Research on Weighted Word2vec Algorithm for Fine-Grained Sentiment Analysis

Xu Qian, Sun Baoshan

Abstract— **The rapid development of the Internet has enriched people's lifestyles. People can buy online desired products and services anytime. More and more people are willing to publish their reviews of a product on Taobao, Douyin and other websites. These comments tend to be subjective and include the commentator's evaluation of the emotional attitude. Users are also accustomed to getting valuable information from various comments to assist their own decision-making. Reviews even produce powerful orientation features. Therefore, how to mine users' real emotional tendency from a large number of comment texts has become one of the current research hotspots. As the main technology, text sentiment classification has been extensively and deeply studied. The first problem to be solved is how to efficiently extract text features and convert unstructured product review text to a text representation that computers can understand. Word2Vec is used widely. On the one hand, it ignores ability of the word to distinguish between categories of text. On the other hand,It does not consider the emotional information contained in the words themselves to distinguish the role of the emotional categories of text.**

Based on this, this paper summarizes the existing word2vec method of emotional feature weighting. It includes TF-IDF-Word2Vec, I-TF-IDF-Word2Vec, CR-Word2Vec and incorporate an emotional dictionary-weighted Word2Vec.

Index Terms—**weighted word2vec, TF-IDF, Emotional characteristics.**

I. Introduction

The penetration and popularization of the Internet has promoted the development of online consumption. Product review texts on the Internet, because they contain consumers' emotional attitudes to products or services, so sentiment analysis of these review texts can help other consumers provide consumption decision-making reference, and play an important role in understanding user needs, improving product functions, and adjusting marketing strategies. As a key technology, sentiment classification mainly includes methods based on sentiment dictionaries and machine learning[1]. The former is based on an artificially constructed emotional dictionary to judge the emotional tendency of the text, simply accumulates the emotion of the sentence, does not consider the semantic information of the context, and the accuracy largely depends on the quality of the emotional dictionary. The latter can consider the semantic information of the context, and the emotional tendency of the text is more accurate, but traditional machine learning methods rely on manual extraction of features and have limited accuracy. Deep neural networks are able to automatically learn the features of text and perform better in sentiment classification tasks.

When using deep neural networks for sentiment classification, the first step is to convert the text into a language that the computer can recognize. At present, the widely used word vector representation model is the Word2Vec model. **When training word vectors, Word2Vec model considers the contextual semantic information of words, but ignores different roles of words in text category differentiation.**

II. TF-IDF-Word2Vec

In the past, the sentiment classification based on word2vec word vector only considered the semantic relationship between the word contexts, and ignored the distribution of words in the entire text, TF-IDF calculated weight of words according to the word frequency (TF) and inverse document frequency (IDF) of words to reflect the relative importance of words in the text, which is widely used in the field of text classification, because these two methods are complementary in word representation, Li Rui et al. use TFIDF value to measure the importance of words to entire text. TFIDF-weighted word2vec word vector sentiment classification method is proposed[2]. **It assigns higher weight to words with category discrimination ability, and achieves higher classification accuracy in sentiment classification tasks.**

TF-IDF feature selection method is a common feature weight calculation method[3]. The frequency of the selected word appearing in the text, f_F, can be used to assess its importance to the text, and f_F is calculated as shown in Equation (1):

$$f_F = \frac{f_{w,\max}}{\sum_k^N f_K} \tag{1}$$

Where: $f_{w,\max}$ represents the highest frequency at which the feature word w appears in the text. f_k indicates how often the kth word appears in the text. N represents the total number of feature words in the text. The reverse file frequency f_{DF} can be used to assess the importance of feature words in the text, which is calculated as shown in Equation (2):

$$f_{DF} = \lg \frac{D}{1+Q} \tag{2}$$

Multiply equation (1) with equation (2) to get the text feature word weight, as shown in equation (3):

$$W = f_F f_{DF} \tag{3}$$

The flow of weighted word2vec algorithm is as follows:
1) Pre-word segmentation of all comment information.

Manuscript received July 11, 2023

Xu Qian, School of Computer Science and Technology, Tiangong university, Tianjin, China
Sun Baoshan, School of Computer Science and Technology, Tiangong university, Tianjin, China

2) Use the SG module in the word2vec model for word vector training, and convert the text data after word segmentation into numerical data.
3) Use PCA to reduce the dimensionality of word vectors.
4) Use the TF-IDF algorithm to calculate the word vector weight after dimensionality reduction.
5) Multiply weights with word vectors to obtain the weighted word vector model

III. I-TF-IDF-Word2Vec

The word weight calculation method based on TF-IDF only considers the inverse document frequency of words in the entire corpus, and ignores the influence of the distribution of words between different categories on the ability of words to distinguish categories[4]. According to the formula (2), IDF treats all texts in the data set as a class, and the more text containing a word, the less weight is given to the word, and the less the ability to distinguish between categories of words. **However, in the text classification problem, the IDF has the following two shortcomings in the way it calculates the weight of words as a class:** first, the IDF does not consider the distribution of words among different categories, when the more text containing words in one class, and the less text containing the word in other classes, it means that the word can better represent the category, has better category discrimination ability, and should be given a higher IDF value; Second, TF-IDF for words in certain categories does not consider the impact of word frequency in other categories of word frequency - inverse document frequency [5]. In fact, if a word appears only in a certain category, then it has the strongest ability to distinguish between categories.

