Analysis and Comparison of Chinese News Text Classification Methods Based on Deep Learning

Jian Chen[1, †], Zekai Feng[2, *, †], Wenxiao Jiang[3, †]

[1] University of Illinois at Urbana Champaign, Illinois, USA

[2] School of Statistics and Mathmatics, Zhongnan University of Economics and Law, Wuhan, China

[3] School of Internet of Things Engineering, Jiangnan University, Wuxi, China

* Corresponding Author Email: fengzekai@stu.zuel.edu.cn

[†]These authors contributed equally.

Abstract. As people in today's world consume an increasing amount of information, the number of Internet News is also vastly increasing. Facing all sorts of different kinds of news, how to accurately distinguished different types of news becomes the direction of many scholars' study.This article uses word cloud to represent keywords used in different domains of news. Moreover, we used two methods: TF-IDF and TextRank, to identify and analyze keywords of different fields of news. To understand the performance using various classification methods, we choose the THUCNews data sets. This data set collects ten fields of news in the history of Weibo. Moreover, we choose nine different kinds of machine learning algorithms, including SVM, XGBoost, RandomForest, GBDT, GRU, LSTM, CNN, RNN, and MLP, to investigate their performance. Among these nine models, GRU has an accuracy of 96.93%, SVM has an accuracy of 96.39%, CNN has an accuracy of 94.72%, and RandomForest has an accuracy of 92.97%, which make them stand out in their similar models. We used word-embedding vectorization for the Neural Network algorithm and TF-IDF vectorization for the others.

Keywords: Natural language processing; Text Classification; Chinese news; Deep learning.

1. Introduction

News has always been one of the important ways for people to understand social dynamics and obtain social information resources. With the vigorous development of technology and big data industries, people are gradually accustomed to using smart mobile terminal devices (such as mobile phones, computers, etc.) to consume news information, which allows them to conveniently access a large amount of information at any time. As news becomes more and more dominant and influential in today's society, it is particularly important to analyze different types of news effectively and accurately. To realize the above analysis process, the news classification is a fundamental and key step, which is beneficial to the management of text content, the realization of news order and the mining of news data. News text classification has attracted a lot of research interests in academia and industry circles.

At present, most of the researches on news classification are in English, whose representative framework mainly includes the classification methods based on machine learning and deep learning. The classic machine learning based methods extract shallow feature information from text, select keywords through algorithms and then use classifiers to complete the classification. However, the feature extraction of machine learning algorithms relies on human processing, whose accuracy is susceptible to subjective effects. The overall performance of the algorithm is not good enough, and the results of different data sets are different, which is not conducive to the promotion of the algorithm. Thanks to the powerful feature representation capability of convolutional neural networks, the classification method of news texts that is based on deep learning has gradually become the primary framework, while there is still a particular room for improvement in the recognition accuracy of relatively complex Chinese news texts.

With the influence of global economic integration and the Belt and Road strategy, as the most widely used language in the world, the importance of Chinese in the world language system has rapidly increased. However, the development of Chinese news text classification is relatively slow. On the one hand, there are few relevant corpora for Chinese text classification; on the other hand, the grammar of Chinese is more complex than that of English, and the formulation of feature extraction rules is more difficult. In this paper, through detailed literature collection and analysis, we firstly introduce the existing research and development of text classification, which is primarily based on machine learning. Second, we use classical machine learning algorithms and also deep learning algorithms to build news text classification models on the THUCNews data sets respectively and quantitatively analyze the classification precision of different methods.

We organized our paper into three parts. We first detail existing researches on traditional machine learning methods in news text classification in Section 2. We then describe our preprocessing method and the different models we applied to the dataset in Section 3. Finally, we analyze and compare the performance of different models in Section 4.

