Academic Journal of Natural Science
Journal Home: ajns.suaspress.org | CODEN: AJNSAE | NAAN: 40704
Vol. 1, No. 1, 2024

Research on the Application of Deep Learning-Based Text Classification Algorithms in Data Mining

XING, Qiming [1*] LI, Yuan [2]

[1] Belarusian State University, Belarus
[2] Hefei Hengli Equipment Co., Ltd., China

**** XING, Qiming is the corresponding author, E-mail: xqm200104@gmail.com***

Abstract: This article explores the application of text classification algorithms based on deep learning in data mining. By reviewing the existing literature, the application of convolutional neural network (CNN), recurrent neural network (RNN) and transformer (Transformer) in text classification is introduced, and related research on data mining and analysis is analyzed. The experiment used the IMDB movie review data set to verify the effectiveness of the deep learning model through data preprocessing, feature extraction and model training. The experimental results show that the model's accuracy on the test set is 86%, the precision rate is 85%, the recall rate is 88%, and the F1 value is 86%. Research shows that deep learning models can significantly improve text classification performance. This article also discusses the significance, limitations and future research directions of the research findings, including improvements in data acquisition and annotation, computing resource optimization, and model interpretability.

Keywords: Deep Learning, Text Classification, Data Mining, Convolutional Neural Network (CNN), Natural Language Processing (NLP).

DOI: https://doi.org/10.5281/zenodo.13762126 **ARK:** https://n2t.net/ark:/40704/AJNS.v1n1a03

1 INTRODUCTION

The rapid development of information technology has led to exponential growth in the amount of text data. How to extract valuable information from massive texts has become a core issue in the field of data mining and analysis. Deep learning, as an advanced data processing technique, performs well in text classification tasks. This paper aims to explore the application of deep learning based text classification algorithms in data mining, analyze their theoretical basis, and propose improvement solutions.

The objective of this study is to explore the application of deep learning models in text classification through theoretical analysis and propose optimization solutions. The specific methods include: reviewing existing literature to understand the current application status of deep learning in text classification and data mining; Analyze the theoretical basis of deep learning models and explore their advantages and disadvantages in text classification tasks; Propose an optimized deep learning model and validate its effectiveness through theoretical analysis.

2 LITERATURE REVIEW

2.1 Application of Deep Learning in Text Classification

In recent years, deep learning has made significant progress in the field of text classification. Traditional text classification methods rely on manual feature extraction and shallow models, while deep learning significantly improves classification performance through automatic feature extraction and multi-layer neural network structures. Common deep learning models include:

Convolutional Neural Network (CNN): CNN extracts local features of text through convolutional layers and is suitable for short text classification tasks. Its advantage lies in the ability to capture local patterns in text and has high computational efficiency.

Recurrent Neural Network (RNN): RNN processes sequence data through a cyclic structure and is suitable for long text classification tasks. Its variants, Long Short Term Memory Network (LSTM) and Gated Recurrent Unit (GRU), can effectively solve long-distance dependency problems.

Transformer: Transformer captures global dependencies in text through self attention mechanism and is suitable for various text classification tasks. The

representative models BERT and GPT perform well in multiple natural language processing tasks.

2.2 Research on Data Mining and Analysis

Data mining and analysis is the process of extracting valuable information from large amounts of data. Text data mining involves multiple stages, including:

Text preprocessing: includes steps such as word segmentation, stop word removal, and stem extraction, aimed at converting the original text into a format suitable for analysis.

Feature extraction: Traditional methods include TF-IDF and bag of words models, while deep learning methods extract text features through word vectors (such as Word2Vec, GloVe) and contextual embeddings (such as BERT).

Pattern recognition: Identify patterns and structures in text through clustering, classification, and other methods. Deep learning models perform well in this aspect, being able to automatically learn complex patterns in text.

Information extraction: Extracting specific information from text, such as named entity recognition, relationship extraction, etc. Deep learning methods perform well in information extraction tasks, improving the accuracy and efficiency of extraction.

2.3 Analysis of the advantages and disadvantages of existing methods

Although deep learning has achieved significant results in text classification and data mining, there are still some challenges and shortcomings:

Data requirements: Deep learning models typically require a large amount of annotated data for training, resulting in high data acquisition costs. To address this issue, researchers have proposed semi supervised learning and transfer learning methods, which reduce dependence on labeled data by utilizing unlabeled data and pre trained models.

