

Enhancing Emotion Classification Through Neural Networks and Data Augmentation

Yichen Xiong

*School of Information Engineering, Minzu University of China, Beijing, China
lkcamellia@163.com*

Abstract: Emotion classification is a key task in natural language processing with lots of applications in our daily life. Traditional methods, such as Logistic Regression, struggle to capture the complex semantic and contextual dependencies in informal and diverse text, often leading to suboptimal performance. To address these limitations, we propose a novel approach that combines BERT with BiLSTM networks and utilizes data augmentation techniques. BERT's pre-trained embeddings contain rich contextual information, while the BiLSTM layer models long-range dependencies within the text. Data augmentation is applied to expand the training dataset, improving the model's generalization and robustness. Experimental results on the ACL IMDB dataset demonstrate that our approach outperforms traditional models, with the BERT + BiLSTM model achieving a testing accuracy of 88%, compared to 82% for BERT alone. This study shows that combining BERT, BiLSTM, and data augmentation significantly enhances the accuracy and robustness of emotion classification models, particularly for complex and informal text.

Keywords: Emotion classification, sentiment analysis, BERT, BiLSTM, data augmentation.

1. Introduction

Emotion classification is becoming more and more important in recent years due to its broad applications in various fields such as sentiment analysis, customer feedback, and social media monitoring. The ability to accurately determine the emotional tone of a text is crucial for understanding human communication and improving services in industries like marketing, healthcare, and social media analysis.

Traditional emotion classification methods primarily rely on shallow machine learning models such as Logistic Regression, which often face limitations in capturing the intricate semantic and contextual relationships in the text. These models struggle with handling complex language expressions, dialects, and informal writing styles commonly seen in social media and online content. Furthermore, the inability to model long-range dependencies in text sequences leads to suboptimal performance in tasks involving complex sentiment analysis.

To address these challenges, we propose a novel approach that combines BERT, BiLSTM, and data augmentation techniques for emotion classification. BERT, with its pre-trained deep contextualized embeddings, captures the nuanced semantic information in the text. The BiLSTM component, designed to model long-range dependencies in sequential data, further improves the model's ability to understand text structures. Additionally, data augmentation is employed to enhance

the model's generalization capability by artificially expanding the training dataset, thus mitigating the overfitting problem and improving the robustness of the model in real-world applications.

Experimental results has shown high accuracy of our model, particularly when dealing with complex and informal expressions. The proposed method shows significant promise in improving emotion classification tasks across various domains, including online reviews, social media posts, and customer feedback.

The remainder of this paper is organized as follows. Section 2 introduces the related work on emotion classification and the challenges faced by traditional methods. In Section 3, we describe the architecture of our proposed model, including the integration of BERT, BiLSTM, and data augmentation. Section 4 presents the experimental setup and evaluation results. Finally, Section 5 concludes the paper with a summary of our findings and potential future directions for research.

2. Related work

2.1. Emotion classification approaches

Emotion classification is a core task in natural language processing (NLP) that aims to categorize textual data into predefined emotion categories, such as joy, sadness, anger, and surprise. Over the years, various methods have been proposed to address this task. Traditional machine learning models, such as Support Vector Machines (SVM) [1], Naive Bayes [2], and Decision Trees [3], have been widely used in early sentiment analysis tasks. These models usually rely on handcrafted features like Bag of Words (BoW) or Term Frequency-Inverse Document Frequency (TF-IDF) to capture the textual information. However, these approaches are limited in their ability to capture complex, context-dependent meanings and long-range dependencies within the text.

With the advent of deep learning, more advanced models have been proposed for emotion classification. Convolutional Neural Networks (CNNs) [4] and Recurrent Neural Networks (RNNs) [5] are among the popular choices due to their ability to automatically learn useful features from the raw text data. Despite these advancements, traditional RNNs often struggle with capturing long-term dependencies and contextual understanding. To address these issues, more sophisticated models like Long Short-Term Memory (LSTM) networks [6] and Bidirectional LSTM (BiLSTM) networks [7] have been introduced, which allow for better representation of sequence data by preserving information from both past and future contexts.

The most recent trend in emotion classification has been the use of transformer-based models like BERT, which have shown state-of-the-art performance on many NLP tasks, including emotion recognition. BERT's pre-trained embeddings have proven particularly effective in capturing contextual relationships in text [8].

2.2. BERT (Bidirectional Encoder Representations from Transformers)

BERT, introduced by Devlin et al. [8], represents a significant breakthrough in NLP by leveraging bidirectional transformer networks. Unlike previous models that read text sequentially (left-to-right or right-to-left), BERT reads the entire sequence at once, capturing both left and right context in parallel. This bidirectional approach allows BERT to better understand the context of words in a sentence, making it highly effective for a range of NLP tasks, including sentiment and emotion classification.

