Research on Automatic Categorization of Product Reviews Based on Short Text Extension and BERT Model

Xiangdong Li
School of Information and Management, Wuhan University, Wuhan, 430072, China, xli_xiao@hotmail.com

Qianru Sun
School of Information and Management, Wuhan University, Wuhan, 430072, China

Jian Shi
School of Information and Management, Wuhan University, Wuhan, 430072, China

Full Research Paper

Research on Automatic Categorization of Product Reviews Based on Short Text Extension and BERT Model

Xiangdong Li[1,2*], *Qianru Sun*[1], *Jian Shi*[1]

[1]School of Information and Management, Wuhan University, Wuhan, 430072, China

[2]Center for E-commerce Research and Development, Wuhan University, Wuhan, 430072, China

Abstract: In view of the fact that the product reviews have the characteristics of short text and nonstandard expression words, this research aims to explore the method of automatic categorization of product reviews by the product categories and its reasons. The core words set of the training set is constructed by TF-IDF and LDA, and short texts are extended by Word2Vec similarity calculation method. After extension, the product reviews are categorized by product categories based on BERT model. The method is compared with the method that based on BERT model without extension and the method of using HowNet similarity calculation to extend based on BERT model. Facing the characteristics of nonstandard expression words, the corresponding experiment is designed to counter test to the effectiveness of the method proposed in this paper. For the product reviews after extension when using BERT classification, the F1 value obtained by the method proposed in this paper is 2.1 percent higher than that when not extended, and it is 0.9 percent higher than that when using Hownet similarity calculation method. The reasons for the effectiveness of the method proposed in this paper are analyzed from the aspects of basic principles, different word similarity calculation methods, and words used methods. The method proposed in this paper can effectively improve the classification performance of product reviews when organizing information by product categories.

Keywords: short text, feature extension, Word2Vec, BERT, product reviews

1. INTRODUCTION

With the rapid development of the Internet, information management systems and e-commerce, the objects of information management have expanded from traditional documents such as books and periodicals to general network resources such as web pages, and have further expanded to product and product reviews. Categorization is one of the two main ways of classification and subject in information management. This research studies the product reviews with nonstandard expression words to be concentrated according to every product, and realizes the automatic categorization method of product reviews by product categories, which provides an effective means for the classification organization of e-commerce information in the field of information management.

Whether product reviews are from an e-commerce platform or a social platform, they are generally only a few words to dozens of words, which are typical short texts. The classification performance of directly applying classical machine learning methods or deep learning methods to short text classification is not good because of short texts with sparse features. Currently, the solution to the problem of sparse features of short texts is to extend the features. Therefore, for the product reviews with nonstandard expression words, this research firstly use TF-IDF and LDA model to construct the core words set of the training set, and then use the Word2Vec similarity calculation method to extend the feature of short texts. After extension, the current mainstream deep learning model BERT is used to realize automatic categorization of product reviews by product categories.

[*] Corresponding author. Email: xli_xiao@hotmail.com(Xiangdong Li)

2. RELATED WORK

With the great development of social productivity and the in-depth application of information management systems and the Internet, information around the product itself and product reviews have become the object of information management. Information management methods related to product are constantly developing and changing, and research on the automatic categorization of products has gradually been carried out. For example, Zhang et al [1]released a paper which reduce the risk of customs transactions by studying the method of automatic categorization of products. With the increase of Internet users and the rapid development of e-commerce, a large number of product reviews have emerged due to online shopping has penetrated into many aspects about people’s daily life. Sentiment analysis of product reviews, which classifies product reviews into two categories, namely positive category and negative category, has become one of the hot spots in e-commerce research. However, the current sentiment analysis of product reviews is basically based on the same product from the same platform, the lack of a certain or more different product reviews from different platforms as the object of sentiment analysis of product reviews, greatly limiting the development of sentiment analysis research of product reviews, further hindering the development of e-commerce research. The main reason is that product reviews are generally short texts with only a few words to dozens of words, and their expressions are not standardized. The reviewers are network users covering the scope from first-tier cities such as Beijing, Shanghai, Guangzhou, etc. to small counties or township areas, so the reviews on the same product have the characteristics of colloquialism, arbitrariness, and locality, making the expression and words used of product reviews extremely nonstandard.

