A classification method for social information of sellers on social network

Check for
updates

Haoliang Cui[1], Shuai Shao[2]* , Shaozhang Niu[1], Chengjie Shi[3] and Lingyu Zhou[1]

* Correspondence: shaoshuaib@163.
com
[2]China Information Technology
Security Evaluation Center, Beijing
100085, China
Full list of author information is
available at the end of the article

Abstract

Social e-commerce has been a hot topic in recent years, with the number of users increasing year by year and the transaction money exploding. Unlike traditional e-commerce, the main activities of social e-commerce are on social network apps. To classify sellers by the merchandise, this article designs and implements a social network seller classification scheme. We develop an app, which runs on the mobile phones of the sellers and provides the operating environment and automated assistance capabilities of social network applications. The app can collect social information published by the sellers during the assistance process, uploads to the server to perform model training on the data. We collect 38,970 sellers' information, extract the text information in the picture with the help of OCR, and establish a deep learning model based on BERT to classify the merchandise of sellers. In the final experiment, we achieve an accuracy of more than 90%, which shows that the model can accurately classify sellers on a social network.

Keywords: User model, Machine learning, Social e-commerce

1 Introduction

With the continuous improvement of social network and mobile payment technology, one kind of commodity trading based on social relations called social e-commerce is in rapid development. According to the 2019 China social e-commerce industry development report released by the Internet society of China, the number of employees of social e-commerce in China is expected to reach 48.01 million in 2019, up by 58.3 percent year on year, and the market size is expected to reach 2060.58 billion yuan, up by 63.2% year on year. Social e-commerce has become a large scale, and the high growth cannot be ignored. Different from e-commerce platforms such as Taobao, social e-commerce is at the end of online retail. It carries out trading activities through social software and uses social interaction, user generated content and other means to assist the purchase and sale of goods. At the same time, sellers on social network use different social software without uniform registration, have no systematic classification of products for sale, and there are no standardized terms for product description. These bring great difficulty to the accurate classification of user portrait. This paper proposes a method based on the NLP classification model, which can realize accurate

business classification of social e-commerce based on social information of social e-commerce. This method analyzes 38,970 sellers on social networks and establishes a deep learning model based on BERT to accurately classify the merchandise of sellers. In addition, we introduced the OCR algorithm to extract the text information in the picture and superimposed it on the social content data, which effectively improved the classification accuracy. The final experiment shows that the measured accuracy is more than 90%.

2 Related work

2.1 Natural language processing

In order to analyze e-commerce business classification based on social data of sellers on a social network, the text needs to be analyzed based on the NLP correlation algorithm. The rapid development of NLP at the present stage is due to the neural network language model (NNLM) Bengio et al. [1] proposed in 2003. Researchers have been trying to realize the end-to-end classification recognition by using a neural network as a classifier in the text classification research based on word embedding. Kim first introduces the convolutional neural network (CNN) into the study of text classification. The network structure is a dropout full connection layer and a softmax layer connected after one convolution layer [2]. Although this algorithm achieves good results in various benchmark tests, it cannot obtain long-distance text dependency due to the limitation of network structure. Therefore, Tencent AI Lab proposed DPCNN, which further enhanced the extraction capacity of long-distance text dependency by deepening CNN [3].

Social content data includes multimedia text data and picture data. With the help of OCR, we extract the text in the picture and convert the picture data into text data. Text is a kind of sequential data, and the classification of it by recurrent neural network (RNN) has been the focus of long-term research in academia [4]. As a variation of RNN, long short-term memory (LSTM) adds control units such as forgetting gate, input gate, and output gate on the original basis, which solves the problem of gradient explosion and gradient disappearance in the long sequence training of RNN and promotes the use of RNN [5]. By introducing the sharing information mechanism, Liu et al. further improved the accuracy of the RNN algorithm in the text multi-classification task and achieved good results in four benchmark text classifications [6].

