Industrial Marketing Management xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Industrial Marketing Management

journal homepage: www.elsevier.com/locate/indmarman

Research paper

Analyzing the impact of user-generated content on B2B Firms' stock performance: Big data analysis with machine learning methods

Xia Liu

Department of Marketing and BIS, William G. Rohrer College of Business, Rowan University, 201 Mullica Hill Rd, Glassboro, NJ 08028, USA

ARTICLE INFO

Keywords:
Big data
User-generated content
LDA
Sentiment analysis
Machine learning

ABSTRACT

Marketing scholars are interested in the big data of user-generated content (UGC) from social media platforms. However, the majority of current UGC studies have been conducted in the business-to-consumer (B2C) context. To fill the knowledge gap in business-to-business (B2B) research, we investigate whether UGC has differential impacts on stock performance for B2B and B2C firms by using big data. We collect a large dataset of 84 million tweets from 20.3 million Twitter accounts and 8 years of stock data for 407 companies from the S&P500 index. The results from machine learning methods are transformed into a monthly panel data. We conduct fixed effects model on the panel data. We find that UGC has a significant impact on firms' stock performance and that its impact on stock performance is much stronger among B2C firms than among B2B firms. While consumers' positive sentiment does not play a significant role in stock performance, consumers' negative sentiment and WOM significantly impact stock prices.

1. Introduction

Unstructured big data is an important byproduct of the digital age. Online reviews, blogs, tweets and Facebook posts are laden with consumer behavioral insights and bring unprecedented potential opportunities to companies as well as academic researchers (George, Haas, & Pentland, 2014). Marketing scholars are interested in the big data of user-generated content (UGC) from social media platforms. Researchers apply new methods of analysis to big data sets to gain a deeper understanding of customers' purchase experiences and brands' reputations. However, the majority of current UGC studies have been conducted in the business-to-consumer (B2C) context. To fill the knowledge gap in business-to-business (B2B) research, marketing scholars are urged to take advantage of increasingly available social media data and newly emerging methodologies from the fields of artificial intelligence and data analytics (P. Chintagunta, Hanssens, & Hauser, 2016; Lilien, 2016; Tirunillai & Tellis, 2014).

This paper investigates the application of big textual data with machine learning methods in B2B settings. To be more specific, we examine two research questions. First, does UGC have a significant impact on firms' stock performance? Second, does UGC have differential impacts on stock performance for B2B and B2C firms? To answer the above questions, we study the relationship between stock performance and social media UGC of Standard and Poor (S&P) 500 companies, whose stock data and Twitter UGC activities are publicly accessible (Heinonen, 2011; Minnick & Noga, 2017).

Major difficulties exist in collecting, preprocessing, merging and analyzing S&P 500 stock and UGC data. The authors used Java programming language and wrote customized code to collect a large dataset of 84 million tweets from 20.3 million Twitter accounts and 8 years of stock data for 407 companies from the S&P500 index. To handle the large volume of data, the ability to process data with scalability and accuracy is critical (Trilling & Jonkman, 2018). Compared to traditional numeric data, tweet messages are considered to be "unstructured" because texts are not encoded in an easily machine-interpretable format. To benefit from the unique characteristics of tweets, natural language processing (NLP) techniques, machine learning models and optimized software tools are applied in this study to give "structure" to unstructured big data (Xiao Liu, Singh, & Srinivasan, 2016).

This paper makes contributions to the current B2B literature in data, methods and findings. First, the final panel data provide a valuable record of the dynamic relationship between stock price and Twitter UGC among 407 companies for a period of 92 months, revealing both inter-company heterogeneity and intra-company dynamics, so they possess distinct advantages over time series or cross-sectional data (Hsiao, 2007). Second, this paper provides a framework (Fig. 1) with detailed steps on how to use NLP and machine learning techniques to process and transform "unstructured big data" into "structured small data". In essence, the framework demonstrates how to analyze big data

E-mail address: liul@rowan.edu.

https://doi.org/10.1016/j.indmarman.2019.02.021
Received 22 July 2018; Received in revised form 15 February 2019; Accepted 28 February 2019
0019-8501/

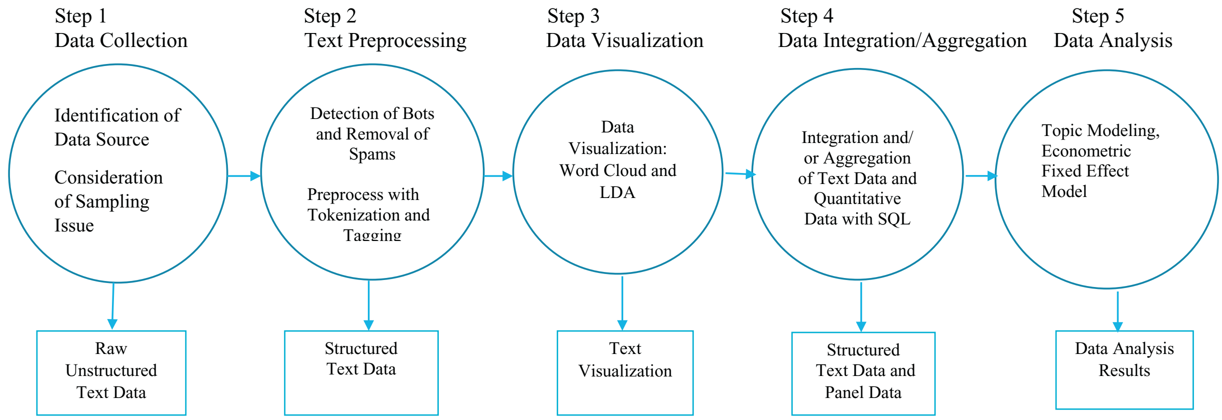


Fig. 1. The steps.

with traditional econometric methods (Varian, 2014b), such as the fixed effects model. The applications of Latent Dirichlet Allocation (LDA) and sentiment analysis help derive more information from UGC. Third, the findings are highly relevant to the B2B environment. We find that UGC has a significant impact on firms' stock performance and that its impact on stock performance is much stronger among B2C firms than among B2B firms. While consumers' positive sentiment does not play a significant role in stock performance, consumers' negative sentiment and WOM significantly impact stock prices.

2. Literature review

2.1. User generated content (UGC)

UGC is created by the general public or consumers instead of by marketing professionals (Kaplan & Haenlein, 2010). With the development of the Internet and social media platforms, a large amount of UGC is posted in textual or other formats, such as tweets (Twitter), product reviews (Amazon) and videos (YouTube). Prior research has demonstrated that UGC has effects on firms' financial performance and sales. For instance, Luo (2009) finds that negative WOM can negatively impact firms' cash flow, and P. K. Chintagunta, Gopinath, and Venkataraman (2010) conclude that online reviews can significantly impact new customer acquisition and movie box office sales.

Others have shown that UGC affects behavioral consumer variables such as buyer intention, product involvement, information sharing and brand engagement (Jin & Phua, 2014; Kim & Johnson, 2016). Goh, Heng, and Lin (2013) compare UGC with marketer-generated content (MGC) and conclude that UGC demonstrates a stronger impact than MGC on consumer purchase behavior. From a different angle, Hewett, Rand, Rust, and van Heerde (2016) examine how UGC interacts with firms' messages and find that negative messages can generate a higher volume of response, so firms need to adjust their communication messages accordingly. However, the majority of current UGC studies use a small amount of data to find support for their research propositions and draw conclusions. Although researchers have called for more applications of machine learning to unstructured textual data of UGC, only a few prior marketing scholars have used text mining and analytics to study big data of UGC on social media (Humphreys & Jen-Hui Wang, 2017; Lee & Bradlow, 2011).