2. Representative text classification methods

Antti [1] optimized the Multinomial Naive Bayes algorithm by evaluating seven ways of optimizing and four searching algorithms from five public datasets. By randomly mixing optimization methods and comparing the result with SVM, the missing rate has decreased by 20% on average. Ogura et al. [2] raises the importance of three types of statistical measures of the key-feature extraction based on the problem of unbalanced samples, improving classification accuracy. Seeing the defects of traditional terminology frequency and IF-IDF, Chen et al. [3] gives a solution of TF-IGM, using new statistical models to strengthen the model's discriminative ability on terminology. After many experiments on SVM and KNN models, they prove their TF-IGM to be effective. KNN is a successful algorithm in many fields of data mining, but the traditional KNN is inefficient by a large number of calculations and highly dependent on training sets' size. Hence, Chen et al. [4] provided a quick k-nearest neighbor to reduce the similarities of different experiments, and with fewer samples in training sets, the searching speed of samples increases markedly. Selvi et al. [5] combined Rocchio algorithms with Random Forest and came up with a mixed-text classification model, using stop word removers and stemmers to avoid the limitation of the Rocchio algorithm. The Simulation-experiments result shows that this new algorithm performs better than Fuzzy correlation clustering, ML-KNN, Naive Bayes, and other text classification models. Seyyedi et al. [6] want to improve models' performance while decreasing the training time, fusion feature extraction, and selection process. Considering different text contains different features, Bidi et al. [7] proposed a feature selection algorithm based on a genetic algorithm to search for feature subsets, then use Naive Bayes, KNN, and SVM to evaluate its performance. Zhao et al. [8] try to improve the KNN algorithm's processing rate of large data sets by improving IF-IDF. Based on Map Reduce, the experiment result shows that the improvement of the IF-IDF not only boosts the accuracy of classification but also increases the convergence rate of training.

Based on the research above, some limitations of traditional machine learning can be drawn: The extraction of features by machine learning algorithms relies on manual processing, and its accuracy is susceptible to subjective effects. The general performance of algorithms is not good enough that the result differs from different data sets, which is not conducive to promoting algorithms. Most Internet texts are not normative, which increases the difficulty of formulating rules for feature extraction. To solve these limitations, researchers turn to deep learning in the study of text classification.

2.1. Deep learning based text classification

For the text classification, deep learning has advantages in adaptively extracting deep features and implementing end-to-end classification, which makes up for the shortcomings of machine learning.

Currently, the mainstream news text classification models are CNN, RNN, and LSTM. Li et al. [9] used the comprehensive expression method and Bi-LSTM-CNN model to accurately express the semantics, which improved the accuracy of text classification. Considering that traditional classier rely on artificial features, Lai et al. [10]. Introduced a recurrent CNN to capture as much context information as possible. Compared with the window-based neural network, less noise is introduced within the process. Liu et al. [11] used a multi-task learning framework for text classification. They used a single-layer structure, a double-layer structure, and a shared-layer structure for training, effectively improving the classification performance. Since traditional text classification is based on statistics and feature selection, and standard neural network models ignore context information, Shi et al. [12] improved the traditional TF-IDF algorithm by calculating the weights of eigenvalues, and proposed a representation method of feature words vectorization based on Word2Vec, avoiding the neglect of contextual semantic connections. Hu et al. [13] studied the model's generalization and applied CNN and RNN to Chinese text classification. Zhao et al. [14] proposed a new neural network called AD-CharCGNN, the main idea of this essay can be outlined into following steps. The model extracts a part of financial articles, then classifies them by dividing and combining temporal and spatial domains. The part to be classified is mapped to the higher dimensions, after that, convolution in the spatial domain is used to calculate the local features. After processing, the features containing temporal information are captured, and finally, the spatial and temporal information is classified by Softmax.

Summing up all the above, the neural network model has been paid close attention by scholars. However, Chinese texts are difficult to classify due to a lack of related corpora. Therefore, this paper researches a specific Chinese news public dataset THUCNews. It aims to use a variety of machine learning and deep learning models to solve related problems and screen out the optimal model.

3. News classification models construction

3.1. Introduction of key theories

In this section, we will introduce the basic theory related to model building, including basic network layers in convolutional neural networks and common machine learning classification methods.

