RESEARCH-IN-PROGRESS

DATA NEVER SLEEP: A MARKETING PERSPECTIVE

Ezgi Akar
Bogazici University, Management Information Systems
Istanbul, Turkey
ezgi.akar@boun.edu.tr

Serkan Akar
Managing Director of Inci Sozluk
Istanbul, Turkey
serkan.akar@incisozluk.com.tr

Abstract

Millions of people share photos, texts, videos, and other types of contents on various social networking sites in their daily lives. It indicates that there is an enormous amount of data generated by those people on the Internet and this data generation continues to grow fast. Businesses collect any data such as consumer preferences, purchases, or trends on the Internet to keep their strategies up-to-date, to take strategic precautions, and to satisfy their consumers. In this sense, this study aims to analyze trending topic data gathered from Twitter that is one of the most popular and publicly available social media data sources. In Twitter, more than 500 million tweets are shared per day, and some of them become trending ones. These trending topics have the power to keep people aware and entertained. This capability also provides e-marketers with a useful tool to get in front of a big and potential audience. In parallel, this study investigates 100 trending topics involving 301.492 tweets and 92.745 unique users, and it clusters these topics considering user-related factors. Thus, this research shows a way for e-marketers how to make a trending topic and to reach new audiences through social networking platforms.

Keywords: clustering, social media, trend topic, Twitter, e-marketing

INTRODUCTION

A new popular term called as *big data* has shown up, and people and academics in information technologies have become more interested in it. Big data is defined as "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze" (Manyika et al., 2011). This definition highlights that big data does not only mean gigabytes or terabytes of large data, but it also refers to datasets that cannot be collected, saved, and analyzed by traditional database systems (Purcell, 2013).

Today, data are not only generated from traditional ways but also from the Internet, social networking sites, multimedia contents, digital images, GPS signals, etc. For example; in 2013, 4.4 zettabytes of data were generated in the world, and by 2020 there will be 44 of zettabytes data (Northeastern University, 2016). Additionally, Google processes 3.5 billion requests per day and stores 10 exabytes data (Deep Web Tech, 2016). Amazon hosts about 1.4 million servers to handle with daily requests. Facebook collects 500 terabytes of daily data including contents, likes, and photos. Moreover, 90% of the data are created within last two years (Gobble, 2013). These statistics indicate that data never sleep.

In parallel to these statistics, the analysis of these data has become beneficial and even crucial for various industries to maintain their status quo and catch up this new this era. 95% of the US businesses state that they prefer to use data to power their business opportunities and 84% of the US businesses say that data have become the part of their business strategies (The Global Data Management Benchmark Report, 2017). Furthermore, investigating lots of data allows businesses to understand consumers' needs and their purchasing habits.

Twitter is one of valuable data sources in where a considerable amount of data is generated. Twitter provides people and academics with application program interface (API) that makes data collection is easy and less effortless for every Internet user (Burgess and Bruns, 2012). Twitter that was released in 2006 is a very popular microblogging platform around the world. It has already become a natural platform where information disseminates severely (Ribarsky et al., 2014). Users send more than 500 million tweets per day (Omnicore, 2017). Tweets are known as text messages including 140 characters. In tweets, words or phrases whether including "# (hashtag)" or not can be a trending topic. For example; both #madonna and "I love Madonna" can be trending topics. These topics can be determined as emerging events, breaking news, and general topics (Ahangama, 2014), and they become visible to all users. In this respect, this study focuses on trending topics on Twitter and tries to find answers to the following questions:

- how a hashtag becomes a trending topic naturally,
- how Twitter users create a trending topic,
- what factors affect topics to be a trending one,
- how e-marketers can benefit from trending topics.

Within the scope of the study, 100 trending topics have been collected over a three-week period. For each trending topic, tweets that included related trending topic were gathered. After that, users who tweeted were obtained, and their total number of tweets, followers, and followings were collected. As a result, 301.492 tweets and 92.745 unique user information have been collected. After that, these trending topics are clustered by taking all related and collected data into considerations.

