International Centre for Education and Research (ICER)
VIT – Bangalore

Sentiment Analysis Using Python and Machine Learning

Submitted by

Devjyot Singh Sidhu

(24MSP3075)

Abstract

Sentiment analysis is an essential tool in natural language processing that determines the sentiment of a given piece of text. This project focuses on analysing sentiments from textual data using Python and machine learning techniques. The IMDB movie review dataset, containing 50,000 reviews, was used to build and evaluate a model that classifies reviews as positive or negative. We employed text preprocessing, feature extraction using the Bag of Words approach, and trained a Naive Bayes classifier. The model achieved 85% accuracy, demonstrating its effectiveness in distinguishing between positive and negative reviews. However, limitations were observed in handling sarcasm and ambiguous text, which could be addressed in future work.

Problem Statement

Sentiment analysis is widely used in industries to to gauge customer satisfaction, monitor brand reputation, and derive actionable insights. Companies rely on user-generated data, such as reviews and social media posts, to understand customer sentiments and improve their products or services.

Literature Review

Traditional Approaches[1]

Early sentiment analysis relied on lexicon-based techniques, where predefined dictionaries of words with positive or negative sentiments were used. While simple, these methods lacked contextual understanding.

Literature Review

Machine Learning Approaches[2][4][5]

Machine learning introduced algorithms like Naive Bayes[3], Support Vector Machines, and Logistic Regression. These models use features extracted from text (e.g., frequency of words) to classify sentiments, offering better accuracy than lexicon-based methods.

Literature Review

Modern Deep Learning Approaches[3]

Recent advancements like Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and transformers (e.g., BERT) enable deeper contextual understanding. However, these methods require substantial computational resources.

