Journal of Computers Vol. 30 No. 3, 2019, pp. 117-129
doi:10.3966/199115992019063003009

An Emotion Feature Highlighting Method for Sentiment Analysis of Social Media Text

Zi-Qiang Shen[1], Tao Song[1], Qi-Rong Mao[1*], Zhen Jiang[1]

[1] School of Computer Science and Communication Engineering Jiangsu University, Zhenjiang, Jiangsu, 212013, China
{mao_qr, 2211508039}@mail.ujs.edu.cn

Received 26 August 2017; Revised 29 December 2017; Accepted 1 March 2018

Abstract. Sentiment analysis of social media text is a very interesting and important study, which has a wide range of applications both in academia and industry. However, most researchers focus only on the study of semantic features without taking the effect of sentiment information into consideration. Motivated by the needs to sentiment information for sentiment analysis, we propose an Emotion Feature Highlighting Method (EFHM), which is able to utilize both the semantic information and emotion information. In the first step, emotion punctuation and word are combined as a new word in our method to enrich the emotion information of the text in the preprocessing stage. In the second step, the emotional words are extended by calculating the emotional relevance on the basis of the emotion dictionary. In the last step, we extract the feature representation of a text using an improved Continuous Bag-of-Words (CBOW) model, in which relation-specific vector offset is updated according to the emotional weight. Our experiments on three social media datasets show the superior performance for sentiment analysis tasks both in Chinese and English text.

Keywords: Emotion Feature Highlighting Method (EFHM), sentiment analysis, text feature extraction

1 Introduction

Online social media platforms, such as Twitter, Facebook, Taobao, have become the mainstream applications due to the popularity of computer applications and the development of Internet technology. Hundreds of millions of people post their own mood, state, opinion, evaluation and other information in these online social media platforms every day. Analyzing these textual data can help organizations understand the sentiment and opinions of the public and provide scientific evidence for product improvement and decision making. In addition, it's beneficial for the government agencies to understand the public's opinions and wills. According to these, the government can timely advocate the correct guidance of public opinion, promote the transfer of positive energy, and actively take the initiative to deal with various crises. Therefore, sentiment analysis of these social media text data becomes particularly important.

Broadly speaking, existing work on the sentiment analysis of text can be divided into two main parts, which are feature extraction and sentiment classification. Most of the early text feature extraction methods rely on relatively simple statistical technique. Other less-obvious methods of features extraction focus on Part-Of-Speech (POS) tagger. Recently, deep learning techniques have been used in the Natural Language Processing (NLP), and more and more studies tend to extract text feature using deep learning models. As for the approaches to sentiment classification, the main techniques can be divided into two categories, including lexicon-based methods and supervised methods [1].

More specifically, text feature extraction methods based on simple statistical technique mainly consist of a bag-of-words model like Term Frequency-Inverse Document Frequency (TF-IDF) or a word co-

* Corresponding Author

occurrence model like n-grams. For instance, Kim et al. [2] used the TF-IDF statistic of each word and phrase as the lexical feature when extracting the lexical features in the comment data. In addition, some text feature extraction methods based on POS tagger predict sentence emotion by extracting emotional phrases as a feature. Peng et al. [3] studied an unsupervised learning algorithm that extracts the emotional phrases of each comment by the rules of the POS tagger model, and then calculated the emotions of unknown emotional phrases by using the emotional vocabulary collected and the known emotional word in the text. More recently, Neural Network Language Model (NNLM) has attracted more and more attention both in academia and industry due to the development of deep learning. Mikolov et al. [4-6] proposed two new models to calculate the continuous vector representation of words in large data sets in 2013. The two models are intended to learn the vector representation of words which are full of semantic information. Since then, other researchers have applied the two models extensively to natural language processing tasks such as machine translation, natural language understanding, and sentiment analysis [7].

Although advances have been made recently in terms of text feature extraction techniques using deep learning models, accurate sentiment classification is still a challenging problem due to complex factors such as the use of informal language and a large number of meaningless information contained in social media data. Furthermore, the feature representation of the text learned by deep learning models only contains semantic feature, but does not significantly emphasize emotion feature. In order to improve the accuracy of text sentiment analysis, we need to make full use of the useful information in these social media data, especially emotion information. For example, Neviarouskaya et al. [8] used repetitive punctuation, words in uppercase letters and standard abbreviations as features for the study of sentiment classification of texts. In addition, some studies have shown that indirect emotional words also have some influence on the study of text sentiment analysis. Indirect emotional words are terms that are emotionally relevant but not directly express emotion, such as sensation, war, and be late [9].

Motivated by the above observations, in this paper we propose an Emotion Feature Highlighting Method (EFHM) for text sentiment analysis. More specifically, our EFHM has three phases. In the first stage, emotion punctuation and word are combined as a new word to enrich the emotion information. In the second step, the emotional words are extended by calculating the emotional relevance based on the emotion dictionary. In the last step, we extract the feature representation of a text using an improved Continuous Bag-of-Words (CBOW) model. The major contributions of this paper are:

(1) Our emotion feature highlighting method enriches emotional information by introducing emotion punctuations and increasing indirect emotional words. Specifically, we combine emotion punctuation and word as a new word and calculate the emotional weights of words using prior emotional dictionary in the process of emotion information.

(2) We learn the feature representation of the text based on the improved CBOW model. Specifically, by updating relation-specific vector offset according to the emotional weights in the model, we can extract the feature representation with semantic information and emphasize emotion features at the same time, which lead to great performance improvement for text sentiment analysis.