2.2. Big data in marketing

Big data has become a hot buzzword in recent years. Marketing professionals from both industry and academia strive to extract valuable insights from big data. A literature review of big data in marketing (Table 1) shows several distinct streams.

First, marketing researchers use machine learning and text mining methods to get insights on sales and firm performance. For instance, Xiao Liu et al. (2016) analyze about 2 billion tweets by using text mining and cloud computing and find that the information content and timeliness of tweets can help predict sales forecast accuracy. Hewett et al. (2016) study 18 million tweets and document the feedback loops among firm messages, news media and UGC.

The second stream of research focuses on the opportunities and challenges in big data research and provides directions for future researchers. For example, Malthouse and Li (2017) present a detailed description of the opportunities and pitfalls of big data in advertising research. Sivarajah, Kamal, Irani, and Weerakkody (2017) provide an analysis of big data analytic methods, applications and challenges. Wang, Kung, Wang, and Cegielski (2018) develop a big data analytics model, defining and providing strategic insights for big data capabilities.

In yet another stream, prior scholars analyze large numbers of online reviews or customer data to extract product- and consumer-related insights. For instance, Tirunillai and Tellis (2014) process 350,000 online reviews with LDA and extract latent dimensions of product quality, including valence, label, validity, dynamics and so on. Similarly, Homburg, Ehm, and Artz (2015) analyze 115,000 online consumer posts and find that consumers demonstrate diminishing returns from firms' customer engagement.

In addition to these streams of research, some scholars have developed big data methods to generate marketing and brand insights. For example, Saboo, Kumar, and Park (2016) investigate temporal variations of marketing mix effectiveness, develop a time-varying effects model, and test it with a set of 250,000 customer data over a period of 36 months. Culotta and Cutler (2016) present a fully automated method to explore more than 200 brands' social connections on Twitter. Xia Liu, Burns, and Hou (2017) develop a framework that automatically extracts latent brand topics and analyzes consumers' sentiments toward brands. They illustrate the application of the framework with 1.7 million brand tweets covering 20 brands across five industries.

In summary, although some researchers have recently studied big data from different angles, there is a lack of research on the impact of big data on B2B firms' financial performance.

2.3. Analyzing unstructured data

The availability of large amounts of online textual data and the increase in computational power make it possible for computer-aided methods to process big data and draw research conclusions. Researchers from diverse disciplines such as communications (Lacy, Watson, Riffe, & Lovejoy, 2015), psychology (Iliev, Dehghani, & Sagi, 2015) and political science (Grimmer & Stewart, 2013) have used automated content analysis methods to answer research questions, but marketing research is still in the early stage of using large-scale

Table 1
Literature review: Big data and text mining in marketing.

Study	Field/Data	Method	Contributions
Lee and Bradlow (2011)	8226 online reviews	Text mining method	Market structure analysis by extracting product attributes
Netzer, Feldman, Goldenberg, and Fresko (2012)	868,000 customer messages	Text mining method	A perceptual map and associative networks for different brands
Grimmer and Stewart (2013)	Political Science	Conceptual	Principles, methods and pitfalls of conducting automated content analysis of political texts
Tirunillai and Tellis (2014)	350,000 customer product reviews	LDA	Extraction of latent dimensions of consumer satisfaction
Homburg et al. (2015)	115,000 online posts	Sentiment analysis	Diminishing returns from customer engagement
Iliev et al. (2015)	Psychology	Conceptual	A review of methods related to automated content analysis, including LSA, LDA and Hybrid methods
Lacy et al. (2015)	Communications	Conceptual	Discussion of issues related to content analysis and suggestions on how to tackle these issues
Büschken and Allenby (2016)	696 restaurant reviews	LDA	Comparison between word-based and sentence-based topics
Culotta and Cutler (2016)	More than 200 brands on Twitter	Automated text mining method	Explore and monitor brands' social connections on Twitter
Hewett et al. (2016)	18 million tweets	VAR model	Feedback loops between firm messages and user-generated content
Liu et al. (2016)	2 billion Tweets	Text mining method	Information content and timeliness help predict forecast accuracy
Mankad et al. (2016)	5830 online reviews	Text mining method	Analysis of online reviews in the hotel industry with a special focus on the impact of negative reviews
Saboo et al. (2016)	250,000 customer data over 36 months.	Time-varying effects model	A time-varying effects model framework and test its effectiveness
Humphreys & Wang (2017)	Marketing	Conceptual	An overview of linguistic theory, statistical methods and issues
Wang et al. (2018)	Healthcare	Case study and content analysis	A big data analytics model and define big data analytics capabilities
Sivarajah et al. (2017)	Conceptual	N/A	Description of big data analytic methods, applications and challenges
This paper	61 million tweets, 20.3 million Twitter accounts and 8 years of S&P 500 stock data	LDA, sentiment analysis fixed effect model	It uses machine learning methods to process unstructured Twitter data and merge it with S&P 500 stock data. It runs a fixed effects model on the final panel data. It finds that big data UGC has a significant impact on firms' stock performance and that the impact on B2C firms is much stronger than that on B2B firms.

automated content analysis (Humphreys & Jen-Hui Wang, 2017).

A study of marketing literature shows that current scholars have used automated text analysis methods to handle both small data and big data in their research designs (Table 1). For instance, Lee and Bradlow (2011) automate the extraction of product attributes and brand positions. Mankad, Han, Goh, and Gavirneni (2016) apply text mining methods to 5830 online reviews in hotel industry. Based on their analysis of 696 restaurant reviews, Büschken and Allenby (2016) draw the conclusion that compared to word-based topics, sentence-based topics are more distinguished. On the other hand, some researchers have applied more complicated text mining methods to draw insights from big data. For example, Netzer, Lattin, and Srinivasan (2008) develop a perceptual map and associative networks for different brands by text mining 868,000 customer messages. After conducting sentiment analysis on 115,000 online consumer posts, Homburg et al. (2015) conclude that from the customer side, there are diminishing returns from customer engagement. From a conceptual perspective, Humphreys and Jen-Hui Wang (2017) provide an overview of automated content analysis and outline the steps of conducting automated text analysis from research question development to validation.

More technically, text analytics makes use of probabilistic models such as topic models or concept mapping models to discover the latent topics of textual content (Blei, 2012; Nunez-Mir, Iannone, Pijanowski, Kong, & Fei, 2016). Text mining uses statistical algorithms to identify topics or themes from a large amount textual data (Blei, 2012). For instance, Lansdall-Welfare et al. (2017) use the n-gram method to identify culturomic changes over 150 years by analyzing 35 million articles, which represent 14% of all the articles published in the United Kingdom during that period. If human beings are given the task of handling and reading these millions of articles, it would take them many years to process and categorize the topics. Therefore, automated content analysis has simplified the processing of large collections of text and is particularly efficient in handling big unstructured data through machine learning methods.

Data mining methods can use a dictionary-based approach, supervised learning or unsupervised learning (van der Meer, 2016). The most common dictionary-based approach is word frequency counting, which uses textual characteristics or words to represent the construct that researchers intend to study (Humphreys & Jen-Hui Wang, 2017). The bag of words approach and word frequency counts are used frequently to analyze content and derive conclusions (van der Meer, 2016). With the supervised learning method, researchers need to manually code a subset of the data to provide a training set. By contrast, unsupervised learning methods can automatically learn the underlying features of the textual content. LDA is an example of an unsupervised learning method. Marketing scholars have utilized LDA in their research to explore customer satisfaction dimensions and brand topics (Xia Liu et al., 2017; Tirunillai & Tellis, 2014). For example, Heng, Gao, Jiang, and Chen (2018) collect more than one million online food reviews to detect the hidden topics of these reviews using LDA and examine the effects of these content topics on review helpfulness and ratings.