This paper is divided into four sections. The related works are explained in the first section. In the second section, the methodology of the study is included. In the third part, the study results are presented. In the last section, the study highlights and implications are discussed.

RELATED WORK

Trending topics have been analyzed in many academic studies for various purposes. In their study, Naaman et al. (2011) characterize the emerging trends on Twitter. They study on two datasets including 8.500 trends and 48.000.000 tweet messages. They focus on the trend detection by using term frequency-inverse document frequency (TF-IDF) weighting as a methodology. In the study of streaming trend detection in Twitter, Benhardus and Kalita (2013) also use this same methodology for the trend detection. They define TF-IDF weighting as "an information retrieval technique that weights a document's relevance to a query based on a composite of the query's term frequency and inverse document frequency."

Additionally, Lee et al. (2011) use TF-IDF weighting technique as a part of their study. They try to classify trending topics based on 18 categories in the study. Firstly, they create 18 categories and then apply two approaches for the topic classification: a bag of words for text classification and network-based classification. In the text-based classification method, they use TF-IDF weighting technique. In network-based classification method, they identify five associated topics for a given topic based on the number of common users. They randomly select 768 trending topics and apply these techniques and compare the accuracy results.

Moreover, Gao et al. (2013) study on the summarization of the Twitter trending topics. They do analyses by using both streams based and semantic-based approaches to detect

important subtopics within a trending topic, and then they propose a sequential summarization. Gao et al. (2013) focus on Latent Dirichlet Allocation (LDA) statistical model and Kurniati et al. (2014) also concentrate on the same statistical model. They compare the effectiveness of LDA and semantic-based Joint Multi-Grain Topic-Sentiment topic modeling techniques in their study. They collect 8.6 million tweets and apply these techniques to detect trend topics from Twitter stream data. Besides, Lau et al. (2012) introduce a novel topic modeling-based methodology to follow emerging events on Twitter based on LDA statistical model. Furthermore, Yang and Rim (2014) expand LDA as TS-LDA which stands for trend-sensitive. This model extracts latent topics from contents.

Wilkinson and Thelwall (2012) make an international comparison. They collect tweets from 6 countries including 0.5 billion tweets based on the top 50 trending keywords. They compare the trending topics based on each country. Lastly, Ma et al. (2013) focus on the predicting the popularity on newly emerging hashtags in Twitter. In their study, they compare five classification models among which the logistic regression model performs the best. Aiello et al. (2013) also compare six topic detection methods by using Twitter stream data.

In addition to this research, Ahangama (2014) presents a new method in his study. This new method finds the trending topics of different social media networks using real-time data that are published on Twitter. Song and Kim (2013) also develop such a system. They call it as "real-time Twitter trend mining system" to process a huge volume of data available on Twitter.

Moreover, Han et al. (2014) study trend topics from a distinct perspective. They try to disambiguate the meanings of the topics in the trending list. They compare and apply key factor extraction, named entity recognition, topic modeling, and automatic summarization methods to extract the contents of trending topics. Giummolè et al. (2013) compare Twitter trends and Google hot queries. They test the relation between comparable Twitter and Google trends by testing three classes of time series regression models.

Furthermore, Ostrowski (2012) makes semantic social network analysis for trend identification. In other words, the methodology focuses on the utilization of semantics and identifies the influence and power of key players in relevant social networks. Zublaga et al. (2015) also classify the trends based on types of triggers such as news, ongoing events, memes, and commemoratives. Lastly, Stafford and Yu (2013) analyze Twitter trend topics and the effects of spam on Twitter's trending topics.

METHODOLOGY

This section explains how the related data are collected, preprocessed for further analyses, and analyzed.