BERT's architecture consists of multiple layers of transformers, where each layer learns to represent increasingly abstract features of the input text. Additionally, BERT is pre-trained on large corpora using a masked language modeling (MLM) task, where random words in a sentence are masked, and the model is tasked with predicting them. This pre-training strategy enables BERT to

capture rich semantic relationships and contextual cues, making it an ideal candidate for downstream tasks such as emotion classification.

Many studies have shown that fine-tuning BERT for emotion classification tasks leads to substantial performance improvements over traditional methods. For example, in a study by Ovidiu-Mihai et al. [9], BERT-based models significantly outperformed earlier SVM-based approaches on sentiment analysis datasets. This success has led to widespread adoption of BERT in various sentiment and emotion analysis applications.

2.3. BiLSTM (Bidirectional Long Short-Term Memory)

BiLSTM, proposed by Schuster and Paliwal [7], extends the traditional LSTM by processing data in both forward and backward directions. This bidirectional approach allows the model to capture information from both past and future contexts within a sequence, which is crucial for tasks that require understanding the full context of a sentence or text.

In sentiment and emotion classification, BiLSTM has been shown to effectively capture long-range dependencies in text, especially when combined with word embeddings like Word2Vec [10] or GloVe [11]. The ability of BiLSTM to maintain long-term memory while avoiding the vanishing gradient problem makes it a powerful tool for sequential data tasks.

BiLSTM has been successfully applied to emotion classification tasks, such as predicting sentiment in movie reviews [12] and classifying emotions in social media posts [13]. Recent studies [14,15] have demonstrated that combining BiLSTM with pre-trained embeddings, such as BERT, further enhances its ability to understand the contextual meaning of words and phrases, improving classification accuracy.

3. Methodology

In this section, we present the methodology used for emotion classification in our research, focusing on data augmentation, and the integration of BERT and BiLSTM networks for effective sentiment analysis.

3.1. Data augmentation

Data augmentation is a crucial technique for enhancing the diversity of training data without the need to collect additional data. This method is especially valuable when labeled data is limited or imbalanced, as it helps improve the model's generalization ability and robustness. In the context of sentiment analysis, particularly when dealing with informal language or non-standard text such as internet slang, data augmentation can significantly enhance model performance by generating synthetic examples that reflect varied linguistic styles and expressions.

A typical approach for data augmentation in text classification tasks involves generating paraphrases or performing transformations that maintain the original meaning while altering the linguistic structure. This can include techniques such as random word deletion, synonym replacement, and sentence reordering. For instance, random word deletion involves removing non-essential words from the sentence, while synonym replacement substitutes certain words with their synonyms, preserving the core meaning.

Consider the original sentence and its augmented version from the text in following image:

Original: "Bromwell High is a cartoon comedy. It ran at the same time as some other programs about school life, such as 'Teachers'."

- **Random Word Deletion:** "Bromwell High is a cartoon comedy. It ran at the same time as some programs about life, such as 'Teachers'."
- **Synonym Replacement:** "Bromwell High is an animated show. It aired alongside other programs about school life, like 'Teachers'."

Figure 1: Sentence transformation examples

In the first example, we randomly deleted the word "other," reducing the sentence's complexity while still preserving its meaning. In the second example, the word "cartoon" was replaced by "animated" and "ran" by "aired," which introduces slight changes in the sentence structure but keeps the overall meaning intact. These transformations ensure that the model is exposed to a greater variety of linguistic patterns, thus helping reduce overfitting and improving the model's robustness to real-world language variations.

Additionally, random word insertion and substitution can further diversify the training data. For example, we might insert a word like "popular" into the sentence: "Bromwell High is a popular cartoon comedy." These variations provide the model with a richer set of examples, improving its ability to generalize to unseen data.

Data augmentation not only increases the size of the training set but also introduces linguistic variability, making the model more robust to different language expressions, idioms, and slang. This is particularly important for sentiment analysis tasks involving non-standard text, where training datasets often lack representation of diverse language forms. By employing these data augmentation techniques, we ensure that the model is better equipped to handle variations in real-world data, leading to improved sentiment classification performance.

3.2. BERT + BiLSTM network

BERT is a pre-trained deep learning model based on transformer architecture that has revolutionized various natural language processing (NLP) tasks, including sentiment analysis. Unlike traditional models that read text sequentially (either left-to-right or right-to-left), BERT processes the input text bidirectionally, allowing it to capture the full context of a sentence. This bidirectional training approach results in more accurate contextualized word embeddings, which have proven to outperform previous state-of-the-art models in various NLP tasks.

