International Journal of Computing Science and Information Technology, **2018**
ISSN: 2278-9669,April 2018 (http://ijcsit.org)

Analysis Of Tweet Sentiments And Causality Extraction For Prediction

P .G.Preethi [1], Sheethal Abraham[2], Dr.V.Uma [3].
[1&2]Indo Asian Academy, Bangalore, India
[3]School of Engineering and Technology, Pondicherry University, Puducherry, India

Received January 2018

Abstract

Nowadays, Microblogging services like Twitter has become an important platform for users to easily express and share their thoughts and views towards products or events. The number of opinions on the web has significantly increased with the emergence of microblogs. Sentiment analysis or opinion mining aims to determine the attitude of a speaker or a writer with respect to some topic. A generalized prediction model based on twitter microblog is presented in this paper. Opinions expressed in a microblogging website usually contain temporal attributes. These temporal attributes are useful for proper ordering of events and better event prediction. More complex type of relationship between the events is causal relation and identifying these relations helps in prediction and reasoning. The proposed work introduces a prediction model based on twitter that can predict the future event sentiment and the possible time period between the events using the temporal attributes and causal relations. The sentiment analysis phase is evaluated using the performance measures precision and recall. The accuracy of prediction phase is calculated using Mean Absolute Percentage Error (MAPE) as a measure.

Keywords: Sentiment Analysis, Causality, Temporal Sentiment Analysis, Prediction

1. Introduction

Nowadays Microblogging has become a very popular communication tool among internet users. Large amount of messages are appearing daily in popular Microblogging web-sites like Twitter, Tumblr, Facebook etc. Because of the simplicity of the microblogging services more and more people use this media to express or share their thoughts and opinion about different product or services.

Sentimental analysis or opinion mining is a field that collects and analyses people's attitude or sentiment towards different entities like products, services, organizations, individuals, issues, events, topics etc. and their attributes. Sentiment analysis and opinion mining mostly concentrate on different attitudes which expresses or implies positive or negative sentiments [1]. Sentiment Analysis can be considered as a way to measure the tone of conversation. Recent studies show that analysis of the social media texts can be useful for predicting trends or events such as election results [2][3][4],movie success [5][6][7],stock market sentiments [8][9],marketing[10] etc.

Opinions expressed in microblogging website usually contain temporal attributes. These temporal attributes are useful for proper ordering of events and analyzing these temporal information is also useful for better event prediction. A more complex type of relationship between events is causal relations. Since causality shows how variations in one variable cause the variations in other variable, it is a more useful method for prediction and reasoning [11].

The proposed work introduces a generalized prediction model that can predict the future event sentiments and time duration between the events. The proposed work uses the concept of temporal sentiment analysis and causal relations for event prediction. In this work, Naive-Bayes classifier is used for sentiment classification and the causal relation is found using the measures support and confidence.
The remainder of the paper is structured as follows. Section 2 provides the survey on the existing methods for sentiment analysis and its relationship with temporal and causal relations. Section 3 explains the architecture of the proposed work and the different modules associated with the development of the system. Section 4 provides the experimental results obtained and Section 5 concludes the work and presents the possible future works.

2. Literature Survey

This section presents the survey done on sentiment analysis, sentiment classification, temporal sentiment analysis and causal rule detection.

2.1. Sentiment Analysis

Sentiment Analysis can be classified into three categories. They are Document Level, Sentence Level and Entity and Aspect Level sentiment analysis. Document Level sentiment analysis mainly concentrates on classifying the whole document as either positive or negative [12]. Sentence Level classification concentrates on each sentence and classifies the sentence as positive, negative or neutral [13]. Entity and Aspect level which is also known as feature level sentimental analysis focuses on the features of a product and provides the summary of the features, the users like or dislike [14].

2.2. Sentiment Classification

Sentiment Classification can be classified into three categories namely supervised learning, unsupervised learning and semi-supervised learning. Vaithyanathan et al. [12] suggests different supervised machine learning approaches (Naïve Bayes, Maximum entropy classification and Support Vector Machine) for sentiment classification. Mullen et al. [15] introduced a supervised sentiment classification technique that assigned values to selected phrases and words, and used the technique for bringing them together to create a model for classification of texts. Gamallo et.al [16] proposed a supervised method for sentiment analysis of tweets written in English language. The system used Naïve Bayes classifier for sentiment classification.