From the perspective of automatic categorization, short text means sparse features, while the expressions of product reviews are not standardized, making the characteristics of expressing the same concept use more different words, resulting in greater difficulties in sparse features. In view of the problem of sparse features of short texts, researchers generally extend the features of short texts firstly, which is the basis for short text classification. There are two main ways to extend the short text features, one is based on external resources and the other is based on the short text itself. Extension based on external resources is mainly through semantic dictionaries such as HowNet, Wikipedia and WordNet. However, there may be feature words in the short text that are not included in external resources, resulting in the inability to calculate words similarity, affecting the classification performance, and extension based on the short text itself doesn’t depend on external knowledge, but it has high requirements for feature extensions and requires the use of classical machine learning methods to build feature engineering for specific problems. Liu et al [2]used TF-IDF feature engineering for short texts feature extension, but TF-IDF doesn’t consider the effect of implied topics on feature words, while LDA probabilistic topic model represents each topic as a multinomial distribution of feature words, compensating for the shortcomings of TF-IDF. Shao et al [3]proposed to combine TF-IDF and LDA probabilistic topic model to extend features of short texts, and acquired good classification performance. Therefore, this research combines TF-IDF and LDA to build the core words set of the training set, and then extended features of the short texts. The basis of feature extension is the word similarity calculation method, and the commonly used word similarity calculation method is based on dictionaries or based on large-scale corpus statistics. Ya et al [4]used HowNet word similarity calculation method to extend features of short texts and improved the short text classification performance. However, there are problems such as poor universality and lack of contextual information in the dictionary-based word similarity calculation method, while the word similarity calculation method based on large-scale corpus statistics can obtain the probability distribution of word context information based on the corpus, avoiding the problems of the dictionary-based word similarity calculation method. Liu et al [5]used Word2Vec to calculate the word similarity for feature extension of short texts and got good classification performance. Therefore, this research uses Word2Vec similarity calculation to extend features of short texts, and compares it with the commonly used method of HowNet similarity calculation to extend the feature of short texts, to verify the effectiveness and superiority of the method used in this paper.

For extended short texts, how to build a classification model is a core step in short text classification. There are two commonly used classification models, namely classification models based on classical machine learning and classification models based on deep learning. Classification models based on classical machine learning include K-Nearest Neighbor, Naive Bayes, Support Vector Machine, etc. These classification models have simple structure but poor classification performance, especially for short text classification [6]. In recent years, with the development of deep learning technology, deep learning-based models have attracted more and more attention from researchers in the field of natural language processing, such as Convolutional Neural Networks(CNN), Recurrent Neural Networks(RNN), and Long Short Term Memory networks(LSTM) etc. However, it’s difficult to solve the problem of sparse features of short texts by directly applying deep learning models to short text classification. Current researchers generally carry out short text classification after improving the deep learning models. Liu at al [7]proposed a multi-channel CNN model with two channels and three cores to improve text accuracy, but due to the locality of convolutional and pooled layers, it requires many layer convolutions to capture long-term dependencies. Zhang eat al [8]proposed a short text classification method based on a semi-supervisory graph neural network, but when classifying new data, the model needs to recompose all the samples and enter them into the graph convolutional network for calculation, resulting in a lack of flexibility in the model. In a word, improving deep learning models is not only time-consuming and laborious, but also difficult to be universal. Some researchers have proposed an idea of using classical machine learning feature engineering to extend features of short texts firstly, and then using deep learning models to categorize short texts. This idea not only solves the problem of sparse features of short texts, but also directly uses existing and highly universal deep learning models, without spending much time and efforts to improve deep learning models. Shao [3]adopted this idea that feature engineering TF-IDF was used to complete short text extension, and then basic deep learning model CNN was used to implement short text classification after the extension. And experimental results show that the F1 value classification based on extension is 2 percent higher than without extension.

