Webology, Volume 12, Number 1, June, 2015

Analysis of tweets in Twitter

Giovanni Borruto
University of Reggio Calabria, Italy. E-mail: borr.giov@gmail.com

Received October 15, 2014; Accepted June 15, 2015

Abstract

Twitter is a microblogging service that commands more than 288 million monthly active users (in 2015) and is growing fast. Twitter users post short messages called tweet about any topic and follow others to receive their tweets. The great amount of data coming from Twitter users is a meaning source of information regarding different aspects of people life. The goal of this paper is to mine such information through the study of the tweets posted in the time interval of one year. The analysis performed on these data allows us to draw meaningful conclusions about the behavior and preference of Twitter users.

Keywords

Social network analysis; Information; User behavior

Introduction

With the term *Big Data* we refer to a great amount of data difficult to process using traditional data processing applications. Social networks are an important and inexhaustible source of big data, and even the mere extraction of such data from social networks is hard. Knowing the behavior of users when they operate in a social network has attracted and attracts the attention of the scientific community. Indeed, better understanding user behavior is a key issue in several contexts: for example, this allows Internet and OSN providers to guide infrastructural and application-level actions; users themselves to enhance awareness in this potentially insecure world; companies and government institutions to make better use of this huge network of people for their finalities; scientists to better understand individuals and communities.

Among the social networks, Twitter, with its 288 million monthly active users in 2015 (Statista, 2015), is one of the most important generator of data stream thanks to the widely-used Tweets, 140-character messages posted by users to express their opinion on a topic, a situation, a news, and so on. These short messages allow users to keep in touch each other easily and to share opinions. Such opinions have a social weight because they are publicly visible and the post of a tweet may discover several aspects of a person life. Consequently, the analysis of tweets has a great importance both from a sociological perspective and to better understand the interaction between Twitter and its users. However, a mere analysis of the more than 500,000,000 tweets per day (InternetLiveStats, 2015) is too time-expensive and meaningless as information is lost in such enormous amount of data.

In this paper, we present the results of a study done on the tweets posted between April 2013 and March 2014. These tweets have been sampled and pre-processed to discard dirty data. Then, we applied several analysis techniques to obtain some important results. We observe that many analyses about Twitter already exist, which mainly regard statistics of Twitter (for instance, active users, tweet number, bandwidth utilization, gender and age of users, and so on (Statista, 2015; Teevan et al., 2011; Java et al., 2007)). However, none of these statistics can be used to mine aspects of user's behavior, as done in this paper. In particular, our results show different behavior between standard users and power users with respect to:

1. tweet typology;
2. location from which tweets are generated;
3. trend of tweet posting;
4. tweet source;
5. tweet language;
6. tweet frequency vs. user's joining date;
7. tweet frequency vs. user's popularity.

The outcomes of these analyses show important and (sometime) unexpected results.

Materials and methods

Data considered in our analysis are the tweets posted in the time interval from April 1, 2013 to March 31, 2014 (one year). They have been extracted via the Twitter Streaming API v1.1 in the Spritzer version, which is freely available. Data are sampled and about 1% of tweets is considered in the analysis. As they are randomly sampled, we expect that no bias has been introduced in the results of the analysis. Each tweet is represented in JSON format, and contains a lot of information, such as timestamp of the generation, identifier, the tweet content, IP

address, user ID, location coordinates, and so on. An example of a fragment of a JSON tweet is reported below, in which we obscured sensitive information:

{"retweet_count":0,"text":"thanks to all #thankstoall","geo":null,
"retweeted":false,"id_str":"154124867866XXX","source":"href="http:
//labs.XX.com","entities":{"hashtags":[],"urls":[],"user_mentions"
:[]},"contributors":null,"place":null,"created_at":"Tue Jan 02 09:
00:06 +0000 2014","user":{"is_translator":false,"statuses_count":
7505,"profile_image_url":"http://XX/image/XX.png","friends_count"
:5,"profile_sidebar_fill_color":"DDEEF6","listed_count":228,
"profile_sidebar_border_color":"C0DEED","screen_name":"XXX",
"followers_count":96,"created_at":"Wed Jan XX XX:XX:56 +0000 2010",
"time_zone":"Tokyo","id":XXXXXX713,"utc_offset":32400}

Results and Discussion

In this section, we show the results of our analysis. They deal with different aspects and different uses of Twitter. For each experiment, we provide a detailed description of the analysis purpose, the involved data and the results.

