IOSR Journal of Engineering (IOSR JEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 64-67

Short-Text Sentiment Analysis by Context-Based Regularization Method

Miss. Sneha P. Jagtap[1], Dr. V. M. Thakare[2]

Student in Department of Computer Science S.G.B.A.U., Amravati, Head of Department of Computer Science, S.G.B.A.U., Amravati.

Abstract: *Sentiment analysis is an important task in natural language processing, which has promises great value to areas of interests such as business, politics and other fields. The prevalence of the internet has caused people to prefer expressing their opinion and sentiment on the Internet via methods such as tweeting on social media and commenting on products.In this paper, we propose a context-based regularization classification method for short text sentiment analysis. Specifically, we use contextual knowledge obtained from the data to improve performance of the sentiment classification. In this paper, the contextual knowledge includes two parts: word-sentiment knowledge and word-similarity knowledge. Moreover, we propose methods to calculate the sentiment of the words and the similarity between words on semantic level. Specifically, on the one side, we use a TRSR method based on the TextRank algorithm to rank words in each sentiment sample to determine the sentiment polarity of each word. On the other, we calculate the similarity between words according to word-embedding. In this way, we can determine the similarity between words and the sentiment polarity of a word.*

Keywords: *Short texts, Text mining, text analysis, sentiment analysis, regularization, contextual knowledge.*

I. Introduction

With the prevalence of the Internet, more and more people prefer to express their opinion and sentiment on the Internet, using conventions such as tweeting on social media, commenting on products, etc. User expressions and their impact on organizations have gained increasing attention in recent years [1]. In general, user's expressions on social media have two important characteristics [2]. Firstly, those expressions are usually short due to the limits of the imposed by social media platform (either implicitly or explicitly) and the flexibility of accepted internet communication. What's more, in the environment of anonymity prevalent on the internet, people are more likely to express their real attitudes and sentiments than in a face-to-face conversation [3]. Thus, short text modes of communication reflect the real sentiments and opinions of users. Additionally, analyzing user sentiment is a popular and valuable task in practice and has an immense impact on a wide variety of fields [4]. A series of works focused on the sentiment analysis of short text has been done in recent years and researchers apply the resultant sentiment analyzing methods into many real-world situations [5].

This paper presents Sentiment analysis through two methods Dataset Pre-processing use datasets to verify our model's effectiveness, English datasets include movie comments collected by Twitter sentiment classification dataset. Context-Based Regularizations Extraction method finds more sentiment words compared with sentiments tagged by general sentiment dictionary.

II. Background

Sentiment classification has been widely studied in recent years. The origin of the classification model is based on sentiment lexicons, using the sentiment of words from sentiment lexicons to decide the over sentence sentiment. As of this writing, many sentiment lexicons have already been developed, examples include SentiWordNet in English and HowNet in Chinese, both have been applied in many sentiment classification problems. In [1] used the unsupervised method, according to word sentiment, to calculate a PMI index to evaluate the sentence sentiment. In [2] assign a sentiment category to a given sentence by combining the individual sentiments of topic keywords and extract the topic of news from sentence structure. and [3] combined grammar rules and spelling styles except lexicons to identify sentiment and sentiment strength. But the word sentiments usually change with different contexts. In the social media field, supervised and unsupervised models are both applied. method is the first researcher who use the supervised model to classify the sentiment [9]. And he compared the effectiveness of three classification algorithms: Naïve Bayes(NB), Maximum Entropy model(ME) and Support Vector Machine (SVM). In [4] propose to extract various kinds of contextual sentiment knowledge from massive unlabeled samples in target domain and formulate them as sentiment relations among sentiment expressions. It propose three constrains: grammar, co-occurrence and lexicon-based method. A graph-based semi-supervised label propagation method to calculate word similarity [5]. However, contextual knowledge imposed by current research mostly relies on the co-occurrence of words, and lexicons, ignoring semantic relationships. And So, in this paper, we join the semantic contextual knowledge to our model, which can fit real-world short text better.

