Network-based Visualization of Opinion Mining and Sentiment Analysis on Twitter

Alemu Molla, Yenewondim Biadgie and Kyung-Ah Sohn*

Department of Computer Engineering, Ajou University
San5, Woncheon-dong, Yeongtong-gu, Suwon 443-749, South Korea

Abstract— **Visualizing the result of users' opinion mining on twitter using social network graph can play a crucial role in decision-making. Available data visualizing tools, such as NodeXL, use a specific file format as an input to construct and visualize the social network graph. One of the main components of the input file is the sentimental score of the users' opinion. This motivates us to develop a free and open source system that can take the opinion of users in raw text format and produce easy-to-interpret visualization of opinion mining and sentiment analysis result on a social network. We use a public machine learning library called LingPipe Library to classify the sentiments of users' opinion into positive, negative and neutral classes. Our proposed system can be used to analyze and visualize users' opinion on the network level to determine sub-social structures (sub-groups). Moreover, the proposed system can also identify influential people in the social network by using node level metrics such as betweenness centrality. In addition to the network level and node level analysis, our proposed method also provides an efficient filtering mechanism by either time and date, or the sentiment score. We tested our proposed system using user opinions about different Samsung products and related issues that are collected from five official twitter accounts of Samsung Company. The test results show that our proposed system will be helpful to analyze and visualize the opinion of users at both network level and node level.**

Keywords— *sentiment analysis, opinion mining; twitter; social networks*

I. INTRODUCTION

Twitter is an online social networking and micro blogging service that enables users to send and read short text messages within 140 characters called "tweets" [1]. Registered users can read and post tweets about any issue. These user-generated comments represent a potential complementary source of essential information about company's brand that can be positive, neutral or negative [2]. Companies also need to listen to the customer's and competitor's voices so that they can analyze what the customers are saying online and identify the events that take place in the business environment.

Opinions can be generated by users with different levels of expertise, backgrounds, and intentions. For example, negative opinions may come from different angles such as from an unsatisfied customer to a disgruntled employee or from competitor to the one who has hidden agenda (enemies) or from loyal customers those are talked about real weakness and achievements of the product. Those involved parties can create a negative and positive impact on a company's brand. Since they have continued to speak their minds when they are happy or disappointed, they might share their negative or positive signal on social media to the rest of the customers. This opinion spread to a large audience can be done at the click of a mouse and have the potential to impact the company's entire brand [3].

Therefore, it would be useful for companies or customers to trace the overall opinion trend in the social network as well as to identify key players. To address this issue, we developed a visualization system that serves as a convenient tool for such purpose. Our system is based on open source libraries such as *LingPipe* for processing raw text files or for classifying the sentiment of tweet messages [4]. Using the output file format of the opinion analyzer and opinion classifier as an input file for social graph construction tools, *NodeXL*, we constructed the social network graph to visualize the communication flow as well as to identify the influential persons or key players in the communication network.

The edges in the social network may reflect different kinds of relations among people: friendship, cooperation, contact, conflict, etc. Such networks are used to find sub-groups of people with different attitudes [5]. Analyzing users' opinions is a helpful process which not only tells us how people think, but also what preferences they have, what products they are likely to buy or what moves they are likely to watch. In this study, we show how this process will be done through both network level and node level analysis.

This paper is organized as follows. Section II discusses the previous related work, section III illustrates the system flow process, section IV describes a brief overview of the methodology we used to accomplish our work, section V presents the analysis and the incurred result. In Section VI, we conclude and address our future work.

II. RELATED WORKS

There has been done a lot of prior research in sentiment analysis, especially in the area of product or movie reviews and blogs. Sentiment analysis is used to determine whether each sentence is expressed in a positive, negative, or neutral way. This is also related to subjective/objective-polarity [4,6]. Automatic Sentiment classification using machine learning is a well studied field [7]. Esuli and Sebastiani [6] finds a method that automatically distinguishes each word as either positive, neutral, or negative using semi-supervised learning. Pang and Lee [4, 8] applied various machine learning techniques to sentiment classification problems.