Since the analysis of unstructured textual data can be conducted with different methods, researchers need to choose a suitable methodological and statistical approach based on their research questions and purpose. In the method section, we will provide a more detailed description of how to handle statistical issues related to large textual data.

To sum up, an extensive literature review reveals that there is an urgent need to use big data to examine the impact of UGC on stock performance in the B2B context. Hence, in the next section, we investigate the following research questions:

1) Does UGC have a significant impact on firms' stock performance?
2) Does UGC have differential impacts on stock performance for B2B and B2C firms?

3. Research methodology

To answer the above research questions, we present a framework with all the steps of the research methodology (Fig. 1): data collection, text preprocessing, data visualization (summarization), data aggregation and analysis. We provide details on each component and describe

Table 2
Summary of Tweets by Sectors.

Sector	Volume	Retweets	Favorites	Replies
Consumer Discretionary	6,442,017	2,427,679	5,285,850	655,257
Consumer Staples	1,017,629	383,200	758,373	91,082
Energy	62,556	25,488	27,882	995
Financials	685,666	239,595	375,120	104,752
Health Care	124,394	53,636	81,688	11,982
Industrials	1,950,729	445,652	1,010,264	625,378
Information Technology	1,579,680	475,916	671,324	179,156
Materials	41,949	18,703	28,048	562
Real Estate	17,552	6927	9128	281
Telecommunications	792,302	194,033	467,840	38,959
Utilities	107,206	27,978	36,639	6925

how big unstructured data can be transformed into structured data and aggregated into a panel format.

3.1. Data collection

The first step is data collection. This paper's stock data and brand-related tweets were retrieved from two main sources. The datasets span from July 2009 to June 2017. We retrieved 8 years of daily stock information of S&P 500 firms from Yahoo Finance (Engelberg, 2018). We also downloaded 84 million brand tweets of UGC from Twitter and built a detailed database of 20.3 million Twitter accounts. Altogether, we have brand-related tweets for 407 companies across 11 sectors of the economy (Table 2).

Although Twitter data have some limitations (Boyd & Crawford, 2012; Sivarajah et al., 2017), Twitter is a viable source of social media data for marketing research. First, Twitter has over 900 million (Statista, 2018) registered users who actively post content, share information and engage with brands (Hennig-Thurau, Wiertz, & Feldhaus, 2015). This platform makes it possible to record behavioral big data on a daily basis. Second, most of the S&P 500 firms have created accounts on Twitter and use it as an integrated marketing communication platform to spread brand messages, attract prospective customers and promote products and brands. Third, companies have acquired large groups of Twitter followers, who actively post their personal experiences with brands and companies. This provides opportunities to study marketing phenomena from both firm and consumer sides. The variables in the Twitter dataset provide ample opportunities to explore many research questions. Fourth, data availability is also an important consideration. Social networks have different data access policies. Twitter has an open platform that provides data access through its Application Programming Interface (API).

Researchers can use web crawling techniques to download data from online websites. Data scraping functionalities are available in existing popular statistical software like SAS (Wu & Lin, 2018). To fulfill more complex data requirements, customized coding is needed with open source programming languages. For instance, the authors of this paper collected 84 million tweets that specifically contain comments on brands, products and services. To facilitate a data download of such magnitude, we wrote specialized Java code to retrieve the data from Twitter. We searched for brand-related tweets through keyword queries using Twitter's API. Twitter provides a powerful feature called “mention”, signified by the @ symbol (Jungherr, 2014). We improved the precision of our crawling algorithm by using this feature. Take AT&T for example: when AT&T customers write about their opinions on products and services, they often put “@ATT” in their tweets to specifically refer to the Twitter account of AT&T. So our Java code only queried tweets that include “@ATT”.

We have taken steps to increase the quality of samples so that the findings of this paper are more generalizable. The brands were selected based on the list of S&P 500 companies (Uotila, Maula, Keil, & Zahra, 2009). This list has frequently been used to conduct marketing research (Aksoy, Cooil, Groening, & Keiningham, 2008; Fornell, Mithas, Morgeson III, & Krishnan, 2006). In the end, consumer tweets for 407 companies were collected. Second, we tried to include all the Twitter accounts for each company. For instance, Comcast has 46 Twitter accounts, such as @Comcast, @Xfinity, @ComcastCares, @NBCUniversal, @ComcastWomen, and @ComcastCareers, which cover Comcast's social media activities on Twitter for divisions, subsidiaries, products, services, jobs, marketing communications and so forth.

3.2. Data preprocessing

After data collection, the next step is data preprocessing. Text preprocessing is an important step, but it is often under-emphasized. For instance, take a look at a typical tweet in Fig. 2. It has an informal language style. The subject “I” is implied but omitted. The preposition “with” is shortened to “w/”. Two mentions (@Comcast and @PhillyBusiness) and two hashtags (#customerservice and #comcast) appear in the sentence. It has a long URL for a link to an external website. All of these aspects make the processing of unstructured text data complicated and post unique challenges. Properly preprocessing the unstructured text data can significantly increase the accuracy of the final results.

First, spam messages and bot-created tweets need to be removed from the dataset, as these messages are not generated by real consumers and may distort real consumers' opinions and sentiments toward the brands. Morstatter, Wu, Nazer, Carley, and Liu (2016) and Varol, Ferrara, Davis, Menczer, and Flammini (2017) provide detailed procedures for detecting social bots. Tweet content and social graph features of Twitter accounts can be used to detect bots and delete spams (Chu, Gianvecchio, Wang, & Jajodia, 2012; A. H. Wang, 2010). Wang (2010) points out that a Twitter account is most likely to be a spam account if it posts a large amount of duplicate tweets regularly or if a tweet only contains a URL link. In this paper, we wrote programs in Java and classified bots and spam accounts by several features, including the content and linguistic styles of tweets, frequency and follower information of the accounts and domain-specific knowledge for economic sectors and industries.

Text	Just had fourth horrible exp w/ @Comcast Congrats on being #2 in bad #customerservice #comcast http://www.philly.com/philly/blogs/comcast-nation/Latest-from-Silicon-Valley-An-online-chat-robot-to-haggle-over-your-Comcast-bill.html via @PhillyBusiness
Creation	2017-01-08 19:14:28
Tweet Id	818249545993822210
Screen Name	Ranjnadas
Full Names	Ranjna Das
URL	https://twitter.com/ranjnadas/status/818249545993822210
Retweets	1
Favorites	3
Replies	1

Fig. 2. An example of user-generated Tweet.

(Chu et al., 2012; Liu et al., 2017; Varol et al., 2017). Altogether, we detected 23 million spam tweets and removed them from the dataset.

Second, the preprocessing of texts refers to NLP procedures that clean and transform data into machine-friendly formats (Bird, Klein, & Loper, 2009; Manning & Schütze, 1999). Hashtags and URLs need to be removed from the tweets.

Third, to achieve high quality results, machine learning methods often require textual data to be tokenized, lemmatized and tagged (Pustejovsky & Stubbs, 2012). Furthermore, certain stop words such as "the", "in" and "with" need to be removed because they appear frequently in English but do not provide much insight to the sentences. We performed tokenization, lemmatization and tagging with a free open source software package called Stanford CoreNLP (Manning et al., 2014).

