Chinese News Text Classification Method Based on Deep Neural Networks (Postprint)

Authors: Zheng Chuangwei, Wang Yong, Xing Gutao, Xie Zhicheng, Chen Yifei

Date: 2023-10-08T00:00:00+00:00

Abstract

[Objective] This article compares multiple Chinese news text classification models based on deep neural networks, aiming to identify methods with higher accuracy for practical application and to provide more efficient approaches for Chinese news text classification. [Methods] The study reviews and summarizes text classification techniques and Chinese news classification, elaborates on the characteristics and preprocessing of Chinese news texts, and provides detailed introductions to the FastText algorithm, Bert classification algorithm, TextCNN algorithm, and TextRNN algorithm. [Results] The four deep neural network algorithms can all be applied to Chinese news text classification, effectively handling information disorder issues and enabling rapid and accurate classification. [Conclusion] Through experiments and comparative analysis of the four deep neural network algorithms, the FastText model is found to deliver the most outstanding text classification performance in practical applications.

Full Text

Preamble

ChinaXiv Collaborative Journal
Chinese News Text Classification Method Based on Deep Neural Networks

Zheng Chuangwei, Wang Yong, Xing Gutao, Xie Zhicheng, Chen Yifei
(Shenzhen Creative Wisdom Port Technology Co., Ltd., Shenzhen, Guangdong 518034)

Abstract

[Objective] This paper compares multiple Chinese news text classification models based on deep neural networks, aiming to identify highly accurate methods

for practical application and provide more efficient approaches for Chinese news text classification. [**Methods**] We systematically review and summarize text classification techniques and Chinese news categorization, elaborate on the characteristics and preprocessing of Chinese news texts, and provide detailed introductions to the FastText algorithm, BERT classification algorithm, TextCNN algorithm, and TextRNN algorithm. [**Results**] All four deep neural network algorithms can be applied to Chinese news text classification, effectively handling information disorder problems and enabling rapid and accurate categorization. [**Conclusion**] Through experimental testing and comparative analysis of the four deep neural network algorithms, we find that the FastText model demonstrates the most outstanding performance for text classification in practical work.

Keywords: deep neural networks; text classification; Chinese news; natural language processing
CLC Number: TP183
Document Code: A
Article ID: 1671-0134(2023)03-147-05
DOI: 10.19483/j.cnki.11-4653/n.2023.03.033
Citation Format: Zheng Chuangwei, Wang Yong, Xing Gutao, Xie Zhicheng, Chen Yifei. Chinese News Text Classification Method Based on Deep Neural Networks[J]. China Media Technology, 2023(03): 147-151.

Introduction

With the rapid development of the information age, network information has experienced explosive growth. Mainstream news websites such as Sina and Toutiao provide millions of news articles daily, presenting enormous challenges for these platforms. News text classification can effectively categorize texts rapidly and accurately, improving website operational efficiency and becoming a research hotspot in recent years. News text classification represents a subtask of general text classification, which finds widespread application across various domains including webpage classification, microblog sentiment analysis, and user comment mining, making it one of the most widely used technologies in natural language processing. The most important function of text classification is its ability to effectively handle information disorder, particularly for massive datasets, as it helps users quickly, efficiently, and accurately locate required information, thereby enabling more effective data analysis [1]. This paper explores and elaborates on news text classification technology, focusing on classification characteristics and identifying the strengths and weaknesses of various algorithms through experiments, while forecasting future trends in news classification development.

2. Chinese News Text Classification Methods

2.1 Overview of Chinese News Classification

Chinese text is unstructured data that cannot be directly processed by computers and must be converted into structured data. This transformation process requires data preprocessing followed by feature extraction methods to enable computational processing [2]. Feature extraction can be summarized into three categories: (1) bag-of-words model, (2) feature weight calculation, and (3) vector space model. The bag-of-words model ignores word order and grammar, treating text merely as a collection of words. If the vocabulary contains N words, each text is represented as an N-dimensional vector where elements are 0/1, indicating whether the text contains the corresponding word. Feature weight calculation generally includes Boolean weights, TF-IDF weights, and entropy-based weights. The vector space model builds upon the bag-of-words model, reducing dimensionality through feature selection and performing secondary calculations using feature weights [3]. Through these methods, unstructured text can be transformed into structured arrays for text classification.