However, Word vectors cannot be constructed in Word embedding to solve the problem of polysemy. Even though different semantic environments are considered during training, the result of training is still one word corresponding to one row vector. Considering the widespread phenomenon of polysemy, Peters et al. propose embeddings from language model (ELMO) to address the impact of polysemy on natural language modeling [7]. ELMO uses a feature-based form of pre-training. First, two-way LSTM is used to pre-train the corpus, and then word embedding resulting from training is adjusted by double-layer two-way LSTM when processing downstream tasks to add more grammatical and semantic information according to the context words.

The ability of ELMO to extract features is limited for choosing LSTM as the feature extractor instead of Transformer [8], and ELMO's bidirectional splicing method is also weak in feature fusion. Therefore, Devlin et al. propose the BERT model, taking Transformer as a feature extractor to pre-train large-scale text corpus [9].
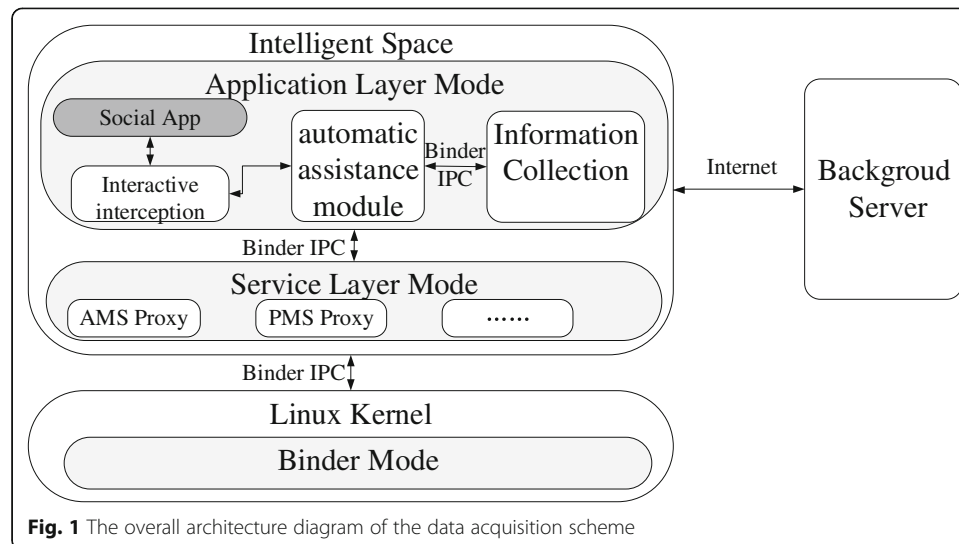
2.2 User analysis of social networks

User analysis is an important part of social network analysis. Most existing studies use user-generated content or social links between users to simulate users. Wu et al. modeled users on the content curation social network (CCSN) in the unified framework by mining user-generated content and social links [10]. They proposed a potential Bayesian model, multilevel LDA (MLLDA), that could represent users of potential interest found in social links formed by text descriptions contributed by users and information sharing. In 2017, Wu et al. proposed a latent model [11], trying to explain how the social network structure and users' historical preferences change over time affect each user's future behavior and predict each user's consumption preferences and social connections in the near future. Malli et al. proposed a new online social network user profile rating model [12], which solved the problem of large and complicated user data. In terms of data analysis platform, Chen et al. [13] developed a big data platform for the study of the garlic industry chain. Garlic planting management, price control, and prediction were realized through data collection, storage, and pretreatment. Ning et al. [14] designed a ga-bp hybrid algorithm based on the fuzzy theory and constructed an air quality evaluation model by combining the knowledge of BP neural network, genetic algorithm, and fuzzy theory. Yin et al. [15] studied two methods of extracting supervisory relations and applied them to the field of English news. One is the combination of support vector machine and principal component analysis, and the other is the combination of support vector machine and CNN, which can extract high-quality feature vectors from sentences of support vector machine. In the social apps, the data we obtain is mostly image data, so we introduced the OCR technology to identify text information in images.

