TWITTER FEEDS SENTIMENT ANALYSIS AND VISUALIZATION

ARLENE R. CABALLERO¹, JASMIN D. NIGUIDULA² & JONATHAN M. CABALLERO³

¹ College of Technology, Lyceum of the Philippines University, Philippines

^{2,3} College of IT Education, Technological Institute of the Philippines, Philippines

ABSTRACT

Twitter contains enormous amount of data whether structured, semi-structured or unstructured which analysts and decision makers in an institution hardly interpret and describe (Sheela, 2016). The study explored the Twitter profile of a select institution in terms of frequency of posts, number of user interaction, and intensity of trending topics. It focused on Twitter sentiment analysis which classifies the polarity of the twitter comments into negative, positive, and neutral categories. Sentiments were visualized using Sentiment Visualizer as a tool which represents opinion in an emotional scatter diagram mapped with pleasure and stimulation. The methods applied include the use of various twitter analytics and twitter sentiment analysis tools such as Twitonomy, Rapid Miner and Tweet Viz. These tools were performed to over 1,500 tweeter feeds from a select institution user from year end of 2011, until mid of year 2016. It reveals that the intensity of trending topics becomes higher, because other users considered the select institution tweets as a basis of information. The twitter sentiments of the select institution in terms of sentiment polarity turned out to be majority positive while the estimated sentiments were generally visualized as pleasant and acceptable.

KEYWORDS: Social Networking, Data Mining, Opinion Mining & Twitter Profile

Received: Jun 20, 2017; **Accepted:** Jul 10, 2017; **Published:** Jul 18, 2017; **Paper Id.:** IJESRAUG20174

INTRODUCTION

Background/ Objectives and Goals

Social networking sites are compilation of services over the web that permits the user to build a profile within the system. It describes a directory of interconnected users with same connections through sorts of interests, concerns, or activities (Boyd & Ellison, 2008) (Fornacciari, Mordonini, & Tomaiuolo, 2015). With the vast data available in the web, micro-blogging services, and social networking sites contribute a lot in making textual data on the Internet to grow fast. One of the most popular micro-blogging services available in the web is Twitter. Twitter contains an enormous amount of data whether structured, semi-structures and unstructured which analysts and decision makers in an institution hardly interpret and describe (Sheela, 2016). As of mid-year 2011, Twitter has posted over 200 million tweets per day (Twitter Engineering, 2011). Known as one of the most popular social networks, Twitter has been the subject of attention of researchers of different organizations (Fornacciari, Mordonini, & Tomaiuolo, 2015).

Microblogging platforms such as Twitter can be used by an organization for many reasons. It can be used by different users to express opinions in a variety of topics and social interest. Thus, it is a valuable source of people's opinion (Sheela, 2016). Almost in all human activities, opinions are fundamental factor because it influences our behaviors. Most often, we intend to know and consider the opinion of other people whenever we need to make a decision (Liu, Sentiment Analysis and Opinion Mining, 2012) (Liu, Web Data Mining: Exploring

Hyperlinks, Contents, and Usage Data, 2011).

The study aims to describe the Twitter profile of a select institution in terms of frequency of posts, number of user interaction, and intensity of trending topics. Visualization of profile is also considered to clearly understand and discover patterns or relationships among data. This study also focuses on Twitter sentiment analysis which classifies the polarity of the twitter comments into negative, positive, and neutral categories. The Twitter feeds were classified using Target-dependent sentiment classification of tweets where the query serves as the target of the sentiments (Jiang, Yu, Zou, Liu, & Zhao, 2011). The terms such as #enrollment, #tuition, and #ncaa were empirically chosen as regard to the target sentiment of the Twitter user of the select institution. Lastly, general sentiments were visualized using sentiment visualizer.

METHODS

Mining Techniques

Data Mining is the process that uses statistical, mathematical, artificial intelligence and machine-learning techniques to extract and identify useful information and subsequently gains knowledge from large databases (Turban, Aronson, Liang, & Sharda, 2007). Data mining twitter feeds to discover user's opinion became the attention of most company and other institution such as in education (Fornacciari, Mordonini, & Tomaiuolo, 2015).