1. Tweet Typology

Each tweet can contain simple text as well as media (i.e., photos or videos), hashtags, other users mentions and URLs. Starting from this consideration, we divide the analyzed tweets into four categories depending on the entities they contain. Hence, we count the occurrences of each entity and we compute the percentage of tweets in each category. In Figure 1, we show the results.

Observe that, all tweets contain text. The most used entities are media (5% of the total amount of tweets). As for mentions, they are used in more than 2% of the total number of tweets. By contrast, hashtags are less utilized than mentions (0,77%). Probably this is because mentions are newer features than hashtags and, for this reason, users are showing a growing interest for them. Finally, URLs are the less used entities in Twitter.

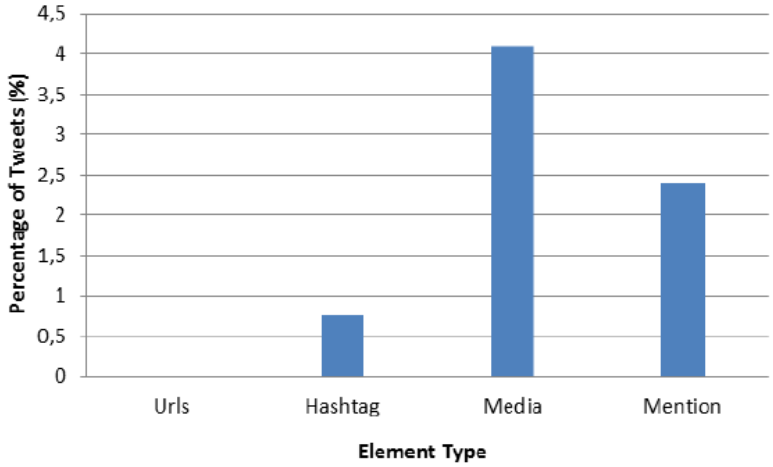

Figure 1. Typology of tweets

2. Tweet Origin

In this experiment, we aim to find the most frequent locations from where tweets are posted. We examine the nation of the user who publishes the tweet. Then, we group the different nations into geographical zones. Note that only a small percentage of users declare their nation in their profile or allow Twitter to localize their tweets automatically. This is probably due to both the lack of explicitly user consent and the unavailability of localization services at the moment in which tweets are published. The results are illustrated in Figure 2.

As attended, USA have the record of published tweets (about 35% of the total amount), whereas Europe is the second area of tweets production with the 25%. This result can be justified by considering that Twitter, together with Facebook, is one of the most used social networks in both USA and Europe.

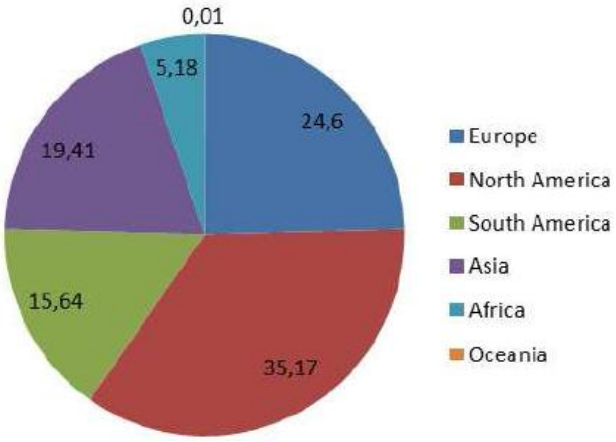

Figure 2. Location of the tweet posting

3. Tweet Trend

This experiment focuses on the period of tweet creation for each day in the period from April 1, 2013 to March 31, 2014. The graphic in Figure 3 depicts the result of this analysis.

