Automation and Machine Learning (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2026.070113
ISSN 2516-5003 Vol. 7 Num. 1

Design and Implementation of a Sentiment Analysis System Based on Deep Learning

Chuwei Wang, Zicheng Wang, Xihe Wang, Mingxing Wang*

School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, China
china205396@163.com

Keywords: Sentiment Analysis, RoBERTa Model, Adversarial Training, Ablation Experiment, Pre-trained Language Model, Text Augmentation, Emotional Attention Mechanism

Abstract: Sentiment analysis, a core NLP task, has important application value in public opinion monitoring, product evaluation, and social media analysis. Traditional methods based on manual feature extraction or shallow learning models have limitations in complex text scenarios. To address these, this paper designs and implements a sentiment analysis system based on an improved RoBERTa model. The system selects three real public datasets, optimizes data quality through text cleaning, normalization, and back - translation augmentation. It integrates an emotional attention mechanism and introduces FGSM adversarial training. Ablation and comparative experiments are conducted. Results show the proposed model achieves an average F1 - score of 94.2% on the three datasets, 3.7 to 5.1 percentage points higher than the baseline model. The system has strong generalization, high robustness, and reliable performance, offering a practical solution for multi - scenario text sentiment analysis.

1. Introduction

With the Internet and social media's rapid development, massive text data with subjective emotions is generated daily, like e-commerce product reviews, video website movie comments, and social network public opinions. Extracting and analyzing emotional tendencies from these data helps enterprises understand consumer demands, assists government departments in monitoring public opinion, and supports social management and market operations. However, natural language's complex text characteristics pose challenges to sentiment analysis. Traditional sentiment analysis methods rely on manual feature engineering and classifiers, which are time - consuming, labor - intensive, and have poor cross - domain adaptability. Pre - trained language models like BERT have improved sentiment analysis performance by capturing contextual semantics. But mainstream pre - trained models have shortcomings in focusing on sentiment features and resisting text noise, leading to performance degradation in complex scenarios[9].

Thus, designing a sentiment analysis system with strong generalization and high robustness based on improved pre - trained models is of theoretical and practical value for emotional text data's intelligent processing.

1.1 Research Background and Significance

The popularization of social media and e-commerce platforms has led to explosive growth of subjective text data. These data carry users' emotional tendencies and are valuable to various fields. Enterprises can analyze product reviews to identify advantages and defects, optimize design and strategies[1]; government departments can monitor public opinion to grasp social trends and respond to concerns[6]; individuals can use sentiment analysis tools to screen valuable information[2].

Although sentiment analysis technology has made progress, there are challenges in practical applications. For example, the same word may have different emotional tendencies in different contexts, and noisy text increases analysis difficulty. Traditional methods and general pre-trained models struggle to balance generalization and robustness, restricting the practical application effect. Therefore, optimizing the model structure and training strategy to improve the system's comprehensive performance has become a key research direction.

1.2 Research Status at Home and Abroad

In foreign research, scholars explored pre - trained model optimization for sentiment analysis. Liu et al.'s RoBERTa optimized BERT's training strategy, increasing data volume, extending training time and removing next - sentence prediction, which improved generalization. Later, attention mechanisms were integrated into RoBERTa to enhance local emotional feature extraction, yet most lacked targeted emotional semantic expression optimization. Adversarial training technologies (FGSM, PGD) were applied to boost model robustness, but their combination with emotional attention mechanisms in sentiment analysis needs more research.

In domestic research, scholars focus on pre - trained models in Chinese sentiment analysis. Technologies for English multi - scenario sentiment analysis are scattered. Some use text augmentation to solve sample shortage, but the impact of augmentation strategies on model performance in different scenarios is unverified[5].

Overall, Transformer - based models are effective in sentiment analysis, but integration optimization of emotional feature enhancement, robustness improvement and data augmentation needs enhancing for multi - scenario applications.