In view of the above defects, **Zhao Ling et al. consider the influence of parts of speech on the classification of text topics on the basis of TF-IDF weighted word2vec word vectors, divide the dataset into j class and non-j class, consider the distribution of words among different categories,** propose an improved TF-IDF algorithm to calculate the weight of text words[6]. And weighted the word vector obtained based on Word2Vec model to represent the text, which not only retains the semantic relationship between word contexts, but also uses the contribution of words themselves to distinguish categories.

$$I-IDF_{i,j} = \log(\frac{m_{i,j}}{m_{i,j}+k_{i,j^-}} \times N) \quad (4)$$

Among them, $m_{i,j}$ is the number of documents that contain the word t_i in class j. $k_{i,j-}$ is the number of documents that contain the word t_i not in class j. The improved IDF value increases with the increase of $m_{i,j}$, and decreases with the increase of $k_{i,j-}$. If the number of documents containing the word t_i in class j is more, and the number of documents containing the word t_i not in class j is less. t_i is more able to represent class j. In the text sentiment classification task, word t_i has more ability to distinguish categories, and the word t_i in class j should be given higher weight, reflecting the idea of improving the first defect. For defect two, the redesigned TF-IDF is shown in Equation (5). The improved TF-IDF weights of the word it in class j are:

$$I-TF-IDF_{i,j} = \frac{TF_{i,j} \times I-IDF_{i,j}}{TF_{i,j^-} \times I-IDF_{i,j}^{-} + 1} \quad (5)$$

I-TF-IDF$_{i,j}$ is an improved TF-IDF weight of the word t_i in class j. To avoid the absence of words not in class j, when the denominator is 0, the denominator is added to 1. If the word t_i only appears in class j, the denominator is 1. If the word t_i appears in class j and non-class j, the denominator is greater than 1. In the former case, the word t_i has stronger classification ability and greater weight. In the latter case, I-TF-IDF$_{i,j}$ decreases as the weight of word t_i in the non-j class increases. It reflects the idea of improvement, that is, if the word t_i only appears in the class j, the word t_i has the strongest class discrimination ability. However, as the weight of the word t_i increases in the non-class j, the ability of the word t_i to represent the weakens of class j, and the weight given to it decreases accordingly. Equation (5) is the improved TF-IDF algorithm proposed.

IV. CR-Word2Vec

The traditional weighted word2vec method does not consider the ratio of the number of words in a certain category of text in the dichotomous sentiment dataset to the number of all texts with the word. For example, the number of words such as "ugly", "stupid", "ugly" and other words in negative texts will generally be much greater than the number of occurrences in positive texts, and some words have a poor ability to distinguish the emotional categories of the text, such as "child", "car" and other words as neutral words, and there is no obvious emotional tendency[7]. The number of occurrences of such words in positive and negative emotional texts is not much different. Therefore, Li Jiajun introduced of a category factor CR (CategoryRatio) to measure the proportional relationship between the number of words in this class and the number of words in all classes[8]. CR is expressed in mathematical formulas (6) as follows:

$$CR = \frac{\max(r_{t_i}, r'_{t_i})}{r_{t_i} + r'_{t_i}} \quad (6)$$

where r_{ti} represents the number of documents in category r that contain the word t_i, and r'_{ti} represents the number of documents that contain the word t_i in non-category r.

CT is the product of CR and TFIDF, and the mathematical expression is shown in (7):

$$CT = CR \times TFIDF \quad (7)$$

The first two words of the three words "recommended", "hehe" and "child" belong to words with obvious emotional tendency, and the number of texts in the positive and negative categories is quite different, indicating that these two words can distinguish the two emotional categories very well, while "child" is a neutral word, and the number of negative and positive texts is basically the same, and there is no obvious emotional tendency. The CT value of the three words can strengthen the bias of different words to emotional tendencies, so that the weight of words without obvious emotional tendencies is reduced, so that the model can pay more attention to those words that have an important impact on emotional tendencies.

V. Incorporate an Emotional Dictionary-Weighted Word2Vec

The emotional dictionary is constructed by referring to a large number of relevant materials and drawing on existing dictionaries, which summarizes and summarizes a large number of words with emotional tendencies and describes the degree of emotion[9]. In the expression of the emotional tendency of the text plays a key role in the emotional words, in which the degree adverb is the key to measure the emotional strength, and the positive and negative words can greatly affect the overall emotional tendency of the text, Hou Min et al. [10] By designing calculation rules and constrcting an emotional dictionary, the common degree adverbs are divided into four categories according to the strength of modification, so as to improve the accuracy of emotional classification. Wu Peng et al. [11] increased the emotional meaning of word vectors by adding emotional factors to the word2vec word vector of words, and used the word vector as input to study the emotional classification. Combined with the emotional dictionary to weight the word2vec word vector, it can more accurately describe the influence of words on the overall emotional tendency of the text, and alleviate the shortcomings of the Word2vec model that ignores the different degrees of influence of specific words on the emotional tendency of the entire text.