Traditional machine learning methods can be summarized as feature engineering + shallow classification models. In machine learning-based classification approaches, datasets are divided into training and test sets according to certain proportions. The classification model parameters are continuously trained and adjusted to achieve higher accuracy, and the model' s classification effectiveness is then evaluated using the test set [4]. During classification, similar corpora can be used to expand the extracted text information to obtain feature vectors, or support vector machines and information gain calculations can be employed for feature selection to improve classification accuracy. Additionally, weighting word vectors can more precisely distinguish the importance of different terms, enhancing classification accuracy and efficiency. Since different tasks require different features, specific problems demand specific analysis, with the primary technology involving classifier construction based on statistical classification methods, including SVM and Naive Bayes classification algorithms [5].

Deep learning-based text classification methods utilize network structures such as CNN/RNN to automatically obtain feature representations and then perform classification, thereby solving problems end-to-end. In deep learning-based classification methods, continuously improving computer performance has enabled rapid development in fields like image recognition and natural language processing. These algorithms simulate neural connections and computations in the human brain, typically comprising input layers, hidden layers, and output layers. Layers are trained and computed through backpropagation algorithms to obtain corresponding training models. Deep learning approaches often involve multiple hidden layers, with each layer responsible for learning different features. These features are ultimately aggregated to accomplish more precise learning tasks [6]. During text classification, structured text feature information can be extracted from perspectives such as user characteristics, text topics, and

comment keywords to achieve better classification results.

From the perspective of text classification, Chinese news has two key characteristics: (1) News requires text classification. With the exponential growth of data in the information age, news volume has also increased exponentially, making it a critical challenge to obtain needed news from these massive datasets. (2) News classification is feasible. Due to the public nature of news data, the internet contains abundant training and testing data. Meanwhile, with the rapid development of classification algorithms, classification performance continues to improve.

2.2 Chinese News Text Preprocessing

Chinese news text preprocessing primarily identifies and removes words without practical meaning, such as numerous stop words or noise, thereby reducing their impact on subsequent processing [7]. The preprocessing procedure mainly includes word segmentation, noise reduction, part-of-speech tagging, and stop word removal.

2.2.1 Word Segmentation Chinese word segmentation lacks the natural spacing features present in English, requiring additional processing such as forward/backward maximum matching algorithms based on dictionaries or statistical methods. Chinese segmentation addresses the challenge of lacking formal delimiters in Chinese text, employing three main techniques: First, string matching technology based on dictionaries, which requires establishing a unified dictionary. When segmenting sentences, they are first split and then matched against the dictionary. Second, understanding-based segmentation methods, where computers simulate human sentence comprehension and expression through neural network algorithms to recognize Chinese words. However, due to the broad semantics of Chinese terms, this approach presents significant difficulty. Third, statistical segmentation technology, which treats segmentation as a probability maximization problem based on constructed corpora. It calculates the probability of adjacent characters forming words and performs segmentation according to these probability values.

2.2.2 Noise Reduction Noise reduction for Chinese news information primarily involves removing 杂乱 text and images from webpages, retaining only well-formatted main content. For short texts, it is also necessary to remove emoticons, forwarding relationships, etc., preserving only pure text for subsequent analysis and processing. During noise reduction, feature extraction or dimensionality reduction may be involved, which can effectively reduce computational overhead, remove noise, and improve model training speed.

2.2.3 Part-of-Speech Tagging After noise reduction, part-of-speech tagging is required for words in Chinese news, including nouns, verbs, adjectives, adverbs, etc. POS tagging primarily enhances efficiency in subsequent text recog-

nition and classification processes, significantly improving processing speed after annotation.

2.2.4 Stop Word/Meaningless Word Filtering The first method uses a predefined stop word list, typically containing modal particles and punctuation marks. After word segmentation and noise reduction, the news information is traversed, and words matching the stop word list are removed. This method offers good controllability and high efficiency, allowing 随时 modification of the stop word list. The second method calculates word frequency in the corpus and removes words with low frequency or occurrence counts. However, this approach involves substantial computation, consumes more resources, and may mistakenly delete low-frequency but influential words.