3 Data collection

In order to analyze the behavior patterns of social e-commerce, we developed an auxiliary tool for social e-commerce. In this tool, sellers on a social network are provided with the independent running environment of social software and the automatic auxiliary ability, and the information acquisition module of the auxiliary process is used to collect the social information published by sellers on a social network, which is uploaded to the background server for model training. We provided this tool to nearly 10,000 sellers on a social network who participated in the experiment to obtain their social information in their e-commerce activities.

3.1 Overall structure

The whole data collection scheme is mainly composed of two parts: intelligent space app and background server. The overall architecture is shown in Fig. 1. Intelligent space app is deployed in the mobile phones of sellers on a social network and implemented based on the application layer of the Android platform, providing sellers on a social network with a secure container for the independent operation of social software. The app contains the automatic assistant module, which provides the automatic assistant capability of various business processes for seller, and collects the social information in the auxiliary process through the information grasping module. The collected information is cached and uploaded locally through the information collection service.

Fig. 1 The overall architecture diagram of the data acquisition scheme

The background server is responsible for receiving the collected data uploaded by the intelligent space, preprocessing the data first, and then classifying the social e-commerce through the data based on the machine learning classification model, and finally storing the classification results.

3.1.1 Security container
The security container is designed to allow social software to run independently without modifying the OS or gaining root privileges. The basic principle of its realization is to create an independent container process; load APK file of social software dynamically; monitor and intercept process communication interface such as Binder IPC through Libc hook, Java reflection, dynamic proxy, and other technical means; and collect social information through an automatic assistant module. The main part of the container is composed of an application layer module and a service layer module.

The application layer module is responsible for the process startup and execution of social software, and its main functions include three parts.

3.1.1.1 *Interactive interception* The application layer module intercepts the interaction between the application process and the underlying system in the container and modifies the calling logic. By hook or dynamic proxy of system library API and Binder communication interface, the application layer module blocks all interfaces that interact with the system during the execution of social software and controls the process boundary of interaction between social applications and system services.

3.1.1.2 *Social information collection* The loading of the automatic auxiliary module by social software is realized when initializing the process of social application. The application layer module injects the corresponding plugins in the automatic assistant module into the social application process. The automatic assistance module provides a number of e-commerce auxiliary functions for sellers on a social network, including customer acquisition, social customer relationship management

(SCRM), group management, sales assistance, and daily affairs. Sellers on social networks publish social information with commercial attributes through auxiliary functions, then the automatic auxiliary module will automatically collect the social information and send it to the information collection service for processing.

3.1.1.3 *Local processing of social information* When the information collection service receives the social information collected by the automatic auxiliary module, the data will be compressed and encrypted in the local cache. The service then uploads the collected data to the background server periodically through the timer, and HTTPS is used to ensure data transmission security.
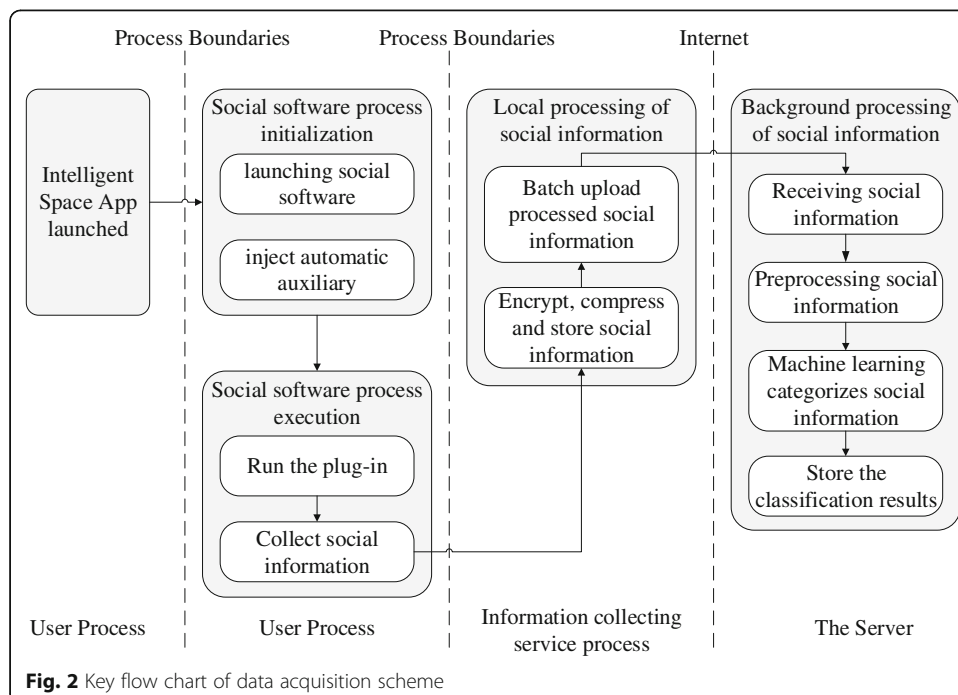
The main function of the service layer module is to take over the call logic modified by the application layer module by simulating the system service modify the parameters in the communication process and finally call the real system service. The service layer module exists in the container as an independent process. It focuses on the simulation of activity manager service (AMS) and package manager service (PMS) and realizes the support of system services in the process of launching and running social software.