In this work, sentiment analysis is performed using twitter data. Since Naïve Bayes classifier has low computational overhead and high performance [16] the proposed work uses Naïve Bayes classifier for sentiment classification..

2.3. Sentiment Analysis, Causal and Temporal Relations

Mishne and Rijke proposed [17] a system called Moodview which tracks and analyses the temporal change of bloggers sentiment. Fukuhara et al.[18] proposed a technique that analyses the temporal trends of sentiments and topics from time stamped text archive.

Web contents including but not limited to tweets, blogs, comments, online news site articles are all used to calculate event sentiment. These methods can easily summarize the events based on time and overall sentiment. In this work, both sentiment analysis and temporal relations have been used for predicting event sentiment and the possible time period between two events using twitter blog.

Causality can be defined as connection between two events or states such that one produces or brings about the other; where one is the cause and the other its effect [11]. Since causal relation shows how variations in one variable can cause changes in the other variable it is useful for prediction and reasoning.

Jiang1 et al. [19] introduced a topic level sentiment analysis method based on probabilistic topic model and language grammar based sentiment analysis technique. The main aim of this work is identifying the sentiment and its changes towards an interested topic over time. This work discusses the topic of sentiment change analysis on web documents and identifies the causality between two events. The proposed work uses twitter social media and identifies causality rule for predicting event sentiments and also identifies the possible time duration between the two events.

Siganos et al. [20] employs Facebook's daily sentiment index and examines its relation to stock returns, trading volume, and stock price volatility across twenty international markets. Since twitter data is simpler than Facebook data the proposed work uses twitter social media domain.

Mirza [11] proposed a system that automatically identifies temporal and causal relation from natural language text. This work presented an annotation scheme that identified different type of causality between events. Smailovic ´et al. [21] [22] used twitter social media to predict the stock market change. This work concentrated only on stock market domain while the proposed work is a generalized method that can be implemented in any domain.

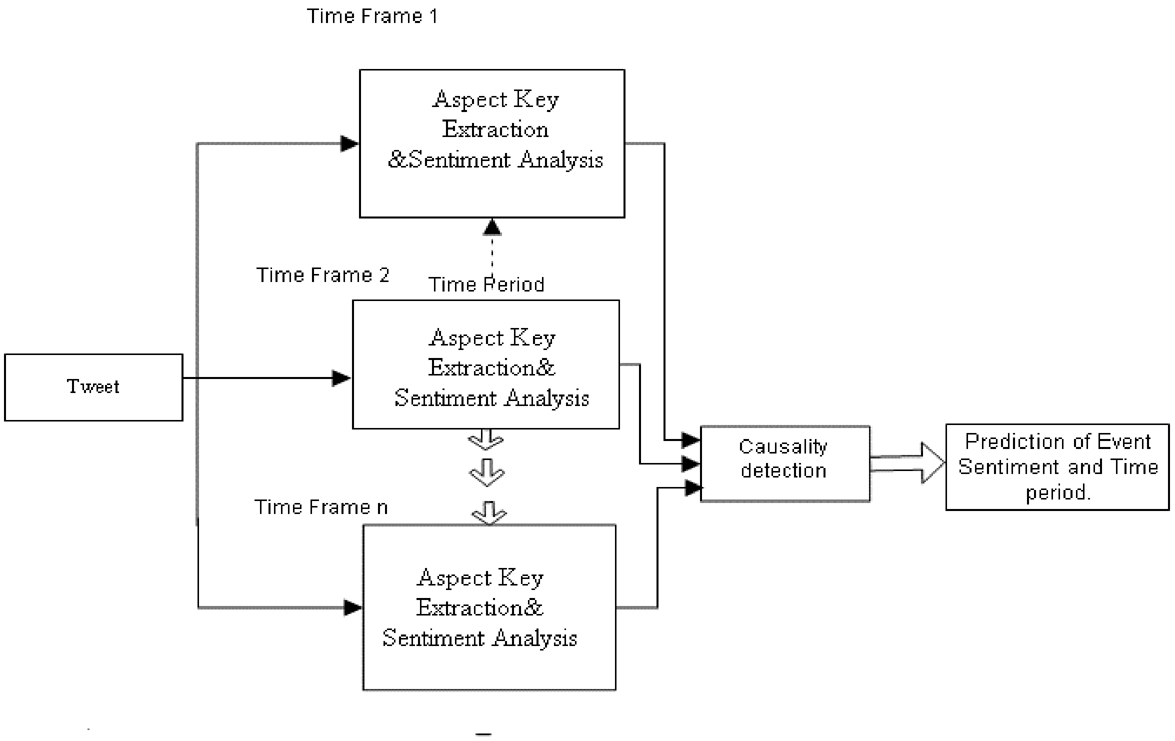
Dehkharghani et al. [23] proposed a methodology for extracting sentimental causal relation from large textual data sources. This concept is very useful for information summarization from large textual data. This work determines sentiment causal rule but no prediction was made while the proposed work includes temporal factor in extracting sentiment causal rules and uses them for better event prediction.

