

Sentiment Analysis and Topic Extraction of the Twitter Network of #Prayforparis

Miyoung Chong

College of Information

University of North Texas, Denton, TX

chongmiyoung@gmail.com

ABSTRACT

Social media includes a copious amount of sentiment-embodied sentences. Sentiment is described as “a personal belief or judgment that is not founded on proof or certainty,” which may depict the emotional state of the user, such as happy, glad, terrified, miserable, or the author’s viewpoint on a topic. In social science, emotions and sentiment make up a significant part of social life and are interconnected with social relationships. When experiencing emotions, people want to reveal those emotions to other people. This study seeks to validate whether the Emotional Contagion social theory holds true in microblogging data. This theory implies that related people tend to have more similar sentiments or opinions. Motivated by this sociological observation, the study explores the sentiment-semantics of the Twitter network of #prayforparis through sentiment analysis and topic extraction. Social Network Analysis was conducted using NodeXL to investigate the research questions. The study implemented R for conducting sentiment analysis and generating word clouds with the collected data. The study also conducted content analysis of tweets through topic extraction by applying the most recent version of SAS Enterprise Miner (13.2). In conclusion, the results confirmed the Emotional Contagion Theory in the Twitter network of #prayforparis.

Keywords

sentiment analysis, text mining, social media, emotional contagion theory, hashtag, Twitter, semantic analysis, R, word clouds, #prayforparis

INTRODUCTION

Web 2.0 users can participate, collaborate, and create virtual online communities. Microblogs, such as Twitter, became user-generated information-abundant resources,

because users began sharing information and opinions in diverse online events and domains, including breaking news, celebrity gossip, product reviews, and discussions about recent incidents – the Orlando massacre in June 2016 was a popular discussion topic, for example. On Twitter, which has 310 million monthly active users (Twitter.com), users vividly exchange information through tweets that are up to 140 characters in length. Twitter can deliver messages in real time from anywhere and anyone despite the diversified user backgrounds such as varying religion or political inclination and the level of education or civilization.

One of most apparent characteristics of tweets are the networked user relationships. Connections among the messages can be demonstrated through a “user-message matrix” and a “user-user interaction matrix” except the content information (Hu, Tang, Tang, & Liu, 2013, p. 537). On Twitter, information or opinions about news and incidents are predominantly organized and shared through hashtags with the # symbol. The hashtag is a community-driven practice for putting on another factors and metadata to tweets by promoting folksonomy. Hashtags are generated by users as a method to classify information and highlight topics. Through substantial adoption of hashtags, Twitter became more expressive of and popular with users (Wang, Wei, Liu, Zhou, & Zhang, 2011).

Social media includes a copious amount of sentiment-embodied sentences. Sentiment is described as “a personal belief or judgment that is not founded on proof or certainty” (WordNet 2.1 definitions), which may depict the emotional state of the user, such as happy, glad, terrified, miserable, or the author’s viewpoint on a topic – for example, “A is wonderful” or “I dislike B,” etc. Sentiment analysis, also called opinion mining, aims to discover whether the polarity of a textual corpus, a collection of written texts, leans towards positive, negative, or neutral sentiments.

In social science, emotions and sentiment make up a significant part of social life and are interconnected with social relationships. When experiencing emotions, people want to reveal those emotions to other people (Keltner & Bonanno, 1997). Furthermore, people are likely to receive others’ emotions through non-verbal reactions, which have

been defined as “emotional contagion” in social sciences (Hatfield, Cacioppo, & Rapson, 1994). Emotional contagion can be important for relationships between individuals, because “it fosters behavioral synchrony and the tracking of the feelings of others moment-to-moment even when individuals are not explicitly attending to this information” (Hatfield et al., 1994). Fowler and Christakis reported the diffusion of happiness – a form of emotional contagion – in a social network (Fowler & Christakis, 2008). Recently, researchers described the appearance of sentiment diffusion (Miller, Sathi, Wiesenthal, Leskovec, & Potts, 2011) in online networks of social media based on the theory of emotional contagion between companions. The results of analysis confirm that the Emotional Contagion theory understood through the perspective of sentiment (Hatfield et al., 1994) could be helpful for sentiment analysis. The theory reveals that the sentiments of two messages posted by friends are more likely to be similar than those of two randomly selected messages, and this conclusion is derived from offline surveys and conversations.