Opinion Mining also called sentiment analysis, it is the field of study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes (Liu, Sentiment Analysis and Opinion Mining, 2012). Sentiment analysis aims to listen and process the data that users post on social media. It is an interdisciplinary field that in recent years has had a significant growth and that makes an extensive use of machine learning techniques (Fornacciari, Mordonini, & Tomaiuolo, 2015). Different approaches, analysis, and techniques in sentiment analysis are explicitly discussed in (Liu, Sentiment Analysis and Opinion Mining, 2012) (Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2011) (Pang & Lee, 2008).

Data visualization converts raw data into images that allow a viewer to "see" data values and the relationships they form. The images allow viewers to use their visual perception to identify features in the data, to manage ambiguity, and to apply domain knowledge in ways that would be difficult to do algorithmically (Healey, Hao, & Hutchinson, Visualizations and Analysts, 2014). In the case of the select institution, the profile was visualized to clearly understand and discover patterns or relationships among data (Shaw, Subramaniam, Tan, & Welge, 2001) (Turban, Aronson, Liang, & Sharda, 2007). It was presented in forms of charts, graphs, and diagrams to describe and discover the underlying nature and activity of the users. This study also visualized sentiment using basic emotional properties embodied in the text, together with an estimated measure of confidence (Healey C., Visualizing Twitter Sentiment, 2016).

The Russel Model of Affect

Sentiment Visualization converts twitter sentiments into images which allow the viewer to see the values and the relationships they form (Healey, Hao, & Hutchinson, Visualizations and Analysts, 2014). The sentiment visualization applied in this study was based on the Russel Model of Affect which proposed the use of valence (or pleasure) and arousal (or stimulation) represented in 2 dimensional plane to build emotional interpersonal circle of affect (Healey C., 2016).

Figure 1: The Russel Model of Affect

The Russel Model of Affect uses emotional dimensions on a 2D plane to position feelings or emotions. Along the horizontal axis represents pleasure with highly unpleasant on left and highly pleasant on the right with different levels of pleasure in between. This model suggested using valence (or pleasure) and arousal (or stimulation) to build emotional interpersonal circle of affect (Healey C., 2016). It applies multidimensional scaling to position 28 emotional states, producing the model shown to the left with valence running along the horizontal axis and arousal along the vertical axes (Healey C., Visualizing Twitter Sentiment, 2016).

To compute for the estimated sentiment, ratings for the common words found in the tools dictionary are combined with a mean rating and a standard deviation of the ratings for each dimension rating. For each word w_i in the tweet that exists in the Tweeter Visualizer dictionary, word's mean valence $\mu_{v,i}$ arousal $\mu_{a,i}$ standard deviation of valence $\sigma_{v,i}$ and arousal $\sigma_{a,i}$ are saved. If a tweet contains less than $n = 2$ words found in the dictionary, it is ignored since it has an insufficient number of ratings to estimate its sentiment. The statistical average of the n means and standard deviations is computed to obtain the tweet's overall mean valence and arousal M_v and M_a (Healey C., Visualizing Twitter Sentiment, 2016).

Figure 2: Calculating Valence and Arousal

As shown in the body of tweet (Figure 2), more than two (2) words were retrieved from the dictionary as highlighted in bold italics. If the tweet's text body contains less than two (2) words found in the dictionary, the tweet will be ignored. The tweet details show each word's frequency, mean and standard deviation of arousal and standard deviation

of pleasure. It further shows how the overall pleasure and arousal of the tweet were calculated.

The Data Mining and Sentiment Analysis techniques applied in this study was built using different twitter analytics and twitter sentiment analysis tools, such as Twitonomy (Diginomy Pty. Ltd., 2016). This analytics tool was utilized to analyze and describe the frequency of post, number of user interaction, and number of trending topics. For sentiment polarity analysis, Rapid Miner (Rapid Miner, 2016) which is a powerful data mining tool was utilized, while visualization was performed using Tweet Viz (Healey C. , 2016), a web-based sentiment analysis tool for analyzing tweeter sentiment. These tools were performed on over 1,500 tweeter feeds from a select institution user from year end of 2011 until mid of year 2016.