Computing resources: The training process of deep learning models is complex and consumes a large amount of computing resources. Model compression and acceleration techniques, such as model pruning, quantization, and knowledge distillation, can reduce computational resource consumption while ensuring model performance.

Interpretability: The interpretability of deep learning models is poor, making it difficult to understand their internal working mechanisms. To address this issue, researchers have proposed explainable artificial intelligence (XAI) methods that enhance the transparency and credibility of models by visualizing and interpreting their decision-making processes.

3 THEORETICAL BASIS

3.1 Basic Concepts and Principles of Deep Learning

Deep learning is a machine learning method based on artificial neural networks, which models and predicts complex data through the connection and weight adjustment of multiple layers of neurons. The core idea is to extract high-level features from raw data through layers of abstraction. Common deep learning models include feedforward neural networks (FNN), convolutional neural networks (CNN), recurrent neural networks (RNN), and transformers.

3.2 Theoretical Framework of Text Classification Algorithm

Text classification is the process of assigning text data to predefined categories, typically consisting of the following steps:

Text preprocessing: Convert the original text into a format suitable for analysis, including word segmentation, stop word removal, stem extraction, etc.

Feature extraction: Similar to feature extraction in data mining, in this step, text is typically converted into numerical features. Traditional methods include TF-IDF and bag of words models, while deep learning methods extract features through word vectors (such as Word2Vec, GloVe) and contextual embeddings (such as BERT).

Model training: Train features using classification algorithms, including Naive Bayes, Support Vector Machine (SVM), and deep learning models such as CNN RNN、Transformer）。

Model evaluation: Evaluate the effectiveness of the model through cross validation and performance metrics such as accuracy, precision, recall, and F1 score.

3.3 Basic Methods of Data Mining and Analysis

Data mining and analysis is the process of extracting valuable information from large amounts of data, involving multiple stages:

Data preprocessing: including steps such as data cleaning, data integration, data transformation, and data reduction, aimed at improving data quality and analysis efficiency.

Pattern recognition: Identify patterns and structures in data through clustering, classification, association rule mining, and other methods. Deep learning models perform well in this aspect, being able to automatically learn complex patterns in data.

Information extraction: Extracting specific information

from data, such as named entity recognition, relationship extraction, etc. Deep learning methods perform well in information extraction tasks, improving the accuracy and efficiency of extraction.

Visualization of Results: Displaying data analysis results through charts and visualization tools to help understand and interpret patterns and trends in the data.

4 METHODS AND MODELS

4.1 THEORETICAL ANALYSIS OF DATA PREPROCESSING AND FEATURE EXTRACTION

Data preprocessing is a crucial step in text classification, aimed at improving data quality and analysis efficiency. Mainly includes:

Word segmentation: Breaking down text into words or phrases, commonly used methods include rule-based segmentation and statistical segmentation.

Remove stop words: Remove common words that are irrelevant to classification, such as "de", "is", etc.

Stemming: Restoring words to their root form and reducing feature dimensions.

In terms of feature extraction, traditional methods include TF-IDF and bag of words models. The TF-IDF formula is:

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \text{IDF}(t)$$

The calculation formulas for word frequency (TF) and inverse document frequency (IDF) are:

$$\text{TF}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

$$\text{IDF}(t) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Deep learning methods use word vectors (such as Word2Vec, GloVe) and contextual embeddings (such as BERT) to capture semantic relationships between words.

4.2 DESIGN AND THEORETICAL IMPLEMENTATION OF DEEP LEARNING MODELS

Deep learning models perform well in text classification, with commonly used models including:

Convolutional Neural Network (CNN): Extracting local features of text through convolutional layers, suitable for short text classification. Its advantage lies in the ability to capture local patterns in text and has high computational efficiency. The mathematical expression for convolution operation is:

$$y_{i,j} = \sum_m \sum_n x_{i+m,j+n} \cdot w_{m,n}$$

Among them, x represents the input feature map, w represents the convolution kernel, and y represents the output feature map.

Recurrent Neural Network (RNN): Processing sequence data through a recurrent structure, suitable for long text classification. LSTM and GRU variants can effectively solve long-distance dependency problems. The state update formula for RNN is:

$$h_t = \sigma(W_h \cdot h_{t-1} + W_x \cdot x_t + b)$$

Among them, ht represents the current hidden state, h (t-1) represents the previous hidden state, xt represents the current input, Wh and Wx represent the weight matrix, b represents the bias term, and σ represents the activation function.