3.1.1 Convolutional layer

Convolutional layer consists of convolutional units with parameters optimized by the backpropagation algorithm.Convolutional layers alternate for convolution and pooling to extract the feature tensor from the input samples. Generally, the convolution kernel and the input images will be convoluted firstly as follows:

$$g(i) = \sum_{x=1}^{m} \sum_{y=1}^{n} \sum_{z=1}^{p} u_{x,y,z} \times \omega_{x,y,z}^{i} + v^{i}, i = 1,2,3\ldots \quad (1)$$

Here, i represents the convolution kernel number, while u and$V$ represents the input and bias. x, Y, and z represents the target dimension, and $g(i)$ is the resulting feature tensor. The calculation processing can be visualized as Figure 1. As shown in Table 1, then the activation function is used to achieve nonlinear transformations, which can increase the models nonlinearity. ReLu is a most widely used activation function in deep learning network due to its more efficient advantages of gradient descent and backpropagation processing. The specific mathematical expressions are as follows:

$$y(i) = f\big(g(i)\big) = \max\{g(i), 0\}, i = 1,2,3\ldots \quad (2)$$

In conclusion, the convolution operation has the following advantages: (1) sparse interactions. Different from the full connection operation, each input unit interacts with all the output units.In the

convolution operation, each input unit only interacts with some input units, and the output unit is the same, which makes the weights and bias parameters used less, and also makes the phenomenon of overfitting improved. (2) Weight sharing. In the full connection, each operation has its own parameters, and each parameter is only used for once, while the convolution operation uses the same parameters for multiple operations, avoiding the difficulty of network training due to the excessive number of parameters. (3) equivariant representations. In the convolution calculation, the rotation and translation operations do not deform.

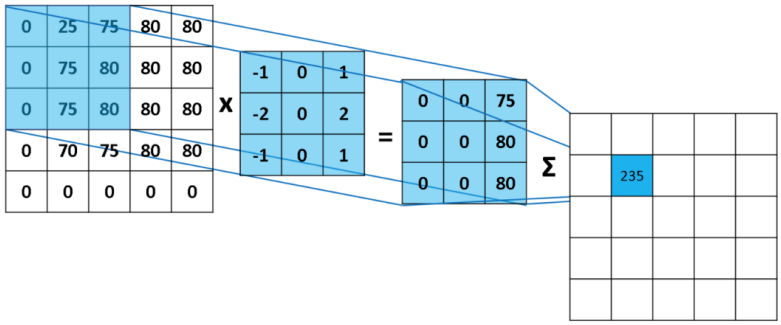

Figure 1. The process of convolution operations.

Table 1. Commonly used activation functions.

function	Sigmoid	Tanh	ReLU
expression	$\frac{1}{1+e^{-x}}$	$\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$	Max(0,x)

3.1.2 The pooling layer

The pooling layer, also known as the sampling layer, scales and maps the data on the upper layer, thus reducing the network parameters, reducing the dimension of the feature map, and avoiding overfitting.Pooling processing generally includes maximum pooling and mean pooling, which output the maximum and average pixels in the region, respectively:

$$X_{max}^{l(i,j)} = max\{u^{l(i,t)}\}, i = 1,2,3.. \quad (3)$$

$$X_{max}^{l(i,j)} = average\{u^{l(i,t)}\}, i = 1,2,3.. \quad (4)$$

Where l represents the number of layers, i is the feature tensor size and j is the number of pooling layers. $u^{l(i,t)}$Represents the neural unit number of the current tensor andω represents the width of the convolutional kernel.

3.1.3 Full connection layer

The fully connected layer is similar to the implicit layer of a neural network, and all the neurons and the upper layer neurons in this layer are fully connected, integrating the extracted feature information. The output layer usually selects the Softmax classifier to classify the output of the results provided by the fully

3.1.4 Random Forest

Random forest integrates multiple decision trees to obtain more accurate results. Within the classification task, the first part is the training process. The random forest model will use random sampling to extract a subset of the original dataset as a new dataset to build a decision tree model as a weak learner. Repeating this process, the model will get a new decision tree after each training. Then comes the classification process, where the final result is voted by the individual decision trees generated during the training process. This voting cannot only improve the accuracy but also

effectively control over-fitting. The random forest model is easy to implement and has also proven successful in many fields.

3.1.5 Support Vector Machine

A support vector machine is a generalized linear classier that performs multi-classification of data based on supervised learning and divides samples of different categories by dividing the hyperplane. In the actual process of dividing the hyperplane, finding a suitable hyperplane in the original space may not be possible, and the samples need to be mapped to higher dimensions using the kernel function. SVM is a black box, and selecting the best kernel function can significantly improve the accuracy and efficiency of the model.

