# Text Classification based on Model of Feature Fusion and Voting Mechanism for Personality Recognition

**Shoryu Teragawa[1,] *, Ming Li[2]**

[1]School of Software, Dalian University of Technology, Linggong Road, Dalian, China

[2]Automotive Electronics, Neusoft Corporation Dalian, Software Park Road, Dalian, China

**Keywords:** User personality recognition; multi-dimensional features; neural network

**Abstract:** Psychology, as a discipline that relies on traditional statistical methods, has its limitations in research methods. Based on this fact, this article attempts to make up for this shortcoming by deep learning method. First, the words in the user text are converted into vectors by frequency through embedding, and the feature information in the user text is extracted through different neural networks. Then feature fusion through voting and other methods enables more accurate user information to be used for classification. According to the experimental results, it is found that the model proposed in this paper is more effective in extracting multi-dimensional features from user texts, and effectively optimizes traditional algorithms. Compared with traditional models, the accuracy is improved.

## 1. Introduction

Psychology is a discipline that studies human behaviours and the laws of psychological activity. People's attention to psychology and behaviour can be traced back to BC. In 1879, Wundt established the first psychology laboratory at the Germany University of Leipzig and makes psychology was separated from philosophy and became an independent science [1] [2]. Because of the limitations of traditional statistical methods, the development of various disciplines that rely on traditional statistical methods has been limited to a certain extent. Psychology, as a discipline that relies on statistical methods, has obvious limitations, such as the traditional algorithm feature extraction is limited for text or picture and music feature extraction. As one of the main artificial intelligence algorithms, machine learning can more effectively extract more data features and is often used in interdisciplinary research [3].

For the above reasons, this article attempts to incorporate deep learning algorithms into the solution of related problems.

## 2. Related Work

With the popularization of the Internet in society, the amount of information on the Internet has exploded. Text classification as a research direction in the field of natural language processing [4], its purpose is to distinguish the differences in texts from the features in texts [5]. The main methods include text library based and deep learning based [6] [7]. For example, Kaiyang Liu [8] proposed a news text classification method based on a combination of Bert word vectors and convolutional neural networks. In the case of using Bert (Bidirectional Transformer) word vectors and convolutional neural networks to classify news, the efficiency is higher than the text classification method combined with word vectors. Tang et al. Proposed text classification based on [9] based on the transformer-capsule integration model. Four single-label datasets and one multi-label Reuters-21578 dataset in the text categorical corpus were selected for experiments, and good experimental results were obtained. Fugang Liu [10] proposed a method for classifying Chinese technology blogs using Naive Bayes classification algorithm. Extract feature vectors from blog titles and calculate the probability of each feature vector on each category. Wen et al [11]. Proposed a feature selection method combining information gain and firefly algorithm. Calculate the information gain

  DOI: 10.25236/icbdem.2020.049

of all feature words and sort them from high to low. Use the firefly algorithm to search for the best feature subset on the feature set with the highest ranking. Xue et al. [12] proposed an LSTM (long short-term memory) A text classification model. The model uses LSTM network to encode the input sequence, and it introduces the attention mechanism to assign different weights to the text features. Finally, the text dataset from incopat is used to verify the validity of the method. Liu et al. [13] proposed the use of genetic algorithms to optimize the selection of text features to make it fit the subsequent text classification algorithms to the greatest extent. While ensuring the accuracy of text classification, the feature dimensions were reduced to reduce prediction time. Although the above text classification algorithms have better results in their respective fields. However, text information based on text personality classification has the characteristics of strong data distribution and high uncertainty. Based on the above characteristics, this paper proposes a method of text classification based on model of feature fusion and voting mechanism for personality recognition. This model takes better into account the feature extraction method of this type of text, so as to obtain a better prediction model.