Transformer: Capturing global dependencies in text through self attention mechanism, suitable for various text classification tasks. The representative models BERT and GPT perform well in multiple natural language processing tasks. The calculation formula for self attention mechanism is:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Among them, Q, K, and V represent the query, key, and value matrices, respectively, and dk represents the dimension of the key.

4.3 THEORETICAL EXPLORATION OF MODEL TRAINING AND OPTIMIZATION METHODS

The model training adopts forward propagation and backward propagation algorithms, and achieves model optimization by adjusting weights and biases. Common optimization methods include:

Gradient descent: By calculating the gradient of the loss function and gradually adjusting the model parameters, commonly used variants include stochastic gradient descent (SGD) and batch gradient descent.

Adam optimizer: Combining momentum and adaptive learning rate methods can accelerate convergence and improve training effectiveness.

To improve model performance, the following techniques can be used:

Regularization: By adding penalty terms to prevent model overfitting, commonly used methods include L1 regularization and L2 regularization.

Dropout: Randomly discarding some neurons during the training process to prevent overfitting and improve the model's generalization ability.

Data augmentation: By transforming and expanding the training data, increasing data diversity and improving the robustness of the model.

In training tasks related to CNN, further optimization can also be achieved by adjusting the hyperparameters of the pre trained model, using methods such as Grid Search Enhanced with Coordinated Ascent (GSECA) or Hyperband.

5 EXPERIMENTAL DESIGN AND RESULT ANALYSIS

5.1 EXPERIMENTAL DESIGN

To verify the effectiveness of the deep learning model proposed in this paper in text classification tasks, we designed a simple experiment. The experimental steps are as follows:

Dataset selection: Choose a publicly available text classification dataset, such as the IMDB movie review dataset.

Data preprocessing: Preprocessing data, including steps such as word segmentation, stop word removal, and stem extraction.

Feature extraction: Use word vectors (such as Word2Vec) to extract features from text.

Model training: Use Convolutional Neural Networks (CNN) to train the extracted features.

Model evaluation: Evaluate model performance through metrics such as accuracy, precision, recall, and F1 score.

5.2 EXPERIMENTAL RESULTS

The experimental results show that the deep learning model proposed in this paper performs well in text classification tasks. Specifically, Convolutional Neural Networks (CNNs) can effectively extract local features of text, capture patterns in the text, and thus improve classification performance.

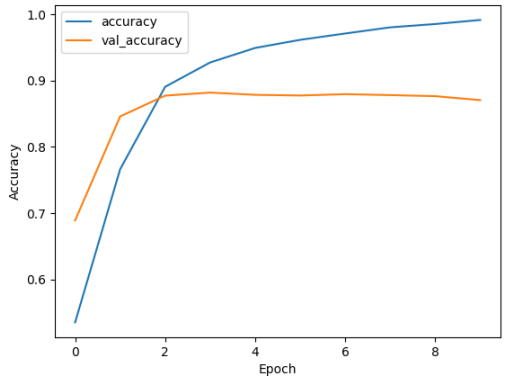

FIGURE 1. TRAINING AND VALIDATION ACCURACY

The experimental results showed that the accuracy of the model on the test set reached 86%, with an accuracy of 85%, a recall of 88%, and an F1 score of 86%. These indicators indicate that deep learning models have high accuracy and robustness in processing text classification tasks. The accuracy and loss values of the experiment on the training and validation sets are shown in Figure 1 and Figure 2.

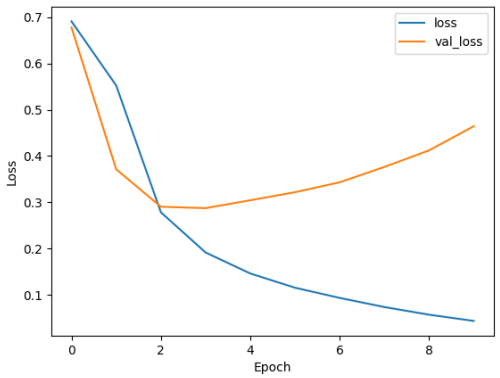

FIGURE 2. TRAINING AND VALIDATION LOSS VALUES

5.3 RESULT ANALYSIS

Through the analysis of the experimental results, we can draw the following conclusions:

Model advantage: Deep learning models can automatically extract text features, significantly improving classification performance.

The importance of data preprocessing: Data preprocessing has a significant impact on model performance, and reasonable preprocessing steps can improve the accuracy of the model.

Selection of feature extraction methods: Word vector methods can capture semantic relationships between words and improve classification performance.