1.3 Main Research Content and Structure

This paper focuses on the design and implementation of a high-performance sentiment analysis system, focusing on solving the problems of poor generalization ability and low robustness of traditional models. The main research contents include: selection and preprocessing of multi-scenario real datasets, construction of an improved RoBERTa model integrating emotional attention mechanism and adversarial training, design of ablation experiments to verify the effectiveness of each module, and comparative analysis with mainstream models to evaluate system performance.

The subsequent structure of this paper is arranged as follows: the second part introduces the core technical foundations including RoBERTa model, emotional attention mechanism, adversarial training, and text augmentation; the third part elaborates on the overall design of the sentiment analysis system, including dataset processing, model architecture design, and training strategy setting; the fourth part presents experimental results and detailed analysis, including ablation experiments, comparative experiments, and generalization ability testing; the final part summarizes the research results, points out the shortcomings of the system, and looks forward to future improvement directions.

2. Related Technical Foundations

2.1 RoBERTa Pre-trained Language Model

RoBERTa is an improved BERT version, retaining BERT's multi - layer Transformer encoder structure and optimizing the training process. It uses a bidirectional attention mechanism to capture contextual semantics and solves the unidirectional information limitation in traditional language models. Compared to BERT, RoBERTa has key optimizations: expanding the training dataset to 160GB, increasing the batch size, removing the next sentence prediction task, and using dynamic mask technology. These enhance the model's ability to learn diverse semantic representations.

The RoBERTa base model has 12 Transformer layers, 12 attention heads per layer, a hidden layer dimension of 768, and about 110 million parameters. It can effectively extract deep text semantic features and support sentiment analysis tasks.

2.2 Emotional Attention Mechanism

The traditional self-attention mechanism in Transformer assigns equal attention weights to all words in the text, which may lead to the model focusing on non-emotional words and ignoring key emotional expressions. The emotional attention mechanism designed in this paper introduces emotional prior knowledge to adjust the attention weight distribution[7]. It first uses a pre-trained emotional dictionary to mark emotional words in the text, then calculates the emotional relevance score between each word and emotional words, and fuses this score into the self-attention calculation process to enhance the attention of the model to sentiment-related words.

The emotional attention mechanism can adaptively capture emotional key features according to the text content, avoiding the interference of irrelevant information, and improving the accuracy of sentiment classification[3].

2.3 FGSM Adversarial Training

Adversarial training is a common technology to improve model robustness. It generates adversarial examples by adding subtle perturbations to the input data, and trains the model with adversarial examples and original examples to enhance the model's ability to resist noise and interference. Fast Gradient Sign Method (FGSM) is a classic adversarial example generation method with low computational complexity and high efficiency.

In this system, FGSM is used to generate adversarial examples for the input text's word embedding vector. By calculating the gradient of the loss function with respect to the word embedding, subtle perturbations are added along the gradient direction to generate adversarial samples that are difficult for the model to distinguish but retain the original emotional tendency. The model is trained with a mixed dataset of original samples and adversarial samples to improve its robustness in complex noisy text scenarios.

2.4 Back-translation Text Augmentation

Text augmentation is an effective method to expand the dataset scale and improve model generalization ability. Back-translation technology converts the original text into an intermediate language (such as French, German) through a machine translation model, and then translates it back to the original language to generate new text samples with the same meaning but different expressions. This method can retain the original emotional tendency while increasing the diversity of the dataset, effectively alleviating the overfitting problem caused by insufficient samples.

In this system, the Google Translate API is used for back-translation, and English text is translated into Spanish first and then back to English to generate augmented samples. The augmented samples are merged with the original dataset after manual verification to ensure data quality.

