ISSN 1846-6168 (Print), ISSN 1848-5588 (Online)
https://doi.org/10.31803/tg-20221228180808

Preliminary communication
Received: 2022-12-28, Accepted: 2022-12-30

Text Classification of Mixed Model Based on Deep Learning

Sang-Hwa Lee

Abstract: At present, deep learning has been widely used many fields, but the research on text classification is still relatively few. This paper makes full use of the good learning characteristics of deep learning, proposes a hybrid model based on deep learning, and designs a text classifier based on the hybrid model. This hybrid model uses two common deep learning models, sparse automatic encoder and deep confidence network, to mix. The hybrid model is mainly composed of three parts, the first two layers are constructed by sparse automatic encoder, the middle layer is a three-layer depth Convolutional Neural Network (CNN), and finally Softmax regression is used as the classification layer. In order to test the classification performance of the classifier based on deep learning hybrid model, relevant experiments were conducted on English data set 20Newsgroup and Chinese data set Fudan University Chinese Corpus. In the English text classification experiment, the classifier based on deep learning hybrid model is used to classify, and a high classification accuracy rate is obtained. In order to further verify the superiority of its performance, a comparative experiment with naive Bayes classifier, K-Nearest Neighbor (KNN) classifier and Support Vector Machine (SVM) classifier demonstrates that the classification effect of the classifier based on deep learning hybrid model is better than that of naive Bayes classifier, KNN classifier and support vector machine classifier. In the experiment of Chinese text classification, the Chinese corpus of Fudan University is tested, and a good classification effect is obtained. The influence of different parameter settings on the classification accuracy is discussed.

Keywords: classification; deep confidence network; deep learning; sparse automatic encoder; softmax

1 INTRODUCTION

Since the widespread application of Internet technology, people are faced with the severe problem of information explosion. The information on the Internet keeps increasing, and its growth momentum is rapid, with geometric magnitude [1-3]. The Internet can carry amazing information, and the world is submerged in information. The Internet has now become a key tool for most people to search for or acquire information, making it an essential tool for people's daily lives and work. The Internet has provided quite a lot of information, among which how to find valuable information accurately and quickly has become an important issue. At present, text information contains a lot of valuable information [4-6]. Classifying published text is one of the important ways to analyze data, and improving the efficiency and quality of text usage makes it possible to organize or manage text effectively. Text classification refers to the analysis of the content of the text, and it is determined that the text belongs to any of the given categories [7][8]. In the early days, people relied on manual classification of texts. This traditional method was time-consuming and laborious, unable to deal with massive text information, and it was difficult to unify the standard because of the unstable classification results caused by human factors. At present, the main methods of text classification are statistics and machine learning, which has made a lot of progress and entered a stage of rapid development [9][10]. So far, text classification is still a hot research topic of many researchers. The main idea is to apply the text classification algorithm, learn the known samples, classify the unknown texts through the learned rules, and finally get the text categories. Text classification can handle a large number of texts, reduce the consumption of manpower and material resources, and enable users to obtain valuable content quickly and efficiently. It provides convenience for the follow-up research work and makes the text information processing rise to a new height. With the continuous in-depth exploration of the field of text classification, text classification has been well promoted in search engines, digital libraries, and email filtering and other fields.

(a) Search engine

Search engine is an indispensable tool in people's life, which can get information from the Internet. The Internet is composed of a huge number of web pages, and it is difficult for people to get the information they want. The function of search engine is to quickly classify the information on the Internet and screen the relevant information from the categories. It involves the classification of texts, which classifies texts according to their contents and then manages them separately. When a user wants to inquire about information, the search engine can provide retrieval service to the user, retrieve the relevant information that the user wants from the relevant classified information, and provide the information in the form of a page.

(b) Digital library

Because information technology has developed steadily and rapidly, digital library has become the development direction of most libraries. The technique of classifying acquired text has also become one of the important techniques for retrieving information. When the library classifies books, it adopts the text classification technology, which can effectively manage books and reduce the tedious work of librarians. The digitalization of the library is convenient for readers, and enables readers to get all kinds of library information in different places.

(c) Mail filtering

With the development of the Internet, e-mail provided great convenience to communicate with other people. And the existence of spam also adds trouble to life. Text classification technology can filter out junk information, so that users can avoid the interference of this information. Text classification technology is trained according to the characteristics of spam, and a spam classifier is obtained. E-

mail classifier can filter out the junk information and keep only the information that users need, so that users' daily life is free from interference. It can be seen from the above functions of text classification in various fields that the research of text classification has important theoretical and practical significance. Text classification, as a basic task, can provide an effective guarantee for deep mining of valuable information in the text.

