A Public Opinion Keyword Vector for Social Sentiment Analysis Research

Yung-Chun Chang
Graduate Institute of Data Science,
Taipei Medical University,
Taipei, Taiwan
changyc@tmu.edu.tw

Fang Yi Lee
Graduate Institute of Data Science,
Taipei Medical University,
Taipei, Taiwan
m946105008@tmu.edu.tw

Chun Hung Chen
Institute of Information Science,
Academia Sinica,
Taipei, Taiwan
hep_chen@iis.sinica.edu.tw

Abstract—**In the Internet era, online platforms are the most convenient means for people to share and retrieve knowledge. Social media enables users to easily post their opinions and perspectives regarding certain issues. Although this convenience lets the internet become a treasury of information, the overload also prevents user from understanding the entirety of various events. This research aims at using text mining techniques to explore public opinion contained in social media by analyzing the reader's emotion towards pieces of short text. We propose Public Opinion Keyword Embedding (POKE) for the presentation of short texts from social media, and a vector space classifier for the categorization of opinions. The experimental results demonstrate that our method can effectively represent the semantics of short text public opinion. In addition, we combine a visualized analysis method for keywords that can provide a deeper understanding of opinions expressed on social media topics.**

Keywords—social media, sentiment analysis, reader emotion

I. Introduction

Due to the booming of social media in the past few years, a spectacular amount of data has been produced. It is a very valuable and important resource for people to understand public opinion [1][2]. Analyzing public opinion is critical to understanding the general impression of a given topic. It can be achieved through an investigation of social media. One of the numerous applications of this technology is to understand the trends in political elections. During the period of the election, a candidate can utilize the public opinions expressed on the social media to capture important issues and make corresponding adjustments in order to gain more support from the general public [3]. For instance, during the Taipei mayoral election, the candidate Dr. Wen-je Ko's campaign used public opinion analysis to determine the public's favoured keywords. Policies and activities were then organized according to the interests of younger voters. The investment in public opinion analysis helped Dr. Ko win the election. This case demonstrates that exploring and analyzing social media can be a powerful means to understand the trends of public opinion.

Sentiment analysis is a significant area of research in natural language processing (NLP). Its purpose is to determine the attitudes and feelings expressed in words and their context. It can be separated into two categories, namely, writer emotion and reader emotion. The former refers to the emotion that the writer (author) wants to express when writing an article. The writer usually expresses emotion toward specific issues through emotive language. On the other hand, reader emotion corresponds to the feelings that may be triggered as one reads the articles [4][5]. One important distinction is that writer and reader can have different perspectives on the same content, so their feelings may not be the same. It is not trivial to determine the reader's emotion directly through the writer's words. So, compared to the research of writer emotion, the research of reader emotion is more challenging [6]. For example, consider a news title mentioning that the oil price will increase tomorrow such as, 'Gas Prices Rising Tomorrow!' Although there is nothing emotional in it, the reaction of the public toward this news title is presumably negative.

Unfortunately, recent emotional analysis research is mostly focused on writer emotion. Only a few researchers, including Lin et al [7], targeted reader emotion. Related research on reader emotion classification mainly focused on the entire composition, e.g., Chang et al. [8]. They aim to learn a language model that identifies reader emotion, and subsequently use the learned model to assist in emotional resonance writing. The experimental results prove that this method can effectively recognize reader emotion through sentence and semantic structures in the compositions. Chang et al. [12] used the whole context of the news article to classify reader emotion, with special focus on products, movies and literatures reviews. In their research, they proposed an innovative reader emotion classification module through an effective use of emotional keywords. The technique was then applied to the news corpus as practical case study. The researchers calculated each candidate's publicity score based on reader emotion classification, and later predicted the voting trend. They were successful in correlating the percentage of votes obtained with publicity score and reader emotion.