In summary, for the product reviews with a few and nonstandard expression words, this research firstly uses TF-IDF and LDA model considering the relationship between implied topics and feature words to construct the core words set of the training set, and then use Word2Vec similarity calculation method to extend features of short texts. After extension, BERT [9], a deep learning model that performs better than previous deep learning models such as CNN in many natural language processing tasks, is used to carry out short text classification.

It is difficult to directly obtain the reviews of specific products through retrieval or crawling from one or more e-commerce platforms and social platforms such as Weibo and WeChat. Merchants are easy to obtain reviews of every product of multiple types such as lipsticks, mobile phones and tablets etc. on their own store management system, but still need to manually collect the multiple types of the reviews into the corresponding product name, and it is difficult to directly obtain the reviews of various products on other platforms due to the barriers of information and limited technology. Directly using the product name as a keyword to crawl or retrieve can get part of the reviews of the corresponding product from other platforms, but there are three deficiencies in this method, which will lead to poor classification results for product reviews. (1) Some product reviews do not mention the product name, so relying on the product name to search may miss the reviews of the corresponding product category, for example, for the review "brighter, very attractive orange color, suitable for autumn and winter", its corresponding product category is lipstick, but its expression does not contain the product name; (2) The reviews crawled by the product name as a keyword may cause categories errors, for example, for the review "I bought a mobile phone on the XX platform before, this time the tablet I also watched on the XX platform for a long time before deciding to buy, very good", may mistakenly categorize this review into a product review of mobile phone or simultaneously product review of mobile phone and product review of tablet, but in fact its corresponding product category tablet; (3) Reviews of a certain type of a certain product under the e-commerce platform can be easily categorized into the corresponding product categories, but on social platforms such as Weibo, WeChat and so on, if you do not carefully trace and explore the background of the dialogue, it's hard to judge the reviews' product categories. For example, in a dialogue about the shopping experience of mobile phones and lipsticks etc. of A, B, C, C reviews that "A's product show temperament" and "The material of B's product is better", if you do not shop with them, it is difficult to know these two reviews' product categories based on the review statement alone. Therefore, categorizing the product reviews by different product categories can help the information management field to carry out efficient information organization for e-commerce information. Additionally, we don't categorize reviews of a specific product from a certain data source or multiple data sources into positive reviews or negative reviews which called sentiment analysis, but categorize the reviews of more than two kinds of products from different data sources into corresponding product categories, realizing the automatic categorization information organization of product reviews by various product categories, which is convenient for merchants to carry out data mining studies such as sentiment analysis of product reviews, competitive intelligence analysis of product reviews and so on.

3. APPROACH OF THIS RESEARCH

3.1 Construction of the core words set

The LDA model is a three-tier probabilistic model that represents text as documents, topics, and feature words, assuming that each document is represented as a random mixed distribution on an implicit set of topics, and each topic is represented as a multinomial distribution of feature words [10].

TF-IDF can be used to assess the importance of a feature word in a category, namely if a feature word appears frequently in a category and rarely appears in other categories, it is considered as a high-frequency word with good class differentiation. The TF-IDF calculation formula is shown in (1), $n_{i,j}$ represents the number of occurrences of a feature word in a document, $\sum_k n_{k,j}$ represents the sum of the number of occurrences of all words in a document, |D| represents the total number of documents in the corpus set and $|\{j: t_i \in d_j\}|$ represents the number of documents that contain a feature word.

$$TF-IDF = TF * IDF = \frac{n_{i,j}}{\sum_k n_{k,j}} * log \frac{|D|}{|\{j: t_i \in d_j\}|} \tag{1}$$

