Sentiment Analysis of IMDB Movie Reviews Using Few-Shot Learning with SetFit

Malak Yasser
malak.eid.2023@aiu.edu.eg

Fares Hassan
fares.ghoniem.2023@aiu.edu.eg

May 2025

Abstract

Sentiment analysis is a critical task in natural language processing, with applications in understanding user opinions across various domains. This paper presents a study on sentiment classification of IMDB movie reviews using the SetFit classifier, a few-shot learning approach based on sentence transformers. We preprocess the IMDB 50k dataset, apply text cleaning and normalization techniques, and train the SetFit model on a small subset of 200 samples. The model achieves a test accuracy of 72% on 100 test samples, demonstrating the efficacy of few-shot learning in sentiment analysis with limited labeled data. We discuss the methodology, evaluate the model's performance using a confusion matrix, and highlight its potential for efficient sentiment classification in resource-constrained settings.

1 Introduction

Sentiment analysis, a subfield of natural language processing (NLP), aims to determine the emotional tone expressed in text, such as positive or negative sentiments. It has widespread applications in areas like product reviews, social media analysis, and customer feedback systems. Traditional sentiment analysis models often require large labeled datasets to achieve high performance, which can be costly and time-consuming to obtain. Few-shot learning, which leverages small amounts of labeled data, offers a promising alternative for scenarios with limited resources.

This study explores the application of the SetFit classifier, a few-shot learning model based on sentence transformers, for sentiment analysis on the IMDB 50k movie reviews dataset. The dataset contains 50,000 reviews labeled as positive or negative, providing a balanced benchmark for binary sentiment classification. We preprocess the data, train the SetFit model on a subset of 200 samples, and evaluate its performance on a test set of 100 samples. Our results demonstrate that SetFit can achieve competitive performance with minimal training data, making it a viable option for efficient NLP tasks.

2 Dataset and Preprocessing

2.1 Dataset

The IMDB 50k movie reviews dataset is used in this study. It consists of 25,000 training and 25,000 test samples, each labeled as positive ("pos") or negative ("neg"). The dataset

is balanced, with an equal number of positive and negative reviews, as confirmed by the value counts showing 12,500 samples per class in the training set.

2.2 Preprocessing

To prepare the data for modeling, we applied the following preprocessing steps:

1. Column Splitting: The original dataset has a single column combining text and sentiment, separated by commas. We split this into separate "text" and "sentiment" columns using the rsplit method with a limit of one split to handle cases where the text contains commas.

2. Text Cleaning: We implemented a clean_text function to:

   - Convert text to lowercase.
   - Remove HTML tags, URLs, and non-alphabetic characters (except spaces and apostrophes) using regular expressions.
   - Replace multiple spaces with a single space and strip leading/trailing spaces.

3. Text Normalization: We further normalized the text by:

   - Tokenizing the cleaned text.
   - Removing English stop words using NLTK's stopword list.
   - Applying lemmatization using NLTK's WordNetLemmatizer to reduce words to their base forms.

4. Label Encoding: Sentiment labels ("pos" and "neg") were encoded as numerical values (1 and 0, respectively) using scikit-learn's LabelEncoder.

5. Data Subsampling: To simulate a few-shot learning scenario, we randomly sampled 200 training samples and 100 test samples using a fixed random seed for reproducibility.

These steps ensured that the text was standardized and suitable for input to the SetFit classifier.

3 Methodology

3.1 SetFit Classifier

The SetFit classifier is a few-shot learning approach that combines sentence transformers with a logistic regression head. It leverages the paraphrase-MiniLM-L3-v2 model to generate dense text embeddings, which are then used to train a lightweight classifier. SetFit is particularly effective for tasks with limited labeled data, as it requires fewer samples to achieve robust performance compared to traditional deep learning models.

3.2 Training and Evaluation

We trained the SetFit classifier on the 200 sampled training examples, using the normalized text and encoded sentiment labels. The model was trained with default hyperparameters, performing one epoch of contrastive learning followed by logistic regression fine-tuning. We evaluated the model on the 100 sampled test examples, computing the accuracy as the proportion of correctly predicted sentiments. Additionally, we calculated the training accuracy to assess potential overfitting. A confusion matrix was generated to analyze the model's performance across positive and negative classes, visualized using a heatmap.

4 Results

The SetFit classifier achieved a test accuracy of 72% on the 100-sample test set, indicating reasonable performance given the small training set size. The training accuracy was significantly higher at 98%, suggesting that the model fits the training data well but may suffer from overfitting due to the limited sample size.

The confusion matrix provides further insight into the model's performance:

- True Negatives (TN): 34 – Correctly predicted negative sentiments.
- False Positives (FP): 15 – Negative sentiments incorrectly predicted as positive.
- False Negatives (FN): 13 – Positive sentiments incorrectly predicted as negative.
- True Positives (TP): 38 – Correctly predicted positive sentiments.

The confusion matrix indicates that the model correctly classified 34 negative and 38 positive reviews, but misclassified 15 negative reviews as positive and 13 positive reviews as negative. This suggests a slight bias toward positive predictions, as the number of false positives (15) is higher than false negatives (13). The balanced accuracy, considering both classes, aligns with the reported 72% test accuracy.

5 Discussion

The results demonstrate that the SetFit classifier can effectively perform sentiment analysis with minimal labeled data, achieving a test accuracy of 72% with only 200 training samples. The confusion matrix reveals that the model performs slightly better at identifying positive sentiments (38 TP) than negative ones (34 TN), with a moderate number of misclassifications (15 FP, 13 FN). The high training accuracy (98%) indicates strong fitting to the training data, but the gap between training and test accuracy suggests overfitting, a common challenge in few-shot learning with small datasets.

The preprocessing steps, including text cleaning and normalization, were crucial in preparing the data for the SetFit model. Removing noise (e.g., HTML tags, URLs) and standardizing word forms through lemmatization likely improved the quality of the text embeddings. However, the reliance on a small subset of the dataset may limit the model's ability to generalize to the full range of linguistic patterns in the IMDB dataset.

Future work could explore the following:

- Hyperparameter Tuning: Adjusting SetFit's training parameters, such as the number of epochs or contrastive learning iterations, could improve generalization.
- Larger Training Sets: Gradually increasing the training set size could help assess the trade-off between data efficiency and performance.
- Advanced Preprocessing: Incorporating techniques like named entity recognition or sentiment-specific tokenization could enhance feature extraction.
- Model Comparison: Comparing SetFit with other few-shot learning models or traditional supervised models could provide a broader perspective on its effectiveness.

6 Conclusion

This study demonstrates the potential of the SetFit classifier for sentiment analysis on the IMDB 50k movie reviews dataset using a few-shot learning approach. With only 200 training samples, the model achieved a test accuracy of 72%, with a confusion matrix showing balanced performance across positive and negative classes. While the high training accuracy indicates strong fitting to the training data, the test performance suggests room for improvement in generalization. These findings underscore the value of few-shot learning for NLP tasks and provide a foundation for further research into optimizing SetFit for sentiment analysis.