3.1.2 Background server

The background server mainly realizes the machine learning model processing of the collected social data, including the functions of data preprocessing, data training, classification, and result storage. The core processing logic will be described in chapter 5.

3.2 Key processes

There are four key processes in the process of social information collection and processing. They are social software process initialization, social software process

Fig. 2 Key flow chart of data acquisition scheme

Cui *et al. EURASIP Journal on Image and Video Processing*   (2021) 2021:4

Page 6 of 12

execution, local processing of social information, and background processing of social information. The complete process is shown in Fig. 2.

3.2.1 Social software process initialization
When launching social software, the intelligent space will first intercept the callback function of the life cycle of all its components, then realize the process loading of the automatic auxiliary module during the process initialization.

3.2.2 Social software process execution
The process execution is completed by the application layer module and service layer module together. Sellers on a social network use automatic auxiliary modules to complete business activities, trigger information capture module to collect social information, and send it to the information collection service for subsequent processing.

3.2.3 Local processing of social information
The local processing of social information is mainly completed by the information collection service. In order to ensure the safe storage and transmission of the collected social information, the information collection service first adopts the encryption and compression method to realize the local security cache and then adopts the HTTPS secure communication and transmission protocol to upload the data.

3.2.4 Background processing of social information
The background processing of social information is completed by the background server. The server first receives the social information uploaded by the intelligent space, next decrypts and decompresses the social information, cleans the plaintext data, uses third-party OCR technology to identify text information in images, and adds it to the user's social information after simple data processing. Then, the classification of sellers on a social network is realized through the data based on machine learning modeling. Finally, the classification results are stored in the target database.

4 Methods
To classify the business attributes of social e-commerce based on the information of sellers on a social network, traditional feature matching scheme and classification clustering scheme based on machine learning can be used to establish the model. In this chapter, we introduce the scheme based on term frequency-inverse document frequency (TF-IDF) clustering and the classification scheme based on BERT.