Data Collection

The corpus of tweeter feeds was collected using specialized tools that download feeds within the date range selected by the user. In the initial stage, unstructured data which includes date, author, and messages were extracted. The data were pre-processed to remove symbols such as quotations, return characters, and other punctuations to make it readable once imported.

Figure 3: Twitter History of Select Institution

Figure 3 is a graph that shows the history of tweets from year 2012-2016. In the case of a select institution, tweeter posts were rarely observed from year 2012 up to year 2014. It shows that the users became active started from year 2015.

Estimating Sentiment

In estimating sentiment, this study utilized Sentiment Visualizer that applies Russell Model of Affect in representing the emotional properties embodied in the text. Machine learning algorithms such as Naïve Bayesian networks were used to estimate sentiment as part of the computational method applied in the tool (Healey C. , Visualizing Twitter Sentiment, 2016). It also uses term frequency - inverse document frequency (TF-IDF) to evaluate the importance of word to a document in a given data set. The number of times a word appears in each tweet is proportional in the increase of importance and is offset by the frequency of word in a data set.

RESULTS

The Twitter Profile

The raw data of the select institution from December 1, 2011 to June 29, 2016 was analyzed to find out the

summary of Tweets Analytics. The result describes the frequency of posts, number of user interaction, and the concentration of trending topics of the select institution.

Figure 4: Tweets Analytics of Select Institution

Figure 4 shows the *User mentions*, *Retweets*, *Replies*, *Tweets favorited*, *Tweets retweeted*, *Links*, *Tweets per day* and *Hashtags*. The Tweets per day is 0.93 which shows the average frequency of tweets posted each day. As the number becomes higher, the more active is the select institution user is on Twitter (Diginomy Pty. Ltd., 2016). On the other hand, the number of user interaction was determined by analyzing the User mentions, Retweets, and number of Replies and showed that there were 655 user mentions with 0.42 average number of mentions per tweet, 83 retweets or 5% retweets and 361 replies or 23% replies in the total of analyzed tweets. This suggested that as the numbers becomes higher, the more that the select institution user has interaction with others (Diginomy Pty. Ltd., 2016).

Tweets favorited, *Tweets retweeted*, and *Links* were also described to determine the intensity of trending topics on the analyzed tweets. There are 1,174 tweets favorited with 75.4% proportion of the user's tweets favorited by others. It has a total of 20,487 times with an average number of 17.45. 967 tweets were retweeted with 62.1% proportion of the user's tweets, over all a total of 29,864 times or 30.88%.

Figure 5: Tweets Analytics for Users Most Mentioned

Figure 6: Tweets Analytics for Users Most Retweeted

Most users mentioned tweets (Figure 5) were @LPUPirates which had 99 tweets or 15.4% of the total number of analyzed tweets. This was followed by @TagaLPU with 59 tweets or 9.2%. The graph further revealed that the users @LYCESGO Manila has 49 tweets or 7.6%, @LPU_Sentinel has 22 tweets or 3.4%, @LPU Manila has 15 tweets or 2.3%, @vsangelapauline has 8 tweets or 1.2%, @ANCALTERS has 7 tweets or 1.1%, while @LPU Cavite, @LPUpepsquad, and @LPUCBAMANILA got an equal number of tweets of 6 tweets with 0.9% of the total number of analyzed tweets.

On the other hand, most users retweeted tweets (Figure 6) were from @LPUPirates which has 22 retweets or 27.2% of the total number of analyzed tweets. This was followed by @LPUMMLmanila with 13 retweets or 16.0%. As illustrated in the graph, the user @KM_Jessica_Soho has 4 retweets or 4.9%, and @NCAAonSports5, a sports news user has 4 retweets or 4.9%. Whereas, @INQUIRERSports, @DepEd_PH, and @LPU_Sentinel were noted as same rank with 3 retweets or 3.7%. These users are news, government and student government users of the select institution. While @TagaLPU, @gmanews, and @AsuncionTrizia has 3 retweets or 2.5% of the total number of analyzed tweets which are news and student government actors of the select institution.

