Advances in Engineering Research, volume 166
3rd International Conference on Automation, Mechanical Control and Computational Engineering (AMCCE 2018)

Emotional Analysis Oriented to Short Texts

Jiewei Luo
University of Electronic Science and Technology of China, Chengdu 610000, China
ljwxqd@163.com

Abstract: With the development of the mobile Internet, short subjective information such as Weibo comments and product reviews has been rapidly increasing, and massive textual information makes manual management more and more difficult. This paper takes the short text as the research object to analyze the emotion. This paper proposes word vector superposition method and weighted word vector method to extract text features based on word vectors, so that the short text features can be further extracted. In the comparison of the performance of the comment emotion analysis model, the effectiveness of the proposed method is illustrated.

Keywords: Emotion analysis, Short text, Weighted word vector.

1. Introduction

With the rapid development of the Internet, the promotion of social networking platforms such as Twitter, Facebook, Weibo, and Taobao, Amazon, Jingdong and other e-commerce websites has promoted the growing number of commentary text resources online. Faced with a large number of unstructured commentary texts from Weibo and Forums, there is an urgent need to analyze and judge the emotional tendencies expressed in texts through natural language processing techniques. For example, identifying the emotional information of the product attributes from the comments can provide decision support to the businesses and other users. In public opinion monitoring, the government can keep abreast of public attitudes to emergencies and social phenomena and guide public opinion trends.

Emotional analysis, also called opinion mining and opinion mining, refers to the excavation of the emotional tendencies contained in the text by analyzing the statistical and semantic information in the text, such as negative, positive and neutral. According to the granularity of processing text, emotion analysis can be divided into word level, phrase level, sentence level [1], chapter level and multi-chapter level. This paper mainly studies the emotional analysis of the commentary, which belongs to the chapter-based research.

The main structure of this paper is as follows: The first part introduces the related work in the field of emotional analysis, the second part introduces the text feature extraction method based on the word vector model, the third part comparative analysis of the experimental results. Finally, the paper summarizes the work.

2. Related research

Emotional analysis methods are mainly divided into two kinds: semantic-based methods and machine-based methods. The language-based method mainly determines the sentiment tendency of the text through the sentiment dictionary and calculating the sentiment value of the texts [2]. The main method based on machine learning is by extracting the features in the text and using the classification algorithm in machine learning, the model is constructed through a certain amount of sample training to predict the emotional tendency of the new text [3,4].

The semantic-based method can make full use of artificial sentiment dictionaries, but the sentiment dictionaries can not include all the sentiment words. Moreover, the diversification of internet terms makes the construction of sentiment dictionaries more difficult. Machine learning-based approaches build models by learning the features of a given training set using machine learning algorithms. Commonly used in text classification machine learning algorithms include

decision tree, KNN, Logistic regression, support vector machine (SVM) and so on. In practical research and experiments, SVM proved to be more effective than other methods in emotion analysis.

In order to solve the sparsity problem in product reviews feature vector space model, this paper presents a new feature extraction method based on word vector model [5]. Bengio proposed a neural network model of NNLM (Neural Network Language Model) is used to predict the probability of the current generation of words in a given context situation [6]. This model also became the word based vector model.

3. Feature extraction based on word vector model

The traditional method of text feature extraction is based on the vector space model, that is, the text is treated as a sequence of unordered words. This vector space model has the disadvantages of sparsity of data and loss of order information. In order to solve the shortcomings of vector space model, the method of adding some complex text features such as lexical and syntactic to the text feature extraction. With more and more features added, the performance of the text analysis method based on machine learning has been greatly improved. This paper takes the word vector as Based on which the feature expression of text is introduced into the vector space of words and the feature extraction method of text in the vector space is carried out in a variety of ways.

3.1 Word2vec word vector model

Word2vec is a high-quality tool open sourced by Google in 2013 that expresses words as real vectors and is an implementation of the word-vector model proposed by Mikolov et al. Word2vec is an unsupervised learning tool that has not been manually tagged As a training set, the words are mapped to a k-dimensional Euclidean space through neural networks. The features of word vectors in K-dimensional European space also reflect the characteristics of words.

Because Word2vec learns the semantic relationship of the text in the corpus, this requires that the corpus used for training be sufficiently large to ensure the quality of the word vector. This article uses the Word2vec tool to train 20 million product reviews, and finally gets a 500MB Word vector model. The similarity of word vector in K-dimensional space and the similarity of words in text can be illustrated by calculating the similarity between words to illustrate the validity of this vector model.

3.2 The method of word vector overlay text

The word vector model can represent each word as a K-dimensional vector. Product reviews can be seen as word serialization. A simple way to vectorize a product comment is to concatenate the word vector, ie, a product comment with n different words is represented as a vector of n * K dimensions. The disadvantage of this approach is that when n is a large value, a particularly high-dimensional vector is obtained, causing a dimensionality disaster. Each product comment contains a different number of terms, which can lead to inconsistencies in the dimensionality of product reviews.