4.1 Feature classification and TF-IDF clustering
4.1.1 Feature classification
We randomly select 5000 sellers on a social network from the data collected by the background server and extracted the text data of their social information for analysis. Each social e-commerce user contains an average of 50 social text data. Based on the content, we manually classify social e-commerce into 11 categories. With the help of e-commerce platforms like JD.COM, 50–100 keywords are sorted out for each category, and these keywords are screened and expanded according to the language habits of
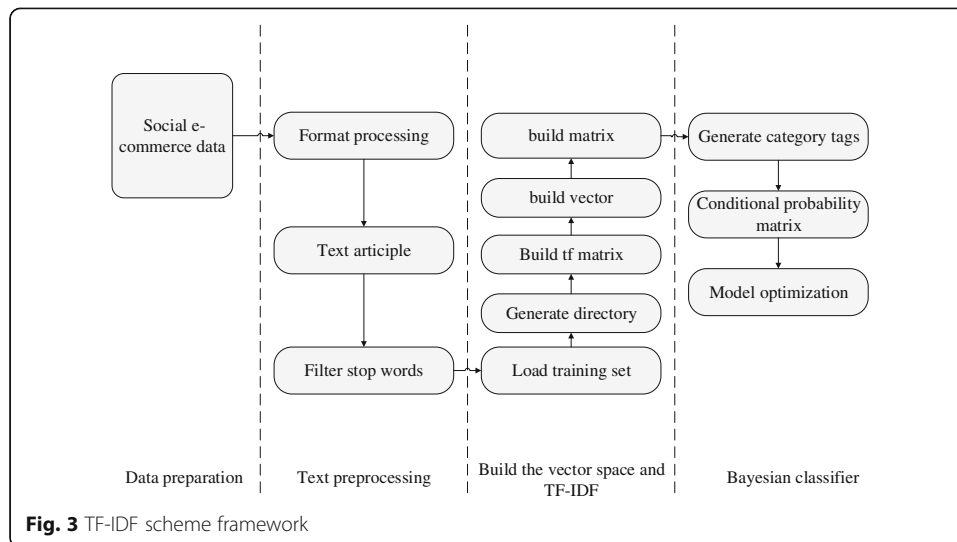
sellers on a social network. On this basis, we collect all the social information of each social network seller, cut and remove word segmentation, and match the results with the keywords of the selected 11 categories. The number of keywords that are matched is counted as the matching degree. According to the situation of different classification, the threshold of matching degree is determined by manual screening of some results, and then all social e-commerce is classified according to the threshold. After optimization and verification, the accuracy of the classical feature matching scheme finally reached 40%. However, due to the simplicity of the rules of the feature matching scheme, the small optimization space, the high misjudgment rate of the scheme, and the large human intervention in the basic word segmentation process, it is difficult to cover various situations of social e-commerce due to the limitation of these basic keywords, thus making it insensitive to the dynamic changes of new hot words of social e-commerce.

4.1.2 TF-IDF clustering

To achieve the goal of accurate classification of social e-commerce, we designed a scheme based on TF-IDF clustering. Term frequency-inverse document frequency (TF-IDF) is a commonly used weighted technique for information retrieval and text mining to evaluate the importance of a single word to a document in a set of documents or a corpus. In this scheme, the social information of each social e-commerce user is mapped as one file set of TF-IDF, and all texts of all sellers on a social network are mapped as the whole corpus. The words with the highest frequency used by each social e-commerce user are the most representative words in this document and become keywords. Category labels can be generated to calculate the probability that a document belongs to a certain category using the naive Bayes algorithm formula. The advantages of TF-IDF clustering to achieve the classification of sellers on the social network include the following: (1) clear mapping; (2) emphasize the weight of keywords and lower the weight of non-keywords; (3) compared with other machine learning algorithms, the characteristic dimension of the model is greatly reduced to avoid the dimension disaster; and (4) while improving the efficiency of classification calculation, ensure that the classification effect has a good accuracy and recall rate. The architecture of the entire solution is shown in Fig. 3.

In the text preprocessing stage, the first thing to do is to format the social information, mainly including deleting the space, deleting the newline character, merging the social e-commerce text, and so on, and finally getting the text to be processed for word segmentation. In this scheme, we choose Jieba's simplified mode for word segmentation, then filter out the noise by filtering the stop words (e.g., yes, ah, etc.).

In the stage of establishing the vector space model, the first step is to load the training set and take the pre-processed social information of each social e-commerce user as a document. The next step is to generate a dictionary, by adding every word that appears in the training set to it, using the complete dictionary to calculate the TF-IDF value of each document. In this scheme, CountVectorizer and TfidfTransformer in Python's Scikit-Learn library are used. CountVectorizer is used to convert words in the text into word frequency matrix, TfidfTransformer is used to count the TF-IDF value of each word in each document, and the top20 words in each document are taken as

Fig. 3 TF-IDF scheme framework

keywords of sellers on a social network. After this step, the keywords with a large TF-IDF value in each document are the most representative words in the document, which become the keyword set of the social e-commerce user. Finally, the naive Bayes method is used to generate the category label, and the document vectors belonging to the same category in the TF-IDF matrix are added to form a matrix of $m*n$, where m represents the number of categories and n represents the number of documents. The weight of each word is divided by the total weight of all words of the class, to calculate the probability that a document belongs to a certain class.