2 RELATED WORK

2.1 Research Status of Text Classification

Text classification was first proposed in the 1960s. Manual classification is the earliest method of text classification, which was manually classified by professional researchers. This classification method would waste much manpower and resources and was limited by the number of professional researchers. Specific classification problems must be formulated and implemented by specific researchers. By the 1890s, the number of texts was exploding, and the proposal and development of machine learning attracted many researchers. Firstly, the text classifier trains a large number of data sets to establish a certain mathematical model, and then automatically classifies other new sample data. In the middle of 20th century, the research on text classification has been carried out abroad. In 1957, Luhn put forward the idea of applying word frequency statistics to text classification, which laid the foundation for text classification. Then, Maron et al. put forward probability model and factorization model algorithm successively, which made the text classification technology develop. In 1970, Salton et al. put forward a vector space model which can represent text well. During this period, the text classification technology mainly uses the method of knowledge engineering, and the method of knowledge engineering depends on the rules formulated by experts. However, the formulation of relevant rules will take a lot of time and energy, which makes this method unable to be popularized. In the 1990s, with the rapid development of the Internet, there was an urgent need to classify more and more different kinds of texts. At this time, machine learning methods emerged and were quickly applied to text classification. Text classification based on machine learning doesn't need manual operation to construct a classifier. It finds different features among texts by learning samples, summarizes these features, and automatically generates a text classifier according to relevant rules. Text classification using machine learning is superior to knowledge engineering in accuracy and efficiency, and it gradually replaces the method based on knowledge engineering and becomes the mainstream. Through a certain experimental analysis, it was found that the new text classifier was comparable to professional researchers in classification accuracy, so it became a common way of text classification technology at that time. In 1971, Rocchio proposed a new linear classifier [11]. In 1979, van Rijsbergen put forward some new concepts in the field of information retrieval and applied them to text classification technology, such as evaluation criteria such as accuracy and recall. In 1995, Vipnik proposed the method of Support Vector Machine. Thorsten Joachims applied linear kernel support vector machine to text classification technology for the first time, so up to now, the theory and application of support vector machine still have great influence on text classification technology. After 1995, Yoav Freund and Robert E. Schapire published a paper on Ada Boost. Robert E. Schapire proposed an Ada Boost algorithm framework and carried out relevant experimental verification. Later, some scholars designed many similar algorithms according to this framework, and these algorithms have made great achievements in the research of text classification [12, 26]. Joachims took the lead in proposing a text classification algorithm based on support vector machine in 1997, which started the upsurge of various theories and researches on the application of support vector machine in text classification. Alfons J. et al. studied the smooth Bayesian text classification algorithm in 2002. Hiroshi O., Hiroshi A., et al. have studied the features in unbalanced texts, and come to the conclusion that different feature selections will affect the results of text classification. Tantreev et al. proposed an improved feature selection method of TF-IDF. Chin H W, Lam H L and others put forward a text classification method based on KNN and support vector, which improves the accuracy of classification. Guptal et al. used rough set method to select features, which greatly reduced the training time of the classifier and achieved good classification results. Hirsch et al. used TF-IDF model for feature selection, and used genetic algorithm as text classification algorithm to classify Reuters data. Arunasalam et al. put forward a thresholding-free association rule classification algorithm for the first time. This algorithm uses a new measure to solve the problem of unbalanced distribution of categories in the text. K. Yi et al. selected the features of the medical field and classified the medical texts by using hidden Markov model classification algorithm. In 2008, Zhou Puxiong used ANN algorithm, KNN and SVM algorithm to classify texts. Compared with the traditional classification methods, using the existing natural language processing tools has the problem of error superposition in the processing process [13]. In 2014, Zeng D J et al. put forward a learning method of text semantic features based on deep convolution neural network. According to the degree of correlation between apparent and potential semantics and the categories of documents, this method can handle the classification of irregular texts such as Chinese network short texts well [14]. In 2018, Li H M, et al. proposed a short text classification model based on dense network as direct expression text [15]. In 2019, Wang Gensheng, Huang Xuejian and Chloe Wang optimized the text classification algorithm by modifying the word vector weight and manually building a dictionary, respectively. However, its learning time complexity is much higher than the traditional method, and it needs further improvement [16]. In 2019, Jin W Z proposed a text classification method based on feature fusion model of deep learning [17]. Although machine learning has made extremely important achievements in the field of text classification, the research on text classification was once stagnant before this, but the characteristics of text classification itself put forward a new development direction for machine learning [18-21], so the

research on text classification is still an extremely important direction in the area of NLP at present.

2.2 Overview of Text Classification

Text classification is a kind of supervised learning. It is known that there is a set of training documents $D = \{d_1, d_2, \ldots, d_m\}$, and each document in the set has a category label. The rules between the category label and the attributes in each document are found through supervised learning, and then the category label is obtained by using the rules for new documents.