The resulting curve describes a particular trend. Indeed, it presents some peaks in correspondence to summer months (i.e., June, July and August) and the end of December (Christmas Holidays). Hence, we can argue that these months are those of major Twitter users' activity. These increased frequency of tweets publishing can be due to different factors, for instance, more free time or a higher number of advertising campaigns.

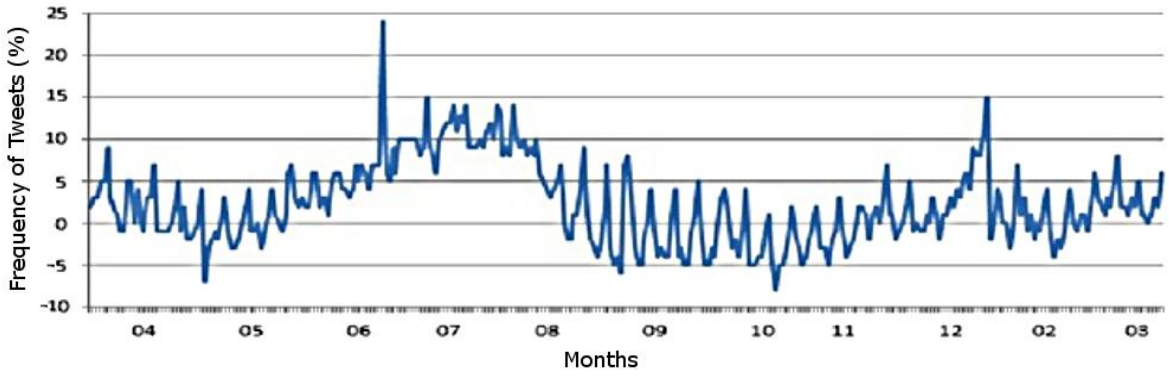

Figure 3. Frequency of tweet publishing from April 2013 to March 2014

4. Tweet Source

Now we consider tweet sources, that is the devices used to send a tweet. To carry out this experiment, we count tweets according to the value of the tweet source. In Figure 4, we show the results of this analysis.

We can observe that the higher percentage of tweets come from Apple IPhone devices. This type of device has the Twitter application installed by default and natively handled by the IOS system. The following three most used sources are Web browsers, Android smartphones and Blackberry smartphones.

It is also possible to send tweets from Facebook via a specific application; this possibility was preferred by the 1,34% of analyzed users. Others interesting sources are Twittbot.net and TweetDeck. The former is a Japanese Web site, which allows multiple users to post to a single Twitter account, and a single user to post to multiple Twitter accounts, whereas the latter is a social media dashboard American application for management of Twitter accounts. Finally, we denote by others all sources with a percentage of use less than 1% (e.g., Windows Phones, UberSocial, Janetter o Echofon, Istagram, Tumblr and Ask.fm).