In the model optimization stage, we optimize the whole scheme model by adjusting the stop word set, adjusting parameters (including CountVectorizer, TfidfTransformer class construction parameters), and adjusting the category label generation method.

The main idea of TFIDF is if a word or phrase appears in an article with a high frequency of TF, and rarely appears in other articles, it is considered that the word or phrase has a good classification ability and is suitable for classification. TFIDF is actually: TF * IDF, TF is term frequency and IDF is inverse document frequency.

In a given document, word frequency refers to the frequency of a given word in the document. This number is a normalization of the number of words to prevent it from being biased towards long documents. For the word t_i in a particular document, its importance can be expressed as:

$$tf_{i,j} = \frac{|D|}{|\{j : t_i \in d_j\}|}$$

among them:

|D|: The total number of files in the corpus

$|\{j : t_i \in d_j\}|$: The number of documents containing the term t_i (i.e., the number of documents in $n_{i,j} \neq 0$). If the term is not in the corpus, it will cause the dividend to be zero, so it is generally used $1 + |\{j : t_i \in d_j\}|$.

and then:

$$tfidf_{i,j} = tf_{i,j} \times idf_i$$

A high word frequency in a particular document and a low document frequency of the word in the entire document collection can produce a high-weight TF-IDF. Therefore, TF-IDF tends to filter out common words and keep important words.

4.2 Classification scheme based on BERT

4.2.1 Data label

We manually classify and mark the data of sellers on a social network according to the characteristics of the products. Classified labels include 38,970 items and 17 categories of data, including 3c, dress, food, car, house, beauty, makeup, training, jewelry, promotion, medicine and health, phone charge recharge, finance, card category, cigarettes, essays, and others. The pre-processing phase removes emojis, numbers, and spaces from the text through Unicode encoding.

4.2.2 Classification scheme

In the BERT model, Transformer, as an encoder-decoder model based on attention mechanism, solves the problem that RNN cannot deal with long-distance dependence and the model cannot be parallel, improving the performance of the model without reducing the accuracy. At the same time, BERT introduced the shading language model (MLM, masked language model) and context prediction method, further enhance the two-way training of the ability of feature extraction and text. MLM uses Transformer encoders and bilateral contexts to predict random masked tokens to pre-train two-way transformers. This makes BERT different from the GPT model, which can only conduct one-way training and can better extract context information through feature fusion. Anaphase prediction is more embodied in QA and NLI. Therefore, we choose the BERT model based on the bidirectional coding technology of pre-training and attention mechanism to classify sellers on a social network.

We chose the official Chinese pre-training model of Google as the pre-training model of the experiment: BERT-Base which is Chinese simplified and traditional, 12-layer, 768-hidden, 12-head, 110M parameters [16]. This pre-training model is obtained by Google's unsupervised pre-training on a large-scale Chinese corpus. On this basis, we will carry out fine-tuning to realize the classification model of sellers on a social network. When dividing the data set, we divided 38,970 pieces of data into training set and verification set according to the ratio of 6:4, that is, 23,382 pieces of training set and 15,588 pieces of verification set.

5 Results and discussion

5.1 TF-IDF clustering scheme

The computer used in the experiment is configured with AMD Ryzen R5-4600H CPU, 16G memory, and windows10 64bit operating system. First, the default construction parameters are used, and the average accuracy of each classification is 45.7%. Next, the parameters are adjusted through a genetic algorithm, and 100 rounds of genetic algorithm optimization are performed, then the average accuracy reached the highest value of 52.5%. In the process of genetic algorithm, statistical estimation of algorithm time is

Fig. 4 Text message token serialization

also carried out. On average, on this data set, the running time of each round of the TF-IDF model is about 28 s.