* Corresponding author (kasohn@ajou.ac.kr)

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education (2012R1A1A2042792).

978-1-4799-6541-0/14/$31.00 ©2014 IEEE

Users on the social network can develop different relationships based on their interests and activities. Finding influential users on the social network and analyzing their activities [9,10,11] is an important issue in this area. Those influential users may play the negative or positive role in the entire brand [12]. Our goal is to visualize a social network using both network-level and node-level analysis techniques to easily identify such potential targets and to analyze overall brand images over time and sentiments. In this work, we demonstrate our system using tweets from Samsung official twitter accounts.

III. METHODOLOGY

A. The Architecture of the Proposed System

The general architecture of the proposed system is shown in Fig.1. The input data is collected directly from Twitter and stored in our tweet database. Next, we perform the data pre-processing, sentiment analysis, and user network construction. The different components of the proposed system are described in detail below.

B. Twitter Crawling

There are libraries [13] accessible in several programming languages to make easy access to the Twitter API. We collected tweets in real-time using Twitter's streaming API crawler, which is made from the Twitter 4J library for Java applications [14]. Our dataset contains about 10,000 tweets related to Samsung between November 15, 2013 to February 19, 2014. The tweets focus on the following five Samsung official twitter accounts: SamsungMobileUS, SamsungSupport, SamsungCanada, SamsungMobile, and Samsungtweets. Each user comment is targeted to different Samsung products and related issues.

C. Data Pre-processing

In addition to the actual opinion of users, tweet messages contain other meta-data information. For example, "Chris 38338359 @ryanjbaird: http://t.co/ I love S4 Samsung camera #troll 43.3108147 -96.4312585 Fri Nov 2902:01:40 KST 2013" represents the typical format of the tweet message. It contains sequentially the user name and ID of the sender, the user name of the receiver which is preceded by @ symbol, the hyperlink to other web contents which is optional, the actual tweet message, the hash tag, location information, creation date, time and year [1]. Hashtags are the words or terms denoting a specific topic proceeded by the # (pound) symbol. Except *the actual opinion*, all the other components of the input files are meta-data. A re-tweet allows a user to repeat a tweet created by another user, usually indicating their support for or interest in that tweet's content. Twitter users can "follow" other individuals receive the messages, and anyone following a particular user is denoted as a "follower" of that user [15].

Because tweets are short, ungrammatical, and very informal [16, 17], we pre-process tweets as follows. We first remove URLs, Hashtags, and other special character. We also removed other punctuations and white space to avoid noise. Capital words are changed to lower case letters.

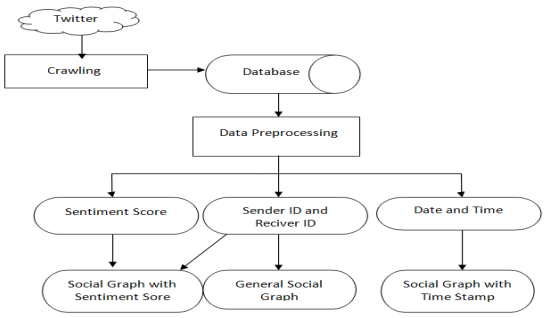


Fig. 1. The Architecture of the Proposed System

D. Sentiment Analysis

After data pre-processing, the sentiment score of the actual tweet message is computed for the identification of subjective opinions. We randomly selected 5,511 tweet messages as training data. We first perform text operations such as parsing, stemming, stop word removal, part of speech tagging and word disambiguation by using a text processing library called *LingPipe* library. Second, each tweet is represented as a vector of feature words by using the standard dictionary of this library. Third, the sentimental label of each tweet (positive, negative or neutral) is assigned. Since these tweet messages are not available in a manually labelled format and it is difficult to label all the tweets manually, we used AlchemyAPI and NLTK applications to label two-third of the data, which may contain some error in the labels. We labelled manually the remaining one-third of the data. This labelled training dataset is used to train the Naïve Bayes classification algorithm in the LingPipe library. Naïve Bayes classifier is very efficient because it is computationally less intensive (in both CPU and memory) and it requires a small amount of training data. Moreover, the training time with Naive Bayes is significantly smaller when compared to alternative methods [18], it produces an oversimplified model, and its classification decisions are accurate [19].