2.3 Main Model Methods for Chinese News Text Classification

Text classification is a process of categorizing texts based on their semantic content. The relationship between the text data collection and category collection can be expressed by the following function (Equation 3-1):

$$\text{Category}(d_i) = \begin{cases} c_j & \text{if } d_i \in c_j \\ \text{null} & \text{if } d_i \notin c_j \end{cases} \quad \text{(Equation 3-1)}$$

Based on the aforementioned news characteristics, applying text classification to the news domain holds significant practical importance. News text classification has three main characteristics [8]: (1) Text analysis must consider the importance of titles: news titles provide a high-level summary of an article and greatly assist in news classification. (2) Text representation must consider news features: thoroughly analyzing the characteristics of news texts and optimizing text representation methods helps improve online news classification effectiveness. (3) Classification criteria tend to focus on themes rather than disciplines. Therefore, this study addresses news data encountered in practical work and employs deep learning classification algorithms using FastText, TextCNN, BERT, and TextRNN models for computation and training. During training, careful attention must be paid to dataset classification, and preset judgment conditions should be as scientific as possible. For example, consider using gradient descent backpropagation algorithms to update weights, thereby gradually improving accuracy and achieving better training results.

2.3.1 FastText Model The FastText model mainly consists of an input layer, hidden layer, and output layer (as shown in [Figure 1: see original paper]). Compared to large neural network structures, it is relatively simple and highly efficient, improving training speed while ensuring classification accuracy [9]. In the input layer, text is treated as a collection of words, generating vectors that represent the text. The key operation involves performing additive averaging on words appearing in the text, and the resulting vector is used to complete

multi-classification tasks. This algorithm' s advantages also include the ability to spontaneously train word vectors without pre-training steps, using word sequences as input, employing hierarchical softmax functions to accelerate classification and predict probability distributions for these categories. This method of establishing hierarchy in Huffman coding tree form significantly reduces computational complexity.

2.3.2 TextCNN Model Selecting appropriate Chinese text classification algorithms is central to Chinese text classification, requiring a thorough understanding of each algorithm and clear cognition of news text classification tasks. Using TextCNN for text processing and classification necessitates data preprocessing operations, including vectorization and word vector initialization, to achieve better analytical results later. In text classification, the TextCNN model is most widely applied, particularly mature in industrial applications, achieving excellent output results. Its network structure is relatively simple, allowing the model to be trained with fewer parameters, effectively saving computational costs and improving training speed. CNN is primarily used in image classification, while TextCNN is a variant applicable to text classification. The structure diagram is shown in [Figure 2: see original paper], where word vectors undergo convolution operations with different kernels to obtain corresponding feature vectors, which then pass through pooling layers to reach fully connected layers. At this point, mapping operations can transform high-dimensional data into low-dimensional data [10].

As seen in [Figure 3: see original paper] TextCNN algorithm flowchart, after inputting text information, data preprocessing begins using word embedding, word vector initialization, and vector dimension transformation. After preprocessing, TextCNN training is performed, outputting classification results through convolution, max pooling, and Softmax. Finally, the output loss value is evaluated. If it exceeds the set threshold, gradient descent backpropagation is used for iterative updates until the loss is less than or equal to the threshold, at which point training ends. The commonly used gradient descent method is batch gradient descent, which requires gradient updates in each iteration. The advantage of gradient descent is its use of matrix operations for all sample data, enabling parallel processing. The disadvantage is that when data volume is large, computing all data in each iteration reduces training efficiency.

TextCNN has weak interpretability and requires manual guidance and intervention, including setting convolution kernel sizes and manual tuning of the model.

2.3.3 BERT Model BERT was originally a language model invented by the Google team, consisting of multiple stacked Transformer Encoders, with the model structure shown in [Figure 4: see original paper]. The Transformer structure employs an attention mechanism that reads text sequences in a single pass, improving reading efficiency and facilitating semantic learning based on word context. This enhances understanding of contextual semantics and aligns

more closely with Chinese language expression. For news text classification, this approach can solve difficulties such as data sparsity and high contextual dependency, making text classification more efficient and meeting demands for greater precision.