These are organized as follows.

Section I Introduction. **Section II** discusses Background. **Section III** discusses previous work. **Section IV** discusses existing methodologies. **Section V** discusses attributes and parameters **Section VI** proposed method and outcome result possible. Finally **section VII** Conclude this review paper.

III. Previous Work Done

In research literature, many models have been studied to provide various schemes and improve the performance in terms ofapproaches that have been proposed to facilitate short text understanding by enriching text.

Zhang Xiangyu et al. (2015) [1] has proposed A context-based regularization method For short-text sentiment analysis. It is use contextual knowledge obtained from the data to improve performance of the sentiment classification. In this, the contextual knowledge includes two parts: word-sentiment knowledge and word-similarity knowledge.

OanaFrunza et al. (2011) [2] has focuses A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts. ML-based methodology for building an application that is capable of identifying and disseminating health care information. It extracts sentences from published medical papers that mention diseases and treatments, and identifies semantic relations that exist between diseases and treatments.

Tao Jiang et al. (2015) [3] have utilizingShort Text Sentiment Entropy Optimization Based on the Fuzzy Sets. A short text sentiment classification model called FECEM base on short text entropy optimization method. This method first selects sentiment features based on expectation cross entropy, and then fuzzy sets is used to correct the degree of the comment words.

Yuling Chen et al. (2018) [4]has Research on text sentiment analysis based on CNNs and SVM. A Convolution Neural Networks (CNNs) model combined with SVM text sentiment analysis is proposed. The proposed method improves the accuracy of text sentiment classification effectively compared with traditional CNN, and confirms the effectiveness of sentiment analysis based on CNNs and SVM.

Lu Ma et. al (2016) [5] has proposed Sentiment Orientation Analysis of Short Text Based On Background and Domain Sentiment Lexicon Expansion. A sentiment orientation analysis method based on background and domain sentiment lexicon expansion, which reflects user’s sentiment orientation of evaluation objects more comprehensively.

IV. Existing Methodologies

Sentiment classification model based on CNNS and SVM

The feature extraction method based on machine learning has been unable to meet the current demand, some scholars tried to use the method of deep learning to solve some problems in Natural Language Processing and achieved good results.

- **Brief introduction of Word2vec**

The first step of translating natural language understanding problem into machine learning problem is to find a way to make these symbols mathematically. The basic idea is: By the language model training, each word in a language is mapped into a fixed-length short vector, all of these vectors constitute a word vector space [1].

- **Convolutional Neural Network**

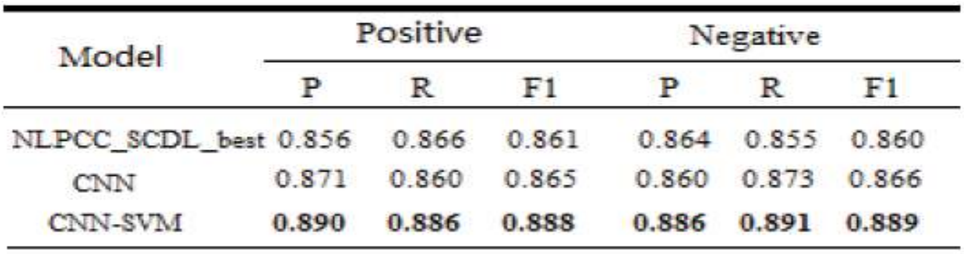
As one of the deep learning models, CNNs is the first supervised learning algorithm to successfully train multilayer network structure. It uses spatial relative relation to reduce the number of parameters to improve training performance. Its essence is multi-layer convolution [2].

- **Text emotion classification model based on CNNs and SVM**

Convolution neural network can extract meaningful feature representation from input samples effectively, but the classification ability of fully connected classification layer is weak for nonlinear separable data. SVM is a supervised machine learning model, which is two-classification model. The SVM method is based on the theory of VC dimension of statistical learning theory and the principle of minimum structural risk. CNNs is good at learning the characteristics of the invariance, and SVM can find the optimal classification surface for the characteristics [4].

