A Literature Review on Twitter Data Analysis

Hana Anber[1*], Akram Salah[2], A. A. Abd El-Aziz[3]

[1] Computer and Information Sciences Department, Institute of Statistical Studies and Research, Cairo University, Giza, Egypt.
[2] Computer Science Department, Faculty of Computers and Information, Cairo University, Giza, Egypt.
[3] Information System and Technology Department, Institute of Statistical Studies and Research, Cairo University, Giza, Egypt.

* Corresponding author. Tel.: +2010 62628474; email: hana.anber@gmail.com
Manuscript submitted December 31, 2015; accepted June 8, 2016.
doi: 10.17706/ijcee.2016.8.3.241-249

Abstract: The widespread and different types of information on Twitter make it one of the most appropriate virtual environments for information monitoring and tracking. In this paper, the authors review different information analysis techniques; starting with the analysis of different hashtags, twitter's network-topology, event spread over the network, identification of influence, and finally analysis of sentiment. Future research and development work will be addressed.

Key words: Big data, data analysis, social media, Twitter.

1. Introduction

The growing phenomena of social media, such as: Facebook, Twitter, Linkedin, and Instgram, with each one has its own characteristics and its usages, are constantly affecting out societies. Facebook, for example, is considered as a social network where everyone in the network has a reciprocated relationship with another one in the same network. The relationship in this case is undirected. Conversely, in Twitter everyone in the network does not necessarily have a reciprocated relationship with others. In this case, the relationship is either directed or undirected.

In this paper, we focus on twitter for data analysis, where twitter is an online networking service that enables users to send and read short 140- character messages called "tweets" [1]. In addition to its publicity, twitter is accessible for unregistered users to read and monitor most tweets, unlike Facebook where users can control the privacy of their profiles. Twitter is also a large social networking microblogging site. The massive information provided by twitter such as tweet messages, user profile information, and the number of followers/ followings in the network play a significant role in data analysis, which in return make most studies investigate and examine various analysis techniques to grasp the recent used technologies.

The rest of the paper proceeds as follows: in Section 2, we discuss various methods used to retrieve twitter data, twitter users rankings, and the network topology. In Section 3, we discuss some techniques used in information diffusion such as the hashtag life cycle, the network toplogy, and the retweet rate. In Section 4, we discuss how other studies gauge the user influence in twitter. In Section 5, we review two approaches for sentiment analysis in twitter, namely "Natural Language Processing" and "Machine Learning". In Section 6, we discuss model evaluation in literature. We conclude in Section 7.

2. Literature Review Methods

To track and monitor different datasets, most studies [2], [3] began with collecting the desired datasets from twitter, and applied filtering techniques to remove redundant data or spam tweets. Then parsed the data into a structured form. Finally analyzed the data. Below we review several types of analyses that most researchers have used.

2.1. Datasets

Analyzing structured data have been widely used. In such case, the traditional Relational Database Management System (RDBMS) can deal with the data. With the increasing amounts of unstructured data on various sources (e.g. Web, Social media, and Blog data) that are considered as **Big Data**, a single computer processor cannot process such huge amount of data. Hence, the RDBMS cannot deal with the unstructured data; a nontraditional database is needed to process the data, which is called NoSQL database.

Most studies focused on tools, such as R (the programming language and the software environment for data analysis). R has limitations when processing twitter data, and is not efficient in dealing with large volume of data. To solve this problem a hybrid big data framework is usually employed, such as Apache Hadoop (an open source Java framework for processing and querying vast amounts of data on large clusters of commodity hardware) [4]. Hadoop also deals with structured and semi-structured data, XML/JSON files, for example. The strength of using Hadoop comes in storing and processing large volume of data, while the strength of using R comes in analyzing the already-processed data.

There are different types of twitter data such as user profile data and tweet messages. The former is considered static, while the latter is dynamic. Tweets could be textual, images, videos, URL, or spam tweets.

Most studies do not, usually, take spam tweets and automatic tweets engines into account as they can, often, affect the accuracy and add noise and bias to analysis results. In [2], the mechanism of FireFox add-on and Clean Tweet filter was employed to remove users that have been on twitter for less than a day and they removed tweets that contain more than three hashtags.