The rest of the paper is organized as follows. Section 2 describes the relevant work. Section 3 describes the architecture of our EFHM and introduces the process of emotion information and how to extract text features by the improved CBOW model in detail. Section 4 introduces and analyzes our experiments. The last section summarizes our research and discusses future research directions.

2 Related Work

For the sentiment analysis of the text, a good text feature is helpful to improve the accuracy of the classifier. Besides, research has shown that emotion information also plays an important role in sentiment analysis task. In the following, we first review the work that concentrates on feature extraction methods for text, and then discuss the literatures about emotion information.

2.1 Feature Extraction Methods for Text

In general, existing work on feature extraction follows two main types of approaches, traditional models and deep learning models. Traditional models used for text feature extraction are generally based on simple statistical techniques, while deep learned model is derived from NNLM.

Traditional models for text feature extraction mainly include TF-IDF, POS tagger and n-gram model.

But beyond that, some studies show that the improved TF-IDF feature weighting method performs better than the traditional TF-IDF method. For instance, Le et al. [10] found that TF-IDF weighting performs better than raw counts in the information retrieval task. Martineau and Finin [11] proposed a novel approach to measure words by calculating the differences between the TF-IDF scores in the positive and negative training corpora. And their experimental results show that the features of Delta TF-IDF obviously work better than TF-IDF or word frequency features. In addition, POS tagger model is also very useful for sentiment analysis. POS tagger aims to label each word with a unique tag that indicates its syntactic role, like noun, verb, adjective etc. It is able to filter out such words that do not contain any sentiment with the help of a POS tagger [12]. Broß et al. [13] preprocessed all text documents by means of the POS tagger in the terminology extraction task.

NNLM was first proposed and studied in depth by Bengio et al. [14]. Their purpose was to fight the curse of dimensionality by learning a distributed representation. Mikolov et al. [15] presented several modifications of the original Recurrent Neural Network Language Model (RNNLM) in 2011, but there are still the problems of the computational complexity. After that, Mikolov et al. [4] proposed two new models that can be used for learning high-quality word vectors from huge data sets with billions of words, and with millions of words in the vocabulary in 2013. Since then, more and more researchers have applied the deep learning model to natural language processing tasks. Li et al. [16] attempted to build a paragraph embedding from the underlying word and sentence embeddings, and then proceeded to encode the paragraph embedding in an attempt to reconstruct the original paragraph. Pennington et al. [17] proposed a global log-bilinear regression model that combines the advantages of the two major model families in the literature: global matrix factorization and local context window methods.

2.2 Emotion Information

Emotion information plays an important role in sentiment analysis. For example, Koto and others [18] did contrast experiments using different textual features in the Twitter sentiment analysis tasks, in which the features with emotion information have been shown to significantly outperform others. The emotion information mainly includes following three parts.

Emotion punctuation. Kiritchenko et al. [19] added punctuation marks (i.e., “,”, “?”, “.”) to the Hashtag Sentiment Base Lexicon, but many of these punctuation did not carry emotional information. Emotion punctuation is generally like exclamation mark and question mark, but when some symbols appear repeatedly, they also express emotion in some degree. For example, Internet users like to use consecutive periods to express helpless or silent mood. Moreover, Chinese Internet users like to use some words with double quotation marks to express the opposite meaning of the word itself. For instance, people can be brazen to such a degree is really “admired”! The phrase “admired” in this sentence expresses the opposite meaning by adding double quotation marks.

Emotion dictionary. Emotion dictionary is generally used to judge the emotional polarity by the corresponding judgment rules. At present, the emotion dictionary resources that can be acquired include the HowNet developed by Professor Dong Zhen-dong [20], the Multiple-Perspective QA (MPQA) library developed by Wiebe et al. [21] and the NTU Sentiment Dictionary (NTUSD) organized by Taiwan University.

Emotion transmission. Emotion transmission is that if a non-emotional word and an emotional word co-occur frequently in the same sentence, it is probably that they convey similar emotion information. Such words are defined as indirect emotional words. Some indirect emotional words can be used to improve the performance of the sentiment analysis system, such as “cheap” and “inexpensive”, which express positive emotion in the product review data [22-23].

From the above we can see that the text feature representation with semantic information learned by the deep learning methods is of the greatest usefulness in sentiment analysis. However, these features have not significantly emphasized the emotion features of the text. In this paper, we propose to combine an improved CBOW and emotion information into a unified framework that is used to extract text feature. Thus, our method can learn a text feature representation with semantic and emotion information.

3 Emotion Feature Highlighting Method for Sentiment Analysis

In this paper, we propose an EFHM for sentiment analysis of social media text, which is shown in Fig. 1. This framework can be broken into the following four steps: sentence inputting, feature extraction, feature representation and sentiment classification.