Text classification can be defined in the following mathematical form: given a set of documents, d_i represents the i^{th} document, and there are m documents in D. Assuming a set of document categories $C = \{c_1, c_2, \ldots, c_m\}$, we can find that there is a certain mapping between the set of documents and the set of categories $f \rightarrow D \times C \rightarrow A$, $A = \{0, 1\}$, the task of text classification is actually to make it equal to F as much as possible. Called a classifier. If $f'\left(\left\langle d_j, c_i \right\rangle\right) = 0$, d_j belongs to the class c_i. If $f'\left(\left\langle d_j, c_i \right\rangle\right) = 1$, it is said that it does not belong to the class c_i.

The text classification process consists of training process and classification process as shown in Fig. 1. In the training process, the training text generally needs to go through the steps shown in Fig. 1, which are the basis of text classification, and then the classifier is continuously trained by the selected classification algorithm. In the process of classification, the test document generally needs to be processed by the steps shown in Fig. 1. After the trained classifier, the classifier will identify the category of the test document.

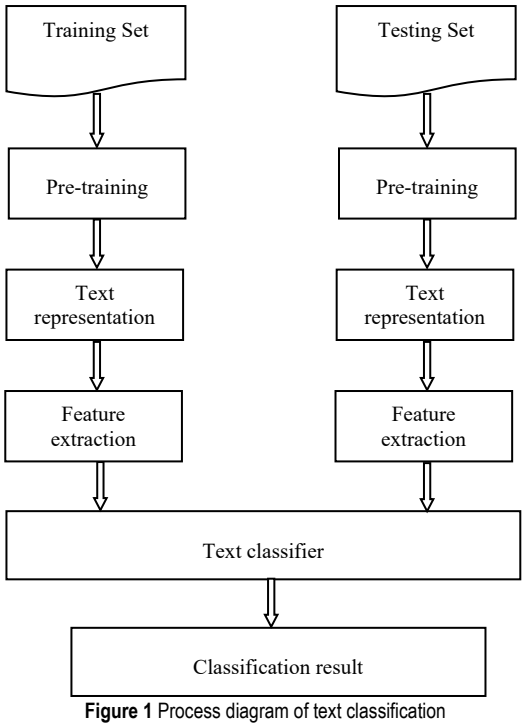

Figure 1 Process diagram of text classification

The text contains a large amount of unstructured or semi-structured information, which is not easily recognized by the classifier. It is necessary to preprocess the text documents to remove these useless information. Text preprocessing refers to processing text information into structured information that can be operated by computer. Text preprocessing is the initial stage of text classification, and the preprocessing results have great influence on the classification results. Text preprocessing includes denoising, word segmentation and stopword removal. In English, spaces and punctuation marks are commonly used for word segmentation. In English, it is necessary to take root, which is to unify words with the same semantics but slightly different forms into one form. It mainly aims at singular and plural forms of nouns, comparative forms of adjectives and adverbs, and various tense forms of verbs in English. To go to stop words is to remove pronouns, prepositions, conjunctions and other features unrelated to classification. These stop words are irrelevant to the meaning of the document.

2.3 Common Models of Deep Learning

2.3.1 Automatic Encoder

Automatic encoder is a kind of unsupervised learning, a new network reconstructed by neural network [22, 23]. The encoder principle of automatically acquiring data features makes input data and output data identical. By constantly adjusting the weight of each layer through training, each hidden layer is another representation of the input data and can be used as the features of the input data. Compared with principal component analysis, automatic encoder relies on the limitation of linear dimension reduction, and it can use nonlinear neural network to reduce the dimension of features. Automatic encoder consists of encoder and decoder. The output of the original data after passing through the encoder is used as the input of the decoder, and finally the output is obtained through the decoder. Then, the original data is printed in another form, as shown in Fig. 2 below.

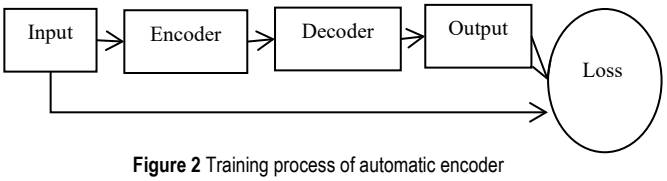

Figure 2 Training process of automatic encoder

2.3.2 Convolutional Neural Network (CNN)

It was put forward by Lecun in 1989, and it is well applied in the field of speech recognition and image recognition. Convolutional neural network is essentially a multilayer perceptron that can recognize images well [27]. Because of its special structure, it can highly perceive other forms of invariance such as translation and scaling of graphics.