2.2. Data Retrieval

Before retrieving the data, some questions should be addressed: What are the characteristics of the data? Is the data static, such as the profile user information “name, user Id, and bio”; or dynamic such as user’s tweets, and user’s network? Why is the data important? How is the data will be used? And how big the data is? It is important to note that it is easier to track a certain keyword attached to a hashtag rather than a keyword not attached to it.

Twitter-API is a widely used application to retrieve, read and write twitter data. Other studies, as in [5], have used GNU/GPL application like YourTwapperKeeper tool, which is a web-based application that stores social media data in MySQL tables. However, YourTwapperKeeper in storing and handling large size of data exhibits some limitations in using, as MySQL and spreadsheets databases can only store a limited size of data. Using a hybrid big data technology might address such limitations as we suggested above.

2.3. Ranking and Classifying Twitter Users

There are different types of user’s networks; a network of users within a specific event (hashtag), a network of users in a specific user’s account, and a network of users within a group in the network, that is, Twitter Lists. Lists are used to group sets of users into topical or other categories to better organize and filter incoming tweets [6].

To rank twitter users, it is important to study the characteristics of twitter by studying the network-topology (number of followers/ followed) for each user in the dataset. Many techniques have been employed in ranking analysis. In [2], twitter users are ranked by identifying the number of followers by studying the PageRank, and by the retweet rate. In that study, 41.7 million user profiles, 1.47 billion social

relations, and 106 million tweets were used. In [6], a new methodology is introduced to rank twitter users by using the Twitter Lists to classify users into the Elite users (Celebrities, Media news, Politicians, Bloggers, and Organizations) and the Ordinary users.

2.4. Homophily

Homophily is defined as the tendency that contacts among similar users occur at a higher rate than among dissimilar users [2], that is, similar users tend to follow each other. It requires studying the static characteristics of twitter data, such as the profile name and the geographic feature of each user in twitter network. [2], [6] studied the homophily in twitter; [2] studied the geographical feature in twitter to investigate the similarity between users based on their location. Additional work had been investigated in [6], homophily was studied using Twitter Lists to identify the similarity between the elite and ordinary users.

2.5. Reciprocity

The characteristic nature of twitter as being both directed and undirected social network has made most studies analyze reciprocity. Reciprocity is the property of following a user and being followed back (mutual relationship). For instance, celebrities tend to follow each other, so are politicians, bloggers, and ordinary users. From [2], [6] we can conclude that homophily and reciprocity have the same logical behavior. In [2], the reciprocal relationship is measured by analyzing the number of followers, PageRank, and retweet rate. Additional methodology is investigated in [6], where the users follower-graph is studied to infer users reciprocities.

3. Information Diffusion

Since there are different kinds of information spread over twitter, there is no agreement on what kind of information is more widely spread than others. There is also no agreement on how messages are spread over twitter network. In this area many studies have attempted to address those questions by studying the First-network topology, and by measuring the retweet rate [5], [7], [8].

3.1. Event Life Cycle

To analyze the life cycle of an event, it is important to choose the measurements of the life cycle such as measuring the number of tweets over a period of time, and the number of users in the network. In [5], the life cycle of five different hashtags were demonstrated and analyzed by tracking the most uprising political events; 45,535 tweets were collected in #FreeIran, 246,736 in #FreeVenzuela, 195,155 in #Jan25, 31,854 in #SpanishRevolution, and 67,620 in #OccupyWallSt. Analysis showed the frequency of messages over a specific period of time.

Regarding the difficulty of tracking a specific event for a long period of time, [9] followed an effective technique by tracking a specific hashtag on different times and employed a comparison between them to examine the fluctuation of the event life cycle as they investigated three metrics to track each hashtag. The first is the contribution metric to examine the activity and the participation of users over a specific hashtag by counting the number of tweets, and to examine the visibility of each user (which is how many times the user is mentioned by other users). The second is the activity metric to examine the activity and contribution of users over a period of time. And the third metric is to combine both the contribution and the activity of users within a specific hashtag over a period of time. We can suggest that the employed method in [9] would benefit in identifying the influential users when analyzing the network-topology and the retweet rate for the most active and contributed users.