3. Design of Sentiment Analysis System

3.1 Dataset Selection and Preprocessing

3.1.1 Dataset Selection

To ensure data authenticity, completeness, and multi-scenario adaptability, this paper uses three public sentiment analysis datasets covering different scenarios and polarities: the IMDB Movie Reviews Dataset, a binary classification dataset with 50,000 balanced positive and negative reviews; the Amazon Product Reviews Dataset, a multi-category dataset with 100,000 reviews labeled as negative, neutral, and positive; and the Twitter Sentiment140 Dataset, a large-scale noisy text dataset with 1.6 million tweets containing abbreviations, emoticons, and typos to test model robustness.

All datasets are randomly split into training, validation, and test sets at a 7:2:1 ratio. The training set is used for parameter learning, the validation set for hyperparameter tuning and overfitting monitoring, and the test set for assessing the final model performance.

3.1.2 Data Preprocessing

To improve data quality and model training efficiency, the dataset undergoes four preprocessing steps: text cleaning, normalization, augmentation, and tokenization. Text cleaning removes URLs, special symbols, and emoticons (except in the Twitter dataset to preserve noise), corrects typos, and converts text to lowercase. Text normalization uses NLTK for stopword removal and stemming to reduce vocabulary size.

Text augmentation applies back-translation to 50% of training samples for the IMDB and Amazon datasets and 30% for the Twitter dataset, then combines augmented and original data. Finally, the RoBERTa tokenizer converts text to token IDs with a maximum sequence length of 256, using padding and truncation for consistent input.

3.2 Model Architecture Design

The sentiment analysis model designed in this paper takes RoBERTa as the base model and integrates an emotional attention mechanism and adversarial training module, with a three-layer architecture: input layer, feature extraction layer, and classification layer.

3.2.1 Input Layer

The input layer receives preprocessed text data, converts it into token IDs, segment IDs, and attention masks through the RoBERTa tokenizer, and inputs them into the feature extraction layer. At the same time, the emotional dictionary is used to mark emotional words in the text, and the marking results are input into the emotional attention mechanism as auxiliary information.

3.2.2 Feature Extraction Layer

The feature extraction layer consists of the RoBERTa base model and the emotional attention mechanism. First, the RoBERTa model extracts text's contextual semantic features, getting a (batch-

size, sequence-length, hidden-dim) feature matrix. Then, the emotional attention mechanism calculates the attention weight between tokens and emotional words, multiplies the weight matrix with the RoBERTa feature matrix to enhance emotional feature representation, and outputs the enhanced matrix.

During training, the FGSM adversarial training module generates adversarial perturbations for the word embedding vector from the RoBERTa embedding layer, adds them to the original vector to form adversarial samples, and inputs them into subsequent layers. The model trains alternately with original and adversarial samples to improve robustness.

3.2.3 Classification Layer

The classification layer takes the [CLS] token feature of the enhanced feature matrix as the global text feature, and maps the feature to the sentiment category space through two fully connected layers. For binary classification tasks (IMDB, Twitter), the Softmax activation function is used to output the probability of positive and negative sentiment; for multi-classification tasks (Amazon), the Softmax function is used to output the probability of three sentiment polarities. The cross-entropy loss function is used to calculate the prediction error and guide model parameter update.

3.3 Model Training and Optimization Strategies

The system is implemented based on the Python programming language, using the PyTorch deep learning framework and the Hugging Face Transformers library for model construction and training. The experimental hardware environment is: Intel Core i9-12900K CPU, NVIDIA GeForce RTX 4090 graphics card (24GB video memory), 64GB memory; the software environment is: Ubuntu 22.04 operating system, Python 3.9, PyTorch 2.0.1, NLTK 3.8.1.

The core training parameters are set as follows: batch size is 32, initial learning rate is 2e-5, training epochs are 20, weight decay is 1e-4 to prevent overfitting. The AdamW optimizer is used for parameter update, which combines the advantages of Adam and weight decay to improve training stability. The early stopping strategy is adopted: if the validation set F1-score does not improve for 3 consecutive epochs, the training is terminated to avoid overfitting, and the model with the best validation set performance is saved.

