IMDB Movie Reviews Sentiment Analysis
Using Logistic Regression and TF-IDF

Muhammad Salman

August 26, 2025

Abstract

This report presents a sentiment analysis project on the IMDB movie review dataset, which contains 50,000 reviews labeled as either *positive* or *negative*. The project involves downloading the dataset, preprocessing the text, converting the reviews into numerical features using TF-IDF, training a Logistic Regression classifier, and evaluating its performance using accuracy, classification metrics, and confusion matrix. Visualization of results is also included.

1 Introduction

Sentiment analysis is an important application of Natural Language Processing (NLP). The IMDB dataset is widely used for benchmarking NLP algorithms, as it contains a balanced set of movie reviews labeled with sentiments. The goal is to build a machine learning pipeline that can classify unseen reviews as **positive** or **negative**.

2 Dataset

The dataset was downloaded from Kaggle using the `kagglehub` library. It consists of 50,000 reviews equally divided between positive and negative classes.

- Dataset size: 50,000 reviews
- Labels: `positive`, `negative`
- Example distribution shown in Figure 1.

3 Methodology

The methodology follows a typical NLP pipeline:

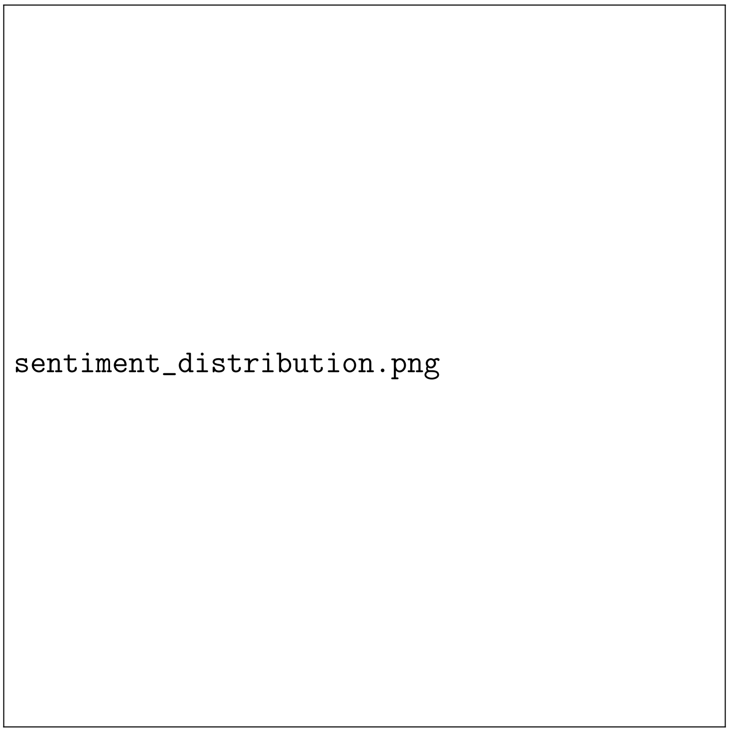

Figure 1: Distribution of sentiments in the dataset (balanced).

3.1 Preprocessing

1. Removal of HTML tags using `BeautifulSoup`.
2. Removal of non-alphabetical characters using Regular Expressions.
3. Conversion to lowercase.
4. Stopword removal using NLTK's English stopword list.

3.2 Feature Extraction

The reviews were transformed into numerical vectors using the TF-IDF (Term Frequency - Inverse Document Frequency) method with a maximum of 5000 features.

3.3 Model Training

A Logistic Regression classifier was trained with:

- `max_iter = 200`
- Train-test split: 80%-20%

3.4 Evaluation

The trained model was evaluated using:

- Accuracy score

- Classification Report (Precision, Recall, F1-score)
- Confusion Matrix

4 Results

The model achieved high accuracy on the test set. A classification report and confusion matrix were generated.