CNN [24] is composed of one or more convolution layers and the top fully connected layer, and includes correlation weights and pooling layers. This structure enables CNN to make use of the two-dimensional structure of input data. The

structure diagram of CNN is shown in Fig. 3. First, the input original features are convolved in C1 layer, and then transformed into feature maps after passing through three filters. Then, the feature maps are weighted, and then biased, and finally Sigmoid function is processed to generate S2 layer feature maps. The obtain feature map is processed as above to successively obtain C3 and S4 layer feature map. The feature mapping mentioned above can well realize the feature that the position is not easy to change, and Sigmoid function is used as the activation function. The middle layer C is the feature extraction layer, and the nodes of neurons in each layer are connected with the local nodes in the front layer to extract local features.

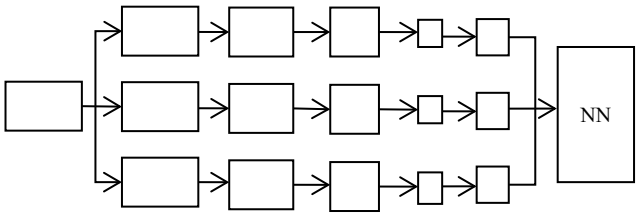

Figure 3 Schematic diagram of convolutional neural network structure

Table 1 20Newsgroups data set

Category	Number of texts
alt.atheism	900
computer.graphics	900
computer.os.ms-windows.misc	900
computer.windows.x	900
misc.forsale	900
computer.system.ibm.pc.hardware	900
computer.system.mac.hardware	900
rec.sport.baseball	900
rec.sport.hockey	900
rec.autos	900
rec.motorcycles	900
science.med	900
soc.religion.christian	900
science.crypt	900
science.electronics	900
talk.politics.guns	900
science.space	900
talk.politics.misc	900
talk.politics.mideast	900
talk.religion.misc	897

3 EXPERIMENTAL RESULT ANALYSIS

3.1 Data Set

The standard foreign language classification database includes: Reuters-21578, 20Newsgroups, OHSUMED, Web KB, etc. Domestic standard Chinese corpus such as Tan Corp, etc. These data sets can be downloaded for free. In this paper, 20Newsgroups data sets such as Tab. 1 are selected for the English text classification experiment. This data set is a text data set compiled by Lang in 1995. It contains the message texts of 20 newsgroups (20 categories) in Usenet, with a total of 1997 articles. Except one newsgroup contains 997 messages, each newsgroup has 1000 message texts. This data set is a typical single-label text classification corpus.

3.2 Text Classification Experiment

A certain number of texts are selected from the random English data set 20Newsgroup to preprocess the texts. Text preprocessing is implemented on Eclipse platform using Java language. For the preprocessed documents, 30% of them are randomly selected as test data sets, and the rest of them are training data sets. The feature dimension of the document is 1500 dimensions. In this paper, the classifier based on the mixed model of deep learning is implemented by MATLAB. Because the original feature dimension is 1500, the number of input nodes in the sparse automatic encoder layer is 1500. After using the sparse automatic encoder with 3000-1500 hidden nodes, the data is compressed by a three-layer deep confidence network with 200-100-20 hidden nodes. Finally, the Softmax layer outputs the test data set with the highest probability that the documents belong to all categories of documents. After text classification, the best accuracy of each category can be obtained as shown in Fig. 5 and Tab. 2.

Table 2 Accuracy of text classification experiment

Category	Number	Accuracy rate
alt.atheism	471	93.49%
computer.graphics	574	96.71%
computer.os.ms-windows.misc	590	80.23%
computer.system.ibm.pc.hardware	590	85.11%
computer.system.mac.hardware	580	88.3%
computer.windows.x	590	93.3%
misc.forsale	585	95.29%
rec.autos	594	94.61%
rec.motorcycles	600	96.37%
rec.sport.baseball	598	94.85%
rec.sport.hockey	600	96.73%
science.crypt	600	97.46%
science.electronics	590	96.45%
science.med	595	89.31%
science.space	594	94.25%
soc.religion.christian	600	95.71%
talk.politics.guns	550	94.68%
talk.politics.mideast	560	95.88%
talk.politics.misc	467	89.77%
talk.religion.misc	380	62.4%
total	11380	91.52%

In order to further verify the performance of text classification based on the hybrid model of deep learning, this paper compares the proposed classifier SDBN with naive Bayes classifier, KNN classifier and support vector machine classifier. In the contrast experiment, the same data set is selected as the training set and the test set, and the text of the training set and the test set is preprocessed. In the experiment of naive Bayes classifier, the naive Bayes classifier [25] of MATLAB is used to get the classification accuracy of the test set. In the experiment of KNN classifier, the KNN-classify classifier built in MATLAB is used to get the classification accuracy of the test set. In the classification experiment of support vector machine, LIBSVM, an open source software package of support vector machine, is used for the experiment, and the classification accuracy of the test set is obtained. Fig. 5 shows the experimental results of comparative experiment, from which it can be seen that the

performance of text classifier SDBN based on deep learning hybrid model is slightly better than other classifiers.