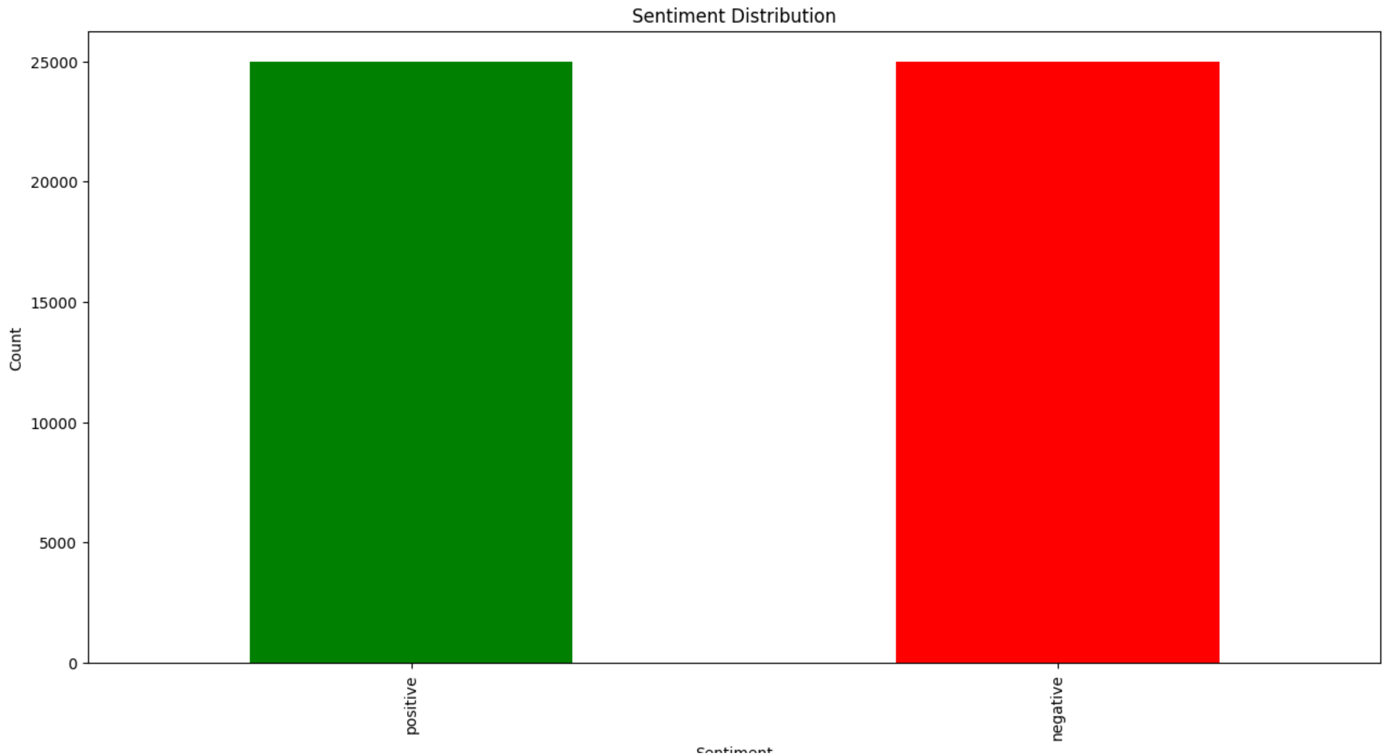
Methodology

Dataset Description

The IMDB dataset contains 50,000 movie reviews equally divided into positive and negative sentiments. Each review is a string of text, varying from a few words to several paragraphs.

Methodology

Review Length Analysis of dataset



Methodology

Feature Extraction

We used the Bag of Words (BoW) technique to convert text into a numerical matrix. The vectorization step transformed each review into a fixed-length representation, where each column represents the frequency of a specific word in the vocabulary.

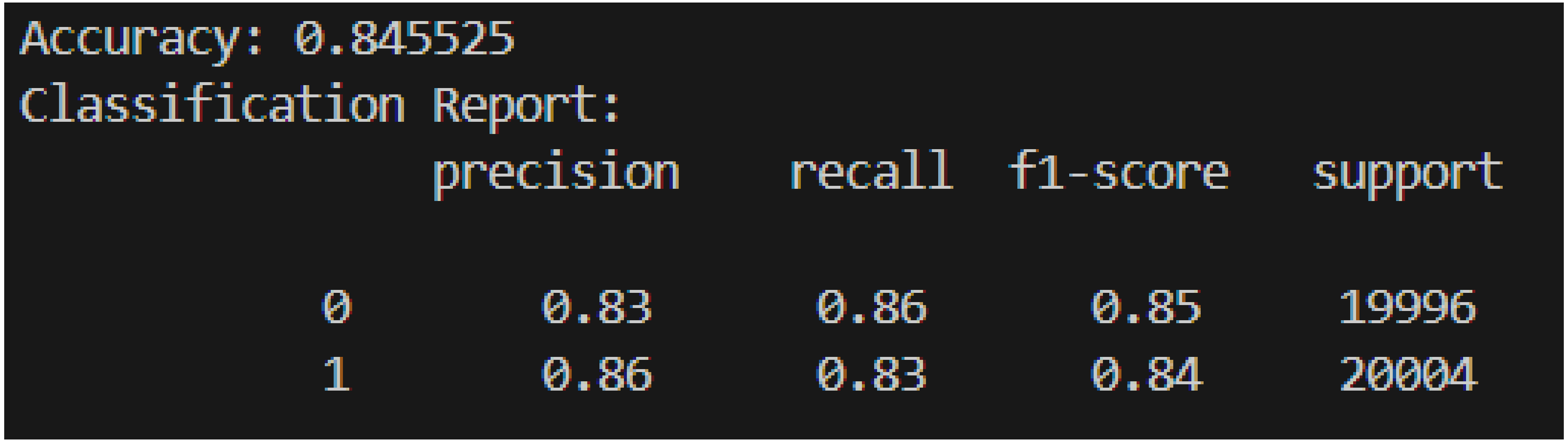
Methodology

Naïve Bayes Classifier

- Ideal for text classification tasks.
- Efficient with small datasets and sparse matrices like BoW.