Feature grain refers to the degree to which feature words portray the content of the text [11]. In text representation, the implied topic is a coarse-grained feature that describes the multinomial distribution of a series of related feature words in the text, such as LDA model. The feature word belongs to the fine-grained feature, directly describing each word in the text, such as the Vector Space Model. We combined coarse-grained with fine-grained methods to acquire the core words set of the training set, namely obtaining the set of topics in the training set by LDA and obtaining the high-frequency word set by TF-IDF. And then merging the set of topics and the high-frequency word set. The specific algorithm steps are as follows:(1) Word segmentation, de stop words and words property filtering are carried out on the training set, and only nouns, verbs and adjectives that have a great impact on classification are retained. (2) Enter the training corpus processed in step (1) into the LDA model, obtain the topic word probability distribution under each topic, and select the top M1 feature words. (3) The training corpus processed in step (1) is counted according to the frequency of feature words by categories, the frequency of each feature word in each category is obtained, the feature words with a frequency less than K in all categories are filtered, and finally the feature words with a proportion greater than the threshold M2 in each category are taken as the high-frequency word set of the training set. (4) Combine the topic sets obtained in step (2) with the high-frequency sets in step (3) and drop duplicate feature words to construct the core words set of the training $W = (W_1, W_2, \dots, W_n,)$set, n represents the number of core words.

3.2 Similarity calculation method based on Word2Vec

Word2Vec [12]uses the context of the word to predict the current word or uses the current word to predict the context of the word, using the contextual information of the word window to convert a word into a low-dimensional real vector, the more similar words are closer in the vector space, so that the context semantic relationship of the word window can better represent the similarity between words. The vector representations u and v of the two words W1 and W2 are obtained through the Word2Vec pretrained language model on the Wikipedia corpus, and the similarity of W1 and W2 is calculated through the cosine similarity, as shown in Figure 1.

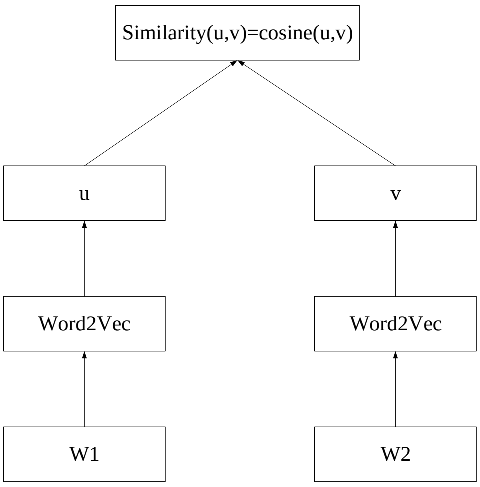


Figure 1. Similarity calculation method based on Word2Vec

3.3 Short text extension based on core words set and similarity calculation

The specific process of using Word2Vec or HowNet to calculate the similarity of words to extend short text features is as follows: (1) Perform word segmentation, deactivation, and words property filtering on the short text T, and retain only the nouns, verbs, and adjectives that have a greater impact on the classification, and obtain the feature words set F, $F = (F_1, F_2, \dots, F_m,)$, m represents the number of feature words after preprocessing. (2) Calculate the similarity between F_i ($F_i \epsilon F$) in the feature word set and W_j ($W_j \epsilon W$) in the core words set based on HowNet, and take the top K core words and the similarity greater than M as the extension words of the feature word F_i, and get the extended T_1 from the short text T by adding the extending words with brackets after the extended words. (3) Calculate the similarity between F_i ($F_i \epsilon F$) in the feature words set and W_j ($W_j \epsilon W$) in the core words set based on Word2Vec, and take the top K core words and the similarity greater than M as the extending words of the feature word F_i, and get the extended T_2 from the short text T by adding the extending words with brackets after the extended words.

For example, for the review "Classic color, very temperamental red", extended using the steps above, the result is shown in the Table 1.

Table 1. Examples of extending text based on HowNet and Word2Vec similarity calculation methods

Extension methods	Extension results
HowNet	Classic color, very (infinite, too, much) temperamental red (progressive, rose orange red, orange red).
Word2Vec	Classic (legendary) color, very temperamental (charming) red (black, white, blue).

3.4 Model

The model of short text classification in this paper is shown in Figure 2, and specific steps are as follows: (1) Preprocess training sets and test sets such as word segmentation, stop word removal, and words property filtering. (2) Use TF-IDF and LDA to handle the pre-processed training set, obtain the topic core words set and high-frequency words set of each category respectively, merge the two and remove the repetitive feature words to obtain the core words set of the training set. (3) Use the Word2Vec similarity calculation method to calculate the similarity of the feature words in the training set, the test set and the words in the core words set respectively, obtaining the extended training set and the extended test set; (4) Perform short text classification on the extended training set and the extended test set based on BERT.