## 3. Text Classification based on Model of Feature Fusion and Voting Mechanism

When analysing a person's personality through speech, the more the number of speeches, the higher the accuracy of the analysis, and the feature fusion can solve the diversity of feature extraction to a certain extent. Based on this fact, this paper proposes a model of feature fusion and voting mechanism. This method uses each model to extract features from the embedded text, then uses SoftMax to predict and average multiple results, and finally uses multiple models to vote as the final output to predict the user's personality. The structure is shown in Figure 1.

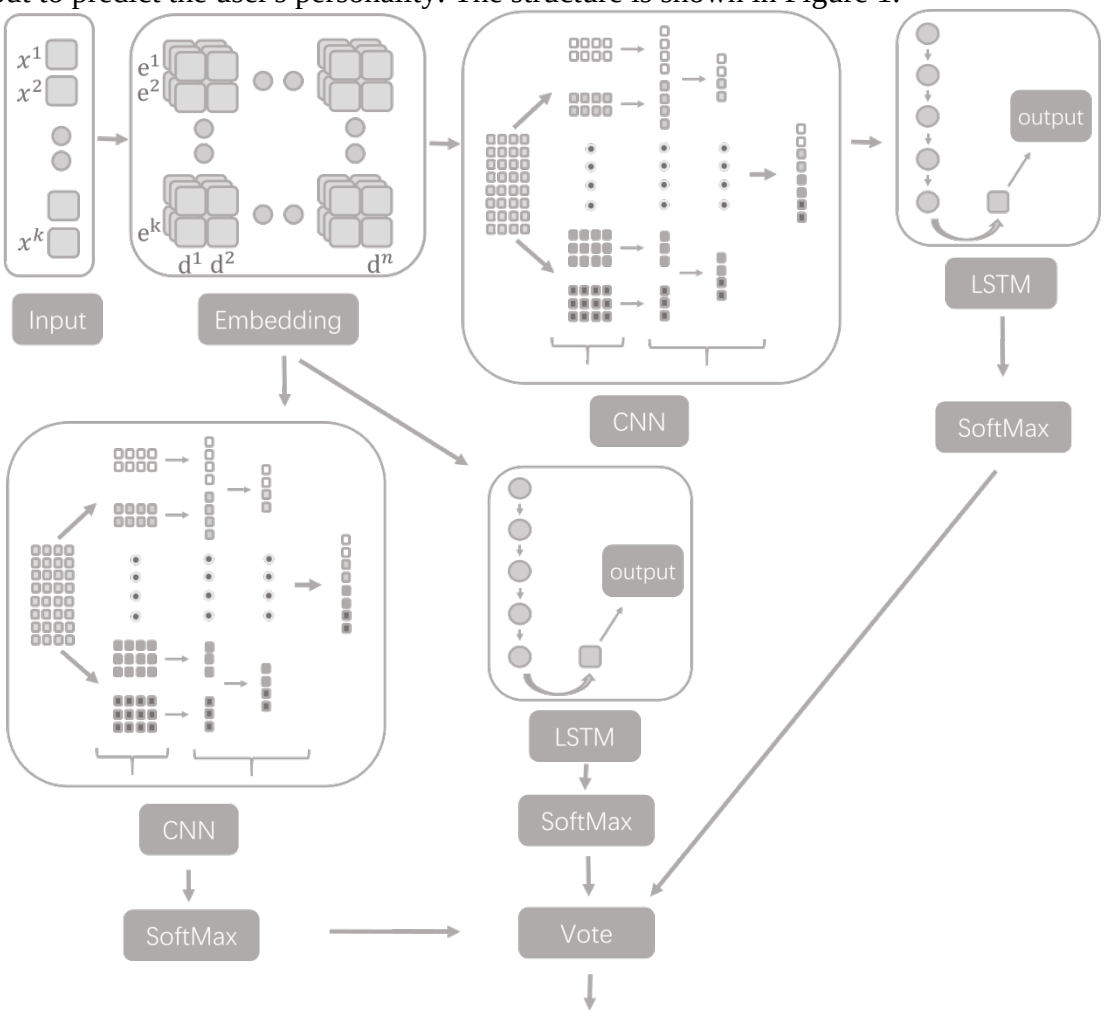


Figure 1 model of feature fusion and voting mechanism

### 3.1. CNN text feature extraction

CNN (convolutional neural network) is a kind of feed-forward neural network, which is widely used in natural language processing, image processing, speech recognition and other fields, and has achieved good results in these areas, it is a kind of basic neural network. The convolutional neural network has a feature extraction function, which can extract features in the data layer by layer through convolution operations and pooling operations through the advantages of weight sharing.

As shown in Figure 2: [14] [15]

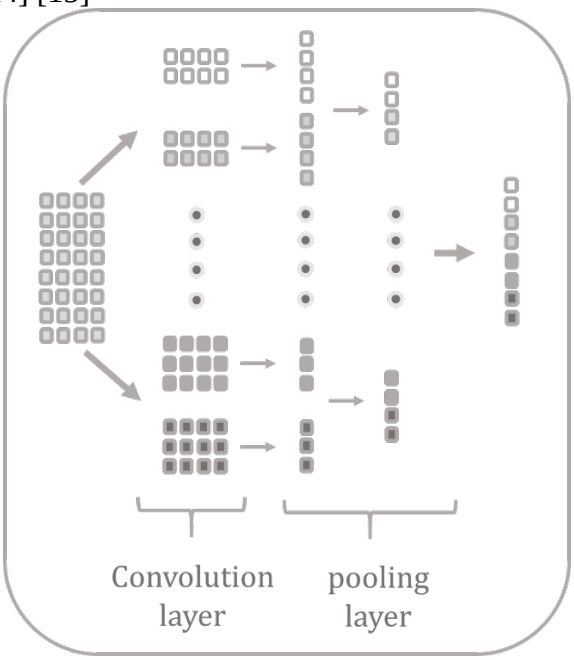


Figure 2 Structure of CNN

The convolution layer obtains new features by convolving with a text vector matrix S of size n × d.