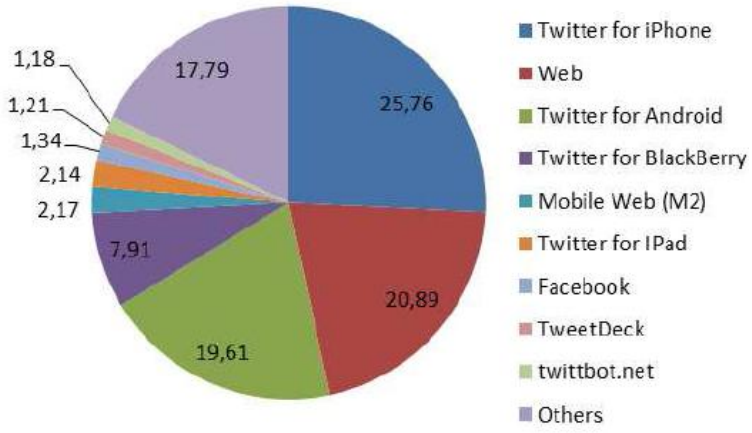

Figure 4. Tweet sources

5. Tweet Language

In this experiment, we put our attention to the most used languages in which the analyzed tweets are written. To do so, we classify tweets depending on the language used. In Figure 5, we present the result of this experiment. It is not surprising that English turns out to be the most used language, followed by Japanese and Spanish. These languages are, indeed, the most spoken in the whole world.

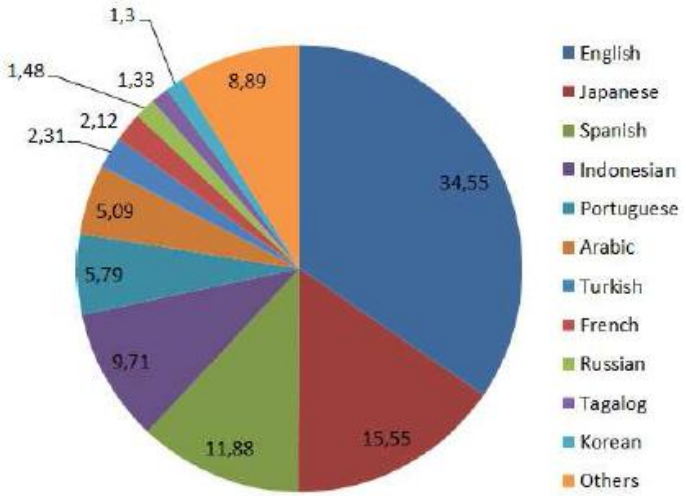

Figure 5. Languages of tweets

6. Tweet vs. User's Joining Date

The aim of this experiment is the estimation of the percentage of tweets classified according to the joining date of the user who sent them (i.e., the date of creation of the twitter account from

which the tweet has been published). For this analysis, we consider years from 2006 to 2013. Figure 6 shows the results of such analysis.

We consider *active* a user who sent at least a tweet per year of life. Figure 6 describes a particular trend. Users who create an account before 2008 are no more active, whereas users newly registered are active users and the more they are younger on the social site, the higher the frequency of tweets they send.

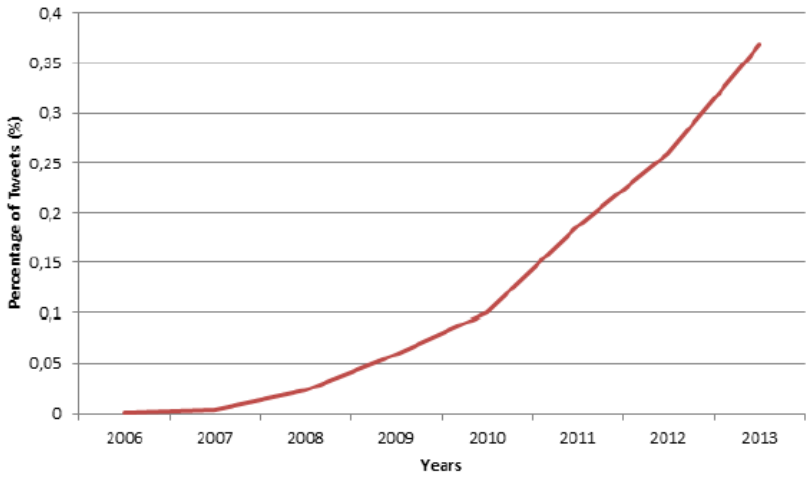

Figure 6. Number of tweets vs. registration year

7. Tweet vs. User's Popularity

This experiment focuses on the rate of tweets per users grouped according to their popularity on the Social Network. User popularity depends on many factors, for instance the number of friends and followers, the number of retweets or the number of mentions. For this analysis, we consider both the number of friends and that of followers of a user to estimate his popularity.