Experiments show that the accuracy of the TF-IDF clustering scheme has been improved after optimization, and it has a certain reference value for the classification of sellers on a social network, but there is still a big gap from the accurate classification. We found three reasons after analyzing the experimental results. (1) Compared to the feature matching scheme, the TF-IDF-based model is improved to some extent. However, the input of the model is still the result of direct word segmentation, and more information is lost in the word segmentation process, such as the semantic information of previous and later texts and the repetition frequency of corpus, which are relatively important in the process of natural language processing. (2) The classification problem of sellers on a social network is complicated. This model does not analyze the correlation between words and is essentially an upgraded version of word frequency statistics, which makes it difficult to improve the accuracy after reaching a certain value. (3) For the optimization of the model, only the parameters of the intermediate function are adjusted, and the method is not upgraded. Therefore, the machine learning scheme based on TF-IDF clustering cannot solve the problem of accurate classification of sellers on a social network. In the next chapter, we will introduce a scheme based on deep learning to achieve the goal of classifying sellers on a social network.

5.2 Classification scheme based on BERT

Text classification fine-tuning is to serialize the preprocessed text information token and input BERT, and select the final hidden state of the first token [CLS] as a sentence vector to output to the full connection layer, and then output the probability of obtaining various labels corresponding to the text through the softmax layer. The experimental schematic diagram is shown in Figs. 4 and 5. The maximum length of the sequence (ma_seq_length) is set to 256 according to the actual text length of the social information data set of the sellers on a social network and
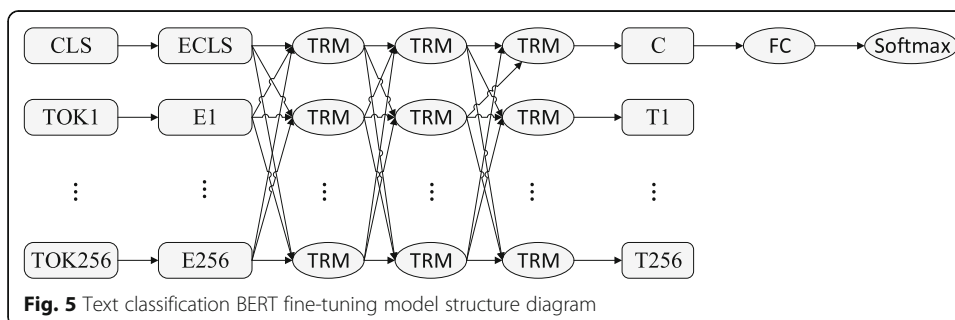

Fig. 5 Text classification BERT fine-tuning model structure diagram

Table 1 Corresponding table of epoch and accuracy

Epoch	eval_accuracy (%)
3	95.84
6	96.05
9	96.2

The training results are shown in Table 2, and the recognition rate is 96.2%

the batch_size and learning rate adopt the official recommended values of 32 and 2e–5. In addition, we also adjust the super parameter num_train_epochs and increase the number of training epochs (num_train_epochs) from 3 to 9 to improve the recognition rate of the model (Table 1). The results are shown in Table 2.

We select an additional 9500 text data of sellers on social networks and test the model after the same preprocessing. The accuracy rate is 90.5%, which is lower than that of the verification set (96.2%). The reason may be that the data of the test set contains a large number of commodity terms not included in the corpus and training set, and the text description of these commodities is too colloquial. Sellers on a social network often use colloquial words in the industry to replace the standard product names when releasing product information, such as "Bobo" instead of "Botox," which to some extent limits the accuracy of text-based classification in the social e-commerce market scene.