Figure 7: Hours and Days Tweets Analytics

Figure 7 shows the summary of Tweets activities of the select institution in terms of days per week and hours of the day. It shows that the user interactions were noticeably high during weekdays and started to be less active on weekends. It also shows that the users were most active between 11:00AM to 1:00PM, and between 5:00PM to 6:00PM. It was

observed that the select institution users were active during regular days and most active during break time.

Sentiment Analysis

Sentiment analysis is a field of study that analyses person's opinions, emotions, reactions and responses concerning a chosen topic (Liu, Sentiment Analysis and Opinion Mining, 2012). The sentiment analysis applied in this study was based on the polarity of the twitter comments into negative, positive, and neutral categories.

Figure 8: Polarity of Tweeter Comments

The tweeter feeds (Figure 8) were classified into negative, positive, and neutral categories based on the given training set. The terms purposively chosen as target of sentiment are #enrollment, #tuition, and #ncaa. As gleaned on the graph, the polarity of tweets of the select institution shows that the sentiments have more positive polarity.

Sentiment Visualization

This study estimates and visualizes the general sentiment recovered from the select institution tweeter feeds. Each tweet's estimated sentiment is represented by a circle mapped by emotions. An unpleasant tweet is plotted in blue circles while pleasant tweets are mapped in green circles. The stimulation or arousal is represented as brighter circles which indicate that the brighter the circle, the more active are the tweets. The confidence in the sentiment estimate is represented by the size and transparency. The larger the sizes of the circle, the more confident are the estimates. Another measure of confidence of the tweet's emotion is the transparency. The more opaque or less transparent tweets, the more confident are the estimates.

Figure 9: Tweeter Sentiment

Figure 9 represented the Tweeter Sentiment of the select institution. The graph shows the general tweeter sentiment visualized in an emotional scatter diagram mapped with pleasure and stimulation represented in horizontal and vertical axis. The 2-dimensional distribution summarizes the overall sentiment of tweets. It depicts that the general estimated sentiment reclined on the positive emotions where majority of the sentiments represented by circles in color green are pleasant (Healey C., 2016).

Figure 10: Common Topics on Twitter Sentiment

Figure 10 shows the common topics on Twitter Sentiment grouped into clusters. The keywords seen above the clusters indicate the topics that were visualized in the tweets. As gleaned in the figure, there were three (3) related topics grouped into cluster. This includes suspension of classes, NCAA games, and endorsement of sites for information concerns. The tweets that were not clustered into the topics were visualized as singletons as gleaned on the right part of the figure (Healey C., Visualizing Twitter Sentiment, 2016). Figure 11 shows the Tag Cloud which visualizes the most frequent occurring words found in the Tweeter feeds of the select institution. These words were mapped on the emotional regions Happy, Unhappy, Upset and Relaxed. The tag cloud indicates that the words which are more frequent were drawn in larger size and how often the words occur over all the tweets in the emotional region (Healey C. , 2016). The figure further revealed that only the terms showing in bright color has an estimated pleasure and stimulation. Terms in grey color has no estimated sentiment. As gleaned on the figure, the words classes and please with both 95 occurrences from the overall tweeter feeds have the most frequent terms.

Figure 11: Twitter Sentiment Tag Cloud

Figure 12: Twitter Sentiment Affinity

Figure 12 illustrated the Tweeter Sentiment Affinity which visualized the important actors in the tweet feeds as well as its relationships to one another (Healey C., Visualizing Twitter Sentiment, 2016). The figure describes the frequent tweets, hashtags, people and URLs which contributed in the corpus of tweeter feeds of the select institution. The orange nodes represent people; yellow nodes signify hashtags, while red nodes represent URLs. As described in the figure, the most frequent actor occurrence is @LPUManila, frequent hashtag is #vivavivapirata, and the frequent URL is facebook.com.

CONCLUSIONS

This study describes a Sentiment Analysis and Visualization performed on the Twitter Feeds of a select institution with over 1,500 tweeter feeds from year end of 2011 until mid of year 2016. It seeks to define the profile of the select institution Twitter in terms of frequency of posts, number of user interaction, and concentration of trending topics. Sentiment Analysis applied in this study classifies the polarity of the twitter comments into negative, positive, and neutral categories. This study also performed Twitter Visualization which aims to view the tweeter sentiments in an emotional scatter diagram mapped with pleasure and stimulation using the Russel Model applied on Twitter Visualizer (Healey C., Visualizing Twitter Sentiment, 2016).