3.2. Network-Topology Analysis

Networks consist of levels of a hierarchal fashion, that is a first-network topology, a second-network topology of the first-network topology, and so on. Most studies have focused on the first-network topology for analyzing information diffusion over twitter. [2], [5], for instance, studied the first-network topology to examine information spread.

In [10], a hybrid methodology had been investigated to analyze the message content, besides analyzing the network-topology, by employing a linear-regression model to predict the speed of message propagation for each crawled hashtag. Furthermore, additional work by [8] has measured the message propagation on-line by studying the first, second, and third-network topologies. As in Fig. 1, for example, if a message *M* propagates through the user U_0, the audiences of U_0 (U_{01}, U_{02}) will receive that message which means that user U_0 is the originator of the message. At this state the message propagates through one hop; that is, in case the message propagates through U_{01}, the audiences of U_{01} will receive the message and at that state the message propagates through two hops and so on until the third hop.

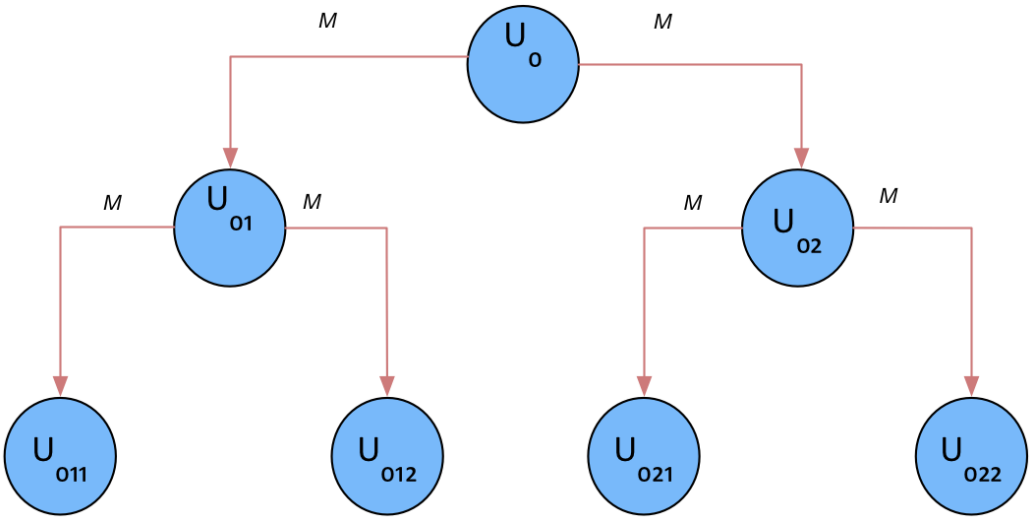

Fig. 1. Example of message diffusion across multiple hops.

3.3. Retweetability

Retweet in twitter is the agreement action to a specific tweet, as in some cases the user passes information to his/her audiences to express their opinion on a particular tweet. The mechanism of retweetability plays a prominent role in information diffusion. [2], [5] studied the retweet rate of the original tweets and the number of mentions related to those tweets to investigate whether the number of retweet and number of mentions are related to the same network-topology. Additional work was done by [2] where the reweetability was studied by deploying two different features the Content (URL and hashtags), and the Contextual feature (age of account and number of followers/ followed) from 74 million tweets.

4. Influence on Twitter

Social influence occurs when an individual's thoughts or actions are affected by other people [8]. Examining the influential users is related by the message propagation by answering on the following questions; who are the originators of the tweets, how many audiences they have, and what is the retweet rate of the original tweet. Most studies agreed on analyzing the network-topology and the retweet rate to identify the influential users. Additional methodology had been used to examine the influence by studying the retweet mechanism through the *"Centrality measures"* technique [11].

Ref. [11] used the *"Degree Centrality"* by counting the number of links attached to the node (user) in case

of directed graph. Also employed the *"Eigenvector Centrality"* by answering the question of *"how many users retweeted this node?"* Furthermore, employed the *"Betweeness Centrality"* which measures the number of the shortest paths to the most important node. As [6], [12] agreed on identifying the influential users by ranking the users using the number of followers, the PageRank, and the retweet rate. Additional method had been employed by [9], which is studying the reply influence metric and identifying the number of replies to the original tweet. In addition to analyze the network-topology, the authors in [13] investigated another methodology by analyzing the number of tweets, the date of joining, and the previous history of the influential users.