E. Social Graph

We visualize the communication network using three types of social network graphs demonstrated in the architecture of the proposed system. All these graphs are constructed using the social graph visualizing tool called *NodeXL*. To get the general view on the communication flow in the social network, a general social graph is first constructed using the sender ID and receiver ID attributes. These attributes are used as nodes to identify each user uniquely. Two nodes are connected if there is a sender and receiver relationship between them. Unfortunately, the receiver ID cannot be extracted by the crawler. We use public software available on http://idfromuser.com to extract the receiver ID. Once the general social network graph is constructed we will see this social graph across the tweeted time stamp and the sentiment score.

1) Social Graph with Time Stamp

We visualize the evolution of the social network across time to see the communication pattern over time. We used tweets about Samsung products from November 15, 2013 to

February 19, 2014. Our system can visualize the communication pattern at a specific time by extracting a sub-graph using the specific time as a filtering mechanism.

2) Social Graph with Sentimental Score

We also visualize the flow of communication in terms of sentimental value. The sentimental score of each tweet in the training data and for the new tweet in real time is computed as described in section III.D. Our system uses this sentiment score as a filtering mechanism and construct the social graph with sentimental score in a specified range.

IV. Experimental Results

In this section, we present the functionality of our proposed system. We demonstrate our system using about 10,000 tweets about Samsung product and related issues from November 15, 2013 to February 19, 2014.

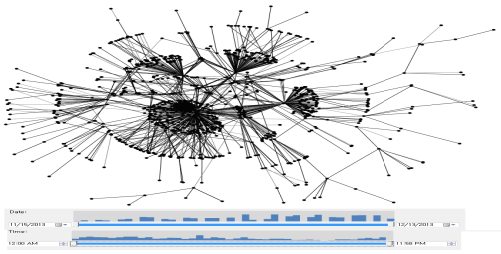
A. Network Level Analysis and Visualization

1) The General Social Network Graph

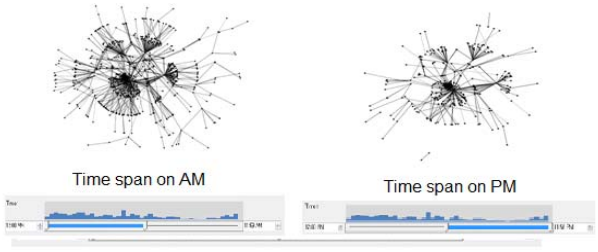
Fig.2 (a) illustrates the general social network graph that shows the *general overview* of the flow of communication. The nodes represent the sender or receiver ID and edges represent simple sender and receiver relationship. The resulting graph contained 2204 vertices and 6695 edges.

2) Social Graph with Time Stamp

While Fig.2 (a) shows the entire social graph for the complete time interval, Fig.2 (b) shows a sub-graph which is obtained when the entire graph is filtered by AM and PM Timestamp. In other words, this graph contains all tweet messages that are tweeted only during AM Timestamp (or during PM Time stamp, respectively). Generally, based on the time stamp of each tweet, we can analyze how the communication network of users is evolving across the date and time using a time bar slider.

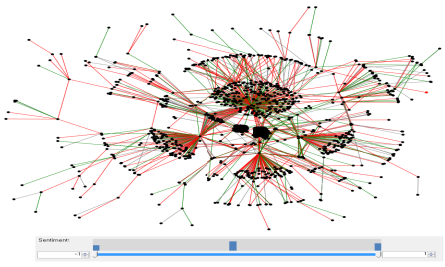


a) The General Social Network Graph with all time stamp

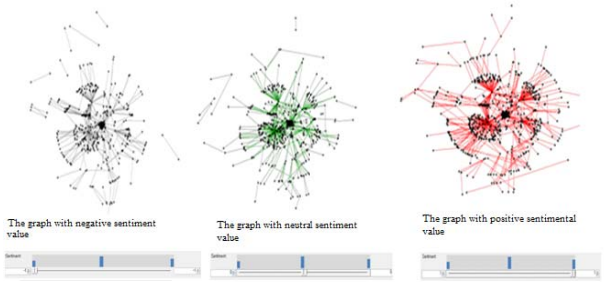


b) Social Graph with AM and PM timestamp

Fig. 2. Evolution of user communication network across the date and time



a) Social Graph with all the sentimental scores



b) Social graph with negative sentiment score

c) Social graph with neutral sentiment score

d) Social graph with positive sentiment score

Fig. 3. Social Graph with Sentimetal Value

3) Social Graph with Sentimental Score

Fig.3 (a) illustrates the Social Graph with three *sentimental scores*. This graph visualizes the flow of communication in the social network by including all the three types of sentimental scores using *sentimental slider bar*. The edges of this graph represent the existence of sender and receiver relationship while the color of edges represents the type of sentimental score of tweets.

From this graph, we extract a sub-graph of the flow of communication with a specific sentimental score. Fig.3 (a) shows the entire social graph with all the three sentimental scores (positive, negative and neutral) while Fig. 3 (b) - (d), shows the sub-graphs filtered by negative, neutral, and positive sentimental scores, respectively.

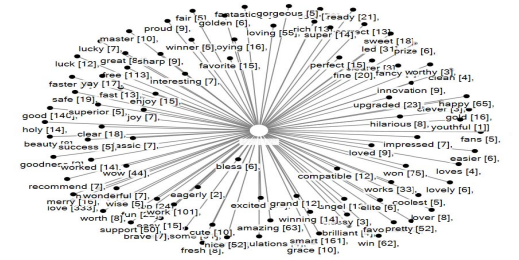
4) Pictorial Representation of Key Words

In addition to visualizing the communication network by extracting sub-graphs in terms of a specific type of sentimental score of the tweet, we can also sub-group tweet messages in terms of the specific type of the sentimental score of feature words that are extracted from the entire dataset. Fig.4 (a) shows the visual representation of positive tweets by using only some positive feature words while Fig. 4(b) shows negative features words. Each black dot represents the name of feature words while the number in the parenthesis represents the frequency of the word in the entire data set.

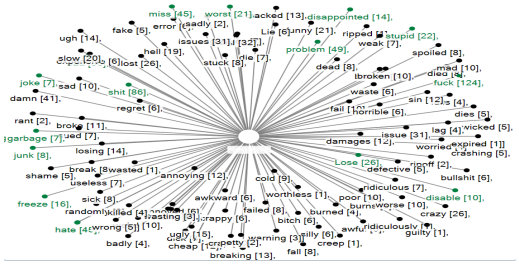
B. Node Level Analysis and Visualization

In addition to network level analysis, our proposed system can also analyze the social communication network at node level to identify influential people using the betweenness centrality, closeness and degree metric. The set of feature words that are used by the identified influential

people can be also identified. For example, the features written by green colors in Fig. 4 (b) show the top 15 most frequently used negative feature words that are used by users whose tweet messages have a negative sentimental score. The distribution of the sentimental score of these feature words is also shown in Fig.5. The association of these words to each influential user is also represented visually in Fig. 6.



a) Positive key words



b) Negative key words

Fig. 4. Visual representation of (a) Positive and (b) Negative keywords extracted from the entire dataset

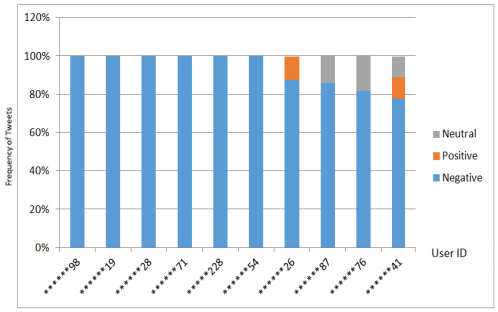


Fig. 5. Sentiment score distribution for top 10 influential users ranked by betweenness centrality using NodeXL