Steps of using HowNet similarity calculation method to extend short texts is the same as the above process.

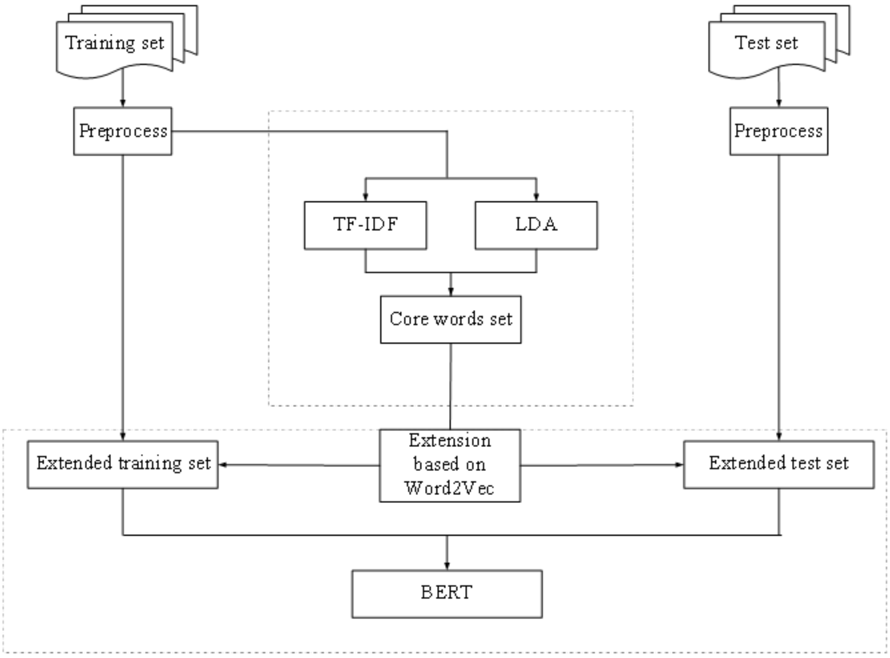


Figure 2. Short text classification framework based on BERT

4. EXPERIMENTS

4.1 Data sets

In order to verify the automatic classification performance of the method proposed in this paper categorizing product reviews by the product categories, this research selects totally 25,500 reviews involved 5 categories of products such as lipstick, mobile phone, pants, shampoo, tablet etc. from e-commerce and social platforms as a self-built corpus. The reviews for these products typically reflect the nonstandard of expression. For example, many reviews of these products involve in color but each has its own focus, reviews of lipstick, pants and shampoo associate color with seasons such as spring and autumn, reviews of lipstick, mobile phone and shampoo associate color with humanistic terms such as temperament, and reviews of mobile phones and tablets associate color with electronic terms such as appearance or screen resolution, etc. The specific performance of these terms in the review varies with the user's writing level, personal usage habits and feelings and regional differences, etc. Reviews that are similar and even completely repetitive in terms of expression are rare. Reviews with nonstandard expression words make the automatic categorization face great difficulties, and the effectiveness of the method proposed in this paper to overcome these difficulties and improve the classification performance can be better confirmed from the positive side by comparing the improvement of the classification performance after the comparison of not extended and other extension method. Contrary to the above self-built corpus, this research also selects the public corpus that expresses standardized as a comparison, and also through the basic unchanged classification performance after the comparison of the not extended and other extension method, verify the effectiveness of the automatic classification of product reviews with nonstandard expression words from the opposite. The selected public corpus is the titles of the news corpus with strict writing specifications, and the text in the corpus is further carefully filtered by experts from many original news, which can better reflect the expression specification of the text. The news is concise and concise, and the expression words are accurate and standardized, and the title is one of the five major elements of the news, which takes the actual content as the outline, specifically refines the essence of the news, and inherits the characteristics of the news, so the news titles have the characteristics of standardized expression words. This research selects 31500 titles from the Toutiao news corpus involved 15 categories such as science, technology, military, etc. and selects 29,400 titles from the Tsinghua news involved 14 categories such as finance, lottery, real state, etc.