6 Conclusion

The classification model proposes in this paper achieves an accuracy of 90.5% in the test data. However, there are still some problems such as non-standard description text. A corpus with a high correlation with a social e-commerce environment will be established in order to further improve the accuracy of social e-commerce classification. At the same time, we will use the knowledge distillation technology to compress and refine the existing model, so as to improve the model recognition rate while simplifying the model and improving the operational performance [16]. In addition, in view of the high labor cost and time cost of large-scale data marking, the next step will be trying to make full use of semi-supervised learning to train unlabeled data and labeled completed data [17]. The full use of large-scale unlabeled data is conducive to further improving the accuracy and generalization ability of the model, as well as the analysis and processing of emerging products, providing strong data support for the model landing. Since the image data have also been studied to profiling the users in a social network [18] and perceptual image hashing schemes are proposed [19], we will improve our model so that the image and text data are combined for analysis.

Table 2 Text information classification results of sellers on social network

Results	Value
eval_accuracy	96.2%
eval_loss	0.25033528
global_step	6024
Loss	0.25023073

Author details
[1]Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China. [2]China Information Technology Security Evaluation Center, Beijing 100085, China. [3]Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100088, China.

References
1. Y. Bengio, R. Ducharme, P. Vincent, et al., A neural probabilistic language model. J. Mach. Learn. Res. **3**, 1137–1155 (2003)
2. Kim Y. Convolutional neural networks for sentence classification arXiv preprint arXiv:1408.5882, 2014.
3. R. Johnson, T. Zhang, *Deep pyramid convolutional neural networks for text categorization [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2017), pp. 562–570
4. Otter D W, Medina J R, Kalita J K. A survey of the usages of deep learning in natural language processing arXiv preprint arXiv:1807.10854, 2018.
5. R. Jozefowicz, W. Zaremba, I. Sutskever, *An empirical exploration of recurrent network architectures [C]//International conference on machine learning* (2015), pp. 2342–2350
6. Liu P, Qiu X, Huang X. Recurrent neural network for text classification with multi-task learning arXiv preprint arXiv:1605.05101, 2016.
7. Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations. arXiv preprint arXiv:1802.05365, 2018.
8. A. Vaswani, N. Shazeer, N. Parmar, et al., *Attention is all you need [C]//Advances in neural information processing systems* (2017), pp. 5998–6008
9. Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding arXiv preprint arXiv:1810.04805, 2018.
10. L. Wu et al., MLLDA: multi-level LDA for modelling users on content curation social networks. Neurocomputing **236**, 73–81 (2017)
11. L. Wu et al., Modeling the evolution of users' preferences and social links in social networking services. IEEE Transact. Knowledge. Data. Eng. **29.6**, 1240–1253 (2017)
12. M. Malli, N. Said, A. Fadlallah, A new model for rating users' profiles in online social networks. Comput. Information. Sci. **10.2**, 39–51 (2017)
13. W. Chen et al., Development and application of big data platform for garlic industry chain. Comput. Mater. Continua **58.1**, 229 (2019)
14. M. Ning et al., GA-BP air quality evaluation method based on fuzzy theory. Comput. Mater. Continua **58.1**, 215–227 (2019)
15. Yin, Libo, et al. Relation extraction for massive news texts. Tech Science Press, CMC,60, no.1(2019), pp.275-285.
16. Sun S, Cheng Y, Gan Z, et al. Patient knowledge distillation for BERT model compression arXiv preprint arXiv:1908.09355, 2019.
17. Yalniz I Z, Jégou H, Chen K, et al. Billion-scale semi-supervised learning for image classification. arXiv preprint arXiv:1905.00546, 2019.
18. Yaqiong Qiao, Xiangyang Luo, Chenliang Li, et al. Heterogeneous graph-based joint representation learning for users and POIs in location-based social network, Inf. Process. Manag., 2020, 57, 102151-1~102151-17
19. Jinwei Wang, Hao Wang, Jian Li, Xiangyang Luo, Yun-Qing Shi, Sunil Kr. Jha, Detecting double JPEG compressed color images with the same quantization matrix in spherical coordinates, IEEE Trans. on CSVT, doi: 10.1109/TCSVT.2019.2922309.

Publisher's Note