A Public Opinion Keyword Vector For Social Sentiment Analysis Research

Jitha Jose[1], Raveena Venu[2], Hridhaya Augustian[3], Amrutha Joseph[4]
[1]Assistant Professor, Dept of Computer Science
[2, 3, 4]Dept of Computer Science
[1, 2, 3, 4] De Paul Institute of Science and Technology, Angamaly

Abstract- *In the Internet era, online platforms are the foremost convenient means for people to share and retrieve knowledge. Social media enables users to simply post their opinions and perspectives regarding certain issues. Although this convenience lets internet become a treasury of knowledge, the overload also prevents user from understanding the whole thing of varied events. This research aims at using text mining techniques to explore popular 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. Additionally, we combine a visualized analysis method for keywords that may provide a deeper understanding of opinions expressed on social media topics.*

Keywords- Social Media, POKE, Reader's emotion, Short text

I. INTRODUCTION

Due to the booming of social media within the past few years, a spectacular amount of knowledge has been produced. It's a very valuable and important resource for people to know popular opinion. Social media enables users to simply post their opinions and perspectives regarding certain issues. It will be achieved through an investigation of social media. Exploring and analyzing social media will be a robust means to know the trends of opinion.

Sentiment analysis could be a significant area of research in natural language processing (NLP). Its purpose is to see the attitudes and feelings expressed in words and their context. It will be separated into two categories, namely, writer emotion and reader emotion. In consideration of the importance of social media analysis and also the fact that no previous work was done on reader emotion analysis supported short text, this research aims at obtaining public opinion through an analysis of reader emotion. Consequently, we proposed a technique which can analyze the reader emotion of short text. Because the experiment result shows, our method can effectively recognize different reader emotion categories. Furthermore, we used the visualization method to know more about the result. Our research can efficiently obtain the general public opinion of related topics and more detailed information about it. Then we are able to control the event of the topic event and also the trend of opinion. We also try to observe the distribution of emotional categories produced by the classifiers and compare it to the particular distribution of articles in each category.

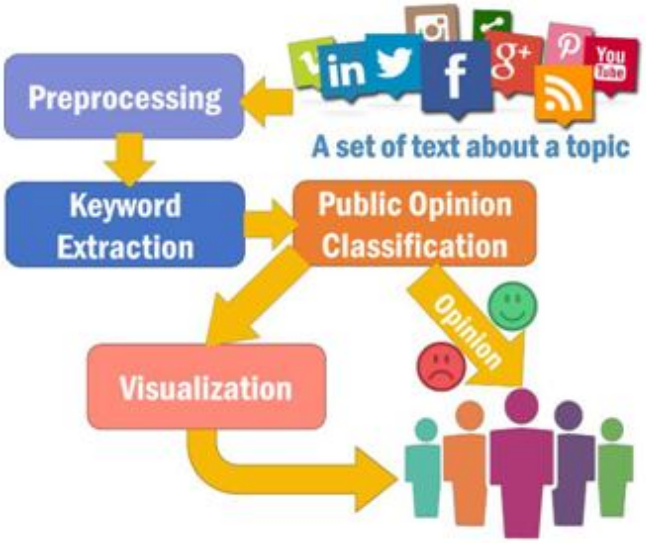
II. EXISTING FRAMEWORK

Analyzing public opinion is critical to understanding the impression of a given topic. The existing system focuses on writer emotions and long articles. The former refers to the emotion that the writer 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. This research aims at using text mining techniques to explore public opinion contained in social media by analysing the reader's emotion towards pieces of short text. 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. Unfortunately, existing emotional analysis research frameworks mostly focus on writer emotions and doesn't perform classification of reader's emotions.

III. PROPOSED SYSTEM

In the proposed framework, 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 is an illustration of the system architecture of the proposed model in this research. probabilities $p(w)$, $p(w|O)$, and $p(w| \wedge O)$ 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.



IV. FOCAL POINTS

A. Pre-processing

In this research, we applied MONPA 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. In this paper, we use log likelihood ratio (LLR) which is an effective feature selection approach to capture keywords in each opinion category. Given a training dataset, LLR employs to calculate the likelihood of the assumption that the occurrence of a word w in opinion O is not random. In LLR, 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 $N(w \wedge O)$ is the number of short text on-topic having w. The

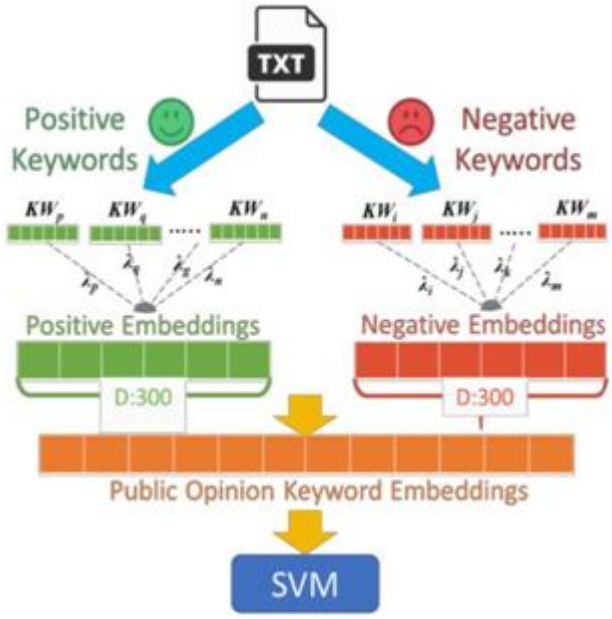


Fig. 2. The short text representative method based on public opinion keyword vector.

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 KWi is represented by 300-dimension vectors. A weight λi 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 Dt. The weighted average of keyword Embeddings is adopted for representing positive and negative opinions, respectively. Finally, we integrate both representations to derive 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.

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. In this system, 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.

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

V. CONCLUSION

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 and E. Mustafaraj, "Social miead 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.