3. Proposed Work

Detailed idea about the proposed work is illustrated in this section. The proposed work is implemented in four phases. Fig. 1 shows the detailed architecture of the proposed system.

Step 1: Aspect Key Extraction & Sentiment Analysis

The first step of the proposed model is Aspect Key Extraction and Sentiment Analysis. In this phase,


Time Frame 1
Aspect Key Extraction &Sentiment Analysis
Time Frame 2
Time Period
Tweet
Aspect Key Extraction& Sentiment Analysis
Causality detection
Prediction of Event Sentiment and Time period.
Time Frame n
Aspect Key Extraction& Sentiment Analysis

Aspect keys are identified from tweets and their corresponding polarity is identified. The aspect keys are identified on the basis of the number of occurrence of the specific word in particular time period. These aspect keys are useful for identifying the causal relations. The text mining API called Alchemy Api is used for this purpose.

Step 2: Sentiment Analysis

After extracting the aspect keys the next step is identifying the polarity of aspect keys. The Naïve Bayes classifier is used for this purpose. Naïve Bayes classifier performs the sentiment analysis process based on conditional probability method [24] that is shown in equation (1).

$$\text{Class}\left(d_i\right) = argmaxp\left(p(c_j)\right)\prod_{i=1}^{d} p\left(\frac{p_i}{c_i}\right) \qquad (1)$$

Class (d_i): Function that calculates the class (ex. positive or negative) of the document d_i.

$p(c_j)$Determines the probability of class c_j .

$p\ \left(\frac{p_i}{c_i}\right)$Determines the probability that the pattern i.e. unigrams, bigrams p_i belongs to class c_j.

Step 3: Temporal Attribute Extraction

The temporal attributes of the aspect keys are identified using the tweet posting time. The time stamp of the tweets is normalized to Greenwich Mean Time (GMT). The time periods between different events are calculated based on the tweets posting time and are used for the prediction of upcoming events and their possible occurrence time.

Step 4: Causality detection

The third step is identifying the causal relation between aspect keys. The causal relations are identified by Support and Confidence method [23] using equations (2) and (3)

$$\text{Support}\ (\ I_1) = \frac{|\omega_1|}{|T|} \qquad (2)$$

$$\text{Confidence}\ (I_1 \rightarrow I_2) = \frac{|\omega_{12}|}{|\omega_1|} \qquad (3)$$

where T is the set of all tweets during a particular time period and ω_1 is a subset of transactions includingI_1 and ω_{12} is a subset of transactions that included I_1 and I_2.

Step 5: Prediction

The final step is prediction based on the causal rules identified in the previous phase. Using the time based analysis of tweets and causal rule identified in the previous step, the sentiment of the upcoming event and time period between the events are predicted. Prediction accuracy is evaluated using (Mean Absolute Percentage Error)[25] MAPE calculated as shown in equation (4).

$$\text{MAPE}=\frac{\sum_{i=1}^{n}\frac{ri-pi}{ri}}{n} \qquad (4)$$

Where r_i is the true value and p_i is the predicted value of i^{th} tweet and n is the total number of observations.

4. Experimental Evaluation

This section describes the experimental setup for the proposed system along with the evaluation.

4.1. Datasets

The input datasets for the proposed work consists of tweets extracted during different time periods. For this proposed work, tweets have been collected using Twitter API and R language for a period of 6 months.

