of the data and the corresponding computation resources needed to apply the most advanced analytic tools. Given industry's keen interest in mining social media to understand the voice of the consumer, we encourage researchers to engage in research that has the potential to be deployed at scale, without sacrificing the rigor of their work.

REFERENCES

Anderson, E. W. (1998). Customer satisfaction and word of mouth. *Journal of Service Research*, 1(1), 5-17.

Anderson, E. T., & Simester, D. I. (2014). Reviews without a purchase: Low ratings, loyal customers, and deception. *Journal of Marketing Research*, 51(3), 249-269.

Anderson, M., & Magruder, J. (2012). Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal*, 122(563), 957-989.

Archak, N., Ghose, A., & Ipeirotis. P.G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science*, 57(8), 1485-1509.

Babic, A., Sotgiu, F., de Valck, K., & Bijmolt, T. H. (2015), The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors. *Journal of Marketing Research*, forthcoming.

Berger, J. (2014). Word of mouth and interpersonal communication: A review and directions for future research. *Journal of Consumer Psychology*, 24(4), 586-607.

Berger, J, & Milkman K. L. (2012). What makes online content viral? *Journal of marketing research* 49 (2), 192-205.

Bird, S., Loper E., & Klein E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993-1022.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1-8.

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 37(1), 179-255.

Büschken, J. & Allenby, G.M. (2015). Sentence-Based Text Analysis for Customer Reviews. working paper.

Chevalier, J.A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3), 345-354.

Culotta, A. (2010) Towards detecting influenza epidemics by analyzing Twitter messages. *Proceedings of the first workshop on social media analytics. ACM*, 115-122.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9), 1375-1388.

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *In Proceedings of the 12th international conference on World Wide Web* (pp. 519-528). ACM.

Decker, R., & Trusov, M. (2010). Estimating aggregate consumer preferences from online product reviews. *International Journal of Research in Marketing*, 27(4), 293-307.

Dellarocas, C. (2006). Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management Science*, 52(10), 1577-1593.

Dellarocas, C., & Narayan, R. (2006). A statistical measure of a population's propensity to engage in post-purchase online word-of-mouth. *Statistical Science*, 21(2), 277-285.

Dellarocas, C., Zhang, X. M., & Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing*, 21(4), 23-45.

Eliashberg, J., Hui, S. K., & Zhang, Z. J. (2007). From story line to box office: A new approach for green-lighting movie scripts. *Management Science*, 53(6), 881-893.

Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. Communications of the ACM, 49(9), 76-82.

Feinerer, I. & Hornik, K. (2015). tm: Text Mining Package. R package version 0.6-2. http://CRAN.R-project.org/package=tm.

Fellbaum, C. (1998). WordNet. Blackwell Publishing Ltd.

Feldman, R., & Sanger, J. (2006). Information extraction. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, 94-130.

Feldman R., Fresko M., Goldenberg J, Netzer O., & Ungar L. (2007). Extracting product comparisons from discussion boards. *Proc. Seventh IEEE Internat. Conf. Data Mining 2007* (IEEE, Piscataway, NJ), 469-474.

Feldman R., Fresko M., Kinar Y., Lindell Y., Liphstat O., Rajman M., Schler Y., & Zamir O. (1998). Text mining at the term level. *In Principles of Data Mining Knowledge Discovery* (pp. 65-73). Springer Berlin Heidelberg.

Feldman R., Fresko M., Goldenberg J., Netzer O., & Ungar L. (2008). Using text mining to analyze user forums. *Proc. Service Systems Service Management 2008 International Conference* (IEEE Systems, Man, and Cybernetics Society, Melbourne, VIC, Australia), 1–5.

Feldman, R., Netzer, O., Peretz, A., & Rosenfeld, B. (2015). Utilizing text mining on online medical forums to predict label change due to adverse drug reactions. *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1779-1788.

Fleming, J. H., Darley, J. M., Hilton, J. L., & Kojetin, B. A. (1990). Multiple audience problem: a strategic communication perspective on social perception. *Journal of Personality and Social Psychology*, 58(4), 593.

Ghose, A., Ipeirotis, P.G. & Li, B. (2012). Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science*, 31(3), 493-520.

Godes, D., & Mayzlin, D. (2009). Firm-created word-of-mouth communication: Evidence from a field test. *Marketing Science*, 28(4), 721-739.

Godes D., & Mayzlin, D. (2004). Using online conversations to study word-of-mouth communication. *Marketing Science* 23(4), 545-560.

Godes, D., & Silva, J. C. (2012). Sequential and temporal dynamics of online opinion. *Marketing Science*, 31(3), 448-473.

Godes, D., Mayzlin, D., Chen, Y., Das, S., Dellarocas, C., Pfeiffer, B., Libai, B.m Sen, S., Shi, M. & Verlegh, P. (2005). The firm's management of social interactions. *Marketing Letters*, 16(3-4), 415-428.

Gopinath, S., Thomas, J. S., & Krishnamurthi, L. (2014). Investigating the relationship between the content of online word of mouth, advertising, and brand performance. *Marketing Science*, 33(2), 241-258.