Results and Analysis

Evaluation Metrices

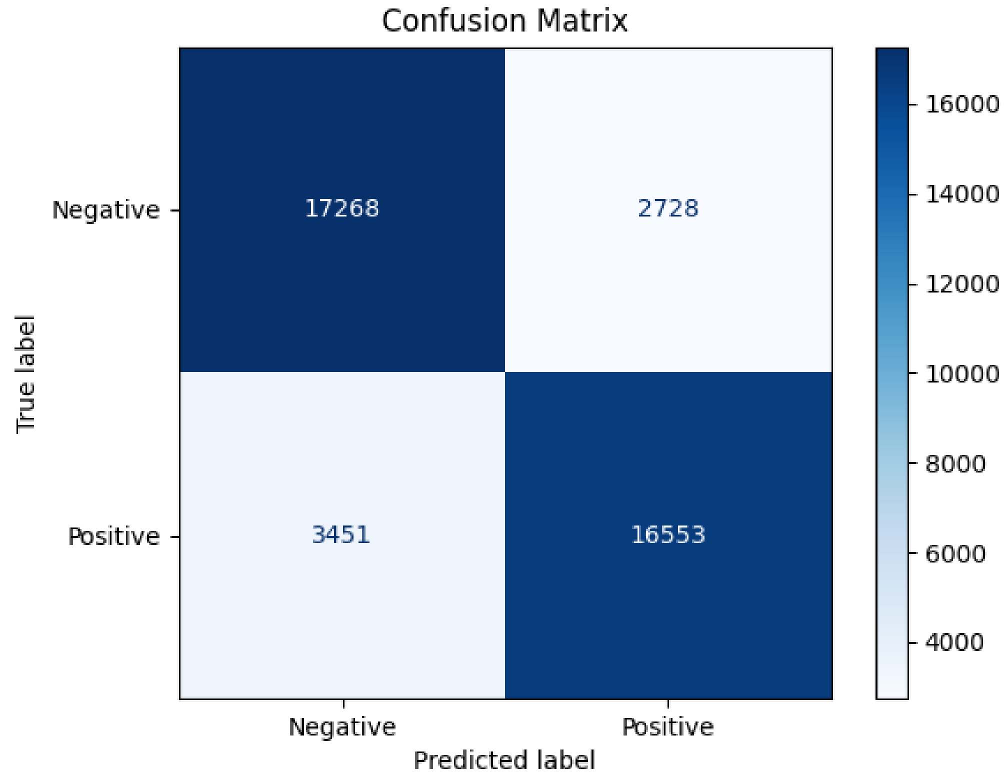


```
Accuracy: 0.845525
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.86      0.85     19996
           1       0.86      0.83      0.84     20004
```

Results and Analysis

Confusion Matrix



Discussion

Strengths

1. High accuracy for classifying text reviews.
2. Scalable preprocessing pipeline.

Discussion

Limitations

1. Struggles with ambiguous sentences and sarcasm.

2. Cannot generalize to languages other than English.

Discussion

Future Improvements

1. Explore pre-trained models like BERT or GPT for contextual understanding.

2. Implement techniques to handle sarcasm and multilingual datasets.

Conclusion

This project successfully demonstrates the implementation of sentiment analysis using the IMDB dataset. The Naive Bayes classifier provides a baseline model with good accuracy and scalability. Future work could explore deep learning models like BERT for better contextual understanding.

References

1. Shaukat, Z., Zulfiqar, A.A., Xiao, C. et al. Sentiment analysis on IMDB using lexicon and neural networks. SN Appl. Sci. 2, 148 (2020)

2. Pang, B., & Lee, L. (2004). "A Sentimental Education: Sentiment Analysis Using Machine Learning Techniques." Presented at the Association for Computational Linguistics (ACL).

3. Zhang, L., Wang, S., & Liu, B. (2018). Deep Learning for Sentiment Analysis: A Survey. Wiley Interdisciplinary Reviews.

4. "Sentiment Analysis Using Machine Learning and Deep Learning Models on IMDB Dataset." IEEE Xplore, 2024.

5. "A Literature Review on Application of Sentiment Analysis Using Machine Learning." SSRN Papers, 2023.

6. "Sentiment Analysis: Machine Learning Approaches Comparison." IEEE Xplore, 2024.

7. Tumasjan, A., et al. (2010). "Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment." Proceedings of the International Conference on Web and Social Media (ICWSM).

8. "Optimization of Sentiment Analysis Using Machine Learning Classifiers." SpringerOpen, 2023.

9. Kaggle Dataset: *https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews*