Correlation Analysis between Sentiment of Tweet Messages and Re-tweet Activity on Twitter

Wonmook Jung, Hongchan Roh and Sanghyun Park

Department of Computer Science, Yonsei University
134, Shinchon-Dong, Seodaemun-Gu, Seoul 120-749, Korea
{jngwnmk,fallsmal,sanghyun}@cs.yonsei.ac.kr

We study the correlation between sentiment of tweet and Re-tweet activity, a term used in the social networking and micro-blogging service (Twitter). To analyze the correlation, we first preprocessed tweet datasets. The preprocessing included extracting tweets that contain Re-tweet activity, the part-of-speech tagging and word-filtering. Second, we exploited the Thayer model as a standard sentiment measure of tweets. The sentiment words were categorized into four groups based on Thayer model. Next, we evaluated sentiment values of each tweet, which express emotion of tweet. Moreover, we visualized degree of sentiment of tweets by mapping each tweet into corresponding coordinate on 2-dimensional space based on sentiment values. The result shows that tweets, which were expressed more positively and more energetically or more negatively and more silently, were more actively spread out on Twitter. Experimental evaluations on datasets indicate that sentiment of tweet has the correlation with Re-tweet activity.

Key Words: social network, data mining, Twitter sentiment analysis

1. INTRODUCTION

In modern society, a social network has an important role in communication among people. Millions of people express and share their emotions or opinions through the social networks such as Twitter and Facebook [5]. Especially, on Twitter, messages (tweets) which are limited to contain only 140 characters include full of opinions and emotions [1]. The tweets are also able to be re-tweeted by users who like the tweets or are eager to spread it. Re-tweet activity is a function of Twitter that enables users to share their tweets with their followers. The Re-tweet activity has a significant role in not only individuals but also businesses or organizations. When Twitter is used for marketing purpose of businesses, marketers will desire spreading out their tweets with Re-tweet to promote their products or services. For this reason, finding key factors to Re-tweet activity is valuable.

In this paper, we present key factors to Re-tweet activity through a sentiment analysis of tweets. By analyzing correlation between sentiment and Re-tweet activity of tweet, we suggested emotion words that affect Re-tweet activity.

To quantify sentiment of tweets, we used the Thayer model which is called Arousal-Valence model [9]. It was used for the standard sentiment measurement to express quantified tweets with Arousal and Valence values.

Although there were studies analyzing sentiment of Twitter [1, 2, 4, 6, 8], we performed not only sentiment analysis for Twitter but also discovered correlation between sentiment and Re-tweet activity.

The remainder of this paper is organized as follows. In the next section, the Part-of-speech tagger and word-filtering for preprocessing is expressed. Section 3 explains the Thayer model to quantify sentiment of tweet and proposes a method to calculate Valence and Arousal of tweet. In section 4, it we provide a visualized

result of correlation between sentiment of tweets and Re-tweet activity. The last section summarizes our work.

2. DATA PREPARATION

When tweets are generated, there are unnecessary components to analyze sentiment of tweet. To achieve a better Twitter sentiment analysis, some lexicons such as personal pronoun, the 3rd person singular present and preposition are excluded for analyzing. For this paper the Stanford Part-of-Speech tagger [7] is used. It reads texts of tweets and assigns parts of speech to each word such as noun, verb, adjective, etc. After tagging process, each word of tweets has annotation. Subsequently, it requires word-filtering process that filters unnecessary words out with referring annotation. We define 14 lexicon lists to be remained after word-filtering process on Table 1.

Annotation	Meaning
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VB	Verb, base form
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
NN	Noun, singular or mass
NNP	Proper Noun, singular
NNPS	Proper Noun, plural
NNS	Noun, plural
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative

Table1. Lexicon lists used for word-filtering process

Operating data preparation helps to reduce data size for remain analyzing processes. For example, Figure 1 shows an example data preparation process applying the Stanford Part-of-Speech tagging and word-filtering.

Figure1. Data preparation process

3. LINKING TWEETS WITH SENTIMENT

3.1 The Standard Sentiment Measurement

After applying the part-of-speech tagging and word-filtering each tweet, it consists of sentiment words and non-sentiment words. To evaluate sentiment value of each tweet, first of all, it is required to make the standard sentiment measurement to quantify sentiment value of each tweet. In this paper, the Thayer model [9] which is called Arousal-Valence model is adopted for the standard sentiment measurement of each tweet. Based on the Thayer model, the tweet is able to be expressed on a two-dimensional space with Arousal-Valence value. As shown in Figure 2, the two-dimensional space is divided into 4 quadrants, and different sentiment groups are placed on the each plane. The right side of the plane refers to the positive sentiment while the left side stands for the negative sentiment. On the other hand, the upper side of the plane refers to energetic sentiment whereas the lower side of the plane stands for the silent sentiment. We define 12 sentiment words and classified those into 4 quadrants as shown in Figure 2.

3.2 Sentiment Values of Tweets

To plot each tweet on the two-dimensional space, each tweet has to have Arousal-Valence value. We define Arousal-Valence value of each tweet as sentiment value. To obtain sentiment value of each tweet, there are three processes calculating each Arousal and Valence value. First of all, it is required to determine Arousal-Valence value of each emotion words used for Thayer model such as "bored", "pleased", and "happy". For instance, "happy" has (0.6, 0.6) as (Valence, Arousal) value while "bored" has (-0.6,-0.6) as (Valence, Arousal) value. In the second process, it observes which emotion words exist in each tweet. Finally, the sentiment value of each tweet is calculated by using previous results. The Valence value of tweet and the Arousal value are calculated as:

$$\text{Valence(tweet)} = \sum_{word \in emotion\_words} I_{(word)} * V_{(word)} \quad \text{Equation(1)}$$

$$\text{Arousal(tweet)} = \sum_{word \in emotion\_words} I_{(word)} * A_{(word)} \quad \text{Equation(2)}$$

where *emotion_words* is a set of 12 defined sentiment words and *word* is the distinct word token in tweet. $I_{(word)}$ is an indicator that indicates whether *word* exists in tweet or not. In other words, if the tweet includes *word,* value of $I_{(word)}$ is 1 whereas value of $I_{(word)}$ is 0 if the tweet doesn't include *word.* Moreover, $V_{(word)}$ is Valance value of *word* and $A_{(word)}$ is Arousal value of *word*.

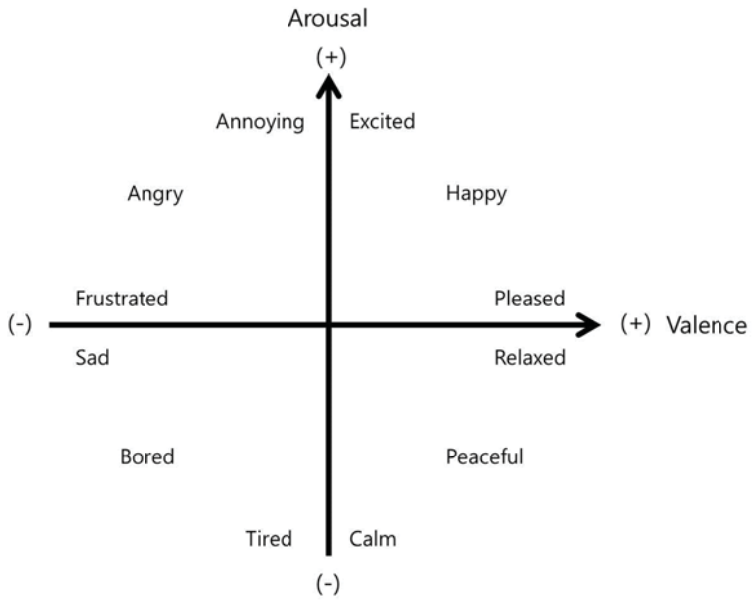

Figure2. Thayer's sentiment model

4. EXPERIMENTAL EVALUATION

The dataset used for our experiments includes tweets, follower/followee relationship data, and event time table of 30,000 twitter users [3]. We created separate tweet data files each of which corresponds to a twitter user, thus making the user account ID the file name. A tweet file includes pairs of a tweet message and the time when the tweet was written, and each pair was differentiated by a line character '\n'. The tweet messages were further classified into re-tweeted and the other tweets by using the "RT" tag included in the tweet messages. We filtered the re-tweeted tweets and they were used for our experiments. Two-dimensional coordinates of the re-tweeted tweets were calculated by using the Valence and Arousal values obtained by equation (1) and (2). Figure 3 presents the visualized

result of the tweets by using the two-dimensional coordinates, where the size of circle is proportional to the number of re-tweeted tweets in the area. The re-tweeted tweets had a tendency to have high values for both valence and arousal or to have low values for both of them. This implies that there is high possibility that tweets can be re-tweeted if the tweets have positive and energetic emotions or if they contain negative and silent emotions.

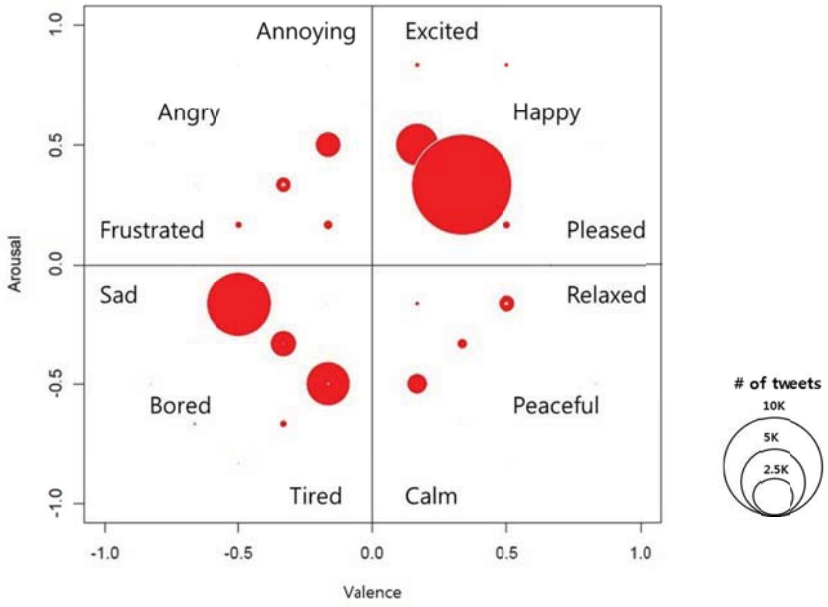

Figure3. Visualized result of the re-tweed tweets based on Thayer's model

5. CONCLUSION

In this paper, we discovered a tendency of re-tweeted tweets with respect to the emotions embedded in tweets. We incorporated Thayer model in order to measure the sentiment degree of tweets as values. We utilized the Arousal and Valence value of Thayer model for our sentiment degree values of tweets. After preprocessing tweets, we plotted the re-tweeted tweets in a two-dimensional space by using the Valence and Arousal values. By doing so, we discovered that tweets can be re-tweeted if the tweets have positive and energetic emotions or if they contain negative and silent emotions.

ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012-0001997).

REFERENCES

[1] Anqi Cui., Min Zhang., Yiqun Liu., and Shaoping Ma., "Emotion Tokens: Bridging the Gap among Multilingual Twitter Sentiment Analysis," *the 7th Asia Information Retrieval Societies Conference, AIRS 2011*, Dubai, United Arab Emirates, Dec, 2011.

[2] Bollen, J., Pepe, A., Mao, H., "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," *ICWSM 2011*, arXiv: 0911.1583, 2011.

[3] Cuckoo project, http://user.informatik.uni-goettingen.de/~txu/cuckoo/index.html

[4] Go, A., Huang, L., Bhayani, R., "Twitter sentiment classification using distant supervision," *CS224N Project Report*, Stanford, 2009.

[5] J. Comm, “Twitter Power 2.0: How to Dominate your Market One Tweet at a Time”, 2010

[6] Katerina Tsagkalidou., Vassiliki koutsonikola., Athena Vakali., and Konstantinos Kafetsios., "Emotional Aware Clustering on Micro-blogging Sources," *the 4th International Conference, ACII 2011*, Memphis, TN, USA,pp.387-396, Oct, 2011.

[7] Kristina Toutanova., Christopher D. Manning.,"Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger," *In Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pp. 63-70, 2000.

[8] O'Connor, B., Balasubramanyan, R., Routledge, B., Smith, N., "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series," *Int. AAAI Conf. on Webblogs and Social Media*, Washington DC, pp. 122-129, 2010.

[9] Thayer, R. E., "The Biopsychology of Mood and Arousal," *Oxford University Press*, 1989.