We draw four different levels of popularity, which are:

- Top Popularity: user with more than 4000 friends and more than one million of followers;
- High Popularity: user with more than 4000 friends and more than 5000 followers;
- Mean Popularity: user with a number of friends between 1000 and 4000 contacts and a number of followers from 1000 to 5000 contacts; and
- Low Popularity: user with less than 1000 friends and less than 1000 followers.

In Table 1, we list the number of users grouped according to popularity. In order to carry out this experiment, first we measure the number tweets created by these users. Then, we compute the percentage of tweets sent by these users in average. The results are shown in Figure 7.

We can observe that top users have published more tweets than the others have. Hence, the more a user is popular, the more his activity level in sending tweets. Although top users represent a very small percentage of Twitter users, they generate more than 80% of Twitter traffic. This trend reflects the trend known as power law (Mislove et al., 2007).

Table 1. Number of users per different levels of popularity

analyzed users	2.5millions
top popularity users	29
high popularity users	21000
mean popularity users	115000
low popularity users	> 2millions

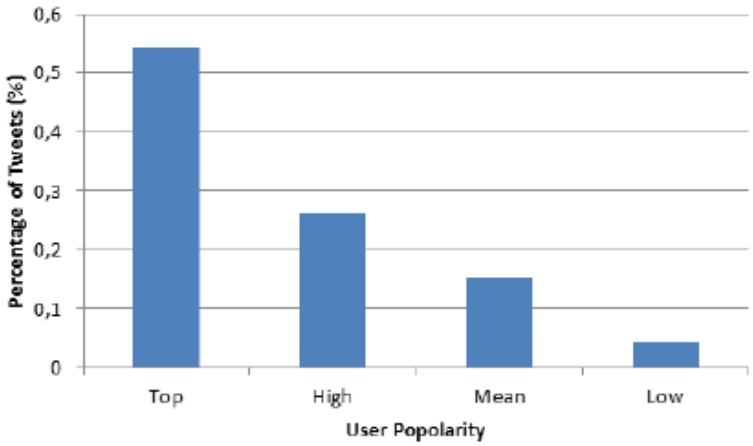

Figure 7. Number of tweets vs. user's popularity

Related Work

In this section, we briefly discuss the literature devoted to the analysis of the behavior of users when they operate in a social network (Buccafurri et al., 2014b; 2014a; 2014c). An analysis of user behavior for OSN workloads forecasting is presented in (Benevenuto et al., 2009). The studies done in (Gill et al., 2007; Cha et al., 2007; Maia et al., 2008; Cheng et al., 2008) focus on the generated content of YouTube users. The issue of the age differences and similarities of users with respect to their activities in MySpace is discussed in (Pfeil et al., 2009). A study on the nature of Facebook is proposed in (Ross et al., 2009), which is based on a sample of undergraduate students. The authors of (Gyarmati & Trinh, 2010) try to characterize user activities and usage patterns in some popular OSNs like Bebo, MySpace, Netlog, and Tagged.

Structural properties of the friendship network of three major systems are compared in (Ahn, et al., 2007). In particular, the authors analyse sample networks from Cyworld, Orkut, and MySpace in terms of degree distribution, clustering coefficient, degree correlation, and average path length.

Several studies have looked at the comparison of the behavior of users among different OSNs (Gyarmati & Trinh, 2010; Zhao et al., 2011; Shen et al., 2013; Forouzandeh et al., 2014; Jalalimanesh & Yaghoubi, 2013; Fogg & Iizawa, 2008; Gao et al., 2012; Buccafurri et al., 2014a; 2012; 2013; 2014d; 2014e; 2015a; 2015b).