4. Experimental Results and Analysis

4.1 Experimental Design

To evaluate the proposed system's performance, three types of experiments are designed: ablation experiments to verify each optimization module's effectiveness, comparative experiments with mainstream models to evaluate overall performance, and cross - dataset generalization experiments to test the model's adaptability to different scenarios. The evaluation indicators include Accuracy, Precision, Recall, and F1 - score, with F1 - score as the main indicator to comprehensively reflect the model's performance in positive and negative samples.

4.2 Ablation Experiment Results and Analysis

Ablation experiments on the IMDB dataset verify the effects of the emotional attention mechanism, FGSM adversarial training and back-translation augmentation. Four groups are set: Group 1 (Baseline: unoptimized RoBERTa), Group 2 (RoBERTa + back-translation), Group 3 (RoBERTa + back-translation + emotional attention), and Group 4 (Proposed Model: RoBERTa +

all three optimizations). Results are in Table 1.

Table 1: Ablation experiment results on IMDB dataset.

Experimental Group	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Group 1 (Baseline)	89.6	90.1	89.0	89.5
Group 2 (RoBERTa + Augmentation)	91.8	92.3	91.2	91.7
Group 3 (RoBERTa + Augmentation + Attention)	93.5	93.8	93.2	93.5
Group 4 (Proposed Model)	94.6	94.9	94.3	94.6

As seen from Table 1, each optimization module can effectively improve model performance. Compared with the baseline model, Group 2's F1-score increases by 2.2 percentage points, showing that back-translation augmentation can expand dataset diversity and alleviate overfitting. Moreover, Group 3's F1-score further increases by 1.8 percentage points compared with Group 2, proving that the emotional attention mechanism can enhance the capture of sentiment key features and improve classification accuracy. Finally, the proposed model in Group 4 achieves the highest F1-score of 94.6%, 5.1 percentage points higher than the baseline model, indicating that FGSM adversarial training can significantly improve the model's robustness and optimize performance.

4.3 Comparative Experiment Results and Analysis

The proposed model is compared with mainstream sentiment analysis models such as BERT-base, ALBERT-base, and DistilBERT on three datasets. All models use the same data preprocessing and training parameters to ensure the fairness of the experiment. The average performance of each model on the three datasets is shown in Table 2.

Table 2: Comparative experiment results on three datasets.

Model	IMDB F1-score (%)	Amazon F1-score (%)	Twitter F1-score (%)	Average F1-score (%)
BERT-base	91.2	88.5	89.8	90.0
ALBERT-base	92.1	89.3	90.5	90.6
DistilBERT	90.5	87.9	89.2	89.2
RoBERTa-base	92.8	90.1	91.3	91.4
Proposed Model	94.6	93.5	94.5	94.2

Table 2 shows the proposed model outperforms other mainstream models on all three datasets. Compared with RoBERTa - base, its average F1 - score increases by 2.8 percentage points, with the most significant improvement (3.4 percentage points) on the Amazon multi - classification dataset, indicating the emotional attention mechanism effectively captures complex emotional features. On the noisy Twitter dataset, it achieves an F1 - score of 94.5%, 3.2 percentage points higher than RoBERTa - base, proving FGSM adversarial training enhances the model's robustness to noisy text[8].The experimental results demonstrate the proposed model performs excellently in both binary

and multi - classification sentiment analysis tasks[10].

4.4 Generalization Ability Test

To verify the model's cross-scenario generalization, the model trained on the IMDB dataset is directly tested on the Amazon and Twitter datasets without retraining and compared with RoBERTa-base. The proposed model achieves F1-scores of 87.2% and 88.5% on the two datasets, 2.1 and 2.3 percentage points higher than RoBERTa-base respectively, demonstrating that back-translation augmentation and adversarial training improve cross-domain and noisy text adaptability with strong generalization.Conclusions and Outlook

4.5 Research Conclusions

This paper designs and implements a sentiment analysis system based on an improved RoBERTa model. Experiments show that text preprocessing and back-translation augmentation effectively expand the dataset and boost generalization, raising the F1-score by 2.2 percentage points over the baseline. The emotional attention mechanism enhances the model's focus on key sentiment features, improving classification accuracy by 1.8 percentage points. FGSM adversarial training strengthens robustness to noisy text, optimizing performance by another 1.1 percentage points. Integrating these three modules, the proposed model achieves an average F1-score of 94.2% across three real datasets, outperforming BERT and RoBERTa-base with high accuracy, strong robustness, and good generalization.