5. Sentiment Analysis

Sentiment analysis is the measure of people's opinions on the level of agreement on a specific topic, a product, or a service, or even elections. Two approaches had been employed to study the sentiment analysis: natural language processing, and machine learning algorithms.

To assess the customers' opinions in the past some paper-based surveys had been used, but it is difficult to monitor and collect all customers' opinions. With the increasing phenomena of social media it has become easier and more accessible to crawl all customers' feedbacks and analyze their sentiments as positive or negative.

5.1. Natural Language Processing Approach

According to [14], natural language processing (NLP) is the interaction between computers and human (natural) languages. To evaluate sentiment of users online, particularly on twitter, effective sentiment annotation should be used. Most studies use the three common sentiment labels: positive, neutral, and negative. In [15], new feature had been used to effectively annotate sentiments of users; "Mixed Sentiment label", it exists in tweets that have two different meanings. For example "I love iPhone, but I hate iPad". "iPhone" entity is annotated with positive sentiment label, and "iPad" entity is annotated with negative sentiment label, that means the tweet has a mixed sentiments.

5.2. Machine Learning Approach

According to [16], machine learning (ML) is a scientific discipline that explores the construction and the study of algorithms that can learn from data. Refs. [17]-[20] used the machine learning approach in analyzing the sentiment of twitter users. [19] Applied a rule-based, supervised, and semi-supervised techniques, and collected tweets about the president (Obama) to measure the sentiment of people's opinion towards his job performance. Furthermore, a cross-correlation analysis of time series was investigated to predict sentiments by labeling 2500 tweets to predict the test dataset of 550,000 unlabeled tweets.

A hybrid method had been used by [18] since an advanced classifier was employed for sentiment analysis "The Latent Dirichlet Allocation Model", in which a topic has probabilities of generating various words; first, the implicit topical structure was extracted from the tweets, second, 32 million tweets were analyzed to predict the US presidential election of 2012. Additional work had been used by [20], [17] where additional features were added to the tweet to improve the accuracy of the sentiment classifier. [20] Added the Semantic feature by adding a semantic concept to each entity in the tweet to predict the sentiments for the collected dataset. [17] Added the emoticons feature beside the twitter messages by employing The Distant Supervised Learning algorithm.

6. Literature Review Model Evaluation

Regarding the homophily and reciprocity analysis, [12] found that the top users by the number of followers are mostly celebrities and mass media, and most of them do not follow back their followers. A low

level of reciprocity had been observed; 77.9% of users pairs are connected one-way, and only 22.1% of users have reciprocal relationship between them. [6] Also showed a low reciprocity in their analysis of the follower graph. Roughly 20% of users have reciprocal relationship. Both [6], [12] agreed that twitter is a source of information than a social networking. [5] Found that bloggers spread information more than other categories like celebrities, media, or organizations. [3], [19] found that using hashtags in tweets improves the accuracy and the performance of the analysis. [5], [13] found that the political hashtags persisted more period of time than other ones, which means a higher frequency of tweets over a long period of time.

The period of time for each hashtag must be consistent. For example, when crawling political hashtags, each hashtag should be measured yearly, monthly, weekly, or daily. Unlike [5]'s ambitious, though flawed, analysis where the time breaks for the five hashtags in study have different time breaks, in which they measured #FreeIran and #FreeVenzuela on a yearly basis, #25Jan and #OccupyWallSt on daily basis, and #SpanishRevolution on monthly basis. Therefore, it was important to set a consistent time measure, as the topic category was the same.

Influential users on twitter may not necessarily be politicians, celebrities, or activists, they can be ordinary users, conversely to [5]'s findings. They resembled the activity on twitter to the Pamphleteering action (a historical term for someone who creates or distributes Pamphlets, where pamphlets used to broadcast the writer's opinions [21]). In pamphleteering the political activists keep pamphleets since they are the only influential people.

Ref. [8] found that it is easier to propagate text messages than photo messages. This means that users are concerned more with information sharing than communicating with other users. Also replying to breaking news messages was significantly more than ordinary messages, that is, users discuss and share information and ideas towards a specific topic more than engaging in conversations. That resulted in increasing the network of users in breaking news events.