A comparative study about the structural properties of social sites is described in (Mislove et al., 2007). In particular, the authors perform a large-scale measurement study and analysis of the structure of four major systems in which they have confirmed the power-law, small world and scale-free properties of the services.

In (Shen et al., 2013), the authors collect objective, privacy-preserved behavior data from user that are active in both Facebook and Gmail. The authors make a comparative analysis on user behavior in OSNs and their way of using email services.

Specific analyses of Twitter are presented in (Teevan et al., 2011; Java et al., 2007). In particular, (Teevan et al., 2011) provides a deep analysis of large-scale query logs and supplemental qualitative data, whereas (Java et al., 2007) focuses on the study of the topological and geographical properties.

From this point of view, our paper provides an original analysis on the use of tweet, which represents an important step head in understanding the behavior of Twitter users.

Conclusions

Twitter is perhaps the most influential online social media platform and its analysis is a great challenge for the research community. In this paper, we carried out a study on the tweets posted between April 2013 and March 2014. Due to the huge amount of data, tweets have been sampled and pre-processed to discard dirty data. Then, we applied several analysis techniques to infer significant results. Our analyses are not only of statistical type, as they concern aspects of user's behavior. We focused on tweet typology, location and language of users, frequency of tweeting, and found important relations between user activity and its time of Twitter registration.

References

Ahn, Y.-Y., Han, S., Kwak, H., Moon, S., & Jeong, H. (2007). Analysis of topological haracteristics of huge online social networking services. *In Proceedings of the 16th international conference on World Wide Web* (pp. 835-844).

Benevenuto, F., Rodrigues, T., Cha, M., & Almeida, V. (2009). Characterizing user behavior in online social networks. *In Proceedings of the 9th ACM SIGCOMM Conference On Internet Measurement Conference* (pp. 49–62).

Buccafurri, F., Foti, V., Lax, G., Nocera, A., & Ursino, D. (2013). Bridge analysis in a social internetworking scenario. *Information Sciences*, 224(1), 1-18.

Buccafurri, F., Lax, G., Nicolazzo, S., Nocera, A., & Ursino, D. (2013). Measuring betweenness centrality in social internetworking scenarios. *In On the move to meaningful internet systems: OTM 2013 workshops* (pp. 666–673).

Buccafurri, F., Lax, G., Nicolazzo, S., Nocera, A., & Ursino, D. (2014c). Driving global team formation in social networks to obtain diversity. In *Proceedings of the international conference on web engineering (ICWE 2014)* (pp. 410-419). Toulouse, France.

Buccafurri, F., Lax, G., Nocera, A., & Ursino, D. (2012). Discovering links among social networks. In *Proceedings of the European conference on machine learning and principles and practice of knowledge discovery in databases (ECML PKDD 2012)* (pp. 467-482). Bristol, United Kingdom: Lecture Notes in Computer Science.

Buccafurri, F., Lax, G., Nocera, A., & Ursino, D. (2014a). Moving from social networks to social internetworking scenarios: The crawling perspective. *Information Sciences*, 256, 126-137.

Buccafurri, F., Lax, G., Nocera, A., & Ursino, D. (2014b). A system for extracting structural information from social network accounts. *Software: Practice and Experience*, 45(9), 1251-1275.

Buccafurri, F., Lax, G., Nicolazzo, S., & Nocera, A. (2014d). A model to support multi-social-network applications. *In Proceedings of the International Conference on Ontologies, Databases and Applications of SEmantics (ODBASE '14)* (pp. 639-656), Amantea, Italy, 2014.

Buccafurri, F., Lax, G., Nicolazzo, S., & Nocera, A. (2014e). Generating K-Anonymous Logs of People-Tracing Systems in Surveilled Environments. *In Atti del Ventiduesimo Convegno Nazionale su Sistemi Evoluti per Basi di Dati (SEBD'14)* (pp. 37-44), Sorrento Coast, Italy, 2014.