In order to eliminate the influence of unbalanced data on experimental results, all experiments in this paper adopted balanced data, and there is no duplication between the training set and the test set, and the average of the multiple groups classification performance was taken as the experimental result. The classification performance is evaluated by F1 value.

4.2 Experimental environment and parameters setting

The experimental running environment is that the operating system is Ubuntu20.04.2, the memory is 16GB. The algorithm is implemented in Python and the deep learning framework is the version 1.8.4 of Torch. The experimental parameters in this paper are set as follows by the preparatory experiment. (1) The core word threshold M1 of the training set is 20, the category specific proportion threshold M2 in the training is 0.5, and the minimum word frequency K is 5. (2) Learning rate and drop rate of all experiments are 2e-5 and 0.1. The batch size, training epochs, and max sequence length of the product reviews without extension are respectively 70, 10 and 128 and these three parameters of the public corpus without extension are 150, 20, and 64 respectively. The batch size, training epochs, and max sequence length of the extended product reviews are 9, 10 and 512 respectively and these three parameters are 100, 20 and 100 respectively.

4.3 Experimental results and analysis

Experiment 1: In the self-built product reviews corpus, multiple groups of data are randomly selected, and in each group of data, the training set is 1500 per category, and the test set is 200 per category, comparing classification performance among the use of Word2Vec similarity calculation to extension based on the BERT model, the use of HowNet similarity calculations to extension based on the BERT model and the classification performance without extension. The results of the experiment are shown in Figure 3.

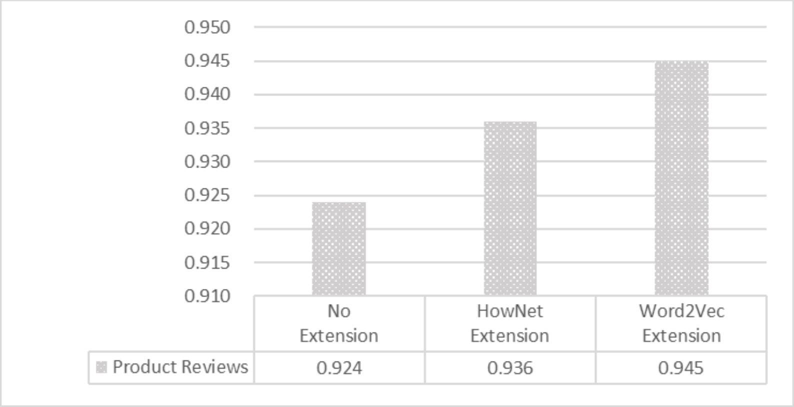
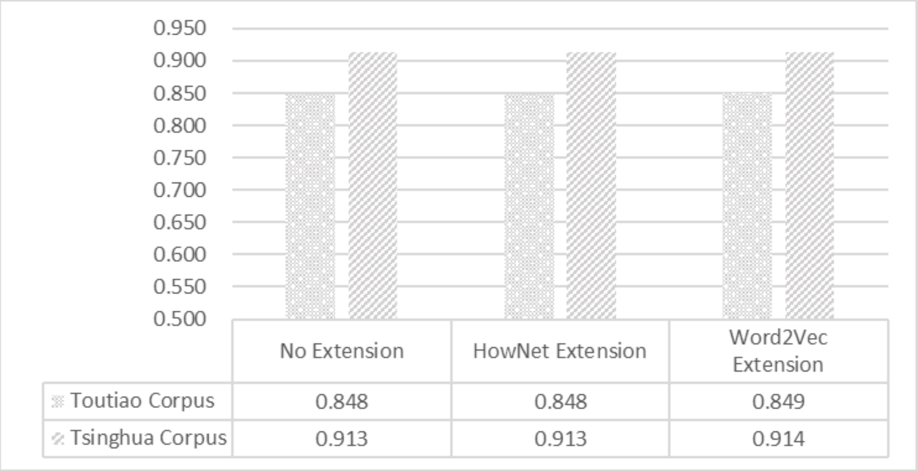


Figure 3. Classification performance among different extension methods for product reviews based on BERT model

As can be seen from Figure 3, for the product reviews, the classification performance when using Word2Vec extension based on BERT model is 2.1 percent higher than that when not extended, and the classification performance when using HowNet extension based on BERT model is 1.2 percent higher than that when not extended, and classification performance is 0.9 percent higher when using Word2Vec extension based on the BERT model than that when using HowNet extension. It demonstrates the effectiveness and superiority of categorizing product reviews by product categories using Word2Vec similarity calculation to extend based on the BERT model proposed in this paper from the front.