Most of the content on social media consists of short texts with about 200 words (e.g., Twitter or Facebook) [13]. Due to the lack of context data, the efficiency of machine learning models are impaired. However, in the past few years, the research focusing on short text has prospered. Like Bharath Sriram et al. [14] extracted specific field's features from the author's profile and written words, then predefined the data as news, opinions, private messages, etc., in order to improve the efficiency of classifying the short text data. In order to improve the efficiency of dealing with the short text, Guo et al. [15] linked the short text and news corpus for expanding the content, letting the machine more easily understand the short text.

In consideration of the importance of social media analysis and the fact that no previous work was done on reader emotion analysis based on short text, this research aims at obtaining public opinion through an analysis of reader emotion. Consequently, we proposed a method which can analyze the reader emotion of short text. As the experiment result shows, our method can effectively recognize different reader emotion categories. Furthermore, we used the visualization method to understand more about the result. Our research can efficiently obtain the public opinion of related topics and more detailed information about it. Then we can control the development of the subject event and the trend of public opinion. We also attempt to observe the distribution of emotional categories produced by the classifiers and compare it to the actual distribution of articles in each category.

II. RELATED WORK

Previous works regarding reader emotion classification are first discussed below. Extracting critical information in the articles can improve the classification efficiency. Chang et al. [8] proposed a flexible principle-based approach (PBA) for reader-emotion classification and writing assistance. PBA can capture variations of similar expressions by generating distinctive emotion templates. For example, "{國家 $_{\text{country}}$}" : [發生 $_{\text{occur}}$] : [地震 $_{\text{earthquake}}$] : {劫難 $_{\text{disaster}}$}" is generated by PBA for the emotion "worried" template. We can observe that this template captures prominent information about natural disasters (earthquake) that have happened in a specific country, and also expresses that it is related to the emotion "worried". PBA can automatically learn emotion templates from raw data and produce the reasonable emotion templates for humans to understand. The writers can then follow the templates and compose articles that resonate with readers.

Bashaddadh et al. [10] used Named Entities to perform topic detection and tracking (TPT), which is a useful method in the information retrieval field. They used keywords and name entities to cluster the vast information from the internet. The importance of keywords has been noticed by many researchers. Further experiments have been conducted based on critical words or keywords. Tang et al. [11] examined the top 200 words with higher relative log frequency ratios in the categories of their target and linked those words to positive and negative reader and writer emotion. After employing relative log frequency ratios to mine sentiment words, they used those words to predict emotion transition by building 4-class and 2-class SVM classifiers. Their results show that using the sentiment keywords method is useful. Furthermore, Chang et al. [12] proposed an innovative document modeling method with emotional keyword embedding, which is called distributed emotion keyword vector (DEKV), to classify reader-emotion. In their research, they treat keywords of each category very seriously. They use keywords to represent articles and take a likelihood ratio as weighting. If there were some articles that couldn't be match with keywords, they calculated the cosine similarity of the word's vector and found the closest keywords to represent those articles. In the case study, they calculated the publicity score based on reader emotion classification for each candidate and successfully predicted the voting trend.

Using hashtags to analyze social media data sounds like a good idea. Hashtags are specified keywords in social media posts. Kouloumpis et al. [9] used Twitter hashtags (e.g., #bestfeeling, #epicfail, #news) to identify positive, negative, and neutral tweets. These were then used for training the three-way sentiment classifiers. They wanted to evaluate their training data with labels derived from hashtags and emoticons. This is useful for training sentiment classifiers for Twitter or other social media platforms. The results proved that using hashtags to collect training data is useful for classification.

A huge number of informal messages are posted every second on social media platforms, which are mostly in short text form. It is more difficult for machines to comprehend and classify short texts when compared to whole paragraphs. (about 200 words) Bharath Sriram et al. [14] bring up a method which is called SentiStrength. It can predict positive emotion with 60.6% accuracy and negative emotion with 72.8% accuracy on short text from social media. Different from the existing emotion classification methods that tend to be commercially used, they focused on the emotion of the users. Their research started from the users' behavior to classify the emotion, like the authors' profile or written words from the past. Guo et al. [15] noticed that the classification efficiency is quite low when the past research methods, which are designed for large context of text, used on the short texts like posts on Twitter. They proposed a method for short texts classification, which is called Linking-Tweets-to-News. It benefits most off-the-shelf NLP tools.