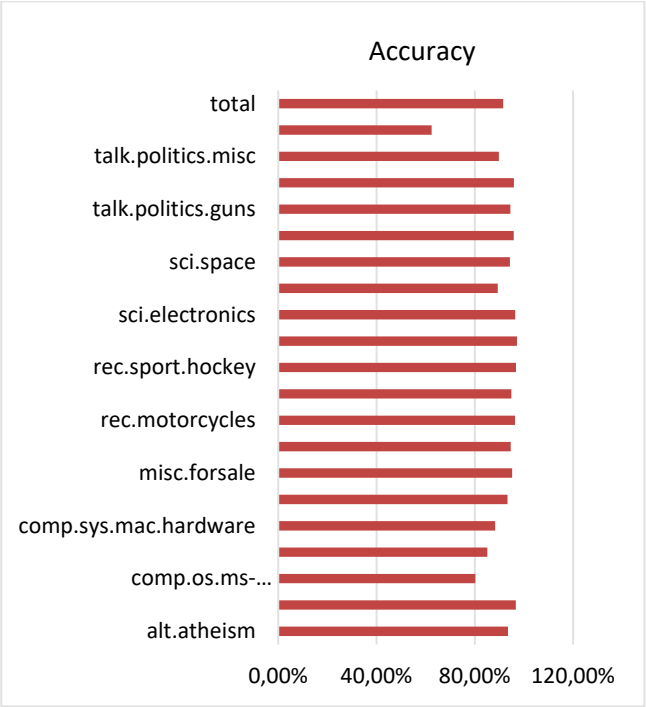

Figure 4 Accuracy rate of text classification experiment

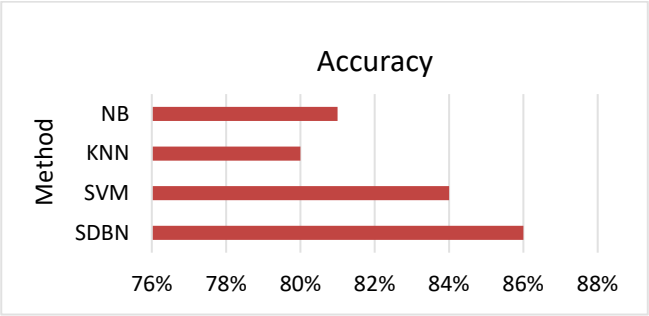

Figure 5 Classification accuracy rate

From Fudan University Chinese Text Classification Corpus, four kinds of documents, namely, economy, sports, computer and agriculture, are selected as the training set and test set of Chinese experiment. In this paper, the classifier based on the mixed model of deep learning is implemented by MATLAB, and 30% of the documents in the preprocessed Chinese data set are randomly selected as the test set. Select features with 1000 dimensions as raw data. Firstly, a sparse automatic encoder with 2000-1000 hidden nodes is used, then a three-layer deep confidence network with 200-100-20 hidden nodes is used to compress the data, and the BP fine-tuning times are 200. Finally, Softmax layer outputs the probability that the documents in the test data set belong to various categories in the documents, and the category with the highest probability is the category to which the documents belong. Tab. 3 is the best correct rate after text classification.

Table 3 Accuracy rate of text classification experiment

Category	Number	Accuracy rate
C34-Economy	1360	86.1%
C39-Sports	1023	87.7%
C19-Computer	1602	92.5%
C32-Agriculture	1255	81.3%

Tab. 4 to Tab. 11 show the comparison of accuracy recall and F1 value of BRCNNs and ACNNNS on MR, Subj, SST, SST2, IMDB, TREC, CR and MPQA data sets respectively. As can be seen from the following accuracy recall and F1 tables, on all data sets used in this paper. There is little difference between the accuracy recall rate and F1 value of BRCNN and ACNN and their variants. This shows that there is no case that a large number of samples of a certain category are predicted as low recall rate caused by other categories, and there is no case that a large number of samples of other categories are predicted as a certain category with low accuracy rate. Therefore, the prediction of the model is relatively average, and even unbalanced data can predict the categories well. For example, the data in the data set TREC is unbalanced, but it can also be predicted well. This shows that the corresponding accuracy recall rate in the table above these models is the value near the "balance point" on the P-R curve. This is mainly because 0.5 is usually used as the threshold for judging categories in the second classification task, and this paper also uses 0.5 as the classification threshold in the experiment. This also shows that BRCNN and ACNN are suitable for text classification tasks.