3.2. Data preprocessing

3.2.1 Data source

The dataset we used in our experiments is the THUCNews dataset. This dataset collects the historical data of the Sina News RSS subscription channel from 2005 to 2011.For the convenience of our experiment, we selected a subset of the entire dataset, in which the training set contains 50,000 samples, the validation set includes 5,000 samples, and the test set includes 10,000 samples. Only ten fields are included in our dataset. And each division takes up the same percentage of the whole dataset.

3.2.2 Word cloud

Before model building, we first remove stop words and punctuation, which have no meaning for classification results. By removing them, the data is shortened and the classification process becomes more efficient. In addition, since there is no natural separator between each word in Chinese, we also perform word segmentation on the data. Chinese word segmentation technology is mature, and there are many open source word segmentation tools, such as SnowNLP, Jieba word segmentation, THULAC, etc. In this experiment, Jieba segmentation is used as a text segmentation tool to segment each article from a different category. When we go through the segmentation step, the word cloud library is used as a data visualization tool. Wordcloud can help organize the data obtained after segmentation into a form that is more logical to the human brain, allowing people to more intuitively see the occurrence and importance of specific phrases in different categories. As shown in Figure 2, we show the visual word cloud after the training set is segmented under the sports, games, and finance categories. The size of different phrases determines their frequency in different categories.

(a) Sports filed. (b) Game field. (c) Finance and Economics field

Figures 2. Word cloud in the different fields.

3.2.3 Keywords extraction and vectorization

A typical natural language processing task is keyword extraction. In this paper, two unsupervised methods are adopted for keyword display: TF-IDF keyword extraction and Textrank keyword extraction. The TF-IDF method is based on discrete weighting, and the Textrank method considers the network association between words. We first perform word segmentation for Chinese text and then score the results.

In this paper, the primary 15 keywords in each field are extracted. According to the result of keywords extraction, although the result extracted by the Textrank method contains noise, this method still precisely reflects the high-frequency words and keywords in various news fields. It is noticeable that the keywords trained by the TF-IDF method vividly reflect the characteristics of the news field. Taking the results of furniture field for instance: design, space, furniture, home, enterprise, China, style, decoration, decoration, market, color, products, industry, industry, brand, life, floor, collocation, company and living room. As long as the relevant information appears in the text, it can quickly distinguish the category to which the news belongs

TF-IDF stands for Term Frequency - Inverse Data Frequency. There are two parts to the algorithm, and the first part is to get the term frequency of each word. It calculates one word's appearance in one article divided by every word. Hence, every word in the article should have its Term Frequency. Then we try to get inverse data frequency by dividing the total number of articles by the number of articles that contain that specific word and then find the log of that result. We used the Tfidf-Vectorizer in the Python Sklearn library in our experiments, then set the max-features to 20,000 to limit the existence of outliers due to a large amount of text. Next, we apply the fit_transform method to the training set to perform TF-IDF-normalization. We use the IDF vector obtained from the training set to transform the test set to complete the vectorization for our experiment. Finally, we used the result to input our Random Forest, XGBoost, and GBDT models.

Tokenizer technology is used to process fake news and real news to digitize the text, which is suitable as the input of neural network models to detect fake news classification problems. This method is based on the bag-of-words model's methods, and the steps are as follows. First, segment the news and remove stop words. This step will remove all irrelevant characters other than Chinese, numbers, and English, such as function words, adverbs, and auxiliary words. Then corpus text with only simple words and phrases is left. Second, use fit_on_texts to perform word frequency statistics on the segmented corpus text, generate a dictionary, and then set the num_words parameter. The Tokenizer function will retain the words with the highest frequency (num_words -1) in the dictionary. This step requires ensuring that meaningless words have been removed while removing the stop words and also ensuring that no keywords are lost while removing the low-frequency words. Finally, through text_to_sequences, each word in the processed corpus text is converted into a numerical value, and a sequence is formed, thereby completing the conversion. This step is the key to digitizing text, and the generated sequence can be directly used as the input of the neural network model after further constraints.