The model' s input layer primarily uses the BERT algorithm for pre-training to represent text as semantic vectors. Markers are required at sentence beginnings and ends, and the read data is processed using mapping index methods to segment text and labels. Each word embedding is then converted into a one-dimensional semantic vector. Through stacked Transformer Encoders, bidirectional semantic feature learning and vector representation are completed. In the feature extraction layer, the BERT model is further fine-tuned, combining attention mechanisms to extract text features. This mechanism focuses more on internal data correlations, improving model computational efficiency through word vector weighting [11]. The BERT algorithm model is a deep network formed by stacking multiple Transformer Encoder layers. This approach reads the entire text sequence at once, enabling contextual semantic learning for words, enhancing understanding of contextual semantics, and approaching human language more closely. The model also performs text feature extraction, as shown in [Figure 5: see original paper], with characteristics such as global temporal optimality that can extract contextual semantic information from text. The implementation process requires using TensorFlow library functions to build bidirectional network operations. In the output layer, the model primarily performs probability prediction for each sample' s labels, enabling efficient extraction of text information. Fully connected layers are then used to improve word segmentation accuracy. This fully connected approach uses activation functions and data linear transformations to improve computational efficiency, employs gradient descent algorithms for parameter learning, and adopts Dropout strategies to prevent model overfitting.

2.3.4 TextRNN Model This recurrent neural network model, also known as TextRNN, can capture longer sequence information when applied to Chinese news text classification. It avoids the limitation of CNN algorithms that cannot extend sequence length and allows simpler parameter adjustment for more accurate expression of contextual information. In RNN algorithms, output results are not merely obtained through matrix and convolution computations. The model calculates a State that continuously influences subsequent computations. After N sample outputs, the results acquire certain sequential characteristics. This allows input data states to be cyclically processed within the neural network itself, generating temporal associations. The special feature of the TextRNN model is that nodes in the same hidden layer are connected, with temporal relationships treated as variables affecting inter-data relationships. The network considers not only current inputs but also 赋予 past memories. In its hidden layers, data output from the first hidden layer may, after adding certain weights, enter the second hidden layer. When inputting to the next layer, the hidden state neurons at a certain moment and the text features at that moment

are input together. After continuous cycling and recursion, connection weights of each layer are adjusted in reverse to obtain optimal parameters. However, precisely because of this structure, TextRNN' s output at a later moment depends on the output of the previous moment, making parallel processing impossible and reducing training efficiency [12].

As shown in [Figure 6: see original paper] TextRNN network structure, after data is expanded in time sequence, a T-dimensional vector can be obtained, where U represents weights from the input layer to the hidden layer (larger weights indicate more input information). The horizontal W represents weights from the previous hidden layer to the next hidden layer, and V represents weights from the hidden layer to the output layer. It is important to note that when RNN processes sequence information, it sometimes biases toward the most recently input information, potentially causing loss of earlier information. Therefore, during weight initialization, extremely large or small values should be avoided, and LSTM (Long Short-Term Memory networks) and GRU (Gated Recurrent Units) should be incorporated.

2.4 Experiments

2.4.1 Dataset Introduction We provide a news and company-related dataset generated by screening and filtering financial data from a certain website, containing 400,000 news articles, all preprocessed into UTF-8 plain text. Based on the original website, the dataset is divided into 1,000 categories, with each category representing a company. Mainstream classification algorithms will be used to test model performance.

Experiments require evaluation of classification result accuracy on the test dataset. If results are not within a reasonable range, the process must return to the feature selection stage to repeat feature extraction until results fall within a reasonable range. Evaluation criteria mainly include accuracy and recall rates. Accuracy represents the precision of the text classification model, but high accuracy with low recall indicates failure to predict labels that should have been predicted, especially for imbalanced samples where minority classes may be predicted as majority classes. Alternatively, in some multi-label classification models, overfitting of features and models may occur, also leading to low recall, which requires careful attention during experiments.