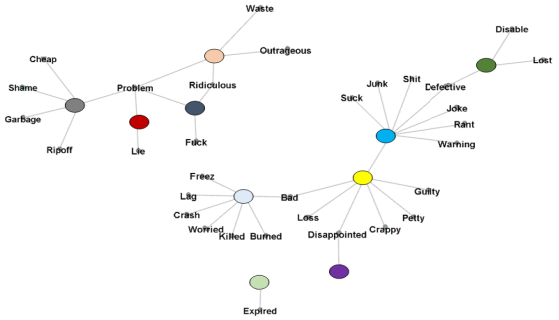


Fig. 6. Negative keyword network for top 10 users ranked by betweenness centrality using NodeXL

V. CONCLUSION AND FUTURE WORK

Opinions may come from different aspects such as from domain expertise to an ordinary man. In this paper, we focused on the opinions contained in the companies' official twitter account. The tweeted opinions have a greater role in distributing the positive and negative impact to the rest of the customers. The main purpose of this system is to identify opinion in the tweet and classify them automatically into positive, negative and neutral sentiments and then visualiz their communication network. We demonstrated different analyses such as how the network has evolved over time, who communicates when, who communicates what.

In this work, we did not include the emoticons of users, which may be helpful for understanding the behavior of users. In our future work, we would focus on the inclusion of emoticons and location information of each tweet.

REFERENCES

[1] A. H. Wang, Don't Follow Me: Spam Detection in Twitter, Proceedings of the 2010 International Conference on Security and Cryptography (SECRYPT), Piraeus, 26-28 July 2010.

[2] M. Tsytsarau, and T. Palpanas, Survey on Mining Subjective Data on the Web. Data Mining and Knowledge Discovery, 2012.

[3] J. S. Brown and P. Duguid. The Social Life of Information. Harvard Business School Press, Boston, MA, USA, 2002.

[4] B. Pang and L. Lee. Opinion Mining and Sentiment Analysis, in Foundations and Trends in Information Retrieval, 2 (1-2), pp. 1–135, 2008.

[5] B. Yang, W. Cheung, and J. Liu, Community Mining from Signed Social Networks. Transactions on Knowledge and Data Engineering 10:1333–1348, 2007.

[6] A. Esuli and F. Sebastiani, SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining, Proceedings of the 5th Conference on Language Re- sources and Evaluation, Genoa, 24-26 May 2006.

[7] C. D. Manning and H. Schutze. Foundations of Statistical Natural Language Processing. MIT Press, 1999.

[8] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: Sentiment Classification Using Machine Learning Techniques," EMNLP-02, vol. 10, pp. 79-86, 2002.

[9] F. Probst, L. Grosswiele, R. Pfleger, Who Will Lead and Who Will Follow.

[10] H. Yuheng, F. Wang, S. Kambhampati, Listening to the Crowd: Automated Analysis of Events Via Aggregated Twitter Sentiment

[11] S. Matthew, Gerbera, Predicting Crime Using Twitter and Kernel Density Estimation.

[12] M. Ellyn, Angelotti Journal of High Technology Law. All Rights Reserved. ISSN 1536-7983, 2013.

[13] Twitter, Twitter Libraries, 6 July 2006. [Online]. Available: https://dev.twitter.com/docs/twitter-libraries.[Accessed 10 July 2012].

[14] Y. Yamamoto (2007), [Online]. Available: http://twitter4j.org/ko/ Sep. 9, 2013.

[15] V. Pandey and C. Iyer, Sentiment Analysis of Micro blogs.

[16] A. Pak and P. Paroubek, Twitter as a Corpus for Sentiment Analysis and Opinion Mining, in Proc. Of LREC 2010.

[17] M. van Meeteren, A. Poorthuis and E. Dugundji, Mapping Communities in Large Virtual Social Networks, Proceedings of the 1st International Forum on the Application and Management of Personal Electronic Information, Cambridge, 12-13 October 2009.

[18] D. Christopher, Manning, P.Raghavan and H. Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

[19] J. Huang, J. Lu*, C. X. Ling, Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy", 2003.