3.3. Model Construction

We used the Tokenizer method to obtain a list of converted sequences during preprocessing. Hence, we can use pad_sequences pads the sequence into a sequence input with a length of 10, which ensures that the length of the input sequence is consistent. The length of each element in input matrix is 600. In processing the news data set, this paper is based on five neural network models for fitting, and compares the accuracy of various models. Some of the parameters of LSTM, GRU and MLP model are same. Due to the large dimension of each layer and conflict structure of these networks, the amount of iteration rounds of these models is only set to 2 and the batch size is 64. The structure of each neural network model is as follows:

(1) LSTM model. This model consists of an Embedding layer, LSTM network unit, fully connected layer, leveling layer, and fully connected layer. The relevant parameters of the models are set as follows: the optimizer is RMSProp, which helps to speed up model convergence and parameter

optimization, the loss function is sparse_categorical_corssentropy, also named as the multi-class cross-entropy which is often used as a regular indicator of multi-classification task. The dropout retention ratio trick is used between the LSTM unit and the following fully connected layer, which stop fifty percent of the output of LSTM unit at random to the next dense in order to prevent overfitting And the accuracy of LSTM model proves that this trick can improve the model.

(2) GRU model. This model consists of an Embedding layer, GRU network unit, fully connected layer, leveling layer, and fully connected layer. The relevant parameters of the models are set as follows: the optimizer is RMSProp, which helps to speed up model convergence and parameter optimization, the loss function is sparse_categorical_corssentropy, also named as the multi-class cross-entropy which is often used as a regular indicator of multi-classification task. The dropout retention ratio trick is used between the GRU unit and the following fully connected layer, which stop fifty percent of the output of GRU unit at random to the next dense in order to prevent overfitting. And the accuracy of GRU model proves that this trick can improve the model.

(3) MLP model. This model consists of an Embedding layer, fully connected layer, fully connected layer, leveling layer, and fully connected layer. The relevant parameters of the models are set as follows: the optimizer is RMSProp, which helps to speed up model convergence and parameter optimization, the loss function is sparse_categorical_corssentropy, also named as the multi-class cross-entropy which is often used as a regular indicator of multi-classification task. The dropout retention ratio trick is used between the GRU unit and the following fully connected layer, which stop fifty percent of the output of GRU unit at random to the next dense in order to prevent overfitting And the accuracy of MLP model proves that this trick can improve the model.

(4) CNN model. This model consists of an Embedding layer, a one-dimensional convolution layer, a global pooling layer, a fully connected layer, dropout, relu, a fully connected layer, and softmax.

The relevant parameters of the models are set as follows: the optimizer is Adam, and the loss function is sparse_categorical_corssentropy. The batch size is 16. The dropout retention ratio is 0.5. The number of iteration rounds is set to 10 rounds.

(5) RNN model. The model consists of an Embedding layer, two layers of GRU with dropout, a fully connected layer, dropout, relu, a fully connected layer, and softmax. The optimizer is the Adam optimizer, which helps to speed up model convergence and parameter optimization, the loss function is sparse_categorical_crossentropy,also named as the multi-class cross-entropy which is often used as a regular indicator of multi-classification task, the dropout retention ratio is 0.8, the batch size is 16, and iteration round is set to 10.

Random Forest, XGBoost, and GBDT are all ensemble learning models, which are based on tree models. We also adjust the tree by adjusting the depth and the size of each tree. Moreover, for GBDT and XGBoost models, the L1 norm value and the L2 norm value can be adjusted. In addition, we also use the support vector machine model where the radial basis function is selected as the kernel function.

4. Experiment

In this section, we quantitatively compare the recognition accuracy of different methods in detail. In-depth analysis of the experimental results in Table 2, we can find that:

(1) The experiment result indicates that GRU has the best accuracy among all neural network models, and SVM is right after. Two models are more accurate than CNN by about 2%. The other three neural network models have similar results, all of which are about 94%, but the RNN model has only 93.05% accuracy, slightly lower than other neural network models.

(2) Keyword extraction and word cloud can effectively filter out words representative of each field. Without machine learning models, a human can use these keywords to determine which field each article is categorized. However, some keyword extraction methods have limitations. For example, segmentation methods will strongly affect the result of keyword extraction methods. Hence, multiple keyword extraction methods' results can be supplemental for better results.