Data Collection

Data collection includes three steps. In the first step, "GET trends/place" standard Twitter API is used to collect trending topics in Turkey. This API is executed in every 60 seconds iteratively due to API execution limit for a developer. 100 unique trend topics and their creation time are collected between 24th May 2014 and 13th June 2014. Table 1 shows the data structure of the collected trending topics.

Table 1. Data Structure of Trend Table

Data	Description
trend id	unique number for each trending topic.
name	word/phrase/hashtag that becomes a trending topic.
trend creation time	the time when the topic becomes a trending topic.

At the second step, shared tweets for each trending topic are collected by "GET search/tweets" Twitter API. As a result, 301.492 tweets are accumulated. Table 2 shows the data structure of the collected tweets. The text form of the tweet, its creation time, retweet count, and user information are gathered.

Table 2. Data Structure of Twitter Table

Data	Description
tweet id	unique number for each tweet.
tweet	text form of the tweet.
tweet creation time	the time when the tweet is sent.
retweet count	the count how many times a tweet is retweeted by other users.
user id	identification number of the user who sends the tweet.
trend id	identification number of the related trending topic.

In the third step, data about users who shared those tweets are collected by "GET search/tweets" Twitter API. As a result, data for 92.745 unique users are accumulated. Table 3 includes the data structure of user-related information. Users' Twitter usernames, account creation time, and the number of tweets, followers, followings, and favorites are collected.

Table 3. Data structure of user table

Data	Description
user id	unique number for each user.
username	Twitter username of the user.
user creation time	the time when the user account is created.
user favorite count	the number of tweets favorited by the user.
user followers count	the number of users following that user.
user tweet count	the number of tweets that the user shared.
user friend count	the number of users that the user follows.
query	query of the trending topic to define specific users who sent tweets for the given trending topic.

Data Preprocessing

Data are preprocessed before performing any analysis on them. Identification of the time of when the word/phrase/hashtag is created and the time of when it becomes a trending topic is essential. The time of when it becomes a trending topic is collected as "trend creation time" as in Table 1. The time when it is created for the first time is taken as the creation time of the first tweet including that topic. In this sense, a new variable called as "trend time" is derived by calculating these two variables. "Trend time" includes the elapsed time from the creation of the topic to the time when it becomes a trending one. For example; #deprem (#earthquake in English) is one of the trending topics. The first tweet including this hashtag is created on 24th May 2014 at 12:26 and it has become a trending topic on 24th May 2014 at 12:31. "Trend time" shows that #deprem has become a trending topic in 5 minutes.

After that, "tweet count" and "retweet count" variables are calculated. All the tweets for the given trending topic are gathered together. The critical point is that tweets including "RT" (Retweet) in their texts are excluded because they are used for the calculation of "retweet count." Then, tweets posted until "trend time" are counted as "tweet count," and retweets are counted as "retweet count." For example; there are 459 tweets and 395 retweets including the hashtag #deprem between 12:26 and 12:31 before the hashtag becomes a trending topic (see Figure 1).

The next step includes the calculations of the average of users' total tweets, followers, and followings for each trending topic. For example; 489 unique users have written tweets including the hashtag, #deprem, up to 12:31 before it becomes a trending topic. It is essential to collect unique users because a user can send more than one tweet including the same trending topic. These 489 individuals follow 578 users and are followed by 3098 users on average, and send 5223 tweets in total. Table 4 shows the final data structure being analyzed.