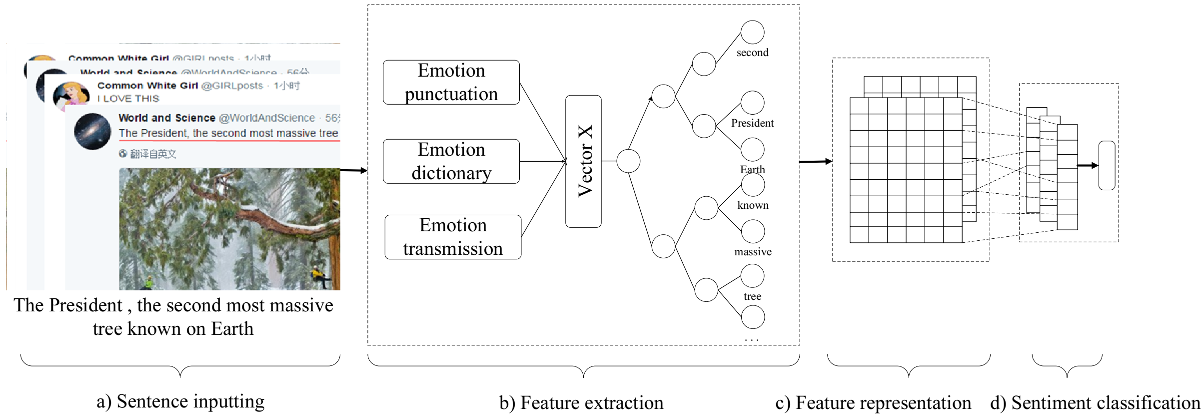

Fig. 1. Framework of emotion feature highlighting method for sentiment analysis

(1) Sentence inputting: In the sentence inputting section, we crawl and sort out a lot of text data as the input samples from social media platforms such as micro-blog, Taobao and Twitter. Each sample is defined as $S(w_1, w_2, \ldots, w_k)$, w_i represents the i –th word and k is the length of each sample.

(2) Feature extraction: In the feature extraction part, we add an emotion feature highlighting method to learn the feature representation of the text based on an improved CBOW model, which is a deep learning model. Specially, emotion punctuation and word are combined as a new word to enrich the emotion features of the text in the process of emotion information. Furthermore, the indirect emotional words are also introduced to extend our emotion dictionary by calculating the emotional relevance between in candidate words and emotional words. In the improved CBOW model, a dictionary with emotional weights obtained by the EFHM is used as an input. Moreover, the relation-specific vector offsets are updated according to the emotional weights in this model.

(3) Feature representation: In the feature representation section, we get a word vector representation of semantic and emotional features through the previous parts. Then we let $x_i \in R^m$ represent the i –th word vector in the sentence, in which m is the dimension of the word vector. The sample data with length n is expressed as

$$X_{1:n} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}. \quad (1)$$

(4) Sentiment classification: In the sentiment classification part, we use Convolution Neural Network (CNN) and Long Short Term Memory (LSTM) networks as the classification model respectively. (CNN) has recently been shown to provide the most advanced performance in the sentiment classification of social data [24-26], which effectively solve the problem of multi-level feature representation. In addition, the LSTM Neural Networks encounter some limitations in performance with multi-level features, but can track the information of time while implementing depth representations in the data [27-28].

The following discussion will focus on the process of emotion information and text feature extraction using an improved CBOW model in our EFHM.

3.1 Processing Emotion Information

Fig. 2 shows the flowchart of the process of emotion information in EFHM. There are two main stages in the process of emotion information. In the first step, emotion punctuation and word are combined as a new word to enrich the emotion features of the text in the process of emotion punctuation. In the second step, the emotional words are extended by calculating the emotional relevance on the basis of the emotion dictionary.

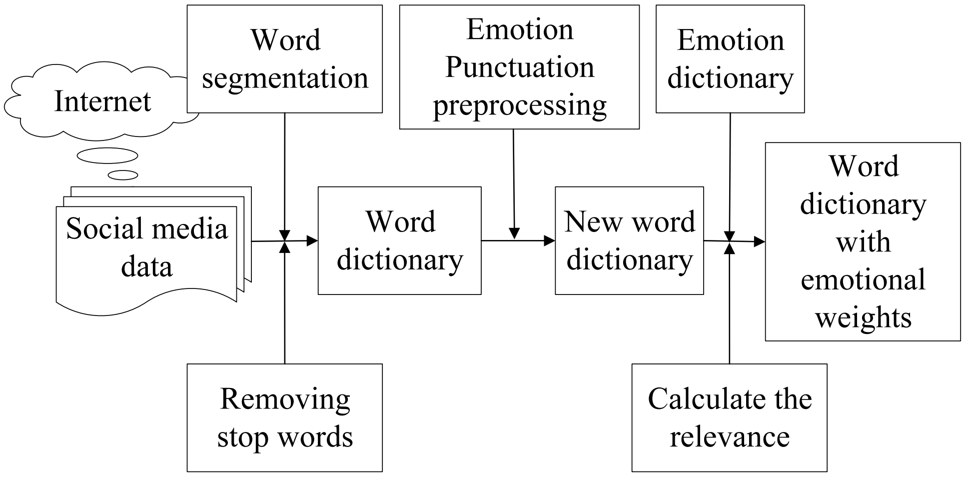

Fig. 2. Flowchart of the process of emotion information

More specifically, to select valid emotion punctuation, we count and analyze the punctuation marks that appeared in datasets from social media platforms. After that, we build an emotion punctuation dictionary using punctuations with emotional information, which is shown in Table 1.

Table 1. A dictionary of emotion punctuation

Symbol	Name	Description
!	Exclamation marks	Exclamation mark can express the feelings of praise, joy, anger, sigh, surprise, mourning and so on.
!!, !!!	Consecutive exclamation marks	Consecutive exclamation marks express more intense emotions when the symbol appears repeatedly.
?	Question mark	Question mark generally represents the doubts and questions.
??, ???	Consecutive question marks	Consecutive question marks can make contribution to emphasis.
。。。	Consecutive Chinese period	Consecutive Chinese period is a non-standard ellipsis, which generally refers to the meaning of speechlessness.
..., ...	Consecutive English period	Consecutive English period is also a non-standard ellipsis.
~~	Consecutive tilde	Consecutive tilde is a bit like an ellipsis but it is lighter than the ellipses in terms of tone.
“ ”	Double quotes	Some words with double quotes are used to express the opposite meaning of the word itself in Chinese.