In summary, data preprocessing for the current study consists of the procedures for removing bot accounts, spam tweets, hashtags, URLs, nonessential grammatical elements and stop words. We preprocessed our 84 million tweets to improve the quality of the data so that the data analysis results can truly reflect consumers' opinions. After necessary data preprocessing and spam removal, we had 61 million brand-related tweets about 407 companies from 11 economic sectors.

3.3. Data visualization

Data visualization is an effective tool to get a general idea of the content of textual data. It is a core technique to summarize big data graphically (Heimerl, Lohmann, Lange, & Ertl, 2014). In this section, we provide a couple of approaches to visualizing big data.

3.3.1. Word cloud

The first type of visualization creates a word cloud by identifying the frequency of the words in the big dataset. Fig. 3 was created as follows. After data preprocessing, only nouns, adjectives, verbs and adverbs from the tagging procedures are kept. We normalize the distribution matrix of words and use the Wordcloud package in the statistical software R to generate the word cloud.

Fig. 3 is created from 1.1 million brand tweets about Comcast, spanning from March 2007 to March 2018. We showcase the word cloud about Comcast tweets for several reasons. First, the telecom industry has been well known for its customer service issues (Izogo, 2017). Second, customers have posted a large amount of complaint tweets about Comcast's products and services in particular. Third, Comcast has a large presence on Twitter. "Service" is the most noticeable word in Fig. 3, indicating that customer service has the highest frequency and is a major concern for Comcast customers. The word cloud also shows Comcast's two major products: Internet and cable. The most dominant adjective in the figure is "bad", indicating that the overall feeling toward Comcast is negative.

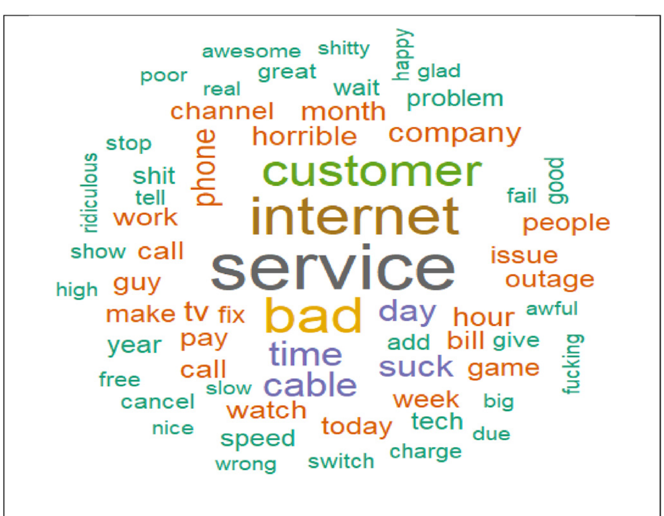


Fig. 3. Comcast Word Cloud.

3.3.2. Latent Dirichlet Allocation (LDA)

Although a word cloud can provide a quick and basic summary of texts with word frequency, more sophisticated methods are needed to demonstrate the themes or topics of the textual big data. LDA is a useful tool to visualize and summarize textual big data (Blei, 2012). As a topic modeling tool, LDA discovers the underlying patterns of the textual content and identifies the hidden topics of user-generated brand tweets. Accordingly, LDA has been increasingly adopted by marketing scholars to handle large amounts of text data (Xia Liu et al., 2017; Tirunillai & Tellis, 2014). Detailed descriptions of LDA and its application in marketing research can be found in Blei (2014), Blei, Ng, and Jordan (2003), and Liu et al. (2017). Researchers need to specify how many topics they intend to identify (Grün & Hornik, 2011). There are implementations of LDA in several programming languages. In this paper, we use a fast implementation of LDA with a Java-based package called Mallet (McCallum, 2002).

Fig. 4a shows the results from running LDA on Comcast's UGC tweets in December 2016. The top 12 words of the most dominant topic are "Internet, pay, outage, shit, service, month, wifi, data, area, time, fix, mbps". This topic covers product, service and billing. The second-most dominant topic consists of "service, customer, worse, call, phone, time, internet, cable, bill, support, company, hold" and conveys consumers' very negative impression of Comcast customer service. These LDA results are consistent with the word cloud in Fig. 3 and show that Comcast customer service is considered to be very poor. Furthermore, they provide much more nuanced information and clearly demonstrate that consumers are dissatisfied with internet outages.

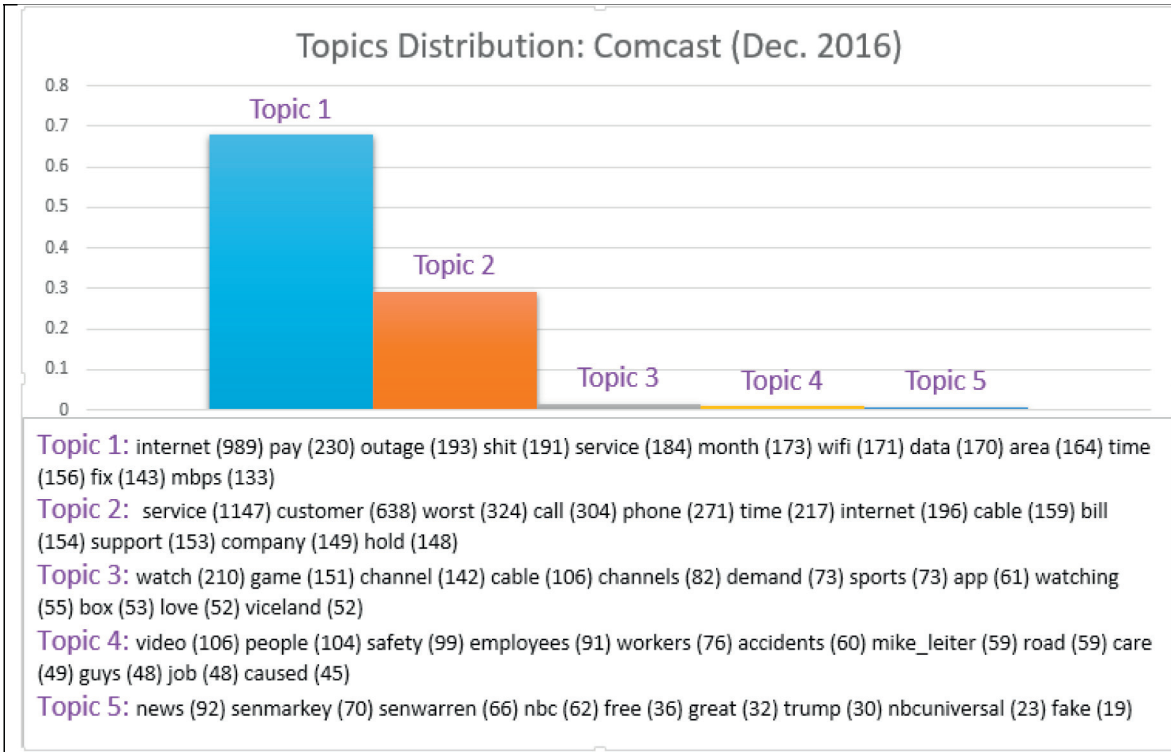
The topics exhibited in Fig. 4a are static. They are extracted from all the Comcast UGC tweets in December 2016 and reveal what consumers talk about in one time period. In a more dynamic portrayal of consumer topics, LDA can be used to visualize the temporal development of topics. In Fig. 4b, we run LDA on UGC tweets for each December from 2013 to 2017. The contents of the topics remain stable and are mostly concerned about customer service and products. A notable exception occurs in December 2017, when topics are mostly related to consumers' reactions to the repeal of the net neutrality regulation by the Federal Communication Commission (FCC).