In order to solve the shortcomings of the word vector splicing method, this paper first proposed the superposition of the word vectors in the product reviews to get the vectorized representation of the product reviews. Word vector superposition will be a dimension and word vector with the same dimension of real comment on the vector of goods. Such as the comments "delicious, cheap, cashier attitude is good, in general, is very good," after the word [delicious, cheap, cashier, attitude, very good, overall, it is, very good]. Each word can be expressed as a K-dimensional vector, and the word vectors of "delicious" and "cheap" are superposed to obtain a K-dimensional vector to express the text in vector. The effectiveness of the extraction is compared with the textual sentiment analysis of the traditional space vector model.

3.3 The method of weighted word vector text direction quantization

TF-IDF is a concept in information retrieval and is also considered as the most important invention in the field of information retrieval. It is widely used in the fields of search and classification. TF or Term Frequency, said a word appears in a document frequency. IDF or Inverse Document Frequency, said the number of documents in the text set contains the word, is the word document frequency. The TF-IDF value is the product of TF and IDF. It considers not only the

appearance frequency of words in a document but also the frequency of appearance of the words in the entire document set. It is a measure of the importance of words in the text.

In order to make full use of the information in the product reviews that play a bigger role in emotion analysis, this article further proposes a method of weighted word vectors, which makes full use of the weight information of words in the product reviews, In the process, the TF-IDF value of the word in the document set is taken as a weight in the process of vectorization.

3.4 Emotional analysis model

In the text classification, there are a large number of classification algorithms, such as KNN, Logistic regression, decision tree and so on. However, a large number of experiments and studies have shown that SVM is better than other classification algorithms in text classification, and a large number of text classification studies are based on SVM [7-9]. In this paper, we construct a text classifier based on SVM algorithm, and compare the validity of the proposed text feature extraction method with the traditional space vector model.

Algorithm pseudo-code:

a) Read manually annotated product reviews
b) Text preprocessing, word segmentation, removal of stop words
c) Product Review Initialization Vector doc2vector=[0,0, ⋯,0]
d) for $word_i$ in [$word_1$,$word_2$, ⋯,$word_n$]
e) if $word_i$ is in a subvector model
f) Take the word vector Wi of $word_i$.
g) Calculate the tf_idf value $weight_i$ of $word_i$ in the document
h) doc2vector= doc2vector+$weight_i$* W_i
i) SVM algorithm for model training to get emotional analysis model

4. Comparison of experimental results

In the last section, we introduced two methods of text feature extraction: word vector superposition and weighted word vector. In order to verify the effectiveness of the proposed method in sentiment analysis, we combine the two feature extraction methods with the linear SVM algorithm to construct the sentiment classification system and compare them with the traditional space vector model.

This article uses 20000 artificially annotated product reviews as experimental data. Among them, 10,000 were rated as good ones and 10,000 were poor ones. Eighty-five comments were taken out of the reviews and bad reviews respectively as training sets and 2,000 reviews were used as test sets for model training and model evaluation. Model evaluation criteria using Precision, Recall and F1-Measure. The following table shows the evaluation results of each model.

Table 1 Experimental result

Method	Positive			Negative		
	Precision	Recall	F value	Precision	Recall	F value
Traditional space vector model	0.8143	0.8685	0.8405	0.8591	0.8020	0.8296
Word vector overlay	0.8490	0.9025	0.8749	0.8959	0.8395	0.8668
Weight word vector	0.8949	0.9325	0.9133	0.9295	0.8905	0.9096

5. Conclusions

The experimental results show that the text vector quantization method based on word vector proposed in this paper has a better effect on the classification effect and fully proves the validity of the method proposed in this article. Commentary text is a kind of text with obvious subjective emotional tendency , The traditional vector space model loses a lot of statistical and semantic information in the feature representation and has the shortcomings of feature sparsity and high

dimensionality. This paper proposes a method of text vectorization based on word vectors, The vector model can control the vector in a smaller dimension and effectively solve the sparseness problem in the traditional vector space model. The weight can preserve the importance of the word in the text.

References

[1] Zhao Yan Yan, Qin Bing and so on. Text Sentiment Analysis, Journal of Software, 2010,21 (8): 1834-1848.

[2] Lin Bin. Semantic Analysis of Chinese Information Based on Semantic Technology [Master's Thesis]. Harbin: Harbin Institute of Technology, 2006.

[3] Cui Zhigang. User Sentiment Analysis Based on Commodity Review Data of E-commerce Website [Master's Thesis]. Beijing: Beijing Jiaotong University, 2014.

[4] Song Jingjing. Analysis and Analysis of Emotional Tendency of Chinese Short Texts [Master's Thesis] .Chongqing: Chongqing University of Science and Technology, 2013.

[5] Zhang Xuegong, et al., Statistical Learning Theory and Support Vector Machines. Journal of Automation, 2000.

[6] Turian J, Ratinov L, Bengio Y. Word representations: A simple and general method for semi-supervised learning. Meeting of the Association for Computational Linguistics. 2010.

[7] Ye Zhigang. Application of SVM in text classification [Master's thesis]. Harbin: Harbin Engineering University, 2006.

[8] Wu Yue. Text Classification Based on SVM [Master's thesis]. Chengdu: University of Electronic Science and Technology, 2014.

[9] Zhang Guoliang, Xiao Chaofeng. Research on News Text Classification Based on SVM. Electronic Technology, 2011.