Then we combine emotion punctuation and word as a new word, which can rich emotion information of text. Some examples are shown in Table 2. For example, Chinese Internet users like to use some words with double quotes to express the opposite meaning of the word itself. The phrase “smart” in first sentence means not smart, which expresses a feeling of irony.

Table 2. Examples in the process of emotion punctuation

Examples	General word segmentation	Our word segmentation
You are “smart” oh!	You/are/smart/oh	You/are/“smart”/oh!
A friend said it looked good, but could not use it! ...	A/friend/said/it/looked/good/but/could /not/Use/it	A/friend/said/it/looked/good/but/could/n ot/use/it! /...

In the second step, we use the HowNet as the emotional benchmark dictionary, which contains 9,193 Chinese evaluation words/phrases and 9,142 English evaluation words/phrases. This dictionary provides richer emotional resources for sentiment analysis of text. Based on 1, we stipulate that the sentiment weights of positive and negative words are 2 and -2 respectively.

Then, on the basis of emotional benchmark dictionary, we judge the emotional tendencies of the words to increase the emotion features of the text by calculating the Point Mutual Information (PMI) between the candidate words and the benchmark words. The PMI is calculated as follows:

$$PMI(w_1, w_2) = \log_2\left(\frac{p(w_1 \,\&\, w_2)}{p(w_1)p(w_2)}\right) = log_2 \frac{freq(w_1, w_2) \times N}{freq(w_1) \times freq(w_2)}, \tag{2}$$

where w_1 and w_2 represent the candidate words and the benchmark words respectively, $p(w_1 \,\&\, w_2)$ is the probability that the two words w_1 and w_2 co-occur, $p(w_1)$ and $p(w_2)$ respectively represent the probability that the two words appear separately, and $freq(*)$ is the frequency of the $*$ that appears in the corpus. N is the total number of words in the corpus. We set a threshold τ, when $PMI(w_1, w_2) > \tau$, the candidate word and the benchmark word express the same emotional tendency, that is, they have the same emotional weight.

By following the steps outlined above, we get a dictionary $D\{(w_1,s_1),(w_2,s_2),\cdots,(w_N,s_N)\}$ with emotional weights by the process of emotion information, where s_i is the emotional weight of the i-th word and N is the total number of words in the corpus. Finally, the dictionary D is used as an input to an improved CBOW to learn the feature representation of text.

3.2 Text Feature Extraction Using Improved CBOW

In the feature extraction part, we add the emotion feature highlighting method to learn the feature representation of the text based on the improved CBOW model. The architecture of the improved CBOW model for text feature extraction is shown in Fig. 3, which includes three parts: input layer, projection layer and output layer.

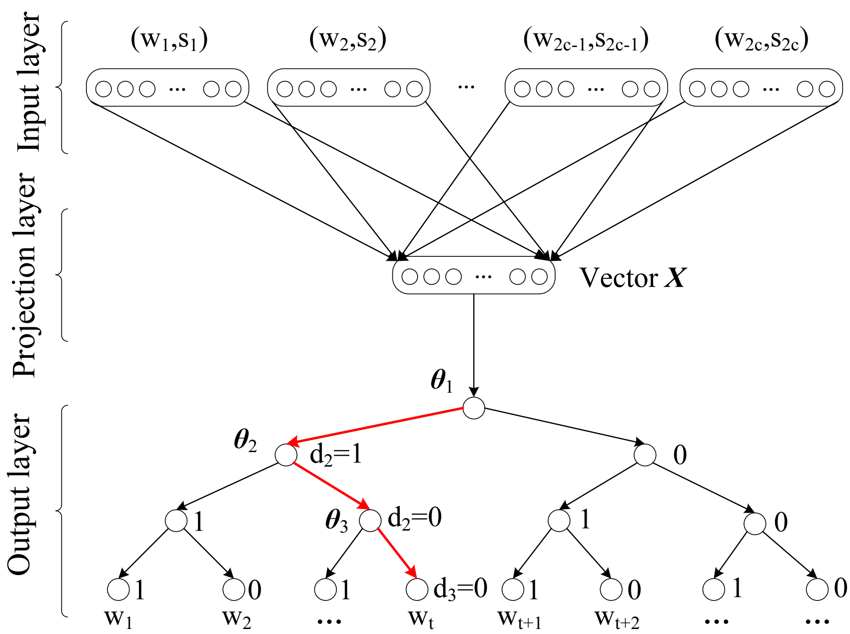

Fig. 3. Architecture of text feature extraction using improved CBOW

The idea of the CBOW model is to predict the current word w_t under the premise that the context $w_{t-c}, w_{t-c+1}, \ldots, w_{t+c-1}, w_{t+c}$ of the current word w_t is known. However, in our method, the difference from the CBOW model is that the dictionary $D\{(w_1,s_1),(w_2,s_2),\cdots,(w_N,s_N)\}$ with different emotional weights learned by the process of emotion information is used as an input to input layer. The effect of the projection layer is to sum the context vector of the current word. The result is denoted as X_w,

