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

This paper explores sentiment analysis on the IMDB 50k movie reviews dataset using the SetFit classifier, a few-shot learning approach based on sentence transformers. We preprocess the dataset, train the model on 200 samples, and evaluate it on 100 samples, achieving a test accuracy of 72%. The results, including a confusion matrix, demonstrate the efficacy of few-shot learning for sentiment classification with limited data. We discuss the methodology, performance, limitations, and future directions for improving generalization in resource-constrained settings.

1 Introduction

Sentiment analysis, a key task in natural language processing (NLP), involves classifying text based on its emotional tone, such as positive or negative sentiments. It is widely applied in domains like product reviews, social media monitoring, and customer feedback analysis. However, traditional sentiment analysis models often require large labeled datasets, which can be costly and time-consuming to collect, posing a challenge in scenarios with limited labeled data.

The objective of this study is to investigate the effectiveness of few-shot learning for sentiment analysis, specifically using the SetFit classifier on the IMDB 50k movie reviews dataset. By training on a small subset of 200 samples, we aim to achieve robust performance with minimal data, addressing the problem of data scarcity in NLP tasks. This work evaluates the model's accuracy and analyzes its performance through a confusion matrix, providing insights into its applicability in low-resource settings.

2 Related Work

Sentiment analysis has been extensively studied, with early approaches relying on lexicon-based methods and rule-based systems. The advent of deep learning introduced models like recurrent neural networks (RNNs) and transformers, which significantly improved performance but required large datasets. The IMDB 50k dataset has been a standard benchmark for binary sentiment classification, often used with models like BERT and LSTM-based architectures.

Few-shot learning has emerged as a solution for data-scarce scenarios. Sentence transformers, which generate dense text embeddings, have been adapted for tasks like sentiment analysis. The SetFit classifier, combining sentence transformers with a logistic regression head, has shown promise in few-shot text classification by leveraging contrastive learning to maximize performance with minimal labeled data. Unlike traditional supervised learning, SetFit requires fewer examples, making it suitable for our study's focus on efficient sentiment analysis.

3 Methodology

3.1 System Description

The system employs the SetFit classifier, a few-shot learning model that uses the 'paraphrase-MiniLM-L3-v2' sentence transformer to generate text embeddings, followed by a logistic regression classifier. SetFit's contrastive learning approach optimizes embeddings for class separation, enabling effective classification with limited data.

3.2 Dataset

The IMDB 50k movie reviews dataset comprises 50,000 reviews, split evenly into 25,000 training and 25,000 test samples, with balanced positive and negative sentiments. For this study, we randomly sampled 200 training and 100 test samples to simulate a few-shot learning scenario, using a fixed random seed for reproducibility.

3.3 Implementation

The implementation involves the following steps:

1. Data Preprocessing:
    - Split the dataset's single column into "text" and "sentiment" using rsplit.
    - Clean text by converting to lowercase, removing HTML tags, URLs, and non-alphabetic characters (except spaces and apostrophes) via regular expressions, and normalizing spaces.
    - Normalize text by tokenizing, removing English stop words using NLTK, and applying lemmatization with WordNetLemmatizer.
    - Encode sentiment labels ("pos" = 1, "neg" = 0) using scikit-learn's LabelEncoder.
2. Model Training: Train the SetFit classifier on 200 samples with default hyperparameters (one epoch of contrastive learning, followed by logistic regression fine-tuning).
3. Evaluation: Assess performance on 100 test samples using accuracy and a confusion matrix, visualized as a heatmap with seaborn.

The implementation is coded in Python using libraries such as pandas, NLTK, scikit-learn, and setfit, executed in a Jupyter Notebook environment.

4 Experiments

4.1 Experimental Setup

The SetFit classifier was trained on 200 randomly sampled training examples and evaluated on 100 test samples from the IMDB dataset. The model used the 'paraphrase-MiniLM-L3-v2' backbone with default settings. Training involved one epoch of contrastive learning followed by logistic regression. The random seed was fixed for reproducibility.

4.2 Metrics

Performance was evaluated using:

- Accuracy: Proportion of correctly predicted sentiments.
- Confusion Matrix: Breakdown of true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP) to assess classification performance across classes.

4.3 Results

The model achieved a test accuracy of 72% and a training accuracy of 98%. The confusion matrix is as follows:

- 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 results indicate reasonable performance for a few-shot setting, with a slight bias toward positive predictions (FP $>$ FN). The high training accuracy suggests strong fitting to the training data.

5 Discussion

5.1 Interpretation

The 72% test accuracy demonstrates that SetFit can effectively classify sentiments with only 200 training samples, highlighting its efficiency in few-shot learning. The confusion matrix shows balanced performance, with 38 true positives and 34 true negatives, though the model slightly favors positive predictions (15 FP vs. 13 FN). The 98% training accuracy indicates excellent learning on the training set but suggests overfitting due to the small sample size.

5.2 Limitations

The primary limitation is the small training set, which may not capture the full linguistic diversity of the IMDB dataset, contributing to overfitting. The default hyperparameters may not be optimal for this specific task, and the model's slight positive bias could affect performance on datasets with different sentiment distributions.

Future Work Future improvements could include:

- Tuning SetFit hyperparameters (e.g., epochs, learning rate) to enhance generalization.
- Increasing the training set size to balance data efficiency and performance.
- Exploring advanced preprocessing, such as sentiment-specific tokenization or named entity recognition.
- Comparing SetFit with other few-shot or supervised models to benchmark its effectiveness.

6 Conclusion

This study demonstrates the potential of the SetFit classifier for sentiment analysis on the IMDB 50k dataset in a few-shot setting. With only 200 training samples, the model achieved a 72% test accuracy, supported by a confusion matrix showing balanced classification. While the high training accuracy suggests overfitting, the results highlight SetFit's efficiency for low-resource NLP tasks. Future work on hyperparameter optimization and larger datasets could further improve performance.