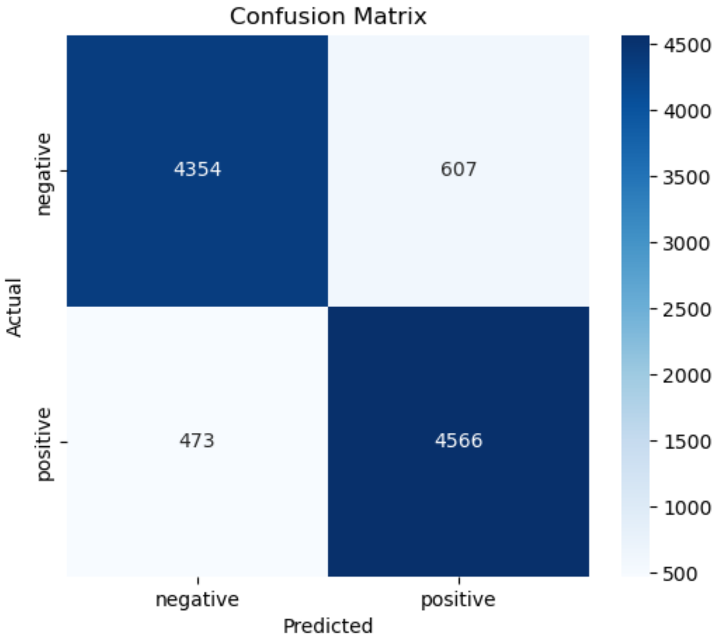

Figure 2: Confusion Matrix of Logistic Regression model on test data.

5 Custom Review Prediction

To test the model on new data, a sample review was used:

> "This movie was absolutely fantastic, I loved it!"

The model predicted the sentiment as **positive**, which matches the expected outcome.

6 Conclusion

This project successfully built a sentiment analysis system using Logistic Regression with TF-IDF features. The model achieved strong performance in classifying IMDB reviews. Future improvements could include experimenting with deep learning models (e.g., LSTM, BERT) to capture contextual information.

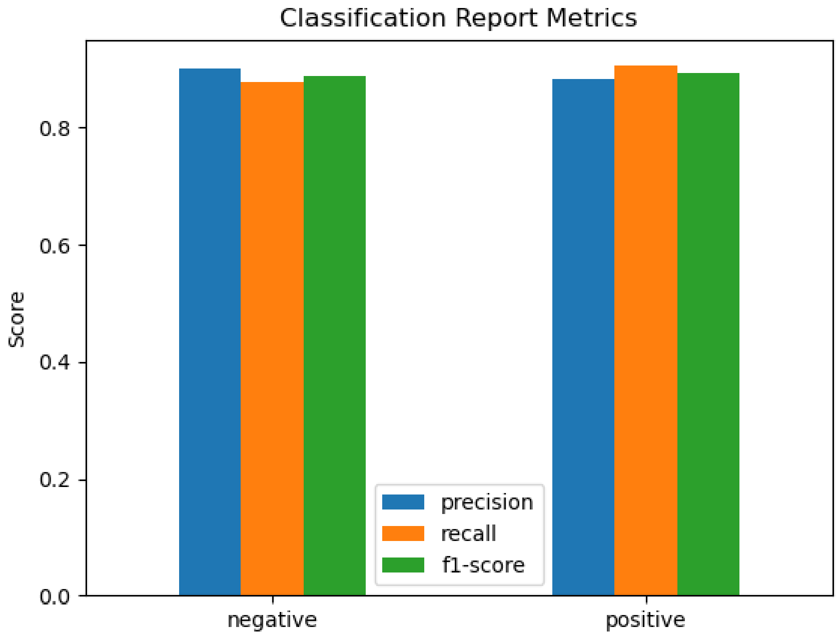

Figure 3: Classification metrics (Precision, Recall, F1-score) for positive and negative classes.

References

- IMDB Dataset: `https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k`
- NLTK Stopwords: `https://www.nltk.org/`
- Scikit-learn Documentation: `https://scikit-learn.org/`