A Public Opinion Keyword Vector For Social Sentiment Analysis Research

Raveena Venu[1], Hridhaya Augustian[2], Amrutha Joseph[3]
[1, 2, 3] Dept of Computer Science
[1, 2, 3] De Paul Institute ofScience and Technology, Angamaly

Abstract- *These days, 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. 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. 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; short text

I. INTRODUCTION

Social media is a very valuable and important resource for people to understand public opinion. Analysing 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. This case demonstrates that exploring and analysing social media can be a powerful means to understand the trends of public opinion.

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 analyse 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. LITERATURE REVIEW

Due to the booming of social media in the past few years, a spectacular amount of data has been produced. This system provides a powerful means to understand the trends of public opinion. This research aims at obtaining public opinion through an analysis of reader emotion. 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 Embedding (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.

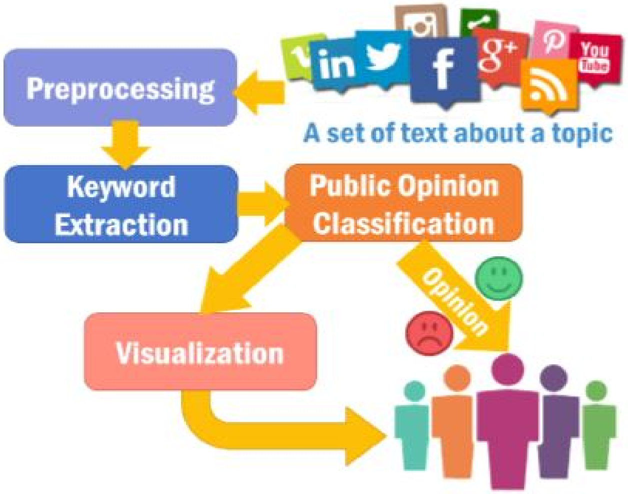

Fig. 1. Systematic architecture of the proposed model for public opinion analysis.

III. STUDIES AND FINDINGS

A. Pre-processing

In this research, we applied MONPA for preprocessing of short texts. 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.

B. Keyword Extraction

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 (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(¬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.

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

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.

C. Public Opinion Classification

Keywords were extracted for each of the public opinion categories, and represented by word embeddings. As shown inFig. 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.

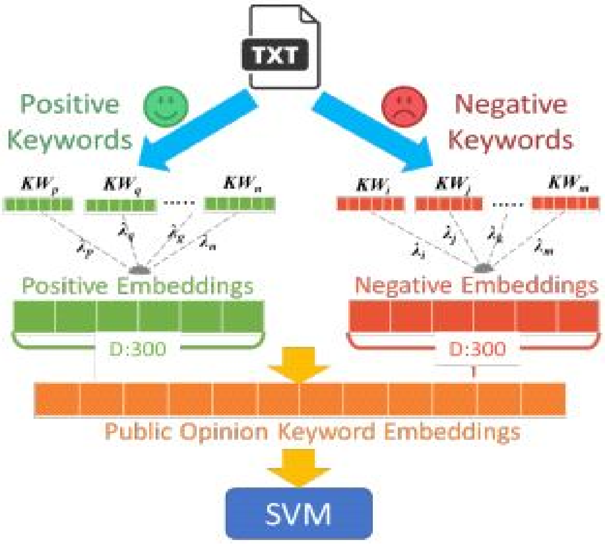

Fig. 2. The short text representative method based on public opinion keyword vector.

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.

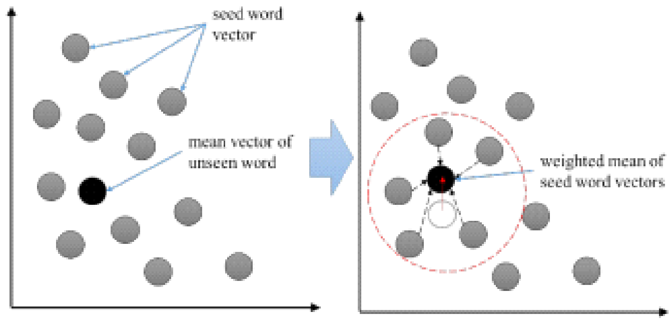
seed word vector
mean vector of unseen word
weighted mean of seed word vectors

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

D. Visualization and Analysis

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.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.

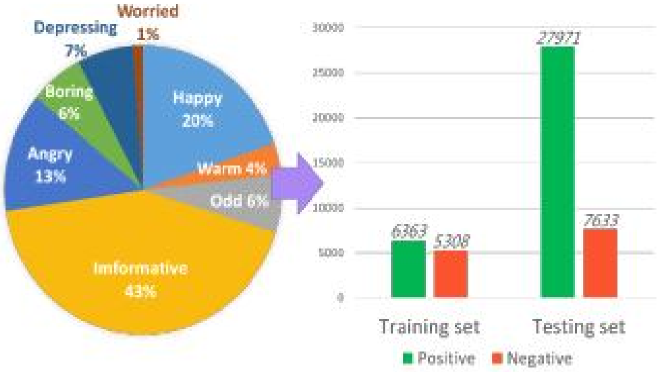

Fig. 4. Positive and negative public opinion news article amount distribution diagram

V. CONCLUSION

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

REFERENCES

[1] [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] [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 media and the elections,” Science, vol. 338, Issue 6106, pp. 472-473 , 2012.