$$c_i = f(w \times S_{i:i+h-1} + b) \quad (1)$$

Where i means the eigenvalue, h means sliding window size in the convolution calculation, w means filter, f means non-linear activation function, and b means the bias. C means feature vector.

$$C=(c_1, c_1, \ldots, c_{n-h+1}) \quad (2)$$

$$c_{max} = Max(c_i) \quad (3)$$

The pooling layer is used to extract the feature map information output by the convolution layer and reduce network parameters. This paper uses the maximum pooling method to take the largest feature value in the pooled area in the feature map.

### 3.2. LSTM Text Feature Extraction

As an improved version of the RNN (Recurrent Neural Network) algorithm, LSTM compensates for the shortcomings of the text sequence feature retention short time in the RNN algorithm, and its accuracy is significantly improved compared to the RNN algorithm. Sequence semantic information in text can be saved for a long time. As shown in Figure 3:

Equation (4) is Forget gate:

$$f_t = \text{sigmoid}\,(U_f \cdot c_{t-1} + W_f \cdot x_t + b_f) \quad (4)$$

Equation (5) is Input gate:

$$i_t = \text{sigmoid}\,(U_i \cdot h_{t-1} + W_i \cdot x_t + b_i) \quad (5)$$

Equation (6) is Output gate:

$$o_t = \text{sigmoid}\left( U_o \cdot h_{t-1} + W_o \cdot x_t + b_o \right) \quad (6)$$

Equation (7) is Memory cell input:

$$\tilde{c}_t = \tanh\left( U_c \cdot h_{t-1} + W_c \cdot x_t + b_c \right) \quad (7)$$

Equation (8) is Memory cell output:

$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \quad (8)$$

Equation (9) is Final output:

$$h_t = o_t \circ \tanh(c_t) \quad (9)$$

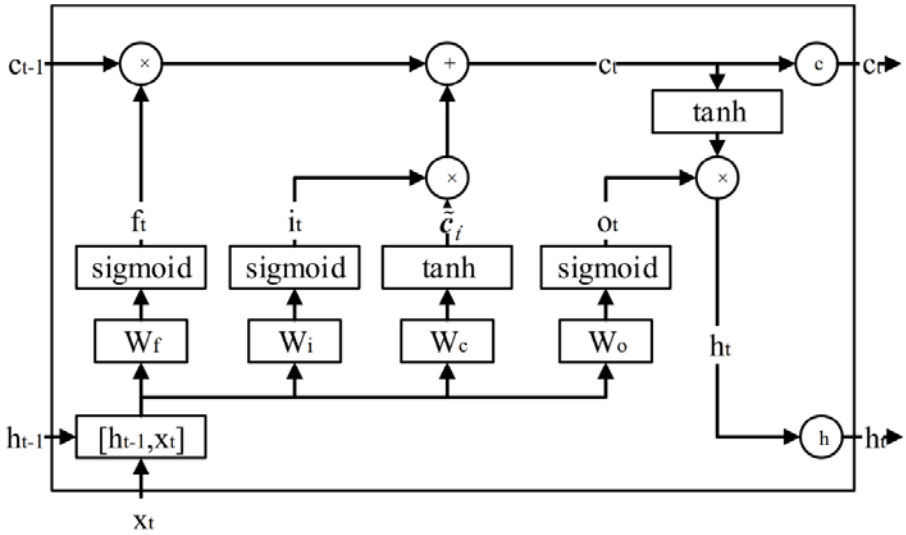


Figure 3 Structure of LSTM [16]

### 3.3. Voting Mechanism

The voting mechanism is determined by the fitting results of each test data, where the first layer of data is determined by the size of each data through each predicted value. Among them, 0 indicates that the fitted data result is an outward type, and 1 indicates that the fitted data result is an inward type. The voting result is determined by the party with the most votes. If the vote is a tie, the total number of votes of the voter and the result after the SoftMax are compared as the final voting result. The second layer of voting is the sum of 10 attributes (10 sentences of a person), which finally determines the fitted value of the user's personality attributes.

## 4. Experiment and Results

### 4.1. Dataset

The source of the classification index of the user's personality is determined by the user's personality tag on the Internet. The text takes the user's text information as training data and the user's personality data as fitting data. The source of the data set in this article is from Zhihu.com, a total of 80 people sent out long texts and each person included 10 comments, a total of 800 as data sets. There are two types of user personality, 40 introverts and 40 extroverts.

Sixty percent of the data is used as the training set, ten percent is used as the test set, and thirty percent is used as the test set.

### 4.2. Evaluation Standard

The This paper uses accuracy as criteria for judging network performance

TP- predicts positive classes as positive.

FN- predicts positive classes as negative classes.

FP- predicts negative classes as positive.

TN- predicts negative classes as negative classes.

$$\text{Accuracy} = (TP + TN) / (TP + FN + FP + TN) \quad (10)$$

### 4.3. Parameter Setting

#### 4.3.1. Text Pre-Processing

The text is segmented by Jieba and the word frequency is counted. Each text varies in length, with a maximum of more than 1000 words. Each sentence is limited to 1000 words in length and each word is converted into a 128-dimensional word vector.

#### 4.3.2. CNN Text Convention Operation

Convolution layer: The number of convolution kernels in this paper is 32. The size of the convolution is related to the length of the text. Convolution kernels of 2, 3, and 4 sizes are used in short text texts, and convolution kernels of 3, 4, and 5 sizes are used in long text texts.

Pooling layer: The pooling size used in this article is 5.

#### 4.3.3. LSTM Text Feature Extraction

Extract the sequence features of the text by LSTM and set the LSTM parameter of return to true, parameter of dropout to 0.2, parameter of recurrent dropout to 0.2.