3.3.3. Sentiment analysis

The LDA results from the previous section show that Comcast has serious customer service issues. We want to further explore these issues with sentiment analysis, which is a supervised learning method and can also be used to visualize and summarize big data. Sentiment analysis can be carried out in various models, such as maximum entropy (Ravi & Ravi, 2015), support vector machines (Devi, Kumar, & Prasad, 2016) and neural network (Zhang, Zou, & Gan, 2018). This paper uses a maximum entropy model that is implemented in a Java-based package called Mallet (McCallum, 2002).

Fig. 5a demonstrates the volume of negative tweets in relation to the overall volume of tweets of 46 Comcast Twitter accounts over an eight-year period from November 2009 to July 2017. The numbers of retweets and favorites, which are great indicators of word of mouth (WOM) activities, are also included on the chart. It is not surprising that the volume of negative sentiments is highly correlated with tweet volume. Overall, the number of favorites is higher than the number of retweets, while the number of retweets has spikes depending on the occurrences of specific events, which greatly influenced the spread of information.

Consumer sentiments can be studied at different levels. Fig. 5b summarizes and compares consumers' negative sentiment ratios among some of Comcast's own accounts. The @Comcast account, which is



a: Topic Distribution (Dec. 2016)

Dec. 2013	Dec. 2014	Dec. 2015	Dec. 2016	Dec. 2017
service	Service	service	service	net
internet	internet	internet	internet	neutrality
customer	customer	customer	customer	internet
cable	cable	cable	time	service
phone	time	worst	cable	verizon
time	call	time	pay	fuck
worst	worst	call	worst	att
call	phone	pay	call	pay
watch	pay	phone	phone	people
back	fix	watch	watch	money
fix	back	bill	fix	customer
pay	bill	fix	people	time

b: Topic Distributions (Dec. 2013-Dec. 2017)

Fig. 4. Topic distributions of Comcast UGC Tweets.

Comcast's corporate portal on Twitter, has attracted a significant number of negative tweets, higher than the negativity ratios of the @Comcastcares and @Xfinity accounts. It is worth pointing out that @ NBCUniversal, the account for one of Comcast's major divisions, has a quite low negative sentiment ratio. Furthermore, there is a large fluctuation of consumers' negative sentiments for @Xfinity. This is mainly due to the scarcity of data at the beginning of the data collection period.

We can also compare consumers' negative sentiments by brands. For illustration purposes, we chose Comcast, Dish Network, United Airlines, Southwest, Michael Kors and Ralph Lauren. Fig. 6a shows that among these brands, consumers have the highest negative sentiments toward United Airlines and Comcast. Compared to Comcast, Dish Network and Southwest Airlines have lower negative consumer sentiments, and Michael Kors and Ralph Lauren have even lower. In Fig. 6b, negative sentiments among different industries over the eight-year period are compared and contrasted. The cable & satellite industry ranks the highest and the electrical components & equipment industry ranks the lowest. In summary, sentiment analysis and time series plots can

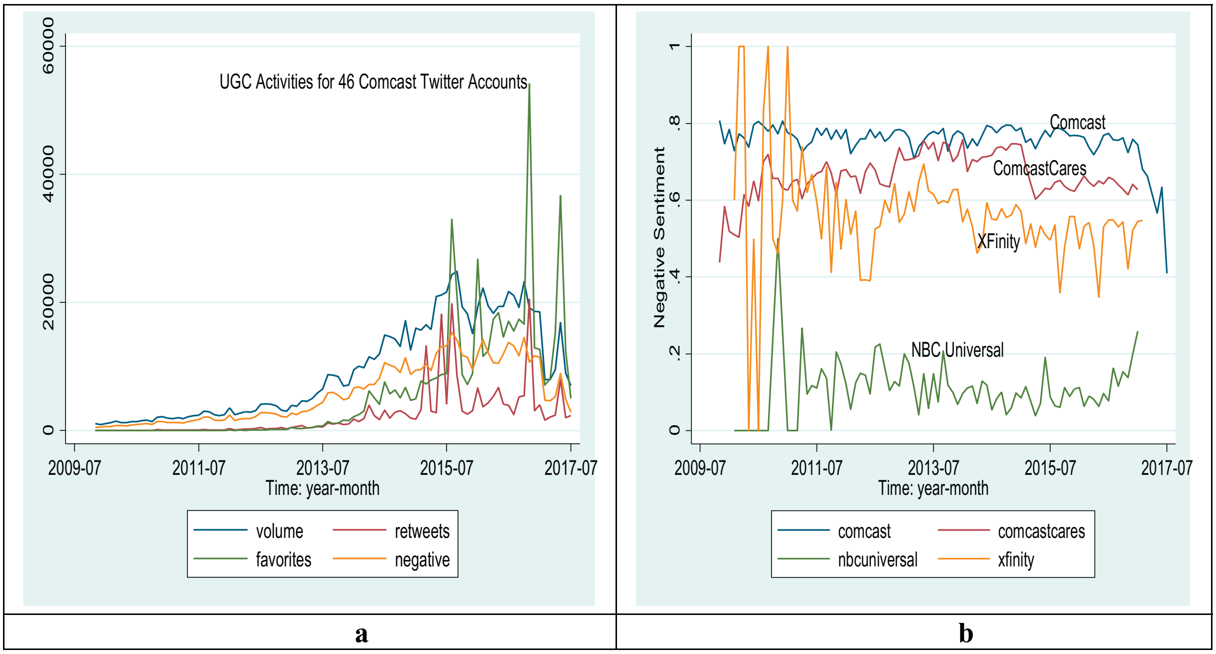


Fig. 5. UGC Activities (5a) and Negative Sentiments (5b).