The unconventional steps we took were focusing on the reader emotion of short texts and using keywords to represent each short text. This is extraordinary progress, because no one has done research focusing on this combination of short text and reader emotion. We realized the important role keywords play in emotion classification and the difference of whole paragraph and short text classification. We used keywords to represent the news titles, which we took as short texts, to train our classifier. As the experiment shows, our method significantly improved the efficiency of reader emotion classification.

III. METHODS

Fig. 1 is an illustration of the system architecture of the proposed model in this research. First, we extract keywords for each opinion category. Then, we propose the Public Opinion Keyword Embeddings (POKE) to represent the document, which then combines support vector machine (SVM) to train our classifier. After training, we can target data from specific topics on social media and recognize public opinion. Finally, we propose a method of visualization, which can reveal more detail of each expressed public opinion. We will provide an in-depth explanation in the following paragraphs.

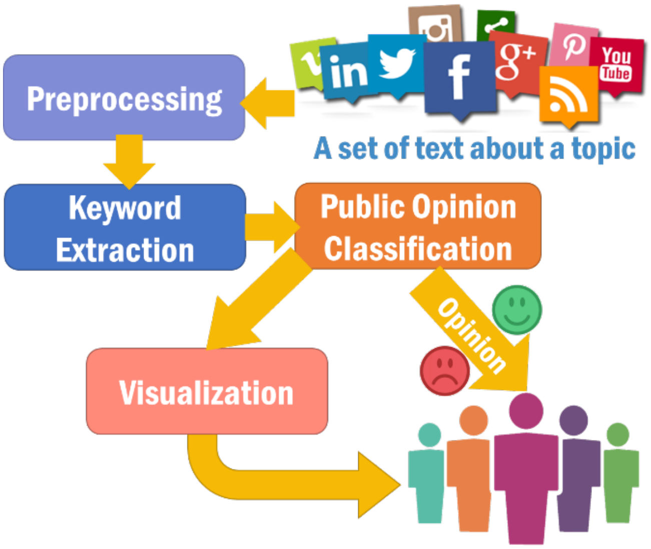

Fig. 1. Systematic architecture of the proposed model for public opinion analysis.

A. Pre-processing

In this research, we applied MONPA [16] for preprocessing of short texts. It is an end-to-end model using character-based recurrent neural network (RNN) to jointly accomplish segmentation, POS tagging and NER of a Chinese sentence. Through this process, we can not only obtain basic information about keywords, but also the named entity recognition which includes personal name, location name and institution name. It helps a lot for the following extraction of keywords. Later on, we removed the stop word in data, and conducting the keywords extraction progress in the remained corpus.

B. Keyword Extraction

According to past text classification research, keywords are effective in improving the performance of classification [8][12]. In this paper, we use log likelihood ratio (LLR) [17] which is an effective feature selection approach to capture keywords in each opinion category. Given a training dataset, LLR employs (1) to calculate the likelihood of the assumption that the occurrence of a word w in opinion O is not random. In (1), O denotes the set of short texts of the opinion in the training dataset; $N(O)$ and $N(\neg O)$ are the numbers of on-topic and off-topic short texts, respectively; and $N(w$ ^ $O)$ is the number of short text on-topic having w. The probabilities $p(w)$, $p(w|O)$, and $p(w|$ ^ $O)$ are estimated using maximum likelihood estimation. A word with a large LLR value is closely associated with the opinion. We rank the words in the training dataset based on their LLR values and select words with high LLR values to compile an opinion keyword list. The opinion keywords are utilized to represent short texts for reducing the dimension.