We tested FastText, BERT, TextCNN, and TextCNN algorithms on the dataset separately, evaluating accuracy and recall rates. Experimental results are shown in the following table:

Model	Accuracy (%)	Recall (%)
FastText		
TextCNN		
TextRNN		

Additionally, this study tested the accuracy and recall rates of several methods on THUCNews, with experimental results shown in the following table:

Model	Accuracy (%)	Recall (%)
FastText		
TextCNN		
TextRNN		

3. Conclusion

Based on the review and research of Chinese text classification, we believe the following directions will become research hotspots: (1) **Unsupervised learning-based news text classification**: The internet contains vast amounts of unsupervised data, and how to effectively utilize this data will become a popular research topic. (2) **Multi-level news text classification**: Fully utilizing hierarchical information of classification systems and adopting layer-by-layer classification thinking for multi-level text classification can effectively reduce algorithm complexity while ensuring classification accuracy, warranting further research. (3) **Cross-modal news text classification**: News text classification primarily considers textual information while ignoring other modal information in news. How to utilize this information to assist classification and effectively fuse textual and image information represents another research hotspot.

This study discussed news text classification and related research, introducing the FastText model, TextCNN model, BERT model, and TextRNN model. Through experiments, the FastText model demonstrated the most outstanding performance for text classification in practical work, while the TextCNN model performed best on THUCNews.

References

[1] Li Zekui, Sun Fei, Chen Jun. Research and Exploration on the Intelligent and Knowledge-Based Path of Chinese Semantic Analysis Technology in the News Media Field[J]. China Media Technology, 2018(8): 35-37.
[2] Li Z, Shang W, Yan M. News text classification model based on topic model[C]// IEEE/ACIS International Conference on Computer & Information Science. IEEE, 2018.
[3] Li Keyue, Chen Yi, Niu Shaozhang. Social E-commerce Text Classification Algorithm Based on BERT[J]. Computer Science, 2021(2): 87-92.
[4] Jia Pengtao, Sun Wei. A Survey of Text Classification Based on Deep Learning[J]. Computer and Modernization, 2021(7): 29-37.
[5] Tan Xin. Practical Exploration of Big Data Analysis Applications for Policy Interpretation[J]. China Media Technology, 2019(3): 22-23.
[6] Liu Meng. Research on the Application of Artificial Intelligence Technology in Media Convergence[J]. China Media Technology, 2021(11): 154-156.

[7] Li Zekui, Sun Fei, Chen Jun. Research and Exploration on the Intelligent and Knowledge-Based Path of Chinese Semantic Analysis Technology in the News Media Field[J]. China Media Technology, 2018(8): 35-37.
[8] Jia Hongyu, Wang Yuhan, Cong Riqing, Lin Yan. Research on Neural Network Text Classification Algorithm Combining Self-Attention Mechanism[J]. Computer Applications and Software, 2020(8): 35-37.
[9] Yang Rui, Chen Wei, He Tao, Zhang Min, Li Ruiling, Yue Fang. Research on Convolutional Neural Network Text Classification Method Fusing Topic Information[J]. Modern Intelligence, 2020(4): 42-49.
[10] Du Sijia, Yu Haining, Zhang Hongli. Research Progress on Text Classification Based on Deep Learning[J]. Chinese Journal of Network and Information Security, 2020(4): 1-13.
[11] Hao Chao, Qiu Hangping, Sun Yi, Zhang Chaoran. Research Progress on Multi-label Text Classification[J]. Computer Engineering and Applications, 2021(10): 48-56.
[12] Wang Mili. Research on Text Classification Based on Machine Learning[J]. Technology Innovation and Application, 2021(26): 70-72.

Author Information:
Zheng Chuangwei (1978-), male, from Shantou, Guangdong, Senior Engineer, research direction: big data, artificial intelligence;
Wang Yong (1977-), female, from Shaoyang, Hunan, Intermediate Engineer, research direction: big data;
Xing Gutao (1984-), male, from Wenchang, Hainan, Intermediate Engineer, research direction: cloud computing;
Xie Zhicheng (1980-), male, from Shantou, Guangdong, Intermediate Engineer, research direction: big data, cloud computing;
Chen Yifei (1981-), from Zhanjiang, Guangdong, Intermediate Engineer, research direction: big data.

(Responsible Editor: Zhang Xiaojing)

Note: Figure translations are in progress. See original paper for figures.