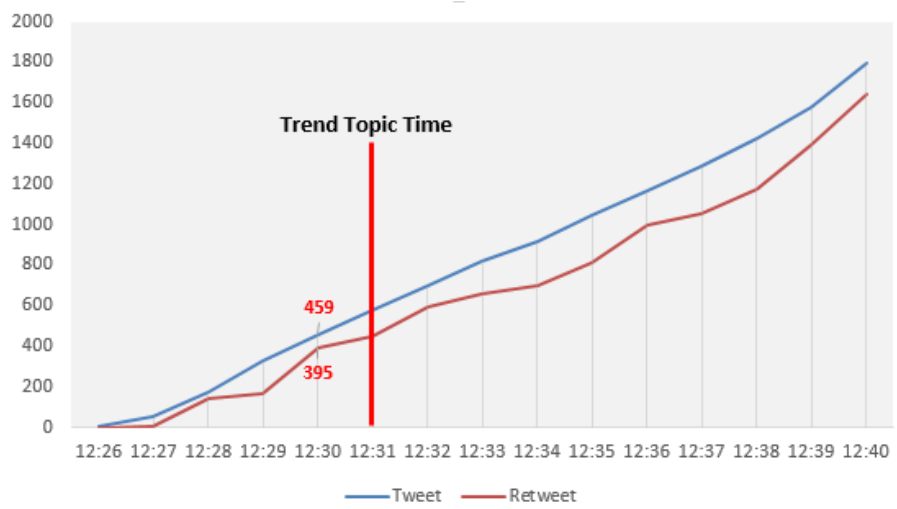

Figure 1. Total tweets and retweets for #deprem before it becomes trend topic

Table 4. Data Structure of The Final Table

Data	Description
hashtag id	unique number for each trending topic.
hashtag name	word/phrase/hashtag that becomes a trending topic.
Date	the date when the trending topic is collected.
Time	the time when the topic becomes a trending topic.
trend time	how much it takes for the topic to become a trending topic.
tweet count	the number of tweets sent by users before the topic becomes a trending topic.
retweet count	the number of retweets sent by users before the topic becomes a trending topic.
average user total tweets	the average number of total tweets shared by users for the trending topic.
average user followings	the average number of total followings of the users tweeted about the trending topic.
average user followers	the average number of followers of the users tweeted about the trending topic.

Data Analysis

To analyze the final data, SPSS 22 is used. Trending topics are clustered by taking the factors of trend time, tweet count, retweet count, average user total tweets, average user followings, and average user followers as shown in Table 4 into consideration. Before the analysis, all variables are standardized to ensure that all of them contribute equally to the similarity between the observations.

Clustering is known as an interdependence technique that variables cannot be classified as independent or dependent variables (Hair et al., 2010). In other words, Hair et al. (2010) state that all variables are examined simultaneously to find an underlying structure to the complete set of variables which is also parallel with the aim of this study. Wald's cluster method is applied to determine the number of clusters. According to Sharma (1996), Ward's method creates clusters by maximizing within clusters homogeneity. It computes the sum of squared distances within clusters and aggregates clusters with the minimum increase in the overall sum

of squares. In other words, this method does not compute distances between clusters and it tries to minimize sums of squares within clusters.

After the determination of the number of clusters, the k-means clustering algorithm that "partitions the observations into a user-specified number of clusters and then iteratively reassigning observations until some numeric goal related to cluster distinctiveness is met" is applied (Hair et al., 2010).

RESULTS

At the first stage, hierarchical clustering analysis is conducted by using agglomerative clustering technique as seen in Figure 2.