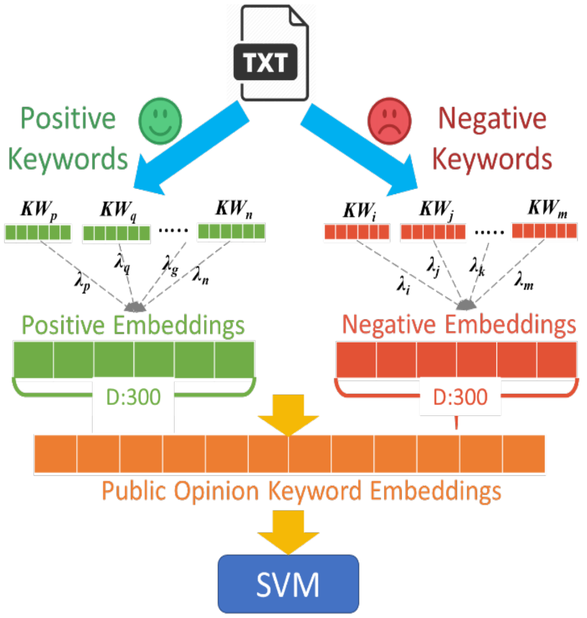

Fig. 2. The short text representative method based on public opinion keyword vector.

C. Public Opinion Classification

Keywords were extracted for each of the public opinion categories, and represented by word embeddings. As shown in Fig. 2, the short text representation method is based on combining opinion keyword vectors from both positive and negative categories. Using LLR, we can collect positive and negative opinion keywords KW, where each keyword KW_i is represented by 300-dimension vectors. A weight λ_i is assigned to each keyword vector, and Eq. 2 is used to combine opinion embedding (OE), which eventually merges OEs from positive and negative sides. We can thus obtain a 600 dimensions distributed opinion keyword vector (DOKV) which can effectively represent the short text D_t.

In Eq. 2, n is the amount of keywords in text, j is word embedding dimension (which is 300 in this study). For instance, a short text composed by 5 words (w_1, w_2, …, w_5), where w_1 and w_2 are positive keywords, and the remaining three words are negative keywords. The weighted average of keyword embeddings is adopted for representing positive and

$$-2\log\left[\frac{p(w)^{N(w\wedge O)}(1-p(w))^{N(O)-N(w\wedge O)}\,p(w)^{N(w\wedge\neg O)}(1-p(w))^{N(\neg O)-N(w\wedge\neg O)}}{p(w\,|\,O)^{N(w\wedge O)}(1-p(w\,|\,O))^{N(O)-N(w\wedge O)}\,p(w\,|\,\neg O)^{N(w\wedge\neg O)}(1-p(w\,|\,\neg O))^{N(\neg O)-N(w\wedge\neg O)}}\right] \qquad (1)$$

negative opinions, respectively. Finally, we integrate both representations to derived the public opinion keyword embeddings which is a 600-dimension vector for short text representation. However, natural language processing and text mining researches usually face the problem of data sparseness, especially for short text. Thus, we propose that if there is no keyword in the text, we can use *kNN* model to infer the word embedding. At first, we turned out text into vector then throw it to the two words pool: Positive keywords' pool and negative keywords' pool. Then we calculate the cosine similarity and find the 5 nearest keywords to represent that text in both pool, as Fig. 3 shows.

$$OE = \sum_{i=1}^{n} \lambda_i KW_{ij} \qquad (2)$$

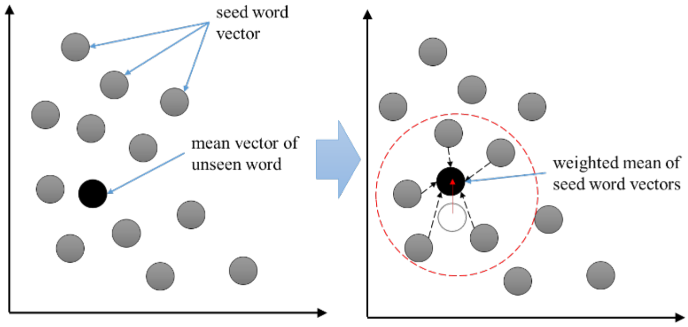
seed word vector
mean vector of unseen word
weighted mean of seed word vectors

Fig. 3. Public opinion keyword vector representative method based on kNN.