$$X_w = \sum_{i}^{2c} s_i v(Context(w)_i) \in R^m, \tag{3}$$

where w is the current word, $v(Context(w)_i)$ is the i-th context vector of the current word w, c is the length of the sliding window, s_i is the emotional weight of the i-th word, and m is the length of the word vector. The output layer corresponds to a binary tree, which is a Huffman tree whose leaf nodes are words that appear in the corpus and the weights are the number of occurrences of the words in the corpus. For any word w in the dictionary D, there must be a path p^w from the root node to the corresponding leaf node of w in the Huffman tree. Assuming that there are $l^w - 1$ branches on path p^w, each branch is treated as a binary classifier, the left child node is 1, and the right child node is 0. The probability of each word is given as follows:

$$P(w \mid Context(w)) = \prod_{j=2}^{l^w} p(d_j^w \mid X_w, \theta_{j-1}^w), \tag{4}$$

$$p(d_j^w \mid X_w, \theta_{j-1}^w) = \begin{cases} \sigma(X_w^T \theta_{j-1}^w), d_j^w = 0 \\ 1 - \sigma(X_w^T \theta_{j-1}^w), d_j^w = 1 \end{cases} = [\sigma(X_w^T \theta_{j-1}^w)]^{1-d_j^w} [1 - \sigma(X_w^T \theta_{j-1}^w)]^{d_j^w}, \tag{5}$$

$$\sigma(X_w^T \theta_{j-1}^w) = \frac{1}{1 + e^{-X_w^T \theta_{j-1}^w}}, \tag{6}$$

where d_j^w is the result of the classification of the binary classifier, θ_{j-1}^w is the model parameter and $\sigma(*)$ denotes a sigmoid function. We usually take the following log likelihood function as the objective function of language model based on the neural network. The objective function is given as

$$L = \sum_{w \in c} \log p(w \mid Context(w)). \tag{7}$$

Then we put $P(w \mid Context(w))$ into Equation 7 and simplify it. The result is as follow:

$$L = \sum_{w \in c} \sum_{j=2}^{l^w} \{(1 - d_j^w) \cdot \log[\sigma(X_w^T \theta_{j-1}^w)] + d_j^w \cdot \log[1 - \sigma(X_w^T \theta_{j-1}^w)]\}. \tag{8}$$

In order to facilitate the calculation, we take the contents of the braces as the objective function, which is written as:

$$\hat{L}(w, j) = (1 - d_j^w) \cdot \log[\sigma(X_w^T \theta_{j-1}^w)] + d_j^w \cdot \log[1 - \sigma(X_w^T \theta_{j-1}^w)]. \tag{9}$$

According to the characteristics of the language model, the objective function is maximized by the random gradient algorithm. Derivations of the parameters θ_{j-1}^w and X_w are calculated as follows:

$$\frac{\partial L(w, j)}{\partial \theta_{j-1}^w} = \eta[1 - d_j^w - \sigma(X_w^T \theta_{j-1}^w)] X_w, \tag{10}$$

$$\frac{\partial L(w, j)}{\partial X_w} = \eta[1 - d_j^w - \sigma(X_w^T \theta_{j-1}^w)] \theta_{j-1}^w, \tag{11}$$

where η denotes a hyper parameter. We know that the CBOW model divides the relationship between each word by capturing syntactic and semantic rules and using relation-specific vector offset. When a sample is inputted, the relevant parameters are updated and the same update is made by using $\frac{\partial L(w, j)}{\partial X_w}$ for $v(Context(w))$. However, in this paper, we update the $v(Context(w))$ according to the emotional weight. The updating function is given as

$$v(Context(w)_i) := v(Context(w)_i) + \frac{s_i}{\sum_k^{2c} s_k} \sum_{j=2}^{l^w} \frac{\partial L(w, j)}{\partial X_w}, \tag{12}$$

where $v(Context(w)_i)$ is the i-th context vector of the current word w and s_i denotes the emotional weight.

During training, emotional words can be quickly dispersed and aggregated because of updating relation-specific vector offset according to the emotional weights. In the end, we learn a text feature representation with semantic and emotional information.

4 Experiments

Based on the above ideas, in this paper we have performed a comparative experiment using different methods in three groups of social media data. Next, we describe our experiments in detail from the five parts of datasets, experimental setup, parameter selection, visualization and performance evaluation.

4.1 Datasets

Our experiments are performed on the three different social media datasets, which include a Chinese comment dataset (CData), a Chinese weibo dataset (WData) and an English twitter dataset (TData).

Table 3. Summary of the datasets from online social media platforms

		CData	WData	TData
TrainSet (50%)	Positive Negative	9287 9164	25000 25000	32730 32718
ValSet (25%)	Positive Negative	4644 4582	12500 12500	16365 16359
TestSet (25%)	Positive Negative	4644 4583	12500 12500	16365 16359

The first dataset is the social media comment data, which consists of three sub-datasets: the Seventh Chinese Opinion Analysis Evaluation-COAE-2015 Task-1 data set, product reviews come from the business platform and Sentiment Classification with Deep Learning data set provided by the conference on Natural Language Processing and Chinese Computing (NLPCC) 2014. It includes 18,575 positive comments and 18,329 negative comments.

The second dataset is from the Data Hall, the first company that focuses on integrated data transactions and services online in China, which is committed to converging and revising all kinds of large data resources to maximize the value of data and promote innovation of related technologies, applications and industries. We select 100,000 data as our data sets, including 50,000 positive data and 50,000 negative data. The last dataset was Twitter data published by Stanford University. And we sort out 65,460 positive data and 65,436 negative data as our third experimental data set.