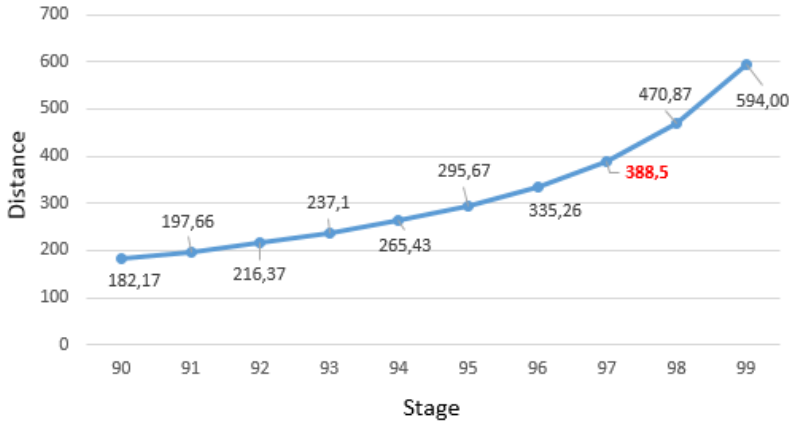

Figure 2. Scree diagram of agglomeration distances

According to Figure 2, it is evident that stage 97 indicates the optimal stopping point for merging clusters, so it is concluded that three clusters are the optimal solution for the given dataset. After this stage, a k-means clustering algorithm is run to obtain three clusters. Table 5 shows the distribution of the observations for each cluster. There are 63, 24, and 13 trending topics in clusters one, two, and three respectively.

Table 5. K-Means Cluster Distribution

Cluster	Number of Cases
1	63
2	24
3	13
Total	100

Table 6 provides ANOVA results for the cluster centers. It is evident that trend time, total tweets, total retweets, average user total tweets, average user total followers, and average user total followings are significant. It indicates that means of all clustering variables differ significantly from each other. Moreover, when F values are considered, it is revealed that average user total followers, average user total followings, and total retweets have the greater F values, respectively. It shows that these variables have the significant influence in the formation of the clusters, whereas trend time with 5.949 F value has the least significant effect.

Table 6. K-Means ANOVA Results

ANOVA						
	Cluster		**Error**		**F**	**Sig.**
	Mean Square	**df**	**Mean Square**	**df**		
Trend time	5.408	2	.909	97	5.949	.004
Total tweets	18.044	2	.649	97	27.820	.000
Total retweets	22.037	2	.566	97	**38.918**	.000
Average User Total Tweets	15.294	2	.705	97	21.684	.000
Average User Total Followers	26.416	2	.476	97	**55.503**	.000
Average User Total Followings	23.371	2	.539	97	**43.382**	.000

Table 7 compares the variables for each cluster. According to Table 7, while topics in the first cluster become trending topics in about 31 minutes, they become trending topics in almost 36 minutes and nearly 54 minutes in the second and third clusters, respectively. One of the outstanding results is that users have more followers and followings in the first and second clusters concerning the third cluster. As it is expected, it indicates that users that have more followers and followings have the power to make a topic as a trending one in between about 31-36 minutes. On the contrary, it takes more time to make a topic as a trending one for users having fewer followers and followings in the third cluster than users having more followers and followings in the first and second clusters.

Furthermore, when three clusters are considered, it is evident that retweeting is more critical than tweeting. It requires more retweets than tweets to be a trending topic for a word/phrase/hashtag in each cluster. Also, considering the average trend time, users that post more tweets or users who are one of the most active Twitter users have more influence to make a topic as a trending one in a short time. Table 7 also indicates that words/phrases/hashtags that become trending topics in a shorter time require fewer tweets than words/phrases/hashtags that

become trend topics in a longer time. The main reason can be that these topics may be diffused more naturally among users. The trending topic algorithm of Twitter may more pay attention to the natural and real contents than the contents that include spam/ad and are shared by bot accounts.

Table 7. Cluster Analysis Results

Cluster	Average Trend Time	Average Total Tweets	Average Total Retweets	Average User Total Tweets	Average User Total Followers	Average User Total Followings
1	31.44	689.47	1171.22	166286.44	76319.71	49332.58
2	36.31	560.86	1686.31	180312.86	85824.02	58670.43
3	53.60	2427.00	5985.20	27430.25	13237.74	12022.79

CONCLUSION

This study analyzes and considers the factors having a role in the creation of trending topics on Twitter. For this purpose, unstructured data from Twitter including 100 trending topics, 301,492 tweets, and 92,745 unique users have been collected over a three-week period and converted into a processable format.

Three clusters are obtained by using Ward's method. After that, cluster analysis is performed by using k-means clustering algorithm. Results indicate that the number of retweets and the number of users' average total followers and their total followings have significant effects on the formation of the clusters. In other words, retweets, the number of followers and followings are vital variables to classify trending topics.