V. Analysis And Discussion

The evaluation criteria of emotional analysis task based on deep learning technology in NLPCC2014 is regarded as the evaluation index of experimental results, according to the evaluation standard, calculate the accuracy rate, and recall rate and F value. The results are shown in Table 1



Model	Positive			Negative		
	P	R	F1	P	R	F1
NLPCC_SCDL_best	0.856	0.866	0.861	0.864	0.855	0.860
CNN	0.871	0.860	0.865	0.860	0.873	0.866
CNN-SVM	**0.890**	**0.886**	**0.888**	**0.886**	**0.891**	**0.889**

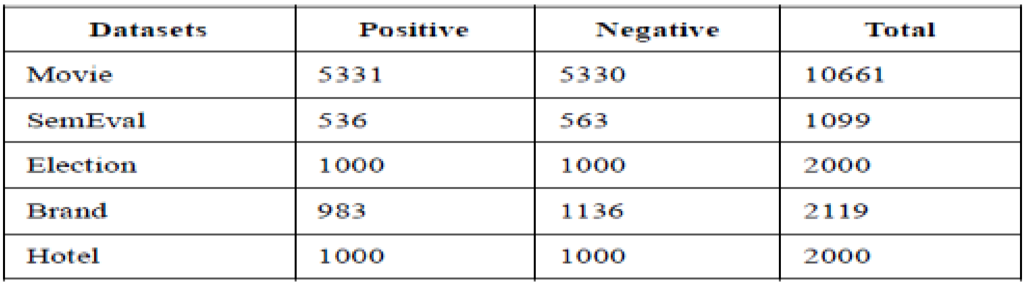
Table 1: The experimental results of the CNN-SVM model and the contrast model on the test set.

According to Table 1, the results show that in the emotion classification task, the accuracy of using the CNN-SVM model is much higher than that of CNN and NLPCC-SCDL-best models, which proves that the model is feasible in dealing with text sentiment classification.

VI. Proposed Methodology

A. Dataset and Pre-processing

In the experiment, we use datasets to verify our model's effectiveness, including three English datasets and two Chinese datasets. English datasets include movie comments collected by Cornell university1, a Twitter sentiment classification dataset. candidates, Hillary Clinton and Donald Trump, from which we sample 1000 positive and 1000 negative comments randomly. We collected user comments of the two candidates. A total of 1000 positive and 1000 negative samples were also collected. All of the datasets we used are concluded two sentiments, positive and negative. The details of the datasets are summarized in table 2.



Datasets	Positive	Negative	Total
Movie	5331	5330	10661
SemEval	536	563	1099
Election	1000	1000	2000
Brand	983	1136	2119
Hotel	1000	1000	2000

Table 2: Statistics of the datasets

B. Context-Based Regularizations Extraction

To establish the model, we need to calculate regularization information. On the one hand, the sentiment polarity of words is decided by TRSR method mentioned above. On the other hand, we use Word2vec to make word-embedding to find similar pairs. In this process, we use extra unlabeled data to train the more accurate word-embedding model.