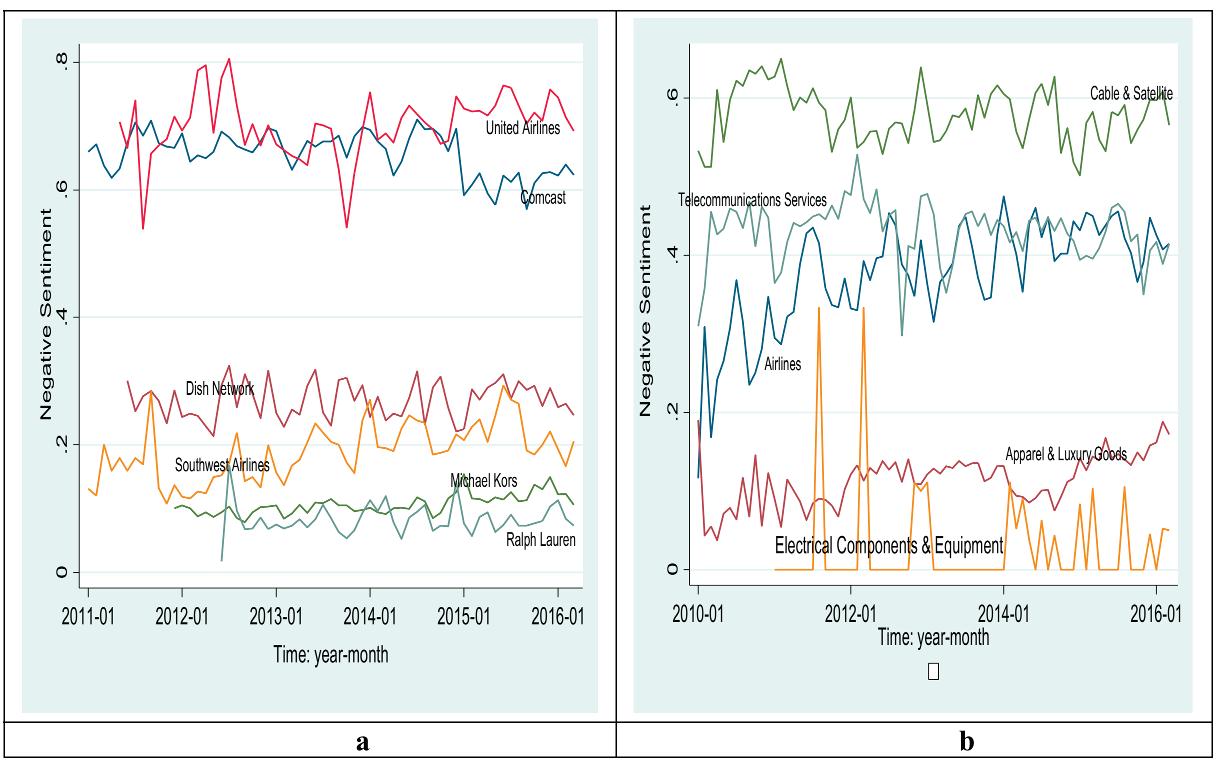


Fig. 6. Negative Sentiments by Brand (6a) and by Industry (6b).

effectively present change and trends in consumers' sentiment over time. Hence, big data visualization provides an efficient way to quickly extract and summarize valuable information from the dataset, enhancing researchers' understanding of the textual content.

We used maximum entropy to conduct sentiment analysis and classified all the tweets into three categories: positive, negative and neutral. We randomly chose 12,000 tweets from the whole dataset to use as a training set for the machine learning method. Two researchers independently coded the tweets and the agreement between the two researchers was 80.3%, which met the text coding standard. Using the training set, we ran the maximum entropy model on 61 million tweets and classified each tweet's sentiment as positive, negative or neutral.

3.4. Data aggregation

When data sets from more than one source are used, they must be integrated. For instance, we combine NLP results of 61 million tweets with stock data for each company. We store these data sets in MySQL database and use union operator to merge tweets, company information and financial data (Wedel & Kannan, 2016). Relational databases are

Table 3
The impacts of user-generated content on stock performance.

Variable	Coefficient	Standard Error	p-value
Volume	1.831	0.6873	0.033*
Negative	−12.9082	1.5897	0.005**
Positive	3.9343	2.1732	0.3346
WOM	0.5664	0.1332	0.037*
Intercept	71.3636	6.7353	0.000**

** $p < 0.01$.
* $p < 0.05$.

invaluable tools in marketing analytics and digital marketing research with big data (George, Osinga, Lavie, & Scott, 2016).

One of the keys to transforming big data into a manageable format is to aggregate it at different temporal intervals. In this paper, we aggregate the textual content variables into multiple-year monthly panel data consisting of 407 different brands. The following Twitter variables are included: number of retweets, number of favorites, negative sentiment, positive sentiment and mentions. Aggregation of datasets is performed with the SQL "group by" statement (Schwartz, Zaitsev, & Tkachenko, 2012). When data sets get even larger, Spark SQL and machine learning libraries can be applied (Armbrust et al., 2015; Meng et al., 2016).

3.5. Data analysis

The final panel data consists of 407 companies, whose stock data and Twitter UGC activities are calculated on a monthly basis. Since these companies started their Twitter accounts at various times, there are some missing periods for UGC data. This means the panel is unbalanced, with 21,407 observations covering 11 sectors of the economy.

Social scientists need to exercise great caution in their attempts to make causal inferences from non-experimental data (Wooldridge, 2010). To properly analyze this valuable panel data set, we ran a fixed effects model, which has some unique advantages. The fixed effects model enables researchers to conduct causal analysis from observational data by controlling for variables which are either unobservable or unmeasurable (Allison, 2009). Then how do we prevent confounding without including these variables in fixed effects model? Allison advises researchers to use each individual as his or her own control" (Allison, 2009, p. 1). Since companies in this dataset are heterogeneous in numerous aspects, such as firm sizes, organizational culture, leadership styles, and competitor landscapes, the ability to use each individual company as its own control is particularly valuable. By controlling for inter-company heterogeneity, the fixed effects model reveals a more accurate relationship between financial performance and UGC within each company.

When regression analysis is conducted on nonstationary time series, the results might appear to be significant when the variables are not even related. Such a regression produces unreliable least squares estimators and *t*-statistics, leads to erroneous statistical results, and is called spurious regression (Shuai, Shen, Jiao, Wu, & Tan, 2017). Therefore, it is important to check whether time series are stationary to avoid the potentially serious threat to our panel data analysis of spurious regression. Stata provides several tests for unit roots (Cameron & Trivedi, 2010). We use the xtunitroot function from Stata 14. Since our panel dataset was unbalanced, a Fisher-type (Choi, 2001) test was conducted on all the variables of interest in the econometric model: stock, tweet volume, negative sentiment, positive sentiment and word of mouth. The test yielded *p*-values less than 0.05 for all the variables. Hence, the panels do not exhibit non-stationarity issues and contain no unit roots.

We then ran the Hausman endogeneity test to decide whether a fixed effects model was indeed preferable to a random effects model (Hausman, 1978; Wooldridge, 2010). The *p*-value from the test was 0.023, so we rejected the null hypothesis and chose the fixed effects (FE) model.

The econometric model is presented as follows:

$$Stock_{it} = \beta_1 Volume_{it} + \beta_2 Negative_{it} + \beta_3 Positive_{it} + \beta_4 WOM_{it} + \alpha_i + \epsilon_{it} \tag{1}$$

where $i = 1, \ldots, N$ ($= 407$) represents companies and $t = 1, \ldots,$ T ($= 92$) represents monthly time intervals; α_i holds the fixed effects of the model; ε_{it} holds the random error for the model. Volume is the total number of brand-related tweets created by Twitter users on a monthly basis. Negative represents the ratio of negative UGC tweets to the volume of UGC tweets for the month *t*. Positive represents the ratio of positive UGC tweets to the volume of UGC tweets for the month *t*. WOM (word of mouth) is operationalized by the sum of monthly retweets and favorites.

The effects of UGC activities on stock performance are provided in Table 3. The modeling results show that three variables significantly impact stock performance: volume ($b = 1.831$, $p = .033$), negative sentiment ($b = -12.9082$, $p = .005$) and WOM ($b = 0.5664$, $p = .037$).

Following the above analysis, we conducted panel data analysis for only B2B firms. B2B firms include companies in industries such as energy equipment and services, oil and gas energy, chemicals, construction materials, packaging, metals and mining, aerospace and defense, construction and engineering and so on. The fixed effects model for B2B firms shows that although negative sentiment still has a significant effect on firms' stock performance, it is much smaller in magnitude ($b = -1.3257$, $p = .002$). In addition, WOM for B2B firms exerts significant influence on stock performance, but with a much smaller coefficient (0.1232, $p = .046$). Volume, however, does not significantly impact stock performance among B2B firms (1.3002, $p = .5289$).