VI. Conclusion

This article summarizes the existing weighted word2vec methods. Because the word2vec method only focuses on the semantic relationship between contexts, it ignores the distribution of words throughout the text. TF-IDF calculates the weight of words according to the word frequency (TF) and inverse document frequency (IDF) of words, so Li Rui et al. use TFIDF value to measure the importance of words to the entire text, and propose a TFIDF-weighted word2vec word vector sentiment classification method. Considering the influence of the distribution of words among different categories on the ability of words to distinguish categories, Zhao Ling et al. consider the influence of parts of speech on the classification of text topics on the basis of TF-IDF weighted word2vec word vector, divide the dataset into j class and non-j class, consider the distribution of words among different categories, and propose an improved TF-IDF algorithm to calculate the weight of text words. Since the traditional weighted word2vec method does not consider the ratio of the number of words in a certain class in the dichotomous sentiment dataset to the number of all texts with the word, a category factor CR (CategoryRatio) is considered to measure the proportional relationship between the number of words in this class and the number of words in all classes, and the CT value can strengthen the bias of different words to emotional tendencies, so that the weight of words without obvious emotional tendencies decreases. This allows the model to focus more on words that have an important impact on emotional tendencies. Considering the influence of degree adverbs, Hou Min et al. improved the accuracy of emotional classification by designing calculation rules and constructing emotional dictionaries, and classifying common degree adverbs into four categories according to the strength of modification.

References

[1] Yanqiu L, Zekun D, Chinese Movie Comment Sentiment Analysis Based on HowNet and User Likes[J]. Journal Physics:Conference Series, 2019, 1229(1):2-9.

[2] Li Rui, Zhang Qian, Liu Jiayong, Microblog sentiment classification based on weighted word2vec[J]. Communication Technology Surgery, 2017, 50(03):502-506

[3] Shi C, Xu C, Yang X, Study of TFIDF Algorithm[J]. Journal of Computer Applications, 2009, 29(S1):167-170.

[4] Alshuraiqi H S, Improved Term Frequency Inverse Document Frequency (TF-IDF) method for arabic text classification[J]. International Journal of Advanced Trends in Computer Science and Engineering, 2020, 9(5): 6939-6946.[36]

[5] Alenezi M., Zagane M., Javed Y, Efficient deep features learning for vulnerability detection using character N-gram embedding[J].Jordanian Journal of Computers and Information Technology (JJCIT), 2021, 7(1): 25-39.

[6] Zhao Ling, Research on Product Review Text Sentiment Classification Model Based on Weighted Word Vectors [D]. Chongqing University of Posts and Telecommunications, 2022.

[7] Tian W F., Li J., Li H G. A method of feature selection based on Word2Vec in text categorization[C]. 2018 37th Chinese Control Conference (CCC). IEEE, 2018: 880-883.

[8] Li Jiajun, Research on text sentiment classification algorithm based on multi-feature weighting and hybrid network [D].Southwest Jiaotong University, 2021.

[9] Li J, Rao Y, Jin F, et al. Multi-label Maximum Entropy Model for Social Emotion Classification over Short Text[J]. Neurocomputing, 2016, 210(19):247-256.

[10] Hou Min, Teng Yonglin, Chen Yuqi, Research on the Tendency Classification of Evaluation Phrases [J]. Chinese Informatics Journal, 2013, 27(06):107-113.

[11] Wu Peng, Ying Yang, Shen Si, Research on Negative Emotion Classification of Netizens Based on Two-way Long Short-Term Memory Model [J]. Journal of Information Science, 2018, 37(8):845-853.

Xu Qian Graduate student, School of Computer Science and Technology, Tianjin Polytechnic University. He graduated from Shandong University of Science and Technology with a bachelor's degree in fine-grained sentiment analysis, and won a third-class academic scholarship.

Sun Baoshan Associate Professor Sun Baoshan, Doctor of Engineering, Master Supervisor, Deputy Head of Department. The country sent abroad to study in the UK visiting scholar, CCF Member of China Computer Society. Graduated from Tianjin University with a master's degree in computer application technology, and graduated from Tianjin Polytechnic University in computer testing Ph.D. graduate. Selected into the "Outstanding Young Teachers Funding Program" of Tianjin Universities. Leading the research team in related research fields and achieved a series of scientific research results in research topics. And has been in foreign SCI journals, EI journals, domestic important journals and international more than 20 high-level academic papers have been published at the conference.