Also, three clusters are compared by considering all related variables. Results reveal that users having more followers and followings have the greatest influence to make the topics as trending ones in a shorter time. For example; when the profiles of the first 20 users who shared tweets including the hashtag #deprem, are examined, they have 332.65 followers, 2,872.95 tweets, and 429.75 followings on average. Besides, as it expected, users having fewer followers and followings on average render a hashtag as a trending topic in a longer time. This result indicates that network of Twitter users plays a significant role to make topics as trending ones.

Retweeting also plays a significant role when the three clusters are compared with each other. It is unexpected that a topic is rendered as a trending topic by more retweeting about the topic than more tweeting about it. It implies that when users begin to retweet the tweets containing the topic, this chain creates an effect on Twitter. In this sense, it can be summarized

that to create a trending topic, a robust social network including more followers and followings, and an organic retweet chain is one of the most critical influential points.

From e-marketers' point of view, they should understand and talk to their consumers in digital platforms such as in Twitter (Linton, 2015). They can offer products and services to their consumers, they can diffuse any product, service, and brand information, and even they can enhance their images on the minds of their consumers, and so they can take advantage of these digital mediums (Chaffey et al., 2006; Sheth and Sharma, 2005; Hutchings, 2012). For example; 70% of consumers use social networking sites to get a product and brand information and to consider other people's recommendations (Kirtiş and Karahan, 2011). In this sense, e-marketers can benefit from trending topics for their brands. Trending topics give insight about what people more care about their lives, the world, politics, marketing, etc. E-marketers can get clues about things such as seasonal trends, purchasing behaviors, or characteristics of users. Additionally, e-marketers can get their brands noticed by creating trending topics and reach lots of their existing and potential consumers. They should pay attention to that contents should be shared organically and retweeted as much as possible by the most active Twitter users. In such a way, they can also start new marketing trends and become highly ranked in front of the eyes of their audiences.

REFERENCES

Ahangama, S. (2014). Use of Twitter Stream Data for Trend Detection of Various Social Media Sites in Real Time. Social Computing and Social Media, 151-159.

Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., ... & Jaimes, A. (2013). Sensing trending topics in Twitter. Multimedia. IEEE Transactions, 15(6), 1268-1282.

Benhardus, J. & Kalita, J. (2013). Streaming trend detection in Twitter. International Journal of Web Based Communities, 9 (1), 122-138.

Burgess, J. & Bruns, A. (2012). Twitter archives and the challenges of "Big Social Data" for media and communication research.0 M/C Journal, 15(5), 1-7.

Chaffey, D., Ellis-Chadwick, F., Johnston, K., & Mayer, R. (2006). Internet Marketing: Strategy, Implementation and Practice. 3rd ed. Pearson, 2006.

Hutchings, C (2012). Commercial use of Facebook and Twitter–risks and rewards. Computer Fraud & Security, 6, 19-20.

Gao, D., Li, W., Cai, X., Zhang, R., & Ouyang, Y. (2014). Sequential Summarization: A Full View of Twitter Trending Topics. Audio, Speech, and Language Processing. IEEE/ACM Transactions, 22 (2), 293 – 302.

Giummolè, F., Orlando, S., & Tolomei, G. (2013). Trending topics on Twitter improve the prediction of Google hot queries. In proceedings of Social Computing (SocialCom),

2013 International Conference, IEEE, 39-44.

Global Data Management Benchmark Report (2017). Retrieved from: https://www.edq.com/resources/data-management-whitepapers/2017-global-data-management-benchmark-report/.

Gobble, M. M. (2013). Big data: The next big thing in innovation. Research-Technology Management, 56(1), 64.

Hair, J.F., Black, W.C., Babin, B.J., and Anderson, R.E. (2009). Multivariate Data Analysis (7th Edition), 2009, Prentice Hall.