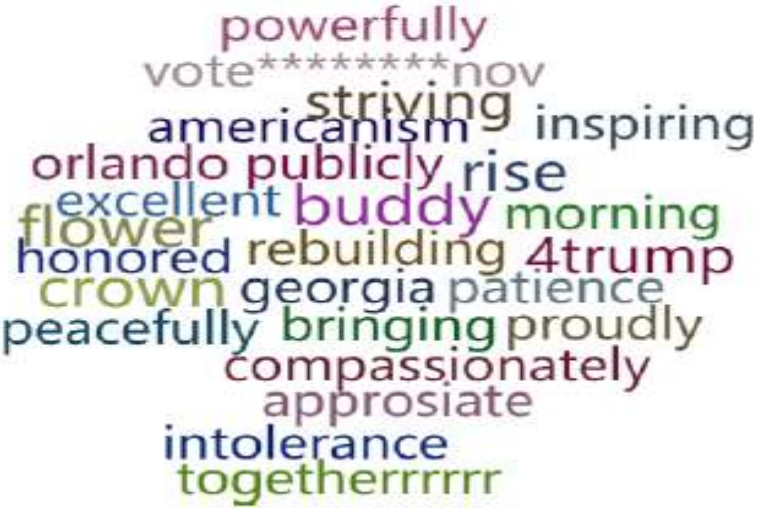


Fig. 1: Example of positive words

From the given example, we can see that our method finds more sentiment words compared with sentiments tagged by general sentiment dictionary. Other than the lexicon words, like positive words: “powerful”, “peaceful”, and negative words: “sexual”, “crooked”, Other sentiment words are discovered, such as “president” or “togetherrrrrr”, and “4trump” as usually being part of the positive polarity. On the contrary, “wikileak”, “video” and “benghazi” are often implied to be a negative sentiment. Based on the above discovery, our method shows effectiveness to calculate the word-sentiment polarity.

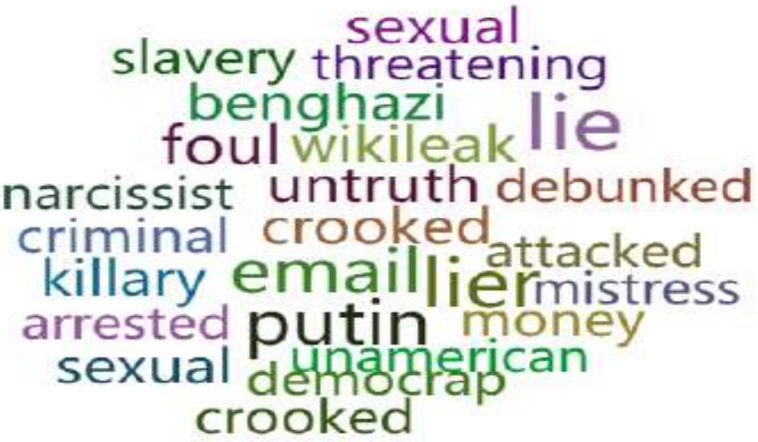


Fig. 2: Example of negative words

VII. Conclusion

In this paper, we propose a context-based regularization method for short text sentiment analysis. More importantly, we use contextual knowledge obtained from the data to improve the performance of the classification. We impose TRSR to calculate the sentiment polarity of a word, and use word-embedding to compute the similarity between words. The contextual knowledge does not only rely on a statistical relationship but also on a semantic level. Last, a unified framework is proposed to combine those two regularizations to a classification model, which converts it to an optimization problem. Then the parameter obtained from training model applies into the logistic regression, and we get the final classification model.

VIII. Future Scope

From observations of the proposed method the future work will include exact short text from large data with the help of methods. one possible research opportunity is to focus on optimization of the algorithm to improve computing performance.

References

[1]. Zhang Xiangyu , Li Hong, Wang Lihong“ A Context-Based Regularization Method for Short-Text Sentiment Analysis”, IEEEInternational Conference, NOVEMBER 2015.

[2]. International Conference, “A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,, VOL. 23, NO. 6, JUNE 2011.

[3]. Tao Jiang, Bin Yuan, Jing Jiang and Hongzhi Yu “Short Text Sentiment Entropy Optimization Based on the Fuzzy sets”, 12th Web Information System and Application Conference, 2015.

[4]. Yuling Chen Zhi Zhang, “Research on text sentiment analysis based on CNNs and SVM”,13th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2018.

[5]. Lu Ma, Dan Zhang, “Sentiment Orientation Analysis of Short Text Based on Background and Domain Sentiment Lexicon Expansion”, 5th International Conference on Computer Science and Network Technology (ICCSNT), 2016.