4.2 Experimental Setup

In our experiments, we divide each dataset into training set, validation set and test set according to the ratio of 2: 1: 1. In the parameter selection, we explore the influence of the parameters. Furthermore, in order to understand the difference between the text features learned by our EFHM and the text feature learned by deep learning algorithms, we visualize the distribution of word vectors learned by word2vec and EFHM. Finally, to further evaluate the performance of our method for the task of text sentiment analysis, we compare it with the following classification methods:

- **N-gram:** In the experiment the bigrams are used as features to train a Support Vector Machine (SVM) [29]. A bigram is an n-gram for $n = 2$. The frequency distribution of every bigram in a corpus is commonly used for simple statistical analysis of text in many applications.
- **TF-IDF:** TF-LDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus [10-11]. It is used as features to train a Native Bayes (NB) classifier in the experiment.
- **POS tagger:** POS tagger aims to label each word with a unique tag that indicates its syntactic role, like noun, verb, adjective etc [12-13].

- **RNNLM:** Recurrent Neural Network Language Model is superior to many competitive language modeling techniques in terms of accuracy, except its high computational complexity [15].
- **Word embedding CNN:** In this method, word vectors are trained by word2vec as initialization features and unknown words are initialized randomly. During the training, all words are kept static and only the other parameters are learned [24-25].
- **Word embedding LSTM:** The input of this method is exactly the same as CNN-static, but it predicts emotional polarity by LSTM [27-28].
- **Sent2Vec:** Sent2Vec can build a sentence embedding from the underlying word, and then proceed to encode the sentence embedding to reconstruct the original sentence [16].
- **GloVe:** GloVe is an unsupervised learning algorithm for obtaining vector representations for words [17]. In the experiment, word embeddings are trained by GloVe as features for sentiment analysis.

4.3 Parameter Selection

The parameters of the improved CBOW model for text feature extraction are tuned by permutation and combination using the validation set. Many different combinations of parameters can give similarly good results. We spend more time tuning the word vector dimension m and threshold τ, since they are the parameters that have the largest impact in the improved CBOW model and the process of emotion information. In Table 4, we show the selected parameter values for our model.

Table 4. Parameters of the improved CBOW model and the process of emotion information for text feature extraction

Parameter	Parameter Name	Value
m	Word vector dimension	{100, 200, 300}
c	Word context window	5
η	Learning rate	0.025
n	Sentence length	140
τ	Emotion threshold	{1.5, 2, 2.5, 3, 3.5, 4, 4.5}

The Choice of Vector Dimension in the Improved CBOW Model. More specifically, to select appropriate parameter m, we compare the accuracy and average accuracy of sentiment analysis with using three different word vector dimensions on three social media datasets. Table 5 shows the accuracy and average accuracy of different dimensions $m \in \{100, 200, 300\}$ respectively. These results show that increasing the word vector dimension can enrich the feature information, yielding better classification performance.

Table 5. The accuracy and average accuracy of sentiment analysis with using three different word vector dimensions and EFHM feature on three social media databases. The classification accuracy is reported in %, and the highest one is highlighted in bold

Parameter	Methods	CData	Datasets WData	TData
$m = 100$	CNN-static	82.0	68.2	71.2
	LSTM-static	77.1	73.4	70.4
	CNN-non-static	84.3	75.4	72.6
	LSTM-non-static	83.0	73.6	**73.3**
average		81.6	72.6	**71.9**
$m = 200$	CNN-static	82.1	74.4	70.3
	LSTM-static	78.1	75.6	71.3
	CNN-non-static	**84.6**	75.4	71.2
	LSTM-non-static	82.5	73.4	72.8
average		81.8	74.7	71.4
$m = 300$	CNN-static	82.1	74.3	70.8
	LSTM-static	80.1	**76.5**	72.4
	CNN-non-static	82.9	76.1	71.1
	LSTM-non-static	82.7	74.5	72.6
average		**81.9**	**75.3**	71.7

On both datasets CData and WData, the average accuracy rate is 81.9% and 75.3% for $m = 300$ respectively, which are higher than that obtained for other values of m. In view of the overall situation, the average accuracy rate increases with the increase of the word vector dimension. Its effects are even well when vector dimension is higher in theory. However, due to limitations experimental equipment, we finally set $m = 300$.

The Choice of the Threshold in the Process of Emotion Information. We set a threshold τ to determine whether the non-emotional word is an indirect emotional word in the process of emotion information. The threshold τ plays an important role in the emotion transmission. Therefore, we analyzed the effects of the threshold for sentiment analysis by changing values of τ with using CNN and LSTM classification method. We draw two curve graphs of the accuracy on CData and TData datasets, which are shown in Fig. 4. Notice that the highest recognition accuracy is achieved when the value of thresholds is close to 4. Thus, we fix it at 4 in all our following experiments.