BERT is pre-trained on a large corpus using the masked language modeling (MLM) objective, where some words in a sentence are randomly masked, and the model is tasked with predicting them. This pre-training allows BERT to learn rich representations of words and their relationships within sentences, which can be fine-tuned on downstream tasks like sentiment classification. The fine-tuning process involves training the model on a task-specific dataset, where the final layer of BERT is adapted to output class predictions (such as positive, negative, or neutral sentiments).

While BERT excels at understanding the contextual relationships between words, it does not inherently model sequential dependencies across long sequences of text. To address this limitation, we combine BERT with a BiLSTM network. BiLSTM is a type of recurrent neural network (RNN) that processes sequences in both forward and backward directions, allowing it to capture long-range dependencies and context from both the past and the future. This combination is particularly useful for sentiment classification tasks, where the context of the entire sentence or paragraph is crucial for determining the sentiment.

In our model, the output of BERT is fed into a BiLSTM layer, which further refines the feature representations by considering the sequential nature of the text. The final output of the BiLSTM layer is passed through a dense layer followed by a softmax activation function to produce the sentiment classification. The integration of BERT and BiLSTM allows the model to leverage both the rich contextual embeddings from BERT and the sequential dependency modeling capability of BiLSTM, leading to enhanced performance in sentiment analysis, especially for complex or noisy text.

The combination of data augmentation, BERT, and BiLSTM enables our model to effectively handle diverse linguistic expressions and long-term dependencies, resulting in a robust and accurate sentiment classification model capable of tackling various types of emotional expression in text.

4. Experiment result

4.1. Experiment setup

To evaluate the effectiveness of our proposed approach, we conducted a series of experiments comparing multiple model configurations. Specifically, we examined the performance of the baseline BERT model, BERT enhanced with BiLSTM layers, and both models with and without data augmentation techniques. The goal of these experiments was to investigate how sequential modeling via BiLSTM and training data expansion via augmentation contribute to overall classification performance. All models were trained and evaluated on a standardized sentiment analysis dataset, with accuracy serving as the primary evaluation metric. The following sections describe the experimental setup, including dataset preparation, model implementation, and training configurations.

4.1.1. Data preparation

In this experiment, we utilized a sentiment analysis dataset, specifically from the ACL IMDB dataset [16], which contains a variety of textual expressions. The dataset was divided into training and testing sets, ensuring a balanced representation of both positive and negative sentiments. The training set consisted of 200 samples, and the testing set consisted of 2000 samples. All text inputs were preprocessed using BERT's tokenizer, which included tokenization, padding, and truncation to ensure that all sequences had a consistent length of 128 tokens.

4.1.2. Model configuration

We conducted experiments with different configurations of the BERT model, incorporating additional layers like BiLSTM for sequence modeling. The models tested were as follows:

- **BERT Only:** The original BERT model was used for sentiment classification, without additional layers.
- **BERT + BiLSTM:** The BERT model was extended with a single BiLSTM layer to capture sequential dependencies in the text.
- **BERT + BiLSTM + BiLSTM:** A more advanced configuration, where two BiLSTM layers were stacked on top of BERT to capture more complex patterns in the data.

The hyperparameters used for training these models are listed in the table below:

Table 1: Hyperparameters of the models

Hyperparameter	Value
Batch Size	32
Learning Rate	2e-5
Epochs	5

Table 1: (continued)

Max Sequence Length	512
Optimizer	Adam
Loss Function	CrossEntropyLoss
Hidden Size for BiLSTM	256

4.1.3. Hardware configuration

The models were trained on an NVIDIA RTX 4090 GPU with the following hardware setup:

- **GPU:** NVIDIA RTX 4090 (24GB)
- **CPU:** 16 vCPUs (Intel® Xeon® Gold 6430)
- **RAM:** 128GB
- **Operating System:** Ubuntu 20.04

his high-performance hardware ensured efficient model training, especially when processing large batches of data during the training phase.

4.2. Experiment result

The training loss comparison between the original and augmented models is shown in Figure 2. The models considered in this study include BERT, BERT+BiLSTM, and BERT+BiLSTM+ BiLSTM, evaluated both in their original configurations and with the application of data augmentation.