Han, S. C., Chung, H., Kim, D. H., Lee, S., & Kang, B. H. (2014). Twitter Trending Topics Meaning Disambiguation. Knowledge Management and Acquisition for Smart Systems and Services, 126-137.

Kirtiş, A.K. & Karahan, F. (2011). To be or not to be in social media arena as the most cost-efficient marketing strategy after the global recession. Procedia-Social and Behavioral Sciences, 24, 260-268.

Kurniati, M.N., Woo-Jong, R., Alam, Md. H., & Lee, S. (2014). Examining the Performance of Topic Modeling Techniques in Twitter Trends Extraction. In proceedings of Information Networking (ICOIN), 2014 International Conference, 10-12 Feb. 2014, Phuket, Thailand, 2014, 364 – 369.

Lau, J., Collier, N., & Baldwin, T. (2012). On-line Trend Analysis with Topic Models: #twitter trends detection topic model online. In proceedings of COLING 2012: Technical Papers, COLING 2012, Mumbai, December 2012, 1519-1534.

Lee, K., Palsetia, D., Narayanan, R., Patwary, M.M.A., Agrawal, A., & Choudhary, A. (2011). Twitter Trending Topic Classification. In proceedings of Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference, 11-11 Dec. 2011, Vancouver, BC, 2011, 251 – 258.

Linton (2017). Six Benefits of Internet Marketing. Retrieved from: http://smallbusiness.chron.com/six-benefits-internet-marketing-31382.html.

Ma, Z., Sun, A., & Cong, G. (2013). On Predicting the Popularity of Newly Emerging Hashtags in Twitter. Journal of the American Society for Information Science and Technology, 64 (7), 1399-1410.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

Naaman, M., Becker, H., & Gravano, L. (2011). Hip and Trendy: Characterizing Emerging Trends on Twitter. Journal of the American Society for Information Science and Technology, 62 (5), 902-918.

Northeastern University. (2016). How Much Data is Produced Every Day. Retrieved from: http://www.northeastern.edu/levelblog/2016/05/13/how-much-data-produced-every-day/.

Omnicore (2017). Twitter by the Numbers: Stats, Demographics & Fun Facts. Retrieved from: https://www.omnicoreagency.com/twitter-statistics/.

Ostrowski, D. A. (2012). Semantic Social Network Analysis for Trend Identification. In

Proceedings of Semantic Computing (ICSC), 2012 IEEE Sixth International Conference, IEEE, 178-185.

Purcell, B. (2013). The emergence of "big data" technology and analytics. Journal of Technology Research, 4, 1-7.

Ribarsky, W., Wang D.X., & Dou, W. (2014). Social media analytics for competitive advantage. Computers & Graphics, 38, 328-331.

Sheth, J.N., Sharma, A. (2005). International e-marketing: opportunities and issues. International Marketing Review, 22 (6): 611-622.

Sharma, S. (1996). Applied Multivarite Techniques, John Wiley & Sons, Inc.S.

Stafford, G., & Yu, L. L. (2013). An Evaluation of the Effect of Spam on Twitter Trending Topics. In Proceedings of Social Computing (SocialCom), 2013 International Conference, IEEE, 373-378.

Song, M., & Kim, M. C. (2013). Real-Time Twitter Trend Mining System. In proceedings of Social Intelligence and Technology (SOCIETY), 2013 International Conference, IEEE, 2013, 64-71.

Wilkinson, D., & Thelwall, M. (2012). Trending Twitter Topics in English: An International Comparison. Journal of the American Society for Information Science and Technology, 63 (8), 1631-1646.

Yang, M. C., & Rim, H. C. (2014). Identifying interesting Twitter contents using topical analysis. Expert Systems with Applications, 41(9), 4330-4336.

Zublaga, A., Spina, D., Martínez, R., & Fresno, V. (2015). Real-time classification of twitter trends. Journal of the Association for Information Science and Technology, 66 (3), 2015, 462-473.