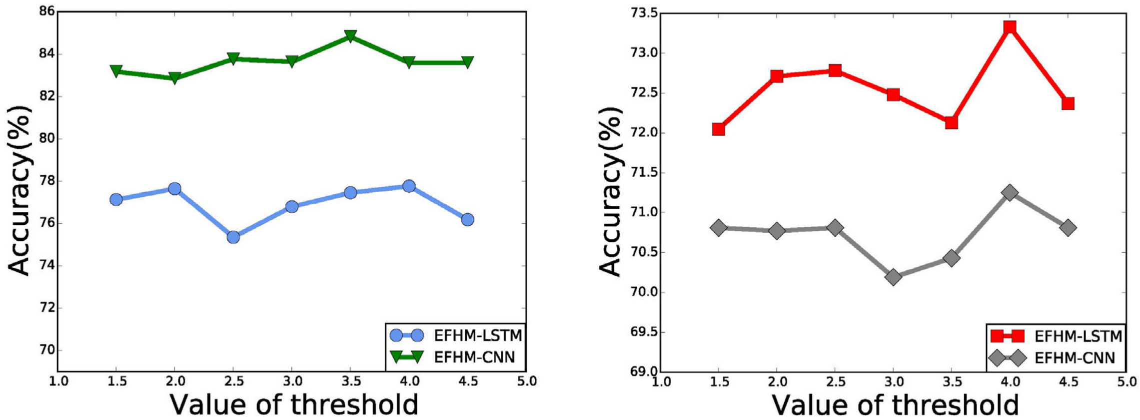

Fig. 4. Accuracy of sentiment analysis with using different thresholds on CData and Tdata datasets

4.4 Visualization

In order to provide intuitive understanding of the learned features by our method, we visualize the distribution of word vectors, which are shown in Fig. 5 and Fig. 6. In Fig. 5 and Fig. 6, circles denote non-emotional words, triangles denote positive words, and squares denote negative words.

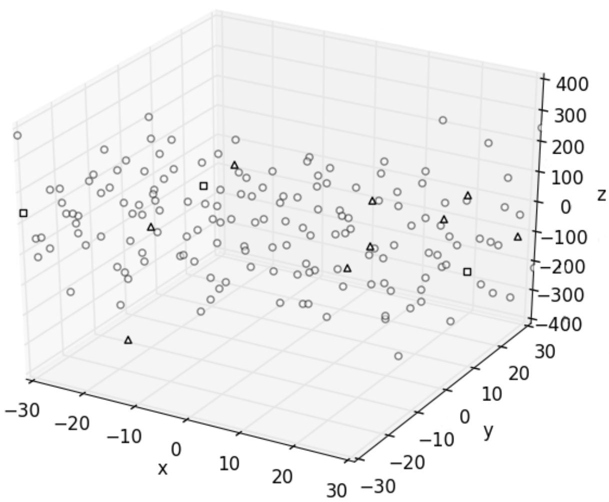

Fig. 5. Word distribution of emotional and non-emotional words with using word2vec

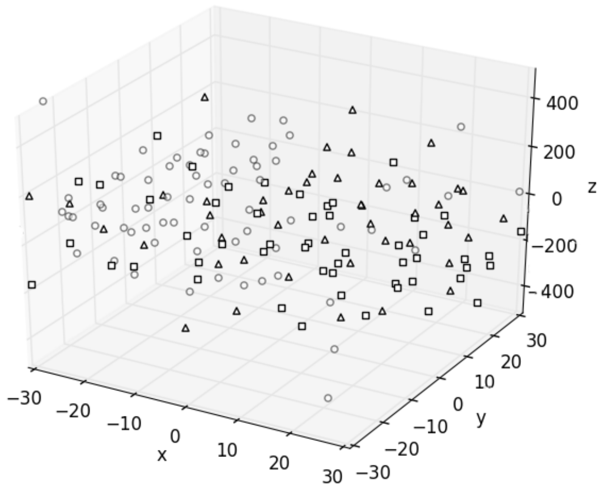

Fig. 6. Word distributions of emotional and non-emotional words with using EFHM

From Fig. 5 and Fig. 6, we find that the proportion of emotional words learned by EFHM is bigger than that learned by word2vec. The emotional words learned by EFHM include not only original emotional words but also emotion punctuation-words and indirect emotional words. Furthermore, words expressing the same emotions are also closer in spatial distribution.

4.5 Performance Evaluation

In this section, we report the classification accuracy on the three social media databases using different classification methods (N-gram, IF-IDF, POS tagger, RNNLM, Word embedding CNN, Word embedding LSTM, Sent2Vec, and GloVe). The results of our methods against other methods are listed in Table 6. The last four results are obtained by our method in which text features are learned by EFHM.

Table 6. The accuracy of different kinds of features on three social media databases using different classification method. The classification accuracy is reported in %, and the highest one is highlighted in bold

Method	CData	WData	TData
N-gram	68.3	69.3	68.5
IF-IDF	70.2	69.2	65.7
POS tagger	55.4	57.7	59.6
RNNLM	67.5	66.9	64.9
Word embedding CNN	79.7	66.5	69.7
Word embedding LSTM	72.9	66.3	69.9
Sent2Vec	67.7	68.6	68.3
GloVe	73.2	75.9	72.3
EFHM -CNN	82.1	74.3	70.8
EFHM -LSTM	80.1	**76.5**	72.4
EFHM -CNN-non-static	**82.9**	76.1	71.1
EFHM -LSTM-non-static	82.7	74.5	**72.6**

These results clearly show that the accuracy obtained by our method is much higher than other methods on three social media databases. More specifically, the first three methods still fall into the traditional model, which have limitations on accuracy and priori knowledge. Furthermore, the RNNLM, Word embedding CNN, Word embedding LSTM, Sent2Vec and GloVe methods with semantic features learned by deep learning algorithms have achieved better results than traditional models. In particular, utilizing both the semantic features and emotion features in our EFHM can further improve the accuracy.

5 Conclusion

In this study, we present a new emotion feature highlighting method for sentiment analysis of social media text. Our main contribution is firstly to train word vectors with semantic information and emotion information simultaneously. In the training process, in order to expand the impact of emotional information, we introduce reinforcement emotion algorithm and extract text feature by an improved CBOW model at the same time. Experimental comparisons to the state-of-the-art methods demonstrate that our method has achieved better performance on the sentiment analysis task of text.