This study seeks to validate whether the Emotional Contagion social theory holds true in microblogging data (Hu et al., 2013). This theory implies that related people tend to have more similar sentiments or opinions. Motivated by this sociological observation, the study explores the sentiment-semantics of the Twitter network of #prayforparis through sentiment analysis and topic extraction. The purpose of this study is to discover the answers of the following research questions:

RQ 1. How does the Twitter network of #prayforparis demonstrate the emotional contagion among connected users?

RQ 2. What are the most prevalent topics in the Twitter network of #prayforparis?

The main contribution of this paper can be summarized as follows: this study analyzed a world event organized by the hashtag feature in a Twitter network and investigated the Emotional Contagion theory through sentiment analysis, using the most recent version of R (3.3.1) for the sentiment analysis and word cloud and latent semantic analysis to analyze contents of the network and provided potential application and methodological challenges by adopting a novel method, SAS enterprise miner (SAS EM), for topic extraction in the Twitter network #prayforparis. The remainder of this paper is features a discussion of related works followed by the presentation of data collection and evaluation measures, and in the final section, results are provided.

RELATED WORKS

Extensive studies have been conducted regarding the various aspects of Twitter networks, including, for example, its network topology, the relationships and kinds of messages among users, the information dissemination in the Twitter network, the trustworthiness of information, and its

potential as a predictor of sentiment. Also, Twitter’s influence based on the semantic differences regarding the tenacity of a hashtag has been discussed.

Sentiment analysis can be conducted at different levels, such as the document, sentence, word, or feature level. In the case of Twitter, word level analysis best fits that environment, which allows users to exchange limited characters of information. Cross-examination of the literature verifies that the evaluation measures of mechanically-annotating sentiment at the word level are categorized into two areas: “dictionary-based approaches” and “corpus-base approaches” (Kumar & Sebastian, 2012, p. 372).

Recently, researchers analyzed Twitter activities and reported that more than 80 percent of users either post status updates to their followers or spread information regarding their daily experiences (Thelwall, Buckley, & Paltoglou, 2011). The study links these to the group of communications organized by the hashtag and uses text-mining techniques to provide semantic characteristics of the #prayforparis network.

DATA COLLECTION

In order to investigate the research questions above, SNA was conducted using NodeXL (Hansen, Shneiderman, & Smith, 2011). NodeXL can create visual network diagrams of collections of actors (vertices), estimate the network impact (e.g., betweenness centrality or page rank) of a single actor on others, and retrieve information on a large scale (Hansen et al., 2011). Tweets were collected by applying #prayforparis through the Twitter application programming interface (API) at 11 am on November 17, 2015, which was four days after the incident.

In this study, a total of 19,592 tweets (vertices) generated a total of 20,295 edges (relations between tweets). By using NodeXL, four different types of Twitter edges, which include retweet, replies- to, mention, and tweet, and following and follower relationships among users, were extracted. The collected Twitter data was unstructured and not suitable for analysis. Therefore, the data was prepared through a cleaning process that removed unnecessary parts of the data such as unusual symbols.

METHOD

The study implemented R for conducting sentiment analysis and generating word clouds with the collected data. R is a functional language for computers and environment for statistical calculation, visual graphics, and unified chambers of software provisions for data manipulation (The R foundation, 2016). The study used the twitterR, the plyr, stringr, ROAuth and the ggplot2 libraries in the R software package version 3.3.1 to conduct the quantitative sentiment score. Word clouds were also created to learn the dominant and emerging topics of the Twitter network of #prayforparis.

The study also conducted content analysis of tweets through topic extraction by applying the most recent version of SAS Enterprise Miner (13.2). SAS EM employs latent semantic analysis (LSA). LSA was introduced in 1990 in the Journal of the American Society for Information Science (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990).

LSA was defined as “as new method for automatic indexing and retrieval” that takes “advantage of implicit higher-order structure in the association of terms with documents (‘semantic structure’) in order to improve the detection of relevant documents on the basis of terms found in queries.” LSA is based on the premise that meaningfully related words will appear in similar text, and subjects or topics can be extracted from the text and a statistical instrument that calculates data with an almost identical method for factor analysis (Evangelopoulos, Ashton, Winson-Geideman, & Roulac, 2015). After the sentiment analysis and the creation of word cloud through R, the study manually filtered the resulting tweets to obtain a set of 10 meaningful topics through the process of text parsing, text filtering, text clustering, and topic extraction in SAS EM.