The profile of the select institution twitter user with respect to frequency of post becomes higher as the number of tweets favorited by other tweeters becomes higher. Common topics that are being showed in the visualization are “suspension of classes”, and “NCAA games”. The tweeter engagement can be traced from 11:00A.M. to 1:00P.M. and between 5:00P.M. to 6:00P.M. In terms of visualization of the tweeter profile, the most frequent actor occurrence is @LPU Manila, most frequent hash tag is @LPu Manila and most frequent URL is facebook.com. The intensity of trending topics increases because other users consider it as source of information. The twitter sentiments of the select institution in terms of sentiment polarity turned out to be major positive while the estimated sentiments were generally visualized as pleasant and acceptable.

ACKNOWLEDGMENTS

This paper is supported by Research Innovation Center of Lyceum of the Philippines University. The researchers would like to thank Dr. Christopher G. Healey of the Department of Computer Science of North Carolina State University for sharing his expertise and allowing us to utilize the TwitViz as a tool applied in this study.

REFERENCES

1. Boyd, D., & Ellison, N. (2008). *Social Network Sites: Definition, History and Scholarship*. *Journal of Computer-Mediated Communication*, 13 (1), 210-230.
2. Diginomy Pty. Ltd. (2016). *Twitonomy*. (Diginomy Pty Ltd) Retrieved August 30, 2016, from Twitonomy Website: <http://www.twitonomy.com/>
3. Fornacciari, P., Mordonini, M., & Tomaiuolo, M. (2015). *A Case Study for Sentiment Analysis on Twitter*. *16th Workshop on From Object to Agents*. Naples, Italy.
4. Healey, C. (2016, May 22). *Visualizing Tweeter Sentiment*. (North Carolina State University) Retrieved August 30, 2016, from North Carolina State University Website: https://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/
5. Healey, C. (2016, May 22). *Visualizing Twitter Sentiment*. Retrieved August 17, 2016, from North Carolina State University Website: https://www.csc.ncsu.edu/faculty/healey/tweet_viz/
6. Healey, C., Hao, L., & Hutchinson, S. E. (2014). *Visualizations and Analysts*. In A. Kott, C. Wang, & R. F. Erbacher, *Cyber Defense and Situational Awareness*. New York: Springer International Publishing.
7. Jiang, L., Yu, M., Zou, M., Liu, X., & Zhao, T. (2011, June 19-24). *Target-dependent Twitter Sentiment Classification*. *49th Annual Meeting of the Association for Computational Linguistics*, (pp. 151-160). Portland, Oregon.
8. Liu, B. (2011). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data* (2nd ed.). Chicago: Springer Link. doi:10.1007/978-3-642-19460-3
9. Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. *Synthesis Lectures on Human Language Technologies*, 5(1), 1-167.
10. Pang, B., & Lee, L. (2008). *Opinion Mining and Sentiment Analysis*. *Foundations and Trends® in Information Retrieval*, 2(1-2), 1-135.
11. RapidMiner. (2016). *RapidMiner*. (RapidMiner) Retrieved August 30, 2016, from <https://my.rapidminer.com/nexus/account/index.html>
12. Shaw, M., Subramaniam, C., Tan, G., & Welge, M. (2001). *Knowledge Management and Data Mining for Marketing*. *Decision Support Systems*.
13. Sheela, L. (2016). *A Review of Sentiment Analysis in Twitter Data Using Hadoop*. *International Journal of Database Theory and Application*, 9(1), 77-86.
14. Turban, E., Aronson, J., Liang, T., & Sharda, R. (2007). *Decision Support and Business Intelligence Systems* (Eight ed.). Pearson Education.
15. Twitter Engineering. (2011, June 30). *200 Million Tweets per day*. Retrieved August 17, 2016, from Twitter: <https://blog.twitter.com/2011/200-million-tweets-per-day>.