Table 4 The precision score, recall score and F1 score of ACNNs and BRCNNs models on the MR dataset

Models	macro-P	macro-FI	macro-R	micro-P	micro-FI	micro-R
BRCNN (RNN)	82.7	82.4	82.5	82.4	85.4	82.4
BRCNN (LSTM)	83.9	83.5	83.5	83.5	83.5	83.5
BRCNN (GRU)	83.0	81.6	81.8	81.6	81.6	81.6
ACNN (RNN)	83.5	80.5	80.5	80.5	80.5	80.5
ACNN (LSTM)	83.9	80.7	80.8	80.7	80.7	80.7
ACNN (GRU)	86.8	86.8	86.8	86.8	86.8	86.8

Table 5 The precision score, recall score and F1 score of ACNNs and BRCNNs models on the Subj dataset

Models	macro-P	macro-FI	macro-R	micro-P	micro-FI	micro-R
BRCNN (RNN)	92.4	92.1	92.0	92.2	92.2	92.2
BRCNN (LSTM)	93.5	93.5	93.6	94.5	95.5	93.5
BRCNN (GRU)	92.7	91.3	92.1	93.4	94.4	94.4
ACNN (RNN)	90.8	90.3	90.1	90.4	90.4	90.4
ACNN (LSTM)	92.0	92.4	92.1	92.6	92.1	92.1
ACNN (GRU)	92.9	92.9	92.8	92.9	92.9	92.9

Table 6 The precision score, recall score and F1 score of ACNNs and BRCNNs models on the ST dataset

Models	macro-P	macro-FI	macro-R	micro-P	micro-FI	micro-R
BRCNN (RNN)	46.3	46.2	46.0	46.3	46.3	46.3
BRCNN (LSTM)	48.0	48.1	48.2	48.1	48.1	48.1
BRCNN (GRU)	47.3	47.7	47.5	47.4	47.4	47.4
ACNN (RNN)	46.6	46.6	46.5	46.6	46.6	46.6
ACNN (LSTM)	48.7	46.5	46.0	48.3	48.3	48.4
ACNN (GRU)	47.9	47.8	47.8	47.7	47.9	47.9

Table 7 The precision score, recall score and F1 score of ACNNs and BRCNNs models on SST2 dataset

Models	macro-P	macro-F1	macro-R	micro-P	micro-FI	micro-R
BRCNN (RNN)	83.3	83.3	83.3	83.3	83.3	83.3
BRCNN (LSTM)	83.0	93.3	83.3	83.3	83.3	83.3
BRCNN (GRU)	83.7	83.7	83.7	83.7	83.7	83.7
ACNN (RNN)	83.7	83.7	83.7	83.7	83.7	83.7
ACNN (LSTM)	87.3	85.8	85.9	85.9	85.9	85.9
ACNN (GRU)	85.7	85.7	85.7	85.7	85.7	85.7

Table 8 The precision score, recall score and F1 score of ACNNs and BRCNNs models on the IMDB dataset

Models	macro-P	macro-F1	macro-R	micro-P	micro-FI	micro-R
BRCNN (RNN)	86.0	86.0	86.l	86.l	86.2	86.2
BRCNN (LSTM)	88.5	88.5	88.5	87.5	88.5	88.5
BRCNN (GRU)	86.9	86.9	86.9	86.9	86.9	86.9
ACNN (RNN)	86.2	86.2	86.2	86.2	86.2□	86.2
ACNN (LSTM)	88.6	87.5	88.5	88.5	88.5	87.5
ACNN (GRU)	88.2	88.3	88.3	88.3	88.3	87.3

Table 9 The precision score, recall score and fl score of ACNNs and BRCNNs models on TREC dataset

Models	macro-P	macro-F1	macro-R	micro-P	micro-F1	micro-R
BRCNN (RNN)	81.2	81.8	83.7	88.6	88.6	88.6
BRCNN (LSTM)	93.6	91.0	90.7	94.0	94.0	94.0
BRCNN (GRU)	91.9	89.6	87.4	91.1	91.1	91.1
ACNN (RNN)	91.4	90.8	89.7	91.9	91.9	91.9
ACNN (LSTM)	94.9	91.3	90.4	93.8	93.8	93.8
ACNN (GRU)	94.4	91.2	90.7	93.4	93.4	93.4

Table 10 The precision score, recall score and F1 score of ACNNs and BRCNNs models on CR dataset

Models	macro-P	macro-F1	macro-R	micro-P	micro-F1	micro-R
BRCNN (RNN)	86.7	86.2	81.7	86.7	86.3	86.7
BRCNN (LSTM)	86.2	86.8	88.7	88.4	88.4	88.4
BRCNN (GRU)	84.9	85.3	85.9	86.3	86.3	86.3
ACNN (RNN)	85.6	85.6	85.6	85.6	85.6	85.6
ACNN (LSTM)	84.3	83.9	83.6	85.3	85.3	85.3
ACNN (GRU)	86.7	86.9	88.2	88.7	88.7	88.7