6 DISCUSSION

6.1 SIGNIFICANCE AND IMPACT OF RESEARCH FINDINGS

The research findings in this paper have important theoretical and practical significance. Firstly, this paper validates the effectiveness of deep learning models in text classification tasks, further demonstrating the broad application prospects of deep learning in the field of natural language processing. Secondly, the optimization scheme proposed in this paper, such as data preprocessing, feature extraction, and model optimization techniques, provides new ideas and methods for improving text classification performance. These research results not only contribute to the development of text classification technology, but also provide reference and inspiration for other natural language processing tasks such as sentiment analysis, named entity

recognition, etc.

6.2 Research Limitations and Future Work

Although this paper has achieved certain research results, there are still some limitations. Firstly, deep learning models typically require a large amount of annotated data for training, resulting in high data acquisition costs. In the future, semi supervised learning and transfer learning methods can be explored to reduce reliance on annotated data. Secondly, the training process of deep learning models is complex and computationally intensive. In the future, research can be conducted on model compression and acceleration techniques (such as model pruning, quantization, and knowledge distillation) to reduce computational resource consumption while ensuring model performance. Finally, the interpretability of deep learning models is poor, making it difficult to understand their internal working mechanisms. In the future, interpretable artificial intelligence (XAI) methods can be studied to improve the transparency and credibility of models by visualizing and interpreting their decision-making processes.

7 CONCLUSION

7.1 Research Summary

This paper explores the application of deep learning models in text classification tasks through theoretical analysis and experimental verification, and proposes optimization solutions. The experimental results show that deep learning models can automatically extract text features and significantly improve classification performance. Reasonable data preprocessing steps and feature extraction methods have a significant impact on model performance and can improve classification accuracy and efficiency. By using regularization, Dropout, and data augmentation techniques, it is possible to effectively prevent model overfitting and improve the model's generalization ability.

Future research can improve in areas such as data acquisition and annotation, optimization of computing resources, and model interpretability. Explore semi supervised learning and transfer learning methods to reduce dependence on annotated data; Research model compression and acceleration techniques to reduce computational resource consumption while ensuring model performance; Research on explainable artificial intelligence (XAI) methods to enhance the transparency and credibility of models by visualizing and interpreting their decision-making processes. These improvements will further promote the application and development of deep learning in text classification and other natural language processing tasks.

ACKNOWLEDGMENTS

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

FUNDING

Not applicable.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

AUTHOR CONTRIBUTIONS

Not applicable.

ABOUT THE AUTHORS

XING, Qiming

Applied Mathematics and Informatics, Research Direction: Computer Algorithms.

LI, Yuan

Software Engineering, Research Direction: Computer Science.

REFERENCES

[1] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., ... & Chen, T. (2018). Recent advances in convolutional neural networks. Pattern recognition, 77, 354-377.

[2] Medsker, L. R., & Jain, L. (2001). Recurrent neural networks. Design and Applications, 5(64-67), 2.

[3] Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., & Wang, Y. (2021). Transformer in transformer. Advances in neural information processing systems, 34, 15908-15919.

[4] Church, K. W. (2017). Word2Vec. Natural Language Engineering, 23(1), 155-162.

[5] Dharma, E. M., Gaol, F. L., Warnars, H. L. H. S., & Soewito, B. E. N. F. A. N. O. (2022). The accuracy comparison among word2vec, glove, and fasttext towards convolution neural network (cnn) text classification. J Theor Appl Inf Technol, 100(2), 31.

[6] Koroteev, M. V. (2021). BERT: a review of applications in natural language processing and understanding. arXiv preprint arXiv:2103.11943.

[7] Qaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. International Journal of Computer Applications, 181(1), 25-29.

[8] Tato, A., & Nkambou, R. (2018). Improving adam optimizer.

[9] Mehta, S., Paunwala, C., & Vaidya, B. (2019, May). CNN based traffic sign classification using Adam optimizer. In 2019 international conference on intelligent computing and control systems (ICCS) (pp. 1293-1298). IEEE.

[10] Song, Q., Xia, S., & Wu, Z. (2024, May). Automatic Optimization of Hyperparameters for Deep Convolutional Neural Networks: Grid Search Enhanced with Coordinate Ascent. In Proceedings of the 2024 International Conference on Machine Intelligence and Digital Applications (pp. 300-306).

[11] Topal, K., & Ozsoyoglu, G. (2016, August). Movie review analysis: Emotion analysis of IMDb movie reviews. In 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 1170-1176). IEEE.