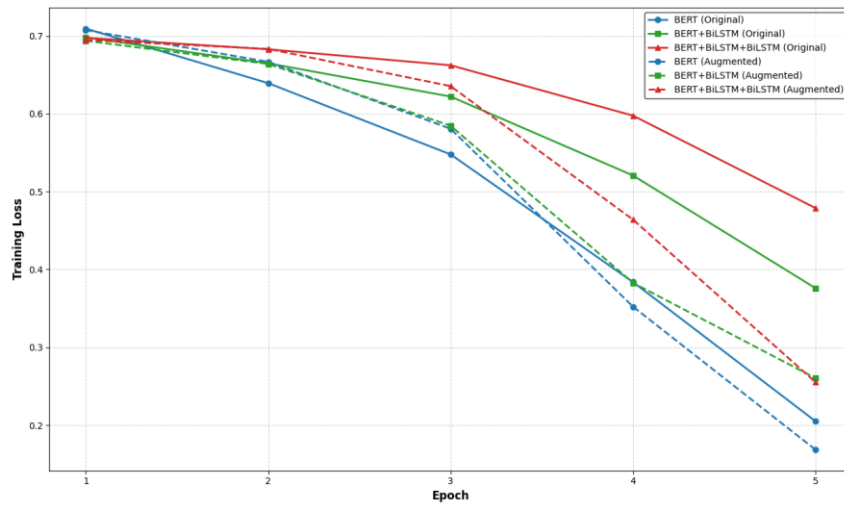


Figure 2: Comparison of training loss curves

The BERT (Original) model exhibits a gradual reduction in training loss over the epochs, with the loss decreasing from 0.7094 in the first epoch to 0.2051 by the last epoch. Although it improves over time, the rate of loss reduction slows significantly after the initial epochs. In contrast, the BERT+BiLSTM (Original) model shows a faster decline in training loss, starting at 0.6980 and ending at 0.3760. This suggests that the incorporation of a BiLSTM layer alongside BERT enhances the model's ability to learn sequential dependencies and results in more efficient training dynamics. The BERT+BiLSTM+BiLSTM (Original) model, which includes two BiLSTM layers, starts with a similar loss as the other models but demonstrates a slower decline, ultimately reaching 0.4789 by the

last epoch. This indicates that while the dual BiLSTM architecture offers greater capacity for capturing complex sequences, it leads to slower convergence compared to the single BiLSTM model.

When data augmentation is applied, all models show significant improvements. The BERT (Augmented) model achieves a faster decline in loss, starting at 0.7075 and dropping to 0.1685 by the last epoch, illustrating the effectiveness of data augmentation in improving model generalization. The BERT+BiLSTM (Augmented) model similarly benefits from augmentation, with a final loss of 0.2604, outperforming its original counterpart. The BERT+BiLSTM+BiLSTM (Augmented) model, combining data augmentation and a more complex architecture, reaches a final loss of 0.2553, which represents a notable improvement over the original model's loss of 0.4789.

The application of data augmentation consistently enhances the performance of all models, with the BERT+BiLSTM+BiLSTM (Augmented) model achieving the best results. The combination of a deeper network architecture and augmented data proves to be an effective strategy for improving model performance in natural language processing tasks.

The results of the experiments are presented in the table below, which compares the performance of various model configurations, both with and without data augmentation. Data augmentation was applied to enhance the generalization ability of the models by artificially expanding the training dataset.

Table 2: Comparison of model performance with and without data augmentation

Model	Without Data Augmentation (Training)	Without Data Augmentation (Testing)	With Data Augmentation (Training)	With Data Augmentation (Testing)
BERT	0.96	0.82	0.94	0.86
BERT+BiLSTM	0.90	0.84	0.90	0.88
BERT+BiLSTM + BiLSTM	0.88	0.85	0.96	0.88

The BERT-only model showed a significant difference in performance between training and testing phases. Without data augmentation, the model achieved a high training accuracy of 96%, but its testing accuracy dropped to 82%. This discrepancy highlights the model's strong performance on the training data, yet it struggles to generalize effectively to unseen data, suggesting overfitting. When data augmentation was introduced, the testing accuracy improved to 86%, demonstrating that augmentation helped the model generalize better and handle a wider variety of input data.

In the case of BERT combined with BiLSTM, the model achieved a testing accuracy of 84% without data augmentation, which represented a slight improvement over the BERT-only model. This suggests that the addition of the BiLSTM layer contributed to better capturing sequence dependencies in the text. With data augmentation, the testing accuracy further improved to 88%, indicating that the BiLSTM layer not only helped the model capture long-range dependencies but also benefitted from the diverse examples generated through augmentation, leading to a noticeable boost in performance.

