Academic Journal of Computing & Information Science
ISSN 2616-5775 Vol. 7, Issue 12: 91-95, DOI: 10.25236/AJCIS.2024.071213

Improving Text Classification by Leveraging Large Language Models for Data Augmentation

Siyun Yu

Tibet National University, School of Information Engineering, Xianyang, Shaanxi, 712082, China

Abstract: *In recent years, with the rapid development of large language models, optimizing BERT's performance on small-scale datasets using large language models has gradually become a research hotspot. To this end, this paper proposes a data augmentation method based on large language models to enhance BERT's performance in text classification tasks. Specifically, we first use large language models to back-translate the training data. By translating the text into other languages and then back into the original language, we generate new samples that are semantically consistent but have diverse expressions, thereby increasing the diversity of the training data. Subsequently, the augmented training set is used to train the BERT model, which significantly improves classification accuracy on the Reuters News Classification dataset. Experimental results show that this method effectively mitigates the limitations of small-scale datasets and significantly enhances the model's generalization ability, providing a novel and efficient solution for text classification tasks.*

Keywords: *GPT-3.5, Bert, Text Classification, Data Augmentation*

1. Introduction

Text classification is a core task in natural language processing (NLP) and is widely applied in various fields such as news categorization, sentiment analysis, and topic modeling. However, the performance of models is highly dependent on high-quality annotated data. In practice, due to the high cost and significant time investment required for data annotation, many scenarios only provide a limited amount of training data. This data scarcity issue limits the learning capacity and generalization performance of models. In recent years, the emergence of pre-trained language models, such as BERT and GPT, has greatly improved the performance of text classification tasks[1]. However, directly using these models on small-scale datasets may still lead to overfitting or insufficient performance due to insufficient data. Therefore, how to enhance model performance by data augmentation on existing small-scale datasets has become an important direction of interest for researchers.Data augmentation techniques are methods that enrich training data by generating additional samples, effectively alleviating the problem of data scarcity [2]. In text classification tasks, traditional data augmentation methods include synonym replacement, random word deletion or insertion, etc. [3]. While these methods are simple and straightforward, they may introduce semantic bias or structural noise, thereby affecting the performance of the model. In contrast, back translation, as a commonly used semantic preservation enhancement method, generates new samples that are semantically consistent but have different expressions by translating text into other languages and then back into the original language, showing excellent performance in multiple NLP tasks. However, the effectiveness of back translation largely depends on the quality of the translation models [4].

With the rapid development of large language models, back-translation techniques based on these models can generate higher quality and more diverse texts, providing strong tool support for data augmentation. This paper proposes a back-translation data augmentation method based on large language models to optimize BERT's performance in text classification tasks. Through experiments on the Reuters News Classification dataset, we have verified the effectiveness of this method in scenarios with small-scale datasets. Therefore, the large language model GPT-3.5 used in this study is based on the Transformer architecture [5], and has excellent text understanding and generation capabilities, especially in handling complex language expressions. This choice provides a solid technical foundation for data augmentation, effectively supporting the research objectives of this paper. The main contributions of this paper include:

(1) Proposing a data augmentation method based on large language models to enhance BERT's

ISSN 2616-5775 Vol. 7, Issue 12: 91-95, DOI: 10.25236/AJCIS.2024.071213

classification performance.

(2) Verifying through experiments on the Reuters News Classification dataset that this method can significantly improve BERT's classification accuracy.

(3) Discussing the optimization effects of data augmentation on small-scale datasets and the advantages of large language models in data generation.

2. Model Details

The model proposed in this paper is based on the large language model GPT-3.5 and BERT, and the overall structure of the model includes the following main components: Data Augmentation Layer, Text Embedding Layer, and Classification Layer.The architecture of the model is depicted in Figure 1.