D. Visualization and Analysis

Because the keyword extraction module can extract keywords in each public opinion category, we can analyze the public opinion of texts through the public classification module. This module can produce a word cloud using the keywords and public opinion after classification. Furthermore, collecting this information helps us to better understand the relationship between various topics and public opinion.

IV. Experiment Results and Analysis

Given this research is focusing on the reader emotion of short text, we collected the Yahoo! kimo news from 2014 to 2016 and the user voted emotional state toward each news article. There are 8 different emotional categories, including happy, warm, odd, informative, angry, boring, depressing and worry. Due to the aim of this research, which is focusing on analysis of positive and negative public opinions, we summarized the data into two categories. Happy, warm, odd, and informative were treated as positive opinions; and angry, boring, depressing and worry as negative opinions. We separated the data into a training dataset (34,334 news) and testing dataset (12,921 news) as the corpus of estimating systematic efficiency. Its distribution is shown in Fig. 4.

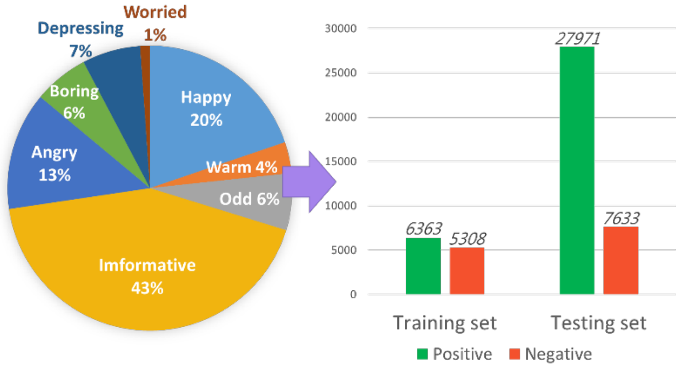

Fig. 4. Positive and nagative public opinion news article amount distribution diagram.

Table 1 shows a comparison of our method to other methods, including Naïve Bayes (NB), Decision Tree (DT), Logistic Regression (LR), and LibShortText (LST). Our method has the highest effectiveness across public opinion categories. In positive public opinion, we obtained 78.7% recall and 84.3% F_1-score. Comparing to DT, our method improves about 43% recall and 33.4% F1-score. In negative reader emotion classification, we achieved 47.4% precision and 56.7% recall, which translates to gaining 21.2% precision and 16.8% F1-score when comparing to DT. For Macro- and Micro-average, our method is able to accomplish the highest score in all classifiers. In consideration of our method, learning positive and negative polarity words of training dataset titles is very useful. As Fig. 2 shows, we created 300 dimension word embeddings and added then to SVM as features. Other methods mainly utilize bag-of-word features to represent the words. It may be the reason the proposed method is more effective at discriminating the emotion contained in short texts.

Interestingly, we observed that the precision scores are commonly lower and the recall scores are generally higher in the negative category. The reason might come from the fact that the amount of negative data is less than positive data in the testing set. There are 27,971 positive news titles and 7,633 negative news titles. The data distribution is very biased, in that the amount of positive news titles to the amount of negative news titles is about 3.7 to 1. We speculate that this bias is the reason why the precision scores are commonly lower and the recall scores in negative perspective are generally higher.