In this work, tweets in English language are alone considered. Nearly 5000 tweets have been collected. SentiWordNet, a lexical resource for opinion mining is used for training the data set.

4.2. Results and Discussion

The results obtained can be considered in three parts. The first part is the identification of the aspect keyword during different time periods. The Alchemy Api is used for this purpose. The Second part is finding the polarity of the Aspect keyword. Twitter API and R language is used for this phase. The result obtained in this section is illustrated in Figure 2.

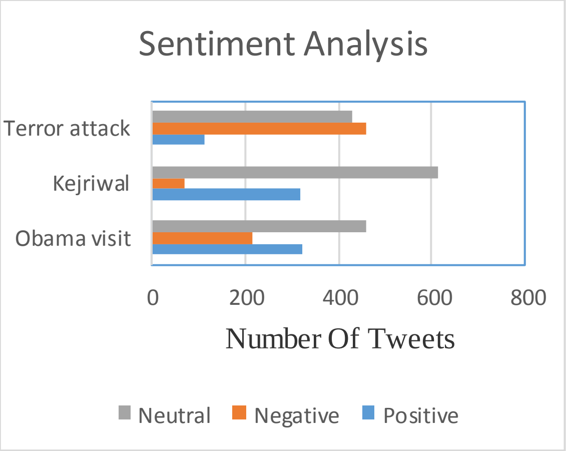


Figure 2. Aspect keywords and their sentiment

Figure 2 shows the tweets expressed on various topics classified based on sentiments (positive, negative, and neutral). The number of tweets containing negative sentiments is found to be more when the aspect keyword the terror attack of Mumbai is considered.

The accuracy of the system is calculated using the Naïve Bayes classifier. The method used in the sentiment analysis phase is evaluated using precision and recall as performance measures. Precision and Recall can be calculated [11] using equations (5) (6) (7) (8) and accuracy is calculated using equation (9).

$$\text{Precision (pos)} =\frac{tp}{tp+fp} \qquad (5)$$

$$\text{Precision (neg)} = \frac{tn}{tn+fn} \qquad (6)$$

$$\text{Recall (POS)} = \frac{tp}{tp+fn} \qquad (7)$$

$$\text{Recall (neg)} = \frac{tn}{tn+fp} \qquad (8)$$

$$\text{Accuracy}=\frac{tp+tn}{tp+tn+fp+fn} \qquad (9)$$

where tp, tn, fp, fn are true positive, true negative,false positive,false negative review for polarity prediction respectively. The precision and recall obtained by the system are shown Figure 3.

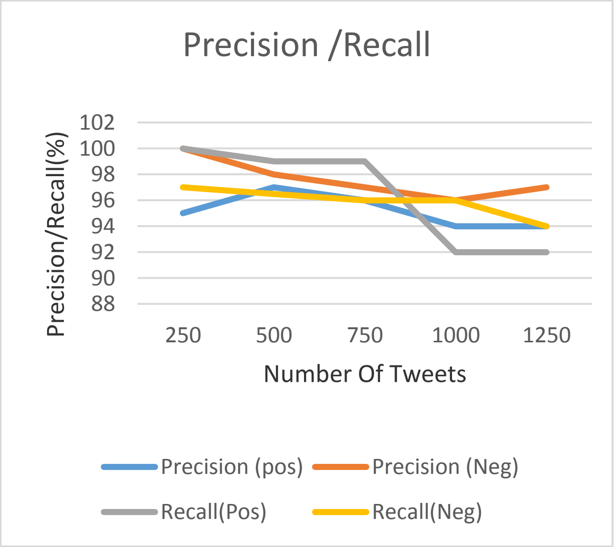


Figure 3. Precision- Recall Graph

The precision value for positive sentiment lies between 94% and 97% and negative sentiment lies between 96% and 100%. The average recall for positive sentiment and negative sentiment are 96.4% and 95.9% respectively Accuracy is calculated on the basis of precision and recall and result obtained is shown in Figure 4.Accuracy of the sentiment analysis system w.r.t number of tweets is found to be between 96% and 97%.