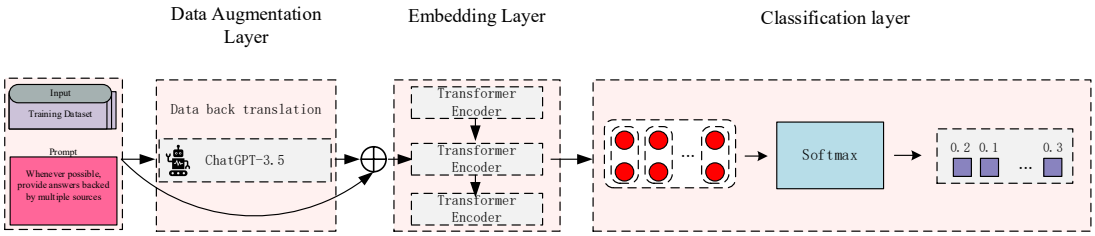

Figure 1: Model architecture diagram

2.1 Data Augmentation Layer Based on Large Language Models

The Data Augmentation Layer is a key component of the model proposed in this paper, designed to leverage the large language model GPT-3.5 for back-translation operations to generate diverse training samples, thereby enhancing the model's generalization capabilities. By expanding the training set, especially in cases where data is limited, it generates new samples that are semantically consistent with the original text but differ in expression, enriching the diversity of training data. In the Data Augmentation Layer, we employ back-translation technology based on the large language model GPT-3.5 to generate augmented samples. The basic process of back-translation is to first translate the original text into another language (such as French, German, etc.), and then translate it back into the original language. This process not only changes the way the text is expressed through cross-language conversion but also maintains the original semantic consistency of the text. Back-translation technology can generate a variety of text samples that are highly consistent with the original text in semantics but differ in language expression. Therefore, back-translation can effectively enrich training data[6], increase sample diversity, and reduce the model's dependence on a fixed expression.For example, when the model processes samples generated through back-translation, it can not only learn different forms of expression but also improve its adaptability to different language structures and vocabulary usage, thereby enhancing its generalization capabilities for various expression variants.

GPT-3.5, as a large-scale language model, has strong generation capabilities and cross-language understanding capabilities. Compared to traditional machine translation models, GPT-3.5 can better capture contextual information during the back-translation process, avoiding common translation errors and semantic loss[7]. After multiple rounds of translation, GPT-3.5 can still maintain semantic consistency and generate diverse and expressive training samples through the richness and flexibility of natural language. This advantage makes back-translation technology significantly effective in enhancing the diversity of datasets, helping to improve model robustness, reduce the risk of overfitting, and especially in the face of small-scale datasets, effectively improving model training results.Next, to ensure the quality of augmented samples, the Data Augmentation Layer performs semantic consistency checks on the texts generated by back-translation.

2.2 Text Embedding Layer

The main task of the text embedding layer is to transform input text, including both original samples and augmented samples generated through back-translation, into dense sentence vector representations using BERT. After undergoing back-translation processing in the data augmentation layer, the original training samples and augmented samples are passed together as input to the text embedding layer to

ISSN 2616-5775 Vol. 7, Issue 12: 91-95, DOI: 10.25236/AJCIS.2024.071213

generate the final text representation.

For each input text, whether it is an original sample or an augmented sample obtained through back-translation, the model first adds a special '[CLS]' token at the beginning of the text to indicate the start of the text and to serve an aggregating function in subsequent classification tasks. Then, the text is directly input into the BERT model, which processes it through the embedding layer to convert it into a high-dimensional dense vector representation.

Through this process, the text embedding layer generates dense sentence vectors for each input text, whether it is an original sample or a back-translation augmented sample. These vectors f_a provide high-quality input for the subsequent classification layer and enhance the model's generalization ability through data augmentation methods.

2.3 Classification Layer

The primary task of the classification layer is to apply the dense sentence vectors derived from the text embedding layer to text classification. The dense sentence vectors generated by the text embedding layer are directly passed as input to the classification layer, and subsequent classification decisions rely on this vector.

The classification layer consists of a fully connected layer and a Softmax function. First, the fully connected layer receives the sentence vector as input and maps it to the category space. Specifically, after processing by the fully connected layer, the input sentence vector is transformed into a vector equal in length to the number of categories, where each element represents the score for a category.

Next, the Softmax function is used to normalize the output of the fully connected layer, converting the scores for each category into corresponding probabilities. The Softmax function exponentiates and normalizes the output vector, ensuring that the probability of each category is between 0 and 1, and the sum of probabilities for all categories equals 1. Ultimately, the category with the highest output probability from the classification layer is taken as the model's predicted result. Through the combination of the fully connected layer and the Softmax function, the classification layer is able to effectively classify text based on the dense sentence vectors provided by the BERT model and make final judgments based on the predicted probabilities for each category.

2.4 Loss Function

In text classification tasks, the loss function is used to measure the difference between the model's predicted results and the true labels, guiding the model to continuously adjust its parameters during the training process to improve classification accuracy. In this model, we use the cross-entropy loss function to calculate the gap between the predicted results and the true labels. The cross-entropy loss function is commonly used for multi-class classification problems, and its definition is as follows:

$$Y_a = soft\max(f_a \cdot W + b) \tag{1}$$

$$L = CrossEntropy(y_{truth}, y_a) \tag{2}$$

For each sample, the cross-entropy loss function calculates the difference between the probability distribution predicted by the model and the true labels. By minimizing this loss function, the model can learn better parameters so that the predicted category probability distribution is as close as possible to the distribution of the true labels. During the training process, the model adjusts the weights based on the gradient information of the loss function through backpropagation, gradually optimizing the parameters of BERT and the classification layer. The cross-entropy loss function has good numerical stability, can effectively handle multi-class classification tasks, and promotes the convergence of the model.