### 4.4. Experiment Results

Table 1 Accuracy of each model.

| Model | Accuracy |
|---|---|
| LSTM | 0.578 |
| CNN | 0.600 |
| CNN-LSTM | 0.633 |
| LSTM-CNN | 0.578 |
| Model of Feature Fusion and Voting Mechanism | 0.833 |

According to the experimental results, it can be concluded that the algorithm proposed in this paper has a significantly higher accuracy rate than traditional algorithms when dealing with text sentiment classification. And because the algorithm proposed in the text takes into account the fusion between different features, it also has a small improvement over other traditional models. As shown in Table 1.

## 5. Conclusion

In order to solve the text-based personality classification problem, this paper proposes a Text classification based on model of feature fusion and voting mechanism for personality recognition based on it. It is found through experiments that this class can be better solved by this method. And compared with the traditional neural network module performance has been improved.

## References

[1] Liu, X., Liu, X., Xiang, Y. and Zhu, T. (2019) Artificial intelligence and big data as applied to Psychology. Science & Technology Review, 37, 105-109.

[2] Zhu, T. (2019) Scenarios of Applying Artificial Intelligence in Psychological Research. Frontiers, 10, 48-53.

[3] Ren, X. (2019) Interdisciplinary research and multidimensional perspective of AI Philosophy. Academic Research, 6, 14-21.

[4] Chen, Z. and Guo, W. (2020) Text Classification Based on Depth Learning on Unbalanced Data. Journal of Chinese Computer Systems, 41, 1-5.

[5] Ding, C., Xia, H. and Liu, Y. (2020) Short text classification model based on knowledge graph and attention mechanism. Computer Engineering, 1-22, 1-8.

[6] Xiao, L., Chen, B., Huang, X., Liu, H., Jing, L. and Yu, J. (2020) Multi-Label Text Classification Method Based on Label Semantic Information. Journal of Software, 1, 1-11.

[7] Wang, H., Liu, Z. and Guo, K. (2020) Capsule Network Model Based on Mixed Word Embedding for Text Classification. Journal of Chinese Computer Systems, 41, 218-224.

[8] Liu, K. (2020) A Chinese news text classification method of combining Bert character vector and Convolutional Neural Networks. Computer Knowledge and Technology, 16, 187-188.

[9] Tang, Z., Wang, Z., Zhou, A., Feng, M., Qu, W. and Lu, M. (2019) Transformer-capsule Integrated Model for Text Classification. Computer Engineering and Applications, 12-19, 1-7.

[10] Liu, F. (2019) Research on Automatic Technology Blog Classification Based on Naive Bayes. Journal of Changchun Normal University, 38, 36-43.

[11] Wen, W., Zhao, C., Zhao, X., Liu, Y. and Fan, R. (2019) Text feature selection based on information gain and firefly algorithm. COMPUTER ENGINEERING AND DESIGN, 40, 3457-3462.

[12] Xue, J., Jiang, D. and Wu, J. (2019) Patent Text Classification based on Long Short-Term Memory Network and Attention Mechanism. Communications Technology, 52, 2888-2892.

[13] Liu, C., Wang, B. and Wu, Y. (2019) Text Feature Selection Based on Genetic Algorithm. Science Technology and Engineering, 19, 302-307.

[14] Chen, R., Ren, C., Wang, Z., Qu, Z. and Wang, H. (2019) Attention based CRNN for text classification. COMPUTER ENGINEERING AND DESIGN, 40, 3151-3157.

[15] Chen, Z., Feng, A. and He, J. (2019) Text sentiment classification based on 1D convolutional hybrid neural network. Journal of Computer Applications, 39, 1936–1941.

[16] Sun, Q. and Guo, Z. (2019) Vehicle following model based on LSTM neural network. Journal of Jilin University (Engineering and Technology Edition), 41, 1-7.