4.6 Limitations and Future Directions

Although the proposed system performs well, it has limitations. The emotional attention mechanism depends on pre - trained emotional dictionaries and is limited for domain - specific text. FGSM adversarial training only perturbs the word embedding layer, with limited robustness improvement. The system only supports English sentiment analysis, and its adaptability to Chinese needs verification. Future research will optimize the emotional attention mechanism through context - aware emotional word recognition to reduce dictionary dependence, explore multi - layer adversarial training for better robustness, and extend the system to Chinese sentiment analysis using pre - trained models like ERNIE.[4]

Acknowledgement

Fund Project: 2026 Liaoning Province College Students Innovation and Entrepreneurship Training Program.

References

[1] Prova I N N ,Ravi V ,Singh P M , et al. Multilingual sentiment analysis in e-commerce customer reviews using GPT and deep learning-based weighted-ensemble model[J].International Journal of Cognitive Computing in Engineering,2026,268.DOI:10.1016/J.IJCCE.2025.10.003.

[2] Tigani S . Next-gen HR interview process parametric AI copilot design based on pretrained LLM and sentiment analysis deep learning models[J].Pattern Analysis and Applications,2025,28(4):205.DOI: 10.1007/S10044-025-01581-Z.

[3] Tarik N ,Jadhav A . DStaL: Multilevel Fusion Classifier Based Analysis of Multimodal Sentiments Using Deep Learning Models[J].Computational Intelligence,2025,41(6):e70154.DOI:10.1111/COIN.70154.

[4] Yarjanli M ,Mahdinasab N . A Hybrid Deep Learning Model for E-Commerce Recommendations: Sentiment Analysis With Autoencoders and Generative Adversarial Networks[J].International Journal of Intelligent Systems,2025,

2025(1):3852068.DOI:10.1155/INT/3852068.
[5] Sitoula S ,Shahi B T ,Wibowo S , et al. Sentiment analysis of Nepali social media text with a hybrid deep learning model[J].Social Network Analysis and Mining,2025,15(1):85.DOI:10.1007/S13278-025-01508-W.
[6] Li L ,Choo C W ,Leong C Y , et al. Predicting purchase intentions in online food delivery using deep learning and AIDA model: Insights from sentiment analysis of user reviews[J].International Journal of Engineering Business Management,2025,17.DOI:10.1177/18479790251375827.
[7] Sharma R ,Kukreja V . Deciphering emotions in comics: analysis of emotion classification of comic characters via attention-based deep learning models[J].Multimedia Tools and Applications,2025,84(39):1-33.DOI:10.1007/S11042-025-21017-4.
[8] Talukder A M ,Uddin A M ,Roy S , et al. A hybrid deep learning model for sentiment analysis of COVID-19 tweets with class balancing.[J].Scientific reports,2025,15(1):27788.DOI:10.1038/S41598-025-97778-7.
[9] Yanping L . Building and Optimizing Deep Learning Models for Sentiment Analysis in English Text[J].Journal of Cases on Information Technology (JCIT),2025,27(1):1-17.DOI:10.4018/JCIT.382380.
[10] Atlas G L ,Arockiam D , Muthusamy A , et al. A modernized approach to sentiment analysis of product reviews using BiGRU and RNN based LSTM deep learning models.[J].Scientific reports,2025,15 (1) :16642. DOI :10.1038/S41598-025-01104-0.