3. Model Experiments

The Reuters Corpus, provided by the Natural Language Toolkit (NLTK) [2], contains 10,788 news articles with a total vocabulary of about 1.3 million. The dataset is divided into a training set and a test set, with the training set containing 7,769 articles and the test set containing 3,019 articles. Each news article can belong to multiple categories, labeled across 90 predefined categories, forming a multi-label

ISSN 2616-5775 Vol. 7, Issue 12: 91-95, DOI: 10.25236/AJCIS.2024.071213

classification task. This corpus provides a rich set of annotated data for text classification tasks, with each article's label numbers ranging from 1 to 15.

To verify the effectiveness of back-translation data augmentation methods, we conducted comparative experiments by dividing the model training into two groups. The first group is the baseline model, which is trained directly using the original training set. The second group uses the training set augmented with back-translation, utilizing GPT-3.5 to generate new samples through back-translation and incorporating these new samples into the training set for training. The performance of the two groups of models on the Reuters dataset was compared before and after the addition of data augmentation samples using metrics such as Accuracy and F1-score.

Table 1: Performance Comparison Before and After Data Augmentation

Dataset	Before		Behind	
	Acc(%)	F1(%)	Acc(%)	F1(%)
Reuters	57.17	49.87	65.63	59.72

As shown in Table 1, the experimental results indicate that the back-translation data augmentation method based on GPT-3.5 significantly improves the model's classification performance on small-scale datasets. The diversified samples generated through back-translation not only expand the size of the training set but also increase its diversity, thereby reducing the model's dependence on specific expressions. The large language model GPT-3.5 is capable of generating high-quality text during the back-translation process, maintaining semantic consistency while varying the expression forms of sentences, which enriches and diversifies the data in the training set. This process effectively enhances the generalization ability of the BERT model, enabling it to better handle unknown samples.

4. Summary and Outlook

In this paper, we proposed a back-translation data augmentation method based on large language models to enhance the performance of BERT in text classification tasks. The diversified samples generated through back-translation effectively enhanced the diversity of the training set, thereby improving BERT's understanding ability for text classification tasks. Through comparative experiments, we demonstrated the effectiveness of this method in small-scale dataset environments, especially on the Reuters News Classification dataset, where the model's classification accuracy was improved.

Large language models like GPT-3.5 indeed possess significant advantages in text data processing, particularly when it comes to enhancing the diversity and flexibility of training sets through back-translation data augmentation methods.This technique not only broadens the scale of the training set but also enriches its variety, which in turn reduces the model's reliance on specific expressions. The powerful text comprehension ability of large language models plays a crucial role in this process, enabling the generation of text that remains semantically consistent while varying in expression.

Furthermore, the use of large language models in back-translation can lead to the generation of new samples that are of high quality, which, when added to the training set, can significantly improve the model's ability to handle unknown samples. The effectiveness of this data augmentation method is evident in its ability to improve the model's performance on text classification tasks, as indicated by metrics such as accuracy and F1-score.Looking ahead, future research can delve deeper into combining large language models with other text generation technologies to further enhance model performance and application potential in classification tasks. This could involve exploring new methods of data augmentation, improving the integration of different models, and leveraging the strengths of various text generation technologies to create more robust and efficient classification models.

References

[1] Zhao H, Chen H, Ruggles T A, et al. Improving Text Classification with Large Language Model-Based Data Augmentation[J]. Electronics, 2024, 13(13): 2535.

ISSN 2616-5775 Vol. 7, Issue 12: 91-95, DOI: 10.25236/AJCIS.2024.071213

[2] Maharana K, Mondal S, Nemade B. A review: Data pre-processing and data augmentation techniques[J]. Global Transitions Proceedings, 2022, 3(1): 91-99.
[3] Mumuni A, Mumuni F. Data augmentation: A comprehensive survey of modern approaches[J]. Array, 2022, 16: 100258.
[4] Lu Z, Zhou A, Ren H, et al. Mathgenie: Generating synthetic data with question back-translation for enhancing mathematical reasoning of llms[J]. arXiv preprint arXiv:2402.16352, 2024.
[5] Nadi F, Naghavipour H, Mehmood T, et al. Sentiment Analysis Using Large Language Models: A Case Study of GPT-3.5[C]//The International Conference on Data Science and Emerging Technologies. Singapore: Springer Nature Singapore, 2023: 161-168.
[6] Ćavar D, Tiganj Z, Mompelat L V, et al. Computing Ellipsis Constructions: Comparing Classical NLP and LLM Approaches[C]//Proceedings of the Society for Computation in Linguistics 2024. 2024: 217-226.
[7] Yang JF, Jin HY, Tang RX et al (2023) Harnessing the power of LLMs in practice: a survey on ChatGPT and beyond. ACM Trans Knowl Discov Data 18(6):1–32. https://doi.org/10.1145/3649506