According to the data analysis, UGC demonstrates differential impacts on firms' financial performance between B2B and B2C firms. Compared to B2B customers, B2C customers are particularly powerful on social media, as the volume of their UGC can significantly impact firms' financial performance. For B2B customers, although WOM is important, the information from B2B customers is not as impactful as that from B2C customers. Moreover, we found that while positive sentiment does not impact firms' financial performance for both B2C and B2B firms, negative sentiments can be harmful to a firm's financial performance, especially B2C firms.

4. Discussion and conclusions

Big data of user-generated content in textual formats contains invaluable information about consumer activities on the Internet. Furthermore, big data analytics can facilitate firm operations and business activities (Y. Wang et al., 2018). However, major challenges in methodology prevent researchers from taking full advantage of these behavioral data. The investigation of big textual data in this paper provides new insights about the impact of UGC on B2B firms' stock performance.

This paper advances current B2B literature by demonstrating how social media can assist B2B firms' marketing activities and how marketing scholars and professionals can utilize big data to extract valuable information to understand customers and predict firms' financial performance. As pointed out by Lilien (2016), the majority of customer data for B2B firms have not been utilized efficiently to provide sufficient marketing information for practitioners. The current findings demonstrate that big data UGC has a significant impact on B2B firms' performance and that consumers' sentiments are particularly important. Hence, B2B firms need to pay particular attention to consumers' satisfaction with their services or products. If not managed properly, negative consumer sentiment can hurt a brand's reputation, leading to a negative impact on firms' financial performance.

Furthermore, marketing managers need to focus more on building customer relationships on social media and directly involving their customers with their marketing messages. Only posting a large number of messages might not be the most effective approach. For instance, personalized messages or direct messages to customers can increase customers' engagement with brands, leading to more WOM and information dissemination on social media.

From a research methodological perspective, this paper provides a framework for all the processes in the text analysis of big data. It describes and illustrates each step with enough detail that researchers can use it on different datasets in their own research. This systematic framework of using both machine learning methods and econometric models provides insightful findings for industrial business research. In addition, the methodology provides procedures, models and software tools that can guide future marketing scholars who are interested in applying big data analysis in B2B research.

The application of unstructured big data and financial information provides unique value to marketing researchers. Special attention is given to the handling of the "big" and "unstructured" aspects of textual user-generated content. To tackle the "bigness" challenge, the application of relational database is adopted. Insights are offered on how to quantify and give structure to unstructured textual data (Liu et al., 2016).

Moreover, details are provided about how to integrate numerical variables with quantitative representations of sentiments. Merging of data happens not only between datasets but also across methods. For instance, results from machine learning are used as input for econometric models. This bridges cutting-edge data analytical techniques and traditional marketing research tools. It offers the best of both worlds and gives much more in-depth insight (Varian, 2014a, 2014b). The particular emphasis on the authenticity of user-generated content is well placed and especially relevant in today's digital world that is full of fake news and junk information.

5. Limitations and future directions

We note that this paper has several limitations. First, although particular attention has been given to sampling representativeness, Twitter is only one of many social platforms on which consumer behavior data abound. Equally valuable textual big data of user-generated content can be found on other digital online sources such as Facebook, Amazon and Yelp as well as online communities, blogs, and discussion forums. Therefore, further research is needed to discover how to best utilize and derive insights from these data channels.

Second, user-generated content exists in drastically different formats, such as videos, pictures and virtual reality. Although these types of data can still benefit from the research methodology in this paper, they might call for new methods. Therefore, future researchers can explore ways to use these data to get marketing insights. Since big data is an evolving concept, it will also be valuable to investigate technologies that can handle data in even bigger volumes, such as parallel computing (Li, Yuan, Ma, & Yao, 2018), No-SQL database (Bjeladinovic, 2018) and large-scale machine libraries provided by Spark (Meng et al., 2016).

References

Aksoy, L., Cooil, B., Groening, C., & Keiningham, T. L. (2008). The long-term stock market valuation of customer satisfaction. *Journal of Marketing, 72*(4), 105–122.

Allison, P. D. (2009). *Fixed effects regression models. Vol. 160*. Thousand Oaks, CA: SAGE.

Armbrust, M., Xin, R. S., Lian, C., Huai, Y., Liu, D., Bradley, J. K., ... Ghodsi, A. (2015). Spark SQL: Relational data processing in spark. *Paper presented at the proceedings of the 2015 ACM SIGMOD international conference on Management of Data, Melbourne, Australia*.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit.* Sebastapol, CA: O'Reilly Media.

Bjeladinovic, S. (2018). A fresh approach for hybrid SQL/NoSQL database design based on data structuredness. *Enterprise Information Systems, 12*, 1202–1220.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84.

Blei, D. M. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application, 1*, 203–232.

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

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679.

Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *Marketing Science, 35*(6), 953–975.

Cameron, A. C., & Trivedi, P. K. (2010). *Microeconometrics using stata. Vol. 2*. College Station, TX: Stata Press.

Chintagunta, P., Hanssens, D. M., & Hauser, J. R. (2016). Editorial—Marketing science and big data. *Marketing Science, 35*(3), 341–342.

Chintagunta, P. K., Gopinath, S., & Venkataraman, S. (2010). The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science, 29*(5), 944–957.

Choi, I. (2001). Unit root tests for panel data. *Journal of International Money and Finance, 20*(2), 249–272.

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of Twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing, 9*(6), 811–824.

Culotta, A., & Cutler, J. (2016). Mining brand perceptions from Twitter social networks. *Marketing Science, 35*(3), 343–362.

van der Meer, T. G. (2016). Automated content analysis and crisis communication research. *Public Relations Review, 42*(5), 952–961.

Devi, D. N., Kumar, C. K., & Prasad, S. (2016). *A feature based approach for sentiment analysis by using support vector machine. Paper presented at the 6th International Advance Computing Conference (IACC), Bhimavaram, India*.

Engelberg, J. (2018). Discussion of "Earnings announcement promotions: A Yahoo Finance field experiment". *Journal of Accounting and Economics, 66*(2–3), 415–418.

Fornell, C., Mithas, S., Morgeson, F. V., III, & Krishnan, M. S. (2006). Customer satisfaction and stock prices: High returns, low risk. *Journal of Marketing, 70*(1), 3–14.

George, G., Haas, M. R., & Pentland, A. (2014). Big data and management. *Academy of Management Journal, 57*(2), 321–326.

George, G., Osinga, E. C., Lavie, D., & Scott, B. A. (2016). Big data and data science methods for management research. *Academy of Management Journal, 59*(5), 1493–1507.

Goh, K.-Y., Heng, C.-S., & Lin, Z. (2013). Social media brand community and consumer behavior: Quantifying the relative impact of user- and marketer-generated content. *Information Systems Research, 24*(1), 88–107.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267–297.

Grün, B., & Hornik, K. (2011). Topicmodels: An R package for fitting topic models. *Journal of Statistical Software, 40*(13), 1–30.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the Econometric Society,* 1251–1271.

Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). Word cloud explorer: Text analytics based on word clouds. *Paper presented at the 47th Hawaii international conference on system sciences (HICSS), Waikoloa, HI*.

Heinonen, K. (2011). Consumer activity in social media: Managerial approaches to consumers' social media behavior. *Journal of Consumer Behaviour, 10*(6), 356–364.

Heng, Y., Gao, Z., Jiang, Y., & Chen, X. (2018). Exploring hidden factors behind online food shopping from Amazon reviews: A topic mining approach. *Journal of Retailing and Consumer Services, 42*, 161–168.