For the BERT + BiLSTM + BiLSTM configuration, the model performed best on the training set, achieving an accuracy of 88% without data augmentation. However, the testing accuracy showed only a marginal improvement to 85% compared to the BERT+BiLSTM model. When data augmentation was applied, both training and testing accuracies saw substantial improvements. The training accuracy surged to 96%, and the testing accuracy increased to 88%. This indicates that the additional BiLSTM layer significantly improved both the model's ability to learn from the training data and its generalization to unseen data. Moreover, this configuration appeared to benefit the most from data augmentation, which likely helped the model capture even more complex patterns in the data.

5. Conclusion

Sentiment classification continues to present significant challenges, particularly when dealing with diverse and non-standardized text data such as social media posts and product reviews. In this study, we introduced an enhanced sentiment classification method that combines BERT with BiLSTM layers, aiming to effectively capture both contextual information and long-range dependencies within the text. Our experimental results show that this approach significantly outperforms traditional methods. Specifically, when BERT was used alone, we observed a notable performance gap between training and testing phases, with the testing accuracy reaching 82%. However, when BiLSTM layers were added, the model's ability to capture sequential dependencies in the text improved, resulting in a 6% increase in testing accuracy, reaching 88%. This demonstrates the effectiveness of BiLSTM layers in improving the model's ability to understand the intricate relationships between words in a sequence, especially for more complex and informal expressions commonly found in social media and product reviews.

Moreover, the application of data augmentation further enhanced the model's generalization ability. With data augmentation, the testing accuracy for the BERT-only model increased from 82% to 86%, showing the positive impact of augmenting the training data with diverse linguistic patterns. This improvement is particularly evident in handling non-standard language variations, such as slang or informal expressions, where data augmentation helps reduce overfitting and provides the model with a broader understanding of language diversity. For the combined BERT + BiLSTM model, the performance boost was even more pronounced. While the model without data augmentation achieved a testing accuracy of 84%, data augmentation helped it reach 88%, further validating the synergy between BiLSTM and data augmentation in improving both sequence modeling and robustness to unseen data.

These results demonstrate that integrating BiLSTM layers and leveraging data augmentation are both effective strategies for improving sentiment classification tasks. However, there is still room for further optimization in model architecture and training strategies, as the accuracy remains somewhat inconsistent across different experimental setups. Additionally, the study was conducted on a limited dataset, and future work could explore testing the model on larger, more diverse datasets to further enhance its generalization and robustness. Future research could also investigate the potential of more advanced data augmentation techniques and refine the model architecture to improve its scalability and performance across different domains.

References

- [1] Joachims, Thorsten. "Text categorization with support vector machines: Learning with many relevant features." *European conference on machine learning*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998.
- [2] McCallum, Andrew, and Kamal Nigam. "A comparison of event models for naive bayes text classification." *AAAI-98 workshop on learning for text categorization*. Vol. 752. No. 1. 1998.
- [3] Quinlan, J. Ross. "Induction of decision trees." *Machine learning* 1 (1986): 81-106.
- [4] Chen, Yahui. *Convolutional neural network for sentence classification*. MS thesis. University of Waterloo, 2015.
- [5] Mikolov, Tomas, et al. "Recurrent neural network based language model." *Interspeech*. Vol. 2. No. 3. 2010.
- [6] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.
- [7] Schuster, Mike, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks." *IEEE transactions on Signal Processing* 45.11 (1997): 2673-2681.
- [8] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019.
- [9] V. Ovidiu-Mihai, I. Tudor-Alexandru and M. Petrescu, "Using BERT to extract emotions from personal journals," *2022 IEEE 18th International Conference on Intelligent Computer Communication and Processing (ICCP)*, Cluj-Napoca, Romania, 2022, pp. 89-94, doi: 10.1109/ICCP56966.2022.10053943. keywords: {Emotion recognition;

Schedules; Mood; Machine learning; Predictive models; Depression; Data models; BERT; emotion detection; machine learning},

- [10] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems* 26 (2013).
- [11] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.
- [12] Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." *Foundations and Trends® in information retrieval* 2.1–2 (2008): 1-135.
- [13] Sajjad, Muhammad, and Soonil Kwon. "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM." *IEEE access* 8 (2020): 79861-79875.
- [14] Al-Omari, Hani, Malak A. Abdullah, and Samira Shaikh. "Emodet2: Emotion detection in english textual dialogue using bert and bilstm models." *2020 11th International Conference on Information and Communication Systems (ICICS)*. IEEE, 2020.
- [15] Gou, Zhinan, and Yan Li. "Integrating BERT embeddings and BiLSTM for emotion analysis of dialogue." *Computational Intelligence and Neuroscience* 2023.1 (2023): 6618452.
- [16] Maas, Andrew L., et al. *Learning Word Vectors for Sentiment Analysis*. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, June 2011, Portland, Oregon, USA, Association for Computational Linguistics*, pp. 142–150.