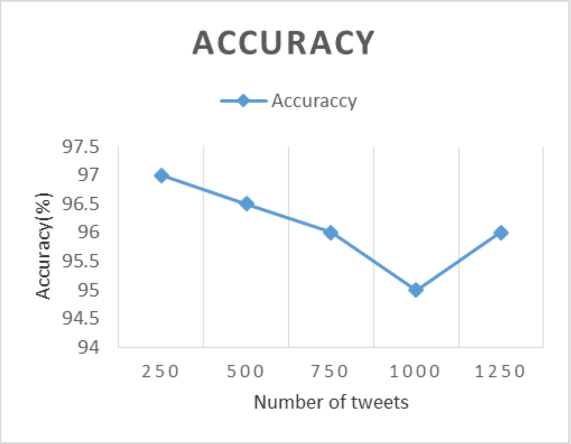


Figure 4. Evaluation of Accuracy

The third part of result concentrates on the causality relation between the aspect keywords. The measures support and confidence are used for finding the causality. The confidence value and the corresponding number of causal relations obtained is illustrated in Figure 5.

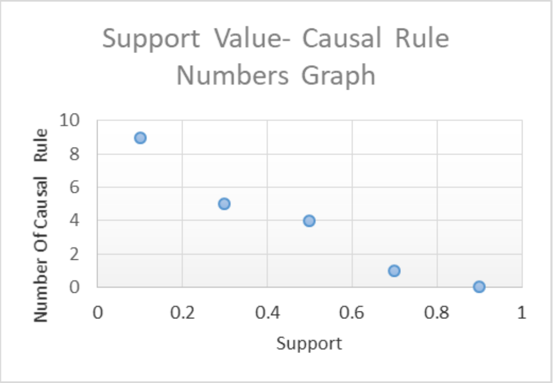


Figure 5. Causal Rule identified for different support values

Number of causal relations decreases as the confidence value increases. Maximum number of causal relations is identified when the support value is 0.1. So this value is taken as the threshold value.

The result of prediction phase is the possible time period between the events and the upcoming events polarity. Mean Absolute Error Percentage for the prediction phase is shown in Figure 6.

Mean Absolute Percentage Error (MAPE) expresses the accuracy with respect to percentage of error. MAPE of proposed model varies between 14.75 % and 21.25% w.r.t the numbers of tweets and this proves that the accuracy of prediction is high.

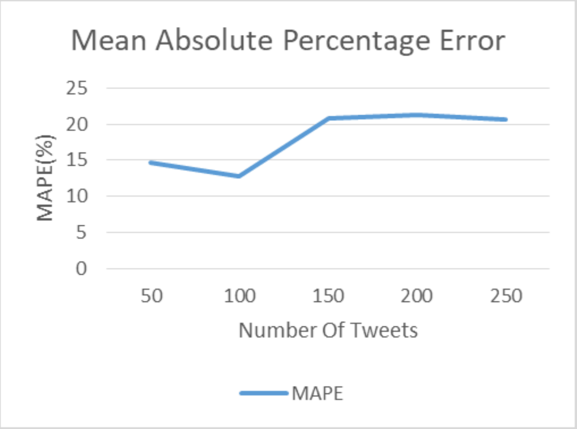


Figure 6. MAPE for Prediction

The proposed work can summarize the Twitter social media sentiment on a particular topic and can predict the sentiment of upcoming event. It can also forecast possible time period between the upcoming events. This prediction model can be used in different domains such as politics, stock market, mass movements, medicine etc.

5. Conclusion

A generalised prediction model based on temporal sentiment analysis and causal rule is presented in this paper. This proposed work can forecast the upcoming event sentiment and possible time duration between the events. Future work is to improve the prediction accuracy and also to extend the work in big data.

REFERENCES:

[1] Liu, B. "Sentiment analysis and opinion mining." *Synthesis Lectures on Human Language Technologies*, vol .5, no. 1, pp. 1-167, 2012.

[2] Park, S., Ko, M., Kim, J., Liu, Y., Song, J. "The politics of comments: predicting political orientation of news stories with commenters' sentiment patterns." In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pp. 113-122. ACM, 2011.

[3] Rill, S., Reinel, D., Scheidt, J., & Zicari, R. V."PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis." *Knowledge-Based Systems vol.69* pp. 24-33, 2014.