Hennig-Thurau, T., Wiertz, C., & Feldhaus, F. (2015). Does twitter matter? The impact of microblogging word of mouth on consumers' adoption of new movies. *Journal of the Academy of Marketing Science, 43*(3), 375–394.

Hewett, K., Rand, W., Rust, R. T., & van Heerde, H. J. (2016). Brand buzz in the echoverse. *Journal of Marketing, 80*(3), 1–24.

Homburg, C., Ehm, L., & Artz, M. (2015). Measuring and managing consumer sentiment in an online community environment. *Journal of Marketing Research, 52*(5), 629–641.

Hsiao, C. (2007). Panel data analysis—Advantages and challenges. *Test, 16*(1), 1–22.

Humphreys, A., & Jen-Hui Wang, R. (2017). Automated text analysis for consumer research. *Journal of Consumer Research, 44*, 1274–1306.

Iliev, R., Dehghani, M., & Sagi, E. (2015). Automated text analysis in psychology: Methods, applications, and future developments. *Language and Cognition, 7*(2), 265–290.

Izogo, E. E. (2017). Customer loyalty in telecom service sector: The role of service quality and customer commitment. *The TQM Journal, 29*(1), 19–36.

Jin, S.-A. A., & Phua, J. (2014). Following celebrities' tweets about brands: The impact of twitter-based electronic word-of-mouth on consumers' source credibility perception, buying intention, and social identification with celebrities. *Journal of Advertising, 43*(2), 181–195.

Jungherr, A. (2014). The logic of political coverage on Twitter: Temporal dynamics and content. *Journal of Communication, 64*(2), 239–259.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons, 53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003.

Kim, A. J., & Johnson, K. K. (2016). Power of consumers using social media: Examining the influences of brand-related user-generated content on Facebook. *Computers in Human Behavior, 58*, 98–108.

Lacy, S., Watson, B. R., Riffe, D., & Lovejoy, J. (2015). Issues and best practices in content analysis. *Journalism & Mass Communication Quarterly, 92*(4), 791–811.

Lansdall-Welfare, T., Sudhahar, S., Thompson, J., Lewis, J., Team, F. N., & Cristianini, N. (2017). Content analysis of 150 years of British periodicals. Proceedings of the

National Academy of Sciences, 201606380.
Lee, T. Y., & Bradlow, E. T. (2011). Automated marketing research using online customer reviews. *Journal of Marketing Research (JMR), 48*(5), 881–894. https://doi.org/10.1509/jmkr.48.5.881.
Li, X., Yuan, J., Ma, H., & Yao, W. (2018). Fast and parallel trust computing scheme based on big data analysis for collaboration cloud service. *IEEE Transactions on Information Forensics and Security, 13*, 1917–1931.
Lilien, G. L. (2016). The B2B knowledge gap. *International Journal of Research in Marketing, 33*(3), 543–556.
Liu, X., Burns, A. C., & Hou, Y. (2017). An investigation of brand-related user-generated content on twitter. *Journal of Advertising, 46*(2), 236–247. https://doi.org/10.1080/00913367.2017.1297273.
Liu, X., Singh, P. V., & Srinivasan, K. (2016). A structured analysis of unstructured big data by leveraging cloud computing. *Marketing Science, 35*(3), 363–388.
Luo, X. (2009). Quantifying the long-term impact of negative word of mouth on cash flows and stock prices. *Marketing Science, 28*(1), 148–165.
Malthouse, E. C., & Li, H. (2017). Opportunities for and pitfalls of using big data in advertising research. *Journal of Advertising, 46*(2), 227–235. https://doi.org/10.1080/00913367.2017.1299653.
Mankad, S., Han, H. S., Goh, J., & Gavirneni, S. (2016). Understanding online hotel reviews through automated text analysis. *Service Science, 8*(2), 124–138.
Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing.* Boston, MA: MIT Press.
Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *Paper presented at the ACL (system demonstrations)*. Retrieved from https://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf.
McCallum, A. K. (2002). MALLET: A machine learning for language toolkit. Retrieved from http://mallet.cs.umass.edu.
Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., ... Owen, S. (2016). Mllib: Machine learning in apache Spark. *The Journal of Machine Learning Research, 17*(1), 1235–1241.
Minnick, K., & Noga, T. (2017). The influence of firm and industry political spending on tax management among S&P 500 firms. *Journal of Corporate Finance, 44*, 233–254.
Morstatter, F., Wu, L., Nazer, T. H., Carley, K. M., & Liu, H. (2016). A new approach to bot detection: Striking the balance between precision and recall. *Paper presented at the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA*.
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.
Netzer, O., Lattin, J. M., & Srinivasan, V. (2008). A hidden Markov model of customer relationship dynamics. *Marketing Science, 27*(2), 185–204.
Nunez-Mir, G. C., Iannone, B. V., Pijanowski, B. C., Kong, N., & Fei, S. (2016). Automated content analysis: Addressing the big literature challenge in ecology and evolution. *Methods in Ecology and Evolution, 7*(11), 1262–1272.
Pustejovsky, J., & Stubbs, A. (2012). *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications.* Sebastapol, CA: O'Reilly Media.
Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems, 89*, 14–46.
Saboo, A. R., Kumar, V., & Park, I. (2016). Using big data to model time-varying effects for marketing resource (re) allocation. *MIS Quarterly, 40*(4), 911–939.
Schwartz, B., Zaitsev, P., & Tkachenko, V. (2012). *High performance MySQL: Optimization, backups, and replication.* Sebastapol, CA: O'Reilly Media.
Shuai, C., Shen, L., Jiao, L., Wu, Y., & Tan, Y. (2017). Identifying key impact factors on carbon emission: Evidences from panel and time-series data of 125 countries from 1990 to 2011. *Applied Energy, 187*, 310–325.
Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. *Journal of Business Research, 70*, 263–286.
Statista (2018). Number of monthly active Twitter users worldwide from 1st quarter 2010 to 2nd quarter 2018 (in millions). Retrieved from https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.
Tirunillai, S., & 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.
Trilling, D., & Jonkman, J. G. (2018). Scaling up content analysis. *Communication Methods and Measures, 12*(2–3), 158–174.
Uotila, J., Maula, M., Keil, T., & Zahra, S. A. (2009). Exploration, exploitation, and financial performance: Analysis of S&P 500 corporations. *Strategic Management Journal, 30*(2), 221–231.
Varian, H. R. (2014a). Beyond big data. *Business Economics, 49*(1), 27–31.
Varian, H. R. (2014b). Big data: New tricks for econometrics. *The Journal of Economic Perspectives, 28*(2), 3–28.
Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. arXiv preprint. arXiv:1703.03107.
Wang, A. H. (2010). Detecting spam bots in online social networking sites: A machine learning approach. *Paper presented at the IFIP annual conference on data and applications security and privacy*.
Wang, Y., Kung, L., Wang, W. Y. C., & Cegielski, C. G. (2018). An integrated big data analytics-enabled transformation model: Application to health care. *Information & Management, 55*(1), 64–79.
Wedel, M., & Kannan, P. (2016). Marketing analytics for data-rich environments. *Journal of Marketing, 80*(6), 97–121.
Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data.* Boston, MA: MIT Press.
Wu, P.-J., & Lin, K.-C. (2018). Unstructured big data analytics for retrieving e-commerce logistics knowledge. *Telematics and Informatics, 35*(1), 237–244.
Zhang, Z., Zou, Y., & Gan, C. (2018). Textual sentiment analysis via three different attention convolutional neural networks and cross-modality consistent regression. *Neurocomputing, 275*, 1407–1415.