Hennig-Thurau, T., Gwinner, K. P., Walsh, G. & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet? *Journal of Interactive Marketing*, 18 (1), 38-52.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177). ACM.

Lee T., & Bradlow E. T. (2011), Automated marketing research using online customer reviews. *Journal of Marketing Research*, 48(5), 881–894.

Li, X., & Hitt, L. M. (2008). Self-selection and information role of online product reviews. *Information Systems Research*, 19(4), 456-474.

Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, 70(3), 74-89.

Liu, B. (2007), *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media.

Liu, B., Hu, M., & Cheng, J., (2005). Opinion observer: Analyzing and comparing opinions on the Web. *Proc. 14th International Conference World Wide Web* (Association for Computer Machinery, Chiba, Japan), 342–351.

Lovett, M. J., Peres, R., & Shachar, R. (2013). On brands and word of mouth. *Journal of Marketing Research*, 50(4), 427-444.

Ludwig, S., De Ruyter, K., Friedman, M., Brüggen, E. C., Wetzels, M., & Pfann, G. (2013). More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. *Journal of Marketing*, 77(1), 87-103.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. 999). Cambridge: MIT press.

Mayzlin, D., Dover, Y., & Chevalier, J. (2014). Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review*, 104(8), 2421-2455.

Moe, W. W., & Schweidel, D. A. (2012). Online product opinions: Incidence, evaluation, and evolution. *Marketing Science*, 31(3), 372-386.

Moe, W. W., & Trusov, M. (2011). The value of social dynamics in online product ratings forums. *Journal of Marketing Research*, 48(3), 444-456.

Nam, H., & Kannan, P. K. (2014). The informational value of social tagging networks. *Journal of Marketing*, 78(4), 21-40.

Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, 31(3), 521-543.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *ICWSM*, 11(122-129), 1-2.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2), 1-135.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count: LIWC 2001*. Mahway: Lawrence Erlbaum Associates, 71, 2001.

Resnik, P., & Zeckhauser, R. (2002). Trust among strangers in internet transactions: Empirical analysis of ebayps reputation system. *The Economics of the Internet and E'Commerce. Advances in Applied Microeconomics*, 11, 127.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. *In Proceedings of the 20th conference on Uncertainty in artificial intelligence*, July 7, 487-494, AUAI Press.

Rzhetsky, A., Iossifov, I., Koike, T., Krauthammer, M., Kra, P., Morris, M., Yu, H., Duboué, P.A., Weng, W., Wilbur, W.J., & Hatzivassiloglou, V. (2004). GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of biomedical informatics*, 37(1), pp.43-53.

Schlosser, A. E. (2005). Posting versus lurking: Communicating in a multiple audience context. *Journal of Consumer Research*, 32(2), 260-265.

Schweidel, D. A., & Moe, W. W. (2014). Listening in on social media: A joint model of sentiment and venue format choice. *Journal of Marketing Research*, 51(4), 387-402.

Schweidel, D. A., Park Y-H., & Jamal, Z. (2014), A multi-activity latent attrition model for customer base analysis," *Marketing Science*, 33 (2), 273-286.

Stephen, A. T., & Galak, J. (2012). The effects of traditional and social earned media on sales: A study of a microlending marketplace. *Journal of Marketing Research*, 49(5), 624-639.

Srinivasan, S., Rutz, O. J., & Pauwels, K. (2015). Paths to and off purchase: quantifying the impact of traditional marketing and online consumer activity. *Journal of the Academy of Marketing Science*, 1-14.

Sun, M. (2012). How does the variance of product ratings matter? *Management Science*, *58*(4),.696-707.

Swanson, D. R. (1988). Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, 31(4), 526-557.

Swanson D. R., & Smalheiser, N. R. (2001). Information discovery from complementary literatures: Categorizing viruses as potential weapons. *Journal of American Society for Information Science and Technology*, 52(10), 797–812.

Tirunillai, S., & Tellis, G. J. (2012). Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Science*, 31(2), 198-215.

Tirunillai, S., and Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. *Journal of Marketing Research* 51(4) 463-479.

Toubia, O., & Netzer, O. (2016). Idea generation, creativity and prototypicality. *Marketing Science*, forthcoming.

Toubia, O., & Stephen, A. T. (2013). Intrinsic vs. image-related utility in social media: Why do people contribute content to twitter?. *Marketing Science*, 32(3), 368-392.

Decker, R., & Trusov, M. (2010). Estimating aggregate consumer preferences from online product reviews. *International Journal of Research in Marketing*, 27(4), 293-307.

Westbrook, R. A. (1987). Product/consumption-based affective responses and postpurchase processes. *Journal of marketing research*, 258-270.

Yu, Y., Duan, W., & Cao, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems*, 55(4), 919-926.

You, Y., Gautham G. V., & Joshi, A. M. (2015). A meta-analysis of electronic word-of-mouth elasticity. *Journal of Marketing*, 79(2), 19-39.

Zhang, H., Kim, G., & Xing, E. P. (2015). Dynamic Topic modeling for monitoring market competition from online text and image data. *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1425-1434). ACM.