However, it is too complex and non-standard of the social media data to dig out enough information from the text data itself. Furthermore, the emotional tendencies of Internet users are also related to personal personality, the surrounding environment and other factors. So, the next step of our plan is to expand the scope of the data, dig out more useful features and information for sentiment analysis.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.61272211, No.61672267 and No.61672268), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR, No.201700022), the general Financial Grant from the China Postdoctoral Science Foundation (No. 2015M570413) and the Natural science fund for colleges and universities in Jiangsu Province (No.15KJB520006), the Innovation Project of Undergraduate Students In Jiangsu University (No.16A240).

References

[1] H. Saif, Y. He, M. Fernandez, H. Alani, Contextual semantics for sentiment analysis of Twitter, Information Processing & Management 52(1)(2016) 5-19.

[2] S.M. Kim, P. Pantel, T. Chklovski, M. Pennacchiotti, Automatically assessing review helpfulness, in: D. Jurafsky, E. Gaussier (Eds.), Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, July 22-23, 2006, pp. 423-430.

[3] T.C. Peng, C.C. Shih, An Unsupervised Snippet-Based Sentiment Classification Method for Chinese Unknown Phrases without Using Reference Word Pairs, in: Proc. Ieee/wic/acm International Conference on Web Intelligence and Intelligent Agent Technology, 2010.

[4] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, in: Proc. ICLR Workshop, 2013.

[5] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in Neural Information Processing Systems 26(2013) 3111-3119.

[6] T. Mikolov, S.W. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in: Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013.

[7] V. John, A survey of neural network techniques for feature extraction from text. <https://arxiv.org/abs/1704.08531>, 2017.

[8] A. Neviarouskaya, H. Prendinger, M. Ishizuka, Textual affect sensing for sociable and expressive online communication, the in: Proc. 2nd international conference on Affective Computing and Intelligent Interaction, 2017.

[9] C. Strapparava, A. Valitutti, O. Stock, The affective weight of lexicon, in: Proc. Proceedings of the Fifth International Conference on Language Resources and Evaluation, 2006.

[10] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: Proc. the 31st International Conference on Machine Learning, 2014.

[11] J. Martineau, T. Finin, Delta TFIDF: An improved feature space for sentiment analysis, in: Proc. 2009 International Conference on Weblogs and Social Media, 2009.

[12] X. Fang, J. Zhan, Sentiment analysis using product review data, Journal of Big Data 2(1)(2015) 1-14.

[13] J. Broß, H. Ehrig, Terminology extraction approaches for product aspect detection in customer reviews, in: Proc. Seventeenth Conference on Computational Natural Language Learning, 2013.

[14] Y. Bengio, P. Vincent, C. Janvin, A neural probabilistic language model, Journal of Machine Learning Research 3(6)(2003) 1137-1155.

[15] T. Mikolov, S. Kombrink, L. Burget, J. Černocký, S. Khudanpur, Extensions of recurrent neural network language model, in: Proc. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing, 2011.

[16] J. Li, M.T. Luong, J. Dan, A hierarchical neural autoencoder for paragraphs and documents, in: Proc. the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015.

[17] J. Pennington, R. Socher, C. Manning, Glove: global vectors for word representation, in: Proc. 2014 Conference on Empirical Methods in Natural Language Processing, 2014.

[18] F. Koto, M. Adriani, A comparative study on Twitter sentiment analysis: which features are good?, in: Proc. 2015 International Conference on Applications of Natural Language to Information Systems, 2015.

[19] S. Kiritchenko, X. Zhu, S.M. Mohammad, Sentiment analysis of short informal texts, Journal of Artificial Intelligence Research 50(2014) 723-762.

[20] HowNet, HowNet's Home Page. <http://www.keenage.com/>, 2011.

[21] J. Wiebe, T. Wilson, C. Cardie, Annotating expressions of opinions and emotions in language, Language Resources and Evaluation 39(2/3)(2005) 165-210.

[22] Y. Lu, M. Castellanos, U. Dayal, C.X. Zhai, Automatic construction of a context-aware sentiment lexicon: an optimization spproach, in: Proc. the 20th International Conference on World Wide Web, 2011.

[23] L. Velikovich, S. Blair Goldensohn, K. Hannan, R. McDonald, The viability of web-derived polarity lexicons, in: Proc. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010.

[24] Y. Kim, Convolutional neural networks for sentence classification, in: Proc. 2014 Conference on Empirical Methods in Natural Language Processing, 2014.

[25] R. Johnson, T. Zhang, Effective use of word order for text categorization with convolutional neural networks. <https://arxiv.org/abs/1412.1058>, 2014.

[26] X. Zhang, Y. Lecun, Text understanding from scratch. <https://arxiv.org/abs/1502.01710>, 2015.

[27] N.K. Nguyen, A.C. Le, H.T. Pham, Deep Bi-directional long short-term memory neural networks for sentiment analysis of social data, in: Proc. Integrated Uncertainty in Knowledge Modelling and Decision Making: 5th International Symposium, 2016.

[28] Y. Wang, S. Feng, D. Wang, Y. Zhang, G. Yu, Context-aware Chinese microblog sentiment classification with bidirectional LSTM, in: Proc. the 18th Asia-Pacific Web Conference, 2016.

[29] J. Silva, L. Coheur, A.C. Mendes, A. Wichert, From symbolic to sub-symbolic information in question classification, Artificial Intelligence Review 35(2)(2011) 137-154.