Buccafurri, F., Lax, G., Nicolazzo, S., & Nocera, A. (2015a). Comparing Twitter and Facebook User Behavior: Privacy and other Aspects. *Computers in Human Behavior*, 52, 87-95.

Buccafurri, F., Lax, G., Fotia, L., Nicolazzo, S., & Nocera, A (2015b). A lightweight electronic signature scheme using Twitter. *In Proceedings of 23rd Italian Symposium on Advanced Database Systems (SEBD 2015)*, Gaeta, Italy, 2015.

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., & Moon, S. (2007). I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *Proceedings of the 7th acm sigcomm conference on internet measurement* (pp. 1–14).

Cheng, X., Dale, C., & Liu, J. (2008). Statistics and social network of YouTube videos. In Quality of service. *16th International Workshop on Quality of Service (IWQoS 2008)*. June 2-4, 2008. University of Twente, Enschede, The Netherlands. (pp. 229–238).

Fogg, B., & Iizawa, D. (2008). Online persuasion in Facebook and Mixi: A cross-cultural comparison. In *Persuasive technology* (pp. 35-46).

Forouzandeh, S., Soltanpanah, H., & Sheikhahmadi, A. (2014). Content marketing through data mining on Facebook social network. *Webology*, 11(1), article 118. Retrieved December 10, 2014, from http://www.webology.org/2014/v11n1/a118.pdf

Gao, Q., Abel, F., Houben, G.-J., & Yu, Y. (2012). A comparative study of users microblogging behavior on Sina Weibo and Twitter. *In User modeling, adaptation, and personalization* (pp. 88–101).

Gill, P., Arlitt, M., Li, Z., &Mahanti, A. (2007). YouTube traffic characterization: A view from the edge. In *Proceedings of the 7th acm sigcomm conference on internet measurement* (pp. 15-28).

Gyarmati, L., & Trinh, T. A. (2010). Measuring user behavior in online social networks. *Network, IEEE*, 24(5), 26-31.

InternetLiveStats. (2015). Twitter usage statistics. Retrieved December 10, 2014, from http://www.internetlivestats.com/twitter-statistics/

Jalalimanesh, A., & Yaghoubi, S. M. (2013). Application of social network analysis in interlibrary loan services. *Webology*, 10(1), article 108. Retrieved December 10, 2014, from. http://www.webology.org/2013/v10n1/a108.html

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th webkdd and 1st sna-kdd 2007 workshop on web mining and social network analysis* (pp. 56-65).

Maia, M., Almeida, J., & Almeida, V. (2008). Identifying user behavior in online social networks. In *Proceedings of the 1st Workshop on Social Network Systems* (pp. 1-6).

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement* (pp. 29-42).

Pfeil, U., Arjan, R., & Zaphiris, P. (2009). Age differences in online social networking–a study of user profiles and the social capital divide among teenagers and older users in Myspace. *Computers in Human Behavior*, 25(3), 643–654.

Ross, C., Orr, E. S., Sisic, M., Arseneault, J. M., Simmering, M. G., & Orr, R. R. (2009). Personality and motivations associated with Facebook use. *Computers in Human Behavior*, 25(2), 578-586.

Shen, J., Brdiczka, O., & Ruan, Y. (2013). A comparison study of user behavior on Facebook and Gmail. *Computers in Human Behavior*, 29(6), 2650-2655.

Statista. (2015). The statistics portal. Retrieved December 10, 2014, from http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Teevan, J., Ramage, D., & Morris, M. R. (2011). # twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth acm international conference on web search and data mining* (pp. 35-44).

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval* (pp. 338-349).

Bibliographic information of this paper for citing:

Borruto, Giovanni (2015). "Analysis of tweets in Twitter." *Webology*, 12(1), Article 131. Available at: http://www.webology.org/2015/v12n1/a131.pdf