[4] Mohammad, S. M., Zhu, X., Kiritchenko, S., & Martin, J "Sentiment, emotion, purpose, and style in electoral tweets." *Information Processing & Management*, 2014.

[5] Basari, A. S. H., Hussin, B., Ananta, I. G. P., & Zeniarja, J. "Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization*." Procedia Engineering*, vol.53, pp. 453-462, 2013.

[6] Lepori, Gabriele M. "Positive mood and investment decisions: Evidence from comedy movie attendance in the US." *Research in International Business and Finance* vol.34, pp. 142-163, 2015.

[7] Schuller, B., Knaup, T. "Learning and knowledge-based sentiment analysis in movie review key excerpts." In *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*, pp. 448-472. Springer Berlin Heidelberg, 2011.

[8] Li, X., Xie, H., Chen, L., Wang, J., Deng, X. "News impact on stock price return via sentiment analysis." *Knowledge-Based Systems* vol .69, pp. 14-23, 2014

[9] Corredor, P., Ferrer, E., & Santamaria, R. "Investor sentiment effect in stock markets: Stock characteristics or country-specific factors? *"International Review of Economics & Finance* vol.27, pp.572-591, 2013.

[10] Mostafa, M. M. "More than words: Social networks text mining for consumer brand sentiments." *Expert Systems with tions* vol.40, no. 10 pp.4241-4251,2013

[11] Mirza, P. "Extracting Temporal and Causal Relations between Events."*ACL 2014*, pp. 10, 2014.

[12] Pang, B., Lee, L., Vaithyanathan, S., "Thumbs up? Sentiment classification using machine learning techniques", *in: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*-Vol 10. Association for Computational Linguistics, pp. 79–86, 2002

[13] Tan, L.K.-W., Na, J.-C., Theng, Y.-L., Chang, K. "Sentence-level sentiment polarity classification using a linguistic approach." In *Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation*, pp. 77-87. Springer Berlin Heidelberg, 2011

[14] Wiebe, Janyce M., Rebecca F. Bruce, and Thomas P. O'Hara. "Development and use of a gold-standard data set for subjectivity classifications." In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 246-253. Association for Computational Linguistics, 1999.

[15] Mullen. T, and Collier.N "Sentiment Analysis using Support Vector Machines with Diverse Information Sources." In *EMNLP*, vol. 4, pp. 412-418. 2004.

[16] Gamallo, Pablo, and Marcos Garcia. "Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets." *SemEval 2014* pp.171.2014

[17] Mishne, G., De Rijke, and M "MoodViews: Tools for Blog Mood Analysis." In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pp. 153-154. 2006..

[18] Fukuhara, T., Nakagawa, H., Nishida, T "Understanding Sentiment of People from News Articles: Temporal Sentiment Analysis of Social Events." In *ICWSM*. 2007.

[19] Jiang, Y., Meng, W., Yu, C., 2011 "Topic sentiment change analysis." In *Machine Learning and Data Mining in Pattern Recognition*, pp. 443-457. Springer Berlin Heidelberg, 2011

[20] Siganos, A., Vagenas-Nanos, E., Verwijmeren, P., "Facebook's daily sentiment and international stock markets." *Journal of Economic Behavior & Organization*, vol 107, 730-743, 2014

[21] Smailović, J., Grčar, M., Lavrač, N., Žnidaršič, M., 2013 "Predictive sentiment analysis of tweets: A stock market application." In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, pp. 77-88. Springer Berlin Heidelberg, 2013.

[22] Smailović, J., Grčar, M., Lavrač, N., Žnidaršič, M. "Stream-based active learning for sentiment analysis in the financial domain*", Information Sciences*, vol.285, pp.181-203, 2014

[23] Dehkharghani, R., Mercan, H., Javeed, A., Saygin, Y." Sentimental causal rule discovery from Twitter." *Expert Systems with Applications*, vol 41, no. 10 pp. 4950-4958, 2014

[24] Kang, H., Yoo, S. J., & Han, D. "Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews." *Expert Systems with Applications* vol.39 no.5 pp. 6000-6010, 2011.

[25] Li, F., Wang, S., Liu, S., & Zhang, M. "SUIT: A Supervised User-Item Based Topic Model for Sentiment Analysis." In *Twenty-Eighth AAAI Conference on Artificial Intelligence*. 2014