RESULTS

Figure 1 displays sentiment score of all tweets in the network of #prayforparis. The histogram presents an asymmetrical graph by indicating more negative sentiment. In the previous analysis of the network, Justin Bieber had the largest number of indegree, which illustrates his popularity, and his indegree (4,196) solely occupies more than 20% of the edge relationships (20,295) in the entire network. Justine Bieber also holds highest betweenness centrality with 51,772,745.861, which means he is located in the most centralized position and reinforces the social connection within the network of #prayforparis.

On November 16, 2015, Justin Bieber posted a tweet about the Paris attack and the loss of his friend, Thomas. His tweet was retweeted more than 50,000 times and received 77,000 likes in the few weeks after his post. By June 2016, the tweet had been retweeted more than 51,000 times and received 81,000 likes. Therefore, it is not surprising that figure 2 displays highly negative sentiment. The Word Cloud of Justin Bieber’s ego network demonstrated the dominant character of Justin Bieber and his related tweets by following the connections with his fans and fan websites.

On the other hand, figure 3 presents an interesting sentiment score. Though the Paris terrorist attack in 2015 was one of the most horrific incidents in Paris, including the Charlie Hebdo shooting, the sentiment scores present strongly neutral and a balanced graph image. The Word Cloud of all tweets with the exclusion of tweets in Justin Bieber’s ego network clearly presents a nearly even distribution of topics when Justin-Bieber-related tweets are excluded.

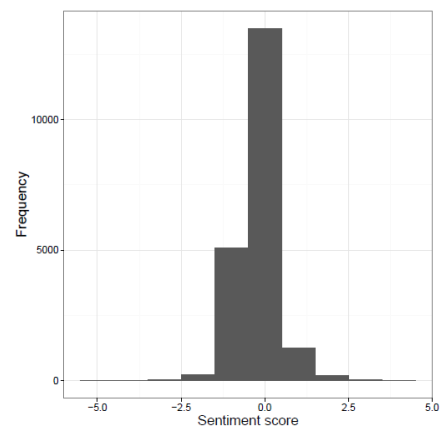


Figure 1. Sentiment score of all tweets

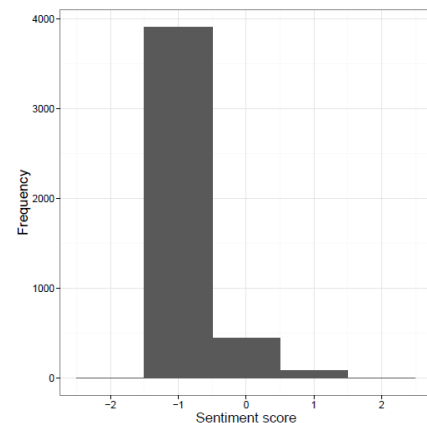


Figure 2. Sentiment score of tweets of Justin Bieber's ego network

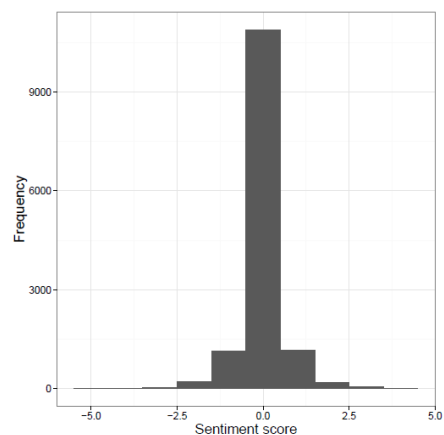


Figure 3. Sentiment score of all tweets with the exclusion of tweets in Justin Bieber's ego network

The procedure of topic extraction in SAS EM includes text parsing, text filtering, and text clustering, and ten topics were extracted from the corpus. The topics provide important detailed information discussion content in the Twitter network of #prayforparis. Figure 5 shows that a large number of users also talked about the massive French fighter jet attack targeting Islamic State’s stronghold in Syria. The last figure describes the number of documents by topics with the exclusion of tweets in Justin Bieber’s ego

network. Topic ID 8 is about the powerful tribute to the victims of the attack during the soccer game between England and France on November 17, 2015.