The analyses in [9] should have been performed to infer the most active and contributing users. However, it would be advantageous if they have identified the retweet rate and the network-topology of the active users to examine their influence, and to address the question of the relation between being active and being influential. Moreover, the methods in [13] lack the conceptual behavior of influence, as the rate of tweets and the date of joining are not indicators of being influential. Also, being influential in the past does not necessarily mean being influential at present or future.

Ref. [2], [19] showed that there is no strong correlation between the retweet rate and the network-topology as a small percentage of retweeted messages and messages with mentions were between interconnected users. [2] Found that in the case of hard-political news (politics, economics, crimes, and disasters) hashtags, the retweet rate was higher between interconnected users. Unlike [2]; [19] found that the network-topology is not the main feature in analyzing retweetability. Additional analysis in [10] showed that the content of messages played a strong role in the message propagation.

Moreover, [18] showed that using the well-known "geo-tagged" feature in twitter to identify the polarity of a political candidates in the US could be done by employing the sentiment analysis algorithms to predict the future events such as the presidential elections results. Comparing to previous approaches in sentiment topics, additional findings by [20] showed that adding the semantic feature produces better Recall (retrieved documents) [22] and F-Score (a measure of a test accuracy as it considers both the precision and the recall of the test to compute the score) in negative sentiment classification [23], see (1), (2), (3). It also produced better Precision (the relevant documents) [24] and F-Score in positive sentiment classification. [6] Found that using machine learning algorithms such as (Naïve Bayes, Maximum Entropy, and SVM) have more accurate results (over 80%) when training the emoticons data along with twitter messages.

Using the weighted F-Measure to measure the accuracy of the sentiment analysis would assist in more accurate results. F_2 measure weighs recall twice as much as precision and $F_{0.5}$ weighs precision twice as much as recall [25]. Ref. [20] Used F-Score to measure the accuracy of their sentiment analysis.

$$\text{Recall} = \frac{\left|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}\right|}{\left|\{\text{relevant documents}\}\right|} \tag{1}$$

$$\text{F1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} \oplus \text{recall}} \tag{2}$$

$$\text{Precision} = \frac{\left|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}\right|}{\left|\{\text{retrieved documents}\}\right|} \tag{3}$$

7. Conclusion

The sheer amount and the different types of data on twitter and the public nature of tweets have allowed exploiting twitter information in data analysis. First, by measuring the life cycle of a specific topic by measuring the number of tweets over a period of time, and second by measuring the sentiment of users towards a specific topic through NLP and ML algorithms. Our aim is to enhance the analysis of twitter data for specific events to measure the effect and the behavior of users towards different events categories. A successive work will focus on studying the data and its attributes, and investigating modeling techniques to identify the frequency distribution for each event.

References

[1] Twitter. From https://en.wikipedia.org/wiki/Twitter

[2] Bastos, M. T., Travitzki, R., & Puschmann, C. (2012). What sticks with whom? Twitter follower- followee networks and news classification. *Proceedings* of *6th International AAAI Conference on Weblogs and Social Media—Workshop on the Potential of Social Media Tools and Data for Journalists in the News Media Industry.*

[3] Hajibagheri, A., & Sukthankar, G. (2014). Political polarization over global warming: Analyzing twitter data on climate change. *Academy of Science and Engineering (ASE), USA.*

[4] Prajapati, V. (2013). *Big Data Analytics with R and Hadoop*. Packet Publishing.

[5] Bastos, M. T., Travitzki, R., & Raimundo, R. (2012). Tweeting political dissent: Retweets as pamphlets in #FreeIran, #FreeVenzuela, #Jan25, #SpanishRevolution and #OccupyWallSt. University of Oxford.

[6] Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Who says what to whom on twitter. *Proceedings of the 20th International Conference on World Wide Web.* ACM New York, NY, USA.

[7] Bongwon, S., Lichan, H., Pirolli, P., & Chi, E. H. (2010). Want to be retweeted? Large scale analytics on factors impacting Retweet in Twitter network. *Proceedings of the 2010 IEEE Second International Conference on Social Computing* (pp. 177-184).