Table 11 The precision score, recall score and F1 score of ACNNs and BRCNNs models on PQA dataset

Models	macro-P	macro-F1	macro-R	micro-P	micro-FI	micro-R
BRCNN (RNN)	87.1	86.1	85.2	88.4	89.4	88.4
BRCNN (LSTM)	88.2	88.3	88.4	91.9	90.9	90.9
BRCNN (GRU)	90.9	87.1	86.2	90.6	90.8	90.6
ACNN (RNN)	91.5	87.3	85.8	90.6	90.6	90.7
ACNN (LSTM)	88.5	87.9	87.4	90.8	90.8	90.8
ACNN (GRU)	90.6	88.1	86.9	92.1	90.1	92.1

In practice, sometimes we pay more attention to the accuracy of classification, and sometimes we may pay more attention to the recall rate of classification. At this time, we can set the classification threshold to different values according to the specific situation to get different accuracy rates or recall rates. For example, reducing the classification threshold will get a higher recall rate, but the accuracy rate will be relatively reduced. On the contrary, increasing the classification threshold will increase the pricision, but the recall rate will decrease. On the whole, the peicision and recall rate of the model are relatively average when it is near the "balance point". In practice, you can draw the P-R curve first, and then set the classification threshold according to the situation analysis.

4 CONCLUSION

To learn the classification performance of classifiers based on deep mixed model, relevant experiments were conducted on English data set 20Newsgroup and Chinese data set Fudan University Chinese Corpus respectively. In the English text experiment, the classifier based on deep learning hybrid model is used for classification, and the classification accuracy rate is 91%. In order to further verify

the superiority of its performance, the comparison experiment with naive Bayes classifier, KNN classifier and support vector machine classifier shows that the classification effect based on deep learning hybrid model is slightly better than that of SVM classifier, KNN classifier and naive Bayes classifier. In the Chinese text experiment, the Chinese corpus of Fudan University is tested and a good classification effect is obtained, and the influence of different parameters on the classification accuracy is discussed.

In the future study, we will further study the work of this paper. We can improve the algorithm in the model of deep learning, and try to use other models in deep learning, such as CNN, to learn the features of text classification. In a word, text classification based on deep learning hybrid model will have a good application prospect in the future development. With the continuous improvement of deep learning theory, more new research results of deep learning will be added to the further research of this paper, which will greatly improve the performance of text classifier based on deep learning hybrid model.

5 REFERENCES

[1] Vaddempudi, S. (2016). Classification of user behaviour in mobile internet. *Asia-pacific Journal of Convergent Research Interchange, SoCoRI, 2*(2), 9-18. https://doi.org/10.21742/APJCRI.2016.06.02

[2] Meng, L., Zhang, S., & Wang, F. (2020). Influence of internet-based social big data on personal credit reporting. *Asia-pacific Journal of Convergent Research Interchange, FuCoS, 6*(7), 39-57. https://doi.org/10.47116/apjcri.2020.07.05

[3] Ko, J.-U., Kim, K.-H., & Jeon, J.-H. (2021). The effects of self-control on internet immersion: The moderating effects of parenting attitude. *Asia-pacific Journal of Convergent Research Interchange, FuCoS, 7*(9), 47-57. https://doi.org/10.47116/apjcri.2021.09.05

[4] Kim, H.-G. & Ko, J.-U. (2022). The effect of adolescents' perceived parenting attitudes on internet immersion - Focusing on the mediating effect of self-control. *Asia-pacific Journal of Convergent Research Interchange*, FuCoS, 8(3), 187-201. https://doi.org/10.47116/apjcri.2022.03.17

[5] Park, W., Jung, D., Jung, H., Ekouka Elvis, T., & Kim, H. (2022). Effecting system quality factors on purchase intention of internet insurance. *Asia-pacific Journal of Convergent Research Interchange, FuCoS, 8*(4), 47-56. https://doi.org/10.47116/apjcri.2022.04.05

[6] Park, W. & Kim, H. (2022). Effecting information quality of information system on corporate performance: Focused on internet insurance. *Asia-pacific Journal of Convergent Research Interchange, FuCoS, 8*(5), 21-30. https://doi.org/10.47116/apjcri.2022.05.03

[7] Kim, D. (2020). Application of deep neural network model for automated intelligent excavator. *Asia-pacific Journal of Convergent Research Interchange, FuCoS, 6*(4), 13-22. https://doi.org/10.21742/apjcri.2020.04.02

[8] Mononteliza, J. (2020). Research on EIoT reservation algorithm based on deep learning. *Asia-pacific Journal of Convergent Research Interchange, FuCoS, 6*(9), 191-205 https://doi.org/10.47116/apjcri.2020.09.16