(3) We used various machine learning models for Chinese news classification. This paper divides nine machine learning models into two parts, the neural network models and other machine learning models. The two categories use different methods of vectorization to reach the best output. The neural network model uses the word-embedding method, and other machine learning models use the TF-IDF method. According to the results, it can be found that among neural network models, the effect of the GRU model is the best compared to the other four models, while among other machine learning models, the effect of the SVM is significantly better than others. In addition, the effect of the CNN model and Random Forest model is also relatively good. In conclusion, these four models on the THUCNews dataset are better than the rest.

(4) When analyzing some models in detail, according to the empiricism, the XGBoost model is an optimization of the GBDT model, the result of exploration on THUCNews dataset proves that XGBoost model has higher accuracy, and its accuracy is also better than that of the GBDT model; and due to the data set itself, the performance of the multi-layer perceptron model is also more brilliant. Since the mechanism of the MLP model is relatively simple, its performance on another dataset may not reach the performance of this particular dataset.

Table 2. Comparison of results of different machine learning models.

	Model	Accuracy
neural network	CNN	94.72%
	RNN	93.05%
	LSTM	94.23%
	GRU	96.93%
	MLP	94.01%
Integrated learning	XGBoost	92.96%
	GBDT	90.64%
	Random Forest	92.97%
Support vector machine	SVM	96.39%

5. Conclusion

In conclusion, this paper introduces the current research progress within the field of news text classification and quantitatively compare the recognition accuracy of different methods in detail. Specifically, we introduce the representative works of news text classification in detail from both machine learning and also deep learning frameworks, including their design ideas, advantages and disadvantages. Second, we analyze the classification results of nine classical machine learning algorithms in the THUCNews dataset in detail.

References

[1] Ogura H , Amano H , Kondo M . Comparison of metrics for feature selection in imbalanced text classification [J]. Expert Systems with Applications, 2011, 38(5):4978-4989.

[2] Chen K , Zhang Z , Long J , et al. Turning from TF-IDF to TF-IGM for term weighting in text classification[J]. Expert Systems with Applications an International Journal, 2016, 66(Dec.):245-260.

[3] Chen S . K-Nearest Neighbor Algorithm Optimization in Text Categorization[J]. IOP Conference Series Earth and Environmental Science, 2018, 108(5):052074.

[4] Selvi S T , Karthikeyan P , Vincent A , et al. Text categorization using Rocchio algorithm and random forest algorithm[C]// 2016 Eighth International Conference on Advanced Computing (ICoAC). IEEE, 2017.

[5] Seyyedi, Seyyed, Hossein, et al. Enhancing Effectiveness of Dimension Reduction in Text Classification[J]. International Journal of Artificial Intelligence Tools: Architectures, Languages, Algorithms, 2017.

[6] Bidi N , Elberrichi Z . Feature selection for text classification using genetic algorithms[C]// International Conference on Modelling. IEEE, 2017.

[7] Yan Z , Yun Q , Li C . Improved KNN text classification algorithm with MapReduce implementation[C]// 2017 4th International Conference on Systems and Informatics (ICSAI). IEEE, 2017.

[8] Li C , Zhan G , Li Z . News Text Classification Based on Improved Bi-LSTM-CNN[C]// 2018 9th International Conference on Information Technology in Medicine and Education (ITME). IEEE Computer Society, 2018.

[9] Lai S , Xu L , Liu K , et al. Recurrent Convolutional Neural Networks for Text Classification[C]// National Conference on Artificial Intelligence. AAAI Press, 2015.

[10] Liu P , Qiu X , Huang X . Recurrent Neural Network for Text Classification with Multi-Task Learning[C]// International Joint Conference on Artificial Intelligence. AAAI Press, 2016.

[11] Shi M , Wang K , Li C . A C-LSTM with Word Embedding Model for News Text Classification[C]// 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS). IEEE, 2019.

[12] Liu C Z , Sheng Y X , Wei Z Q , et al. Research of Text Classification Based on Improved TF-IDF Algorithm[C]// IEEE International Conference of Intelligent Robotic and Control Engineering. 0.

[13] Li H , Zou P , Han W H . Chinese Text Classification Based on Neural Network[C]// International Conference on Advances in Neural Networks. Springer-Verlag, 2013.

[14] Zhao W , Zhang G , Yuan G , et al. The Study on the Text Classification for Financial News Based on Partial Information[J]. IEEE Access, 2020, PP(99):1-1.