[8] Ye, S., & Wu, F. (2013). Measuring message propagation and social influence on Twitter.com. *International Journal of Communication Networks and Distributed System, 11(1),* 59-76.

[9] Bruns, A., & Stieglitz, S. (2013). Towards more systematic Twitter analysis: Metrics for tweeting activities. *International Journal of Social Research Methodology, 16(2)*, 91-108.

[10] Tsur, O., & Rappoport, A. (2012). What's in a Hashtag? Content based prediction of spread of ideas in microblogging communities. *Proceedings of the Fifth ACM international Conference on Web Search and Data Mining* (pp. 643 – 652). ACM New York, NY, USA.

[11] Kumar, S., Morstatter, F., & Liu, H. (2014). *Twitter Data Analytics*. Springer, New York.

[12] Kwak, H., Lee, C., & Park, H. (2010). What is twitter, a social network or a news media? *Proceedings of the 19th International Conference on World Wide Web.* Raleigh, North Carolina, USA.

[13] Romero, D. M., Medeer, B., & Kleiberg, J. (2011). Differences in the mechanics of information diffusion topics: Idioms, political Hashtags, and complex contagion on twitter. *Proceedings of the 20th International Conference on World Wide Web* (pp. 695-704).

[14] Natural language processing. From https://en.wikipedia.org/wiki/Natural_language_processing

[15] Saif, H., Fernandaz, M., & Alani, H. (2013). Evaluation datasets for twitter sentiment analysis: A survey and a new dataset, the STS-Gold. *Proceedings of 1st International Workshop Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI* (ESSEM 2013). Turin, Italy.

[16] Machine learning. From https://en.wikipedia.org/wiki/Machine_learning

[17] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. Technical Report, Stanford Digital Library Technologies Project.

[18] Jahanbakhsh, K., & Moon, Y. (2014). The predictive power of social media: On the predictability of U.S presidential elections using Twitter. *arXiv preprint arXiv: 1407.0622.*

[19] Johnson, C., Shukla, P., & Shukla, S. (2012). On classifying the political sentiment of tweets. *Cs.utexas.edu.*

[20] Saif, H., He, Y., & Alani, H. (2012). Semantic sentiment analysis of twitter. *The Semantic Web* (pp. 508–524). ISWC.

[21] Pamphleteer. From https://en.wikipedia.org/wiki/Pamphleteer

[22] Recall. From https://en.wikipedia.org/wiki/Precision_and_recall#Recall

[23] F-Score. From https://en.wikipedia.org/wiki/Precision_and_recall#F1_score

[24] Precision. From https://en.wikipedia.org/wiki/Precision_and_recall#Precision

[25] Japkowicz, N., & Shah, K. (2011). *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge: Cambridge University Press.

Hana Anber obtained a high graduate studies degree in computer science in 2013 from the Institute of Statistical Studies and Research, Cairo University, Egypt. She is currently a master's student. Her current research is centered on enhancing the analysis of twitter data for specific events to measure the effect and the behavior of users towards different events categories. Hana's research interest is in social network analysis, big data, and machine learning.

Akram Salah graduated from mechanical engineering and worked in computer programming for 7 years before he got his M.Sc. (85) and Ph.D. degrees from University of Alabama at Birmingham, USA in 1986 in computer and information sciences.

He taught in the American University in Cairo, Michigan State University, Cairo University, before he joined North Dakota State University where he designed and started a graduate program that offers Ph.D. and M.Sc. in software engineering. Dr. Salah's research interest is in data knowledge, and software engineering. He has over than 100 published papers.

Currently, he is a professor in the Faculty of Computer and Information, Cairo University. His current research is in knowledge engineering, ontology, semantics, and semantic web.

A. A. Abd El-Aziz has completed the Ph.D. degree in June 2014 in information science & technology from Anna University, Chennai-25, India. He has received B.Sc., and master of computer science degrees in 1995 and 2006, respectively from Faculty of Science, Cairo University. Now, he is an Assistant Professor in the ISSR, Cairo University, Egypt. He has 12 years' experience in teaching at Cairo University, Egypt. His research interests include database system, database security, XML security, cloud computing, big data, data mining, and social network analysis. He has published about 30 research papers in international conferences and journals.