TABLE I. PERFORMANCE EVALUATION

Method	NB	DT	LR	LST	POKE
	Precision / Recall / F_1-score (%)				
Positive	85.8/11.8/20.7	88.6/35.7/50.9	**91.1**/26.7/41.3	90.3/74.3/81.6	90.7/**78.7**/**84.3**
Negative	22.4/**92.9**/36.1	26.2/83.3/39.9	25.3/90.5/39.5	43.0/70.9/53.5	**47.4**/70.4/**56.7**
Macro avg.	54.1/52.4/28.4	57.4/59.5/45.4	58.2/58.6/40.4	66.7/72.6/69.6	**69.1**/**74.6**/**70.5**
Micro avg.	72.2/29.2/24.0	75.2/45.9/48.5	77.0/40.4/40.9	80.2/73.6/75.5	**81.4**/**76.9**/**78.4**

Moreover, we visualized the public opinion keywords with a word cloud. As Fig. 5 shows, the green colored words represent positive public opinion and red colored words negative public opinion. The font sizes express their LLR score, where a bigger font size denotes a stronger relationship with this specific public opinion category. Consequently, through the visualization, we can discover more hidden relations between public opinion and keywords. For example, it concluded a lot of words related to sports in the positive public opinion, like NBA, MLB, Miami Heat, French open (tennis) and so on. Topics related to social welfare promotion, like volunteering, love, social benefit, vulnerable groups and so forth also correlate to positive public opinion. Some words which also make people feel positive are about graduation from schools, like graduation ceremony, graduates, senior high schools and universities.

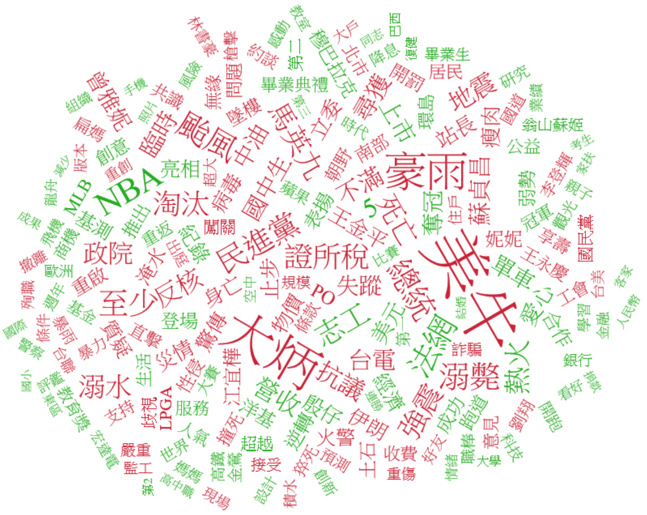

Fig. 5. Word cloud of positive and nagetive public opinion keywords.

There is something worth mentioning, which is that political events frequently make people feel bad. We can clearly tell by the word cloud that the events related to 'American beef' creates negative public opinion. The reason for this phenomenon is that at that time, the discussion on whether or not to import American beef into Taiwan upset a lot of people and created extreme controversy and political unrest. Some policy words and government officials' names are related to some negative public opinions as well, such as income tax on capital gains from stock ownership. At that time, there was discussion about raising this kind of income tax and it caused a surge of negative public opinion.

In addition, an actor named Tony Fish also brought about noticeable negative public feelings, even though he passed away in 2012, which is not in the period of our data collection. Our data was collected from 2014 to 2016. However, in 2014, Tony's brother was nominated for Golden Melody Awards and his niece received the Mayor's award. They both expressed the sadness toward Tony's passing away, which may have caused reluctance and sadness of the public again. Also, the words related to natural disasters usually come up with negative public opinion, as it is imaginable that Taiwanese people are deeply concerned with extreme weather disasters.