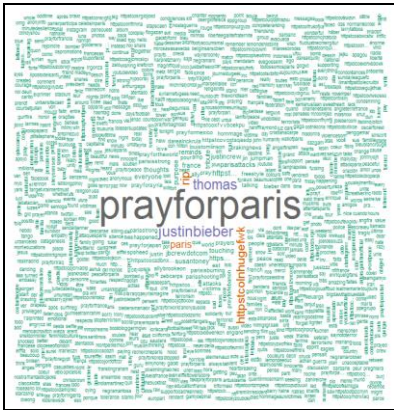


Figure 4. Word Cloud of all tweets

Topic
https://t.co/l4nhugefwk ,rip,thomas,justinbieber,rt
purpose,justincrew,+remind,susantoney,jbcrewdotcom
https://t.co/vcqtqsasdp ,allybrookeon,talking,+ally,+happen
louis_tomlinson,paris,+pray,+prayer,rt
freestyle,zebcarps215,spsheed,darealrickrula,jumpman
parisisburning,ajmee,rt, https://t.co/l4nhugefwk ,rip
syria,+time,raqqa,bombed,allied
kepada,umat,dp,bendera,umat islam yang
+pray, https://t.co/jqt8o5cow ,victim,rxhtba,john
de,à,le,pour

Figure 5. The ten topics from all tweets

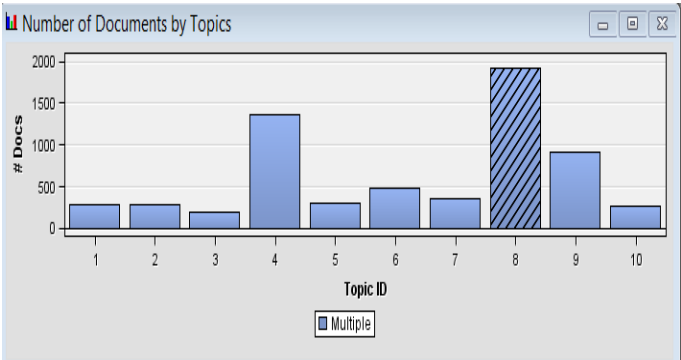


Figure 6. Number of documents by topics with the exclusion of tweets of Justin Bieber's ego network

CONCLUSION

In conclusion, the results confirmed the emotional contagion theory in the Twitter network of #prayforparis, because the grief Justin Bieber shared and diffused through his online community of fans and fan websites strongly connected followers and following relationships on Twitter. In addition, the most prevalent topics of the network included condolences to the victims of the attack (e.g. Justin Bieber's friend), John (Lennon), peace, love, pray, and Paris is burning.

REFERENCES

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6), 391.

Evangelopoulos, N., Ashton, T., Winson-Geideman, K., & Roulac, S. (2015). Latent Semantic Analysis and Real Estate Research: Methods and Applications. *Journal of Real Estate Literature*, 23(2), 353-380.

Fowler, J. H., & Christakis, N. A. (2008). Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. *Bmj*, 337, a2338.

Hansen, D.L., Shneiderman, B., & Smith, M.A. (2011). *Analyzing social media networks with NodeXL: Insights from a connected world*. Boston, MA: Morgan Kaufmann.

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1994). *Emotional contagion*. Cambridge university press.

Hu, X., Tang, L., Tang, J., & Liu, H. (2013, February). Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the sixth ACM international conference on Web search and data mining* (pp. 537-546). ACM.

Keltner, D., & Bonanno, G. A. (1997). A study of laughter and dissociation: distinct correlates of laughter and smiling during bereavement. *Journal of personality and social psychology*, 73(4), 687.

Kumar, A., & Sebastian, T. M. (2012). Sentiment analysis on twitter. *IJCSI International Journal of Computer Science Issues*, 9(4), 372.

Miller, M., Sathi, C., Wiesenthal, D., Leskovec, J., & Potts, C. (2011, July). Sentiment Flow Through Hyperlink Networks. In *ICWSM*.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, 62(2), 406-418.

Wang, X., Wei, F., Liu, X., Zhou, M., & Zhang, M. (2011, October). Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 1031-1040). ACM.