Journal of Machine Learning Research 1 (2024) 1-20 Submitted 4/24; Published –/–

Improving Sentiment Analysis Accuracy using RoBERTa with Synonym-Based Data Augmentation

Rahul Krishnan A A.RAHULKRISHNAN14@GMAIL.COM
Hariharan V
Kavin M
Department of Computer Science Engineering
Dr. Mahalingam College of Engineering and Technology
Pollachi, India

Editor: Mr. P. Boopathi Rajan

Abstract

Sentiment analysis is a fundamental task in natural language processing with applications spanning various industries. This study investigates the efficacy of employing synonym-based data augmentation using WordNet to enhance the accuracy of the RoBERTa model trained on the IMDB movie reviews dataset. Empirical results demonstrate a noticeable improvement in model accuracy, from 90.81% without augmentation to 93% with augmentation. The study highlights the potential of lexical data augmentation techniques in improving transformer model performance.

Keywords: Sentiment Analysis, RoBERTa, Data Augmentation, WordNet, Transformer Models, IMDB Dataset

1 Introduction

Sentiment analysis has gained significant attention due to its widespread applicability in understanding consumer opinions, market trends, and social media monitoring. Transformer models, particularly RoBERTa, have emerged as state-of-the-art solutions. However, their performance is heavily influenced by training data quantity and diversity. Data augmentation methods such as synonym replacement offer viable strategies for enhancing data richness. This paper evaluates the effectiveness of synonym-based augmentation on RoBERTa's sentiment analysis capabilities Liu et al. (2019); Wei and Zou (2019).

2 Background

Recent advancements in natural language processing have shifted from traditional machine learning approaches to transformer-based deep learning models. Among them, BERT and its derivatives like RoBERTa have shown superior contextual understanding by leveraging attention mechanisms and large-scale pretraining. However, these models are still data-hungry and susceptible to overfitting in low-resource or domain-specific tasks. Data augmentation addresses this gap by synthetically increasing training diversity.

3 Methodology

3.1 Dataset

The widely-used IMDB dataset comprising 50,000 movie reviews equally split between positive and negative sentiments was utilized. The reviews were preprocessed to remove HTML tags and unnecessary characters.

3.2 Data Augmentation Technique

WordNet, a comprehensive lexical database, was used for synonym replacement in text augmentation. For each review, nouns, verbs, and adjectives were randomly selected and replaced with their synonyms, ensuring contextual relevance. The augmented data was merged with the original to form a training set twice the size.

3.3 RoBERTa Model and Training

The RoBERTa-base model was fine-tuned using binary cross-entropy loss, the AdamW optimizer (learning rate = 3e-5, batch size = 32, epochs = 3). Tokenization was handled using Huggingface's AutoTokenizer with a maximum sequence length of 128 tokens.

4 Experimental Results

4.1 Baseline Model Performance

Without augmentation, RoBERTa achieved an accuracy of 90.81%. Precision, recall, and F1-score were 89%, 94%, and 91%, respectively, with ROC AUC at 0.9686. The confusion matrix showed high accuracy in predicting both positive and negative classes.

4.2 Augmented Model Performance

Synonym-based augmentation improved accuracy to 93%. Precision increased to 92%, recall to 93%, F1-score to 92%, and ROC AUC to 0.97, indicating significant performance enhancements. Augmented models also showed more stable validation loss across epochs.

4.3 Visualization

Training curves, ROC plots, and a bar chart comparing the baseline and augmented models were generated using Matplotlib and Seaborn. The augmented model demonstrated both higher peak performance and more stable learning dynamics.

5 Discussion

The performance improvement underscores the importance of data augmentation in mitigating overfitting and improving generalization capabilities of transformer-based NLP models. The use of synonym replacement is particularly effective due to its simplicity and linguistic soundness. However, care must be taken to avoid introducing semantic drift.

6 Conclusion and Future Work

The study confirms that synonym-based data augmentation significantly boosts RoBERTa's sentiment analysis accuracy. Future research could investigate more sophisticated augmentation strategies such as back-translation, contextual word embedding replacement, and adversarial training. Further, cross-domain testing could validate the generalizability of the improved model.

Acknowledgments and Disclosure of Funding

We thank Mr. P. Boopathi Rajan, Department of Computer Science Engineering, Dr. Mahalingam College of Engineering and Technology, for his guidance and insightful discussions.

References

Yinhan Liu, Myle Ott, Naman Goyal, et al. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.