Experiment 2: In the public corpus of titles from the Toutiao news and the Tsinghua news, multiple groups of data are randomly selected, in each group of data, the training set is 500 per category, and the test set is 200 per category, comparing classification performance among the use of Word2Vec similarity calculation to extension based on the BERT model, the use of HowNet similarity calculations to extension based on the BERT model and the classification performance without extension. The experimental results are shown in Figure 4.



	No Extension	HowNet Extension	Word2Vec Extension
▩ Toutiao Corpus	0.848	0.848	0.849
▨ Tsinghua Corpus	0.913	0.913	0.914

Figure 4. Classification performance among different extension methods on news titles based on BERT model

As can be seen from Figure 4, for news titles from public corpus, the classification performance when using the HowNet extension based on BERT model is the same as that when not extended; the classification performance when using the Word2Vec extension based on BERT model is 0.1 percent better than that when it is not extended. It demonstrates the robustness of the news titles from the public corpora using the Word2Vec similarity calculation to extend based on the BERT model proposed in this paper, and proves the effectiveness of classifying product reviews by products categories in this paper from the opposite.

Overall, the method proposed in this paper and the method that extending product reviews using HowNet similarity calculations and then categorizing them by the product categories both improve the classification performance of self-built corpus of product reviews. However, the classification performance of both methods on news titles from the public corpus remained stable (up to 0.1% improvement). The two different word similarity calculation method, the self-built corpus and the public corpus, are different in the improvement of classification performance. The reasons for this can be analyzed from three aspects: the basic principle of classification performance improvement, the difference in extension words brought about by different word similarity calculation methods, and the way in which corpus text authors use words.

Firstly, this research proposes to use Word2Vec and HowNet to calculate the similarity of the extended word and the extending word, and add the extending words with brackets after the extended words to achieve short texts extension. The way has three advantages. The first is that the brackets act as an explanation, which is a further annotation on the extended word; the second is that it can make the extended word and the extending word in the close syntactic position, playing the same grammatical function, increasing the richness of the semantic features of the sentence while retaining the original syntactic dependency of the sentence, so that the constructed new sentence adapts to BERT pretraining knowledge; the third is that it can increase the co-occurrence of words in the training set and the test set, which constitutes a co-occurrence relationship when the extending word and the extended word appear frequently together [13]. When using BERT to classify short texts after extension, classification performance is no lower than short texts without extension. Secondly, there are deficiencies such as poor universality and lack of contextual information when using HowNet calculates words similarity, so there are some noisy words in the extended texts. Noise words are shown in two aspects, one is a false extension, such as extending "red" to a context-independent word "progressive"; the other is a valueless extension, such as extending "very" to "infinite", "too", "much", etc., although basically accurate in word meaning, these adverbs are heavily used in different categories of corpus, but have little value for category differentiation. The use of Word2Vec similarity calculation for short texts extension can obtain the probability distribution of word contextual information based on the corpus, which to a certain extent overcomes the shortcomings of HowNet similarity calculation for short texts extension, and reduces the text noisy words after extension, such as all of the extending words of "red" are words that represent color. Therefore, the classification performance of short texts extension using Word2Vec similarity calculation method is better than that using HowNet similarity calculation. Thirdly, for the product reviews, because the writer is various Internet users, the description or writing words are ever-changing, and there are differentiated word expressions such as loose or arbitrary word expressions in the training set and test set, for example, for the word "legendary", it is expressed as "legendary" in the training set, and in the test set as "classic", the extension method proposed in this paper can increase the co-occurrence of the words of the training set and the test set while extending the product reviews, so the classification performance has been greatly improved. For news titles of the public corpus, because the writer has a certain professional ability and a high level of writing and there are usually certain normative requirements for the news titles, so the words of the training set and the test set are usually more rigorously and the expression method is basically consistent with the co-occurrence of words, such as the extending word of "improve" is "enhance", the extending word of "enhance" is "improve". The extension of the short texts does not substantially help the improvement of classification performance, so the short texts extension method proposed in this paper has a small improvement and even continues to be consistent in the classification performance of news titles from the public corpus.