V. CONCLUSION AND FUTURE PERSPECTIVE

This paper proposes a novel research method that is different from existing work that mostly focus on classification of writer emotion analysis and long articles. We devise a model to analyze public opinion derived from reader emotion and classification mechanism which focuses on social media. We extract the keywords and take advantage of the Public Opinion Keyword Embedding (POKE) proposed in this research to represent keywords from different perspectives in vector form. Then, we use the Support Vector Machine (SVM) technique to train the classifier. As the experimental results show, this method can effectively classify public opinion, as well as analyze and visualize public opinion. In future research, we are going to enforce the following tasks. (1) Increase the comparison with the mechanism of this research. (2) Continuously collect more social media data and conduct the analysis of public opinion.

REFERENCES

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? : sentiment classification using machine learning techniques," In Proceedings of the ACL02 Conference on Empirical Methods in Natural Language Processing, pp. 79-86, 2002.

[2] P.D. Turney, "Thumbs up or thumbs down? : semantic orientation applied to unsupervised classification of reviews," In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 417-424, 2002.

[3] P.T. Metaxas, E. Mustafaraj, "Social media and the elections," Science, vol. 338, Issue 6106, pp. 472-473, 2012.

[4] K.H.-Y. Lin, C.-H. Yang, and H.-H. Chen, "Emotion classification of online news articles from the reader's perspective," In Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 220–226, 2008.

[5] Y.-J. Tang and H.-H. Chen, “Mining sentiment words from microblogs for predicting writer- reader emotion transition,” In Proceedings of the 8th International Conference on Language Resources and Evaluation, pp. 1226–1229, 2012.

[6] K. H.-Y. Lin, C.-H. Yang, and H.-H. Chen, “What emotions do news articles trigger in their readers? ,” In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 733– 734, July 2007.

[7] K. H. Y Lin, C. Yang, and H. H. Chen, “Emotion classification of online news articles from the reader's perspective,” In Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 220-226, 2008.

[8] Y.-C. Chang, C.-C. Chen, Y.-L. Hsieh, C.-C. Chen, and W.-L. Hsu, “Linguistic template extraction for recognizing reader-emotion and emotional resonance writing assistance,” In Proceedings of 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, pp. 775-780, 2015.

[9] E. Kouloumpis, W. Theresa, and D.M Johanna, “Twitter sentiment analysis: the good the bad and the omg!,” In Proceedings of the 5th International Conference on Weblogs and Social Media, pp. 538-541, 2011.

[10] A.O.M Bashaddadh, and M. Masnizah, “Topic detection and tracking interface with named entities approach,” In Proceedings of the International Conference on Semantic Technology and Information Retrieval, pages 215–219, 2011.

[11] Y.Tang and H.-H. Chen, “Mining sentiment words from microblogs for predicting writer-reader emotion transition," In Proceedings of the 8th International Conference on Language Resources and Evaluation, pp. 1226-1229, 2012.

[12] Y.-C. Chang, C.-H. Chu, and W.-L. Hsu, “How do I look? publicity mining from distributed keyword representation of socially infused news articles,” The 4th International Workshop on Natural Language Processing for Social Media in EMNLP 2016, pp. 74-83, 2016.

[13] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, “Sentiment strength detection in short informal text,” Journal of the American Society for Information Science and Technology, vol. 62, Issue 2, pp. 2544-2558, 2010.

[14] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, “Short text classification in twitter to improve information filtering,” In Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 841-842, 2010.

[15] W. Guo, H. Li, H. Ji, and M.T. Diab, “Linking tweets to news: a framework to enrich short text data in social media,” In Proceedings of 51st Annual Meeting of the Association for Computational Linguistics, pp. 239-249, 2013.

[16] Y.-L. Hsieh, Y.-C. Chang, Y.-J. Huang, S.-H. Yeh, C.-H. Chen and W.-L. Hsu, “Monpa: multi-objective named-entity and part-of-speech annotator for chinese using recurrent neural network,” In Proceedings of 8th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing pp. 80-85, 2017.

[17] C.D. Manning, P. Raghavan, and H. Schütze, “Introduction to information retrieval,” Cambridge University Press., 2nd edn, 2008.