[9] Jang, S.-B. (2021). Deep neural network structure design for equipment failure prediction in smart factory. *Asia-pacific Journal of Convergent Research Interchange, FuCoS, 7*(12), 1-10. https://doi.org/10.47116/apjcri.2021.12.01

[10] Lee, Y. G. (2022). Developing an automatic floor plan generation and evaluation technology using deep learning in the architectural design process. *Asia-pacific Journal of Convergent Research Interchange*, FuCoS, 8(5), 1-10. https://doi.org/10.47116/apjcri.2022.05.01

[11] Xie, R., Yuan, X., & Liu, Z. (2017). Lexical sememe prediction via word embeddings and matrix factorization. *The 26th International Joint Conference on Artificial Intelligence*, 4200-4206. https://doi.org/10.24963/ijcai.2017/587

[12] Liu, Z. W., Ding, D., & Li, C. W. (2015). A Chinese short text word segmentation method based on conditional random fields. *Journal of Tsinghua University (Science and Technology), 55*(8), 906-910+915.

[13] Du, L. P., Li, X. G., & Yu, G. (2016). New word detection based on an improved PMI algorithm for enhancing segmentation system. *Journal of Peking University (Natural Science), 52*(1), 35-40.

[14] Hu, J. & Zhang, J. C. (2017). Bidirectional recurrent network for Chinese word segmentation. *Journal of Chinese Computer Systems, 38*(3), 522-526.

[15] Zhang, Y. N., Xu, J. N., & Miao, G. Y. (2018). Improving neural Chinese word segmentation using unlabeled data. *IOP Conference Series: Materials Science and Engineering, 435*(1). https://doi.org/10.1088/1757-899X/435/1/012032

[16] Yang, G. B., Yang, F. H., & Mao, G. J. (2019). Chinese word segmentation technology based on group hash and variable length matching. *Computer Age*, 4, 52-55.

[17] Chen, X. C., Shi, Z. & Qiu, X. P. (2017). Adversarial multi-criteria learning for Chinese word segmentation. ACL2017: *Computation and Language*. https://doi.org/10.18653/v1/P17-1110

[18] Yu, C. H., Wang, S. P., & Guo, J. J. (2019). Learning Chinese word segmentation based on bidirectional GRU-CRF and CNN network model. *International Journal of Technology and Human Interaction (IJTHI), 15*(3). https://doi.org/10.4018/IJTHI.2019070104

[19] Darshan, S. L. & Jaidhar, C. D. (2018). Performance evaluation of filter-based feature selection techniques in classifying portable executable files. *Procedia Computer Science*, (125), 346-356. https://doi.org/10.1016/j.procs.2017.12.046

[20] Hancer, E. (2019). Differential evolution for feature selection: A fuzzy wrapper–filter approach. *Soft Computing, 23*(13), 5233-5248. https://doi.org/10.1007/s00500-018-3545-7

[21] Maldonado, S. & López, J. (2018). Dealing with high-dimensional class-imbalanced datasets: Embedded feature selection for SVM classification. *Applied Soft Computing*, (67), 228-246. https://doi.org/10.1016/j.asoc.2018.02.051

[22] Cheong, Y. G. (2021). Analysis of autoencoders for network intrusion detection. *Sensors*, (21). https://doi.org/10.3390/s21134294

[23] Li, X. Liu, Z., & Huang, Z. (2020). Denoising of radar pulse streams with autoencoders. *IEEE Communications Letters*, (99), 1-1. https://doi.org/10.1109/LCOMM.2020.2967365

[24] Davis, P., Aziz, F. B., & Newaz, T. (2021). The classification of construction waste material using a deep convolutional neural network. *Automation in Construction*, 122. https://doi.org/10.1016/j.autcon.2020.103481

[25] Thong, W. & Snoek, C. G. M. (2021). Feature and label embedding spaces matter in addressing image classifier bias. arXiv: 2110.14336. https://doi.org/10.48550/arXiv.2110.14336

[26] Barak, F. & Kaplan, K. (2021). The study of handwriting recognition algorithms based on neural networks. *International Journal of Hybrid Innovation Technologies, 1*(2), 63-67. https://doi.org/10.21742/ijhit.2653-309X.2021.1.2.04

[27] Chayangkoon, N. & Srivihok, A. (2019). Feature Reduction of Short Text Classification by Using Bag of Words and Word Embedding. *International Journal of Control and Automation, NADIA, 12*(2), 1-16. https://doi.org/10.33832/ijca.2019.12.2.01

Author's contacts:

Sang-Hwa Lee
Department of Webtoon Contents, Seowon University,
377-3 Musimseo-ro, Seowon-gu, Cheongju-si,
Chungcheongbuk-do, 28674, Republic of Korea
gomawooi@naver.com