5. CONCLUSIONS

For product reviews, this research firstly obtains the core words set of the training set through feature engineering TF-IDF in classical machine learning methods combined with the LDA probabilistic topic model, and then uses Word2Vec similarity calculation method to obtain the extension words to extends the feature of the short text by adding the extending word after the extended word with brackets, after extension, implements text classification based on BERT model. The experimental results show that the method proposed in this paper can effectively improve the classification performance of product reviews categorized by the product categories, so that the product reviews with a wide range of sources and a large number can be automatically categorized and organized by the product categories, which is convenient for further data mining studies of the product reviews such as sentiment analysis and competitive intelligence analysis. The reason why the method proposed in this paper is effective is that it is more suitable for classifying short texts nonstandard expression words, and the method makes the extended short text more in line with the characteristics of the BERT classification. In future research, consideration can be given to designing appropriate calculation methods and indicators, measuring and explaining the characteristics of the diversity of words used, and studying the influence of the types of words and the distribution of word frequency on text classification.

REFERENCES

[1] Zhang Zixuan, Wand Hao, Zhu Liping, Deng Sanhong. (2019). Identifying risks of HS codes by China customs. Data Analysis and Knowledge Discovery, 3(01):72-84 (in Chinese)

[2] Liu Huiqing, Guo Yanbu, Li Hongling, Li Weihua. (2019). Short text feature extension method based on Bayesian networks. Computer Science, 46(S2):66-71 (in Chinese)

[3] Shao Yunfei, Liu Dongsu. (2019). Classifying short-texts with class feature extension. Data Analysis and Knowledge Discovery, 3(09):60-67 (in Chinese)

[4] Ya H N, Li Z, Ya R J, Wei J W, Shun Q L. (2014). Using semantic correlation of HowNet for short text classification. Applied Mechanics and Materials, 513-517:1931-1934.

[5] Liu W S, Cao Z W, Wang J, Wang X, Wang X Y. (2016). Short text classification based on Wikipedia and Word2vec. 2016 2nd IEEE International Conference on Computer and Communications. Chengdu, China: IEEE, 1195-1200.

[6] Guo Q. (2010). An effective algorithm for improving the performance of naïve bayes for text classification. 2010 Second International Conference on Computer Research and Development. Kuala Lumpur, Malaysia: IEEE, 1678-1684.

[7] Liu Y, Li P, Hu X. (2022). Combining context-relevant features with multi-stage attention network for short text classification. Computer Speech and Language, 71(1): 10.1016/j.csl.2021.101268.

[8] Zhang Binyan, Zhu Xiaofei, Xiao Zhaohui, Huang Xianying, Wu Jie. (2021). Short text classification based on semi-supervised graph neural network[J]. Journal of Shandong University (Natural Science), 56(05):57-65 (in Chinese)

[9] Devlin J, Chang M W, Lee K, Toutanova K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[10] Blei D M, Ng A, Jordan M I. (2003). Latent Dirichlet Allocation. The Journal of Machine Learning Research, 3:993-1022.

[11] Si Xiance. (2010). Research on Content-based Social Label Recommendation and Analysis. D Thesis. Beijing China: Tsinghua University, (in Chinese)

[12] Jin X L, Zhang S W, Liu J. (2019). Word semantic similarity calculation based on Word2Vec. 2018 International Conference on Control, Automation and Information Sciences (ICCAIS), Hangzhou, China: IEEE, 12-16.

[13] Li Wei, Jia Caiyan. (2018). Micro-blog topic detection in frequent word networks. Journal of Data Acquisition and Processing, 33(01):188-194 (in Chinese)