Final Report: Sentiment Analysis on IMDb Movie Reviews

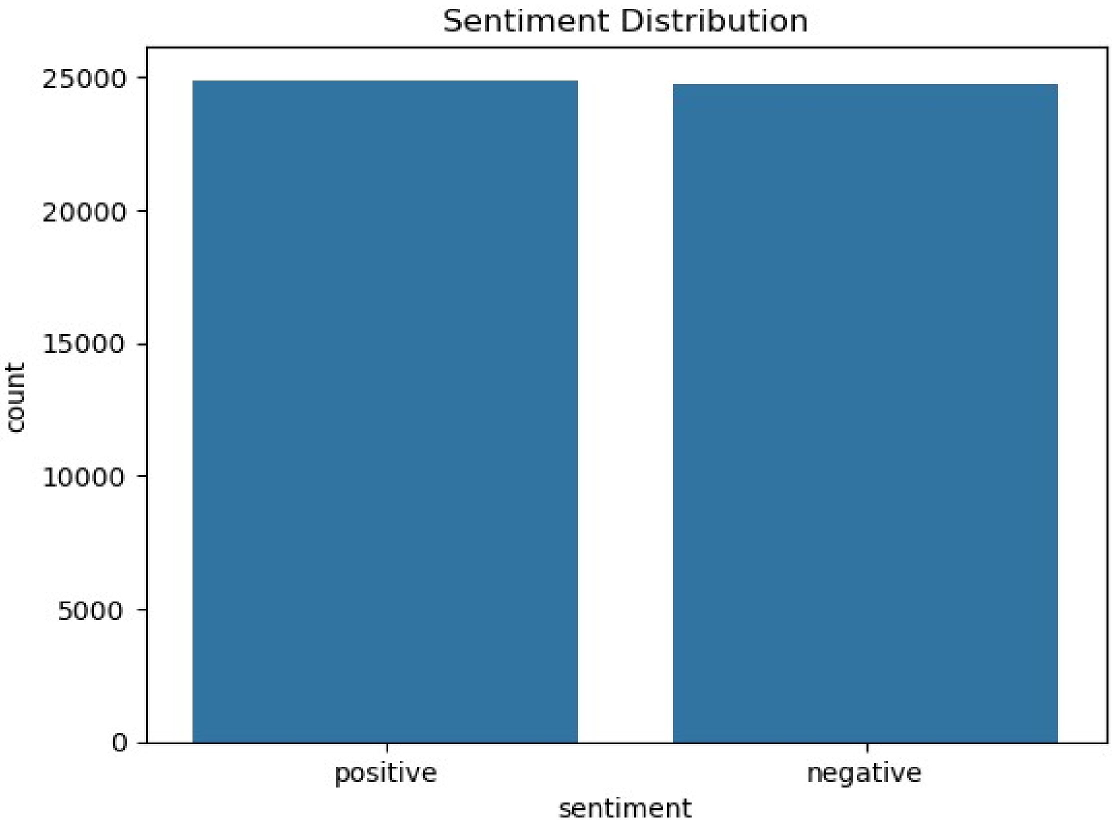
1. Introduction

This project focuses on performing sentiment analysis on IMDb movie reviews. The goal is to classify reviews as either positive or negative based on their content. Natural Language Processing (NLP) techniques are used to clean and preprocess the data, followed by model training and evaluation.

2. Data Exploration

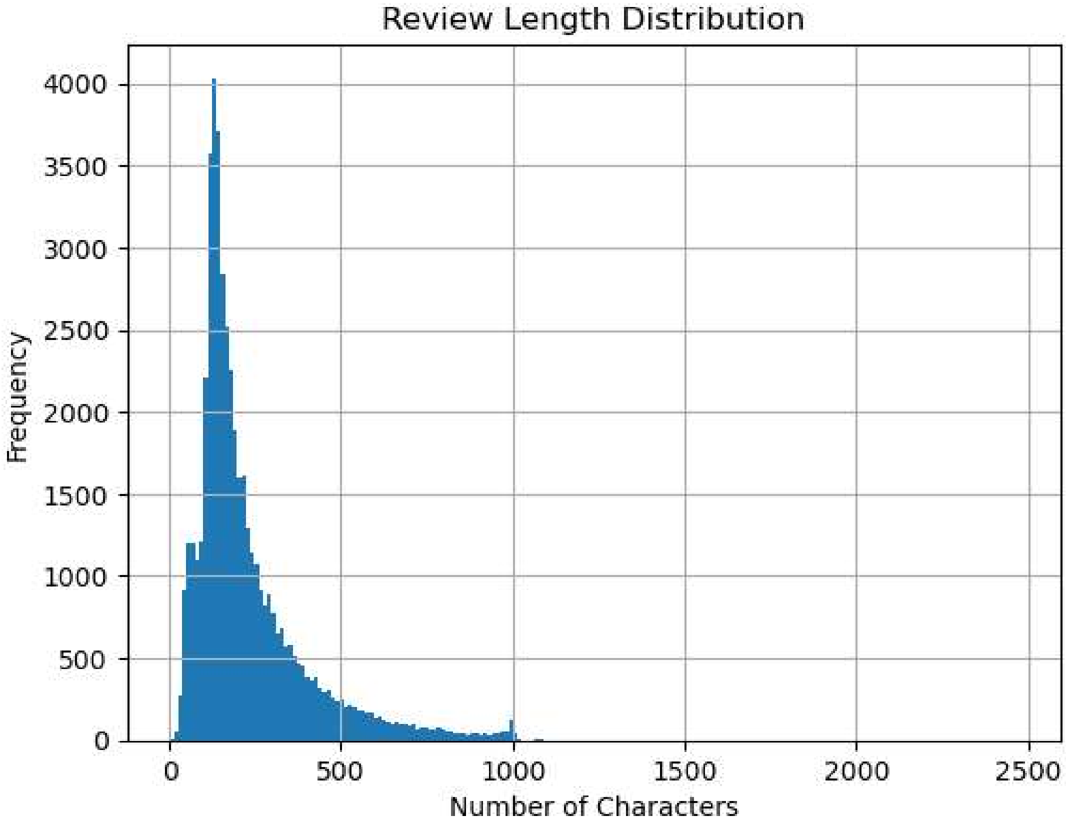
The dataset used contains two columns: **review and sentiment.** Key exploration steps included:

- Displaying the first few entries to understand structure of the data set using pandas "**read_csv**" method.
- Calculating total number of rows and columns or the **shape** of the dataframe.
- Identifying number of unique reviews and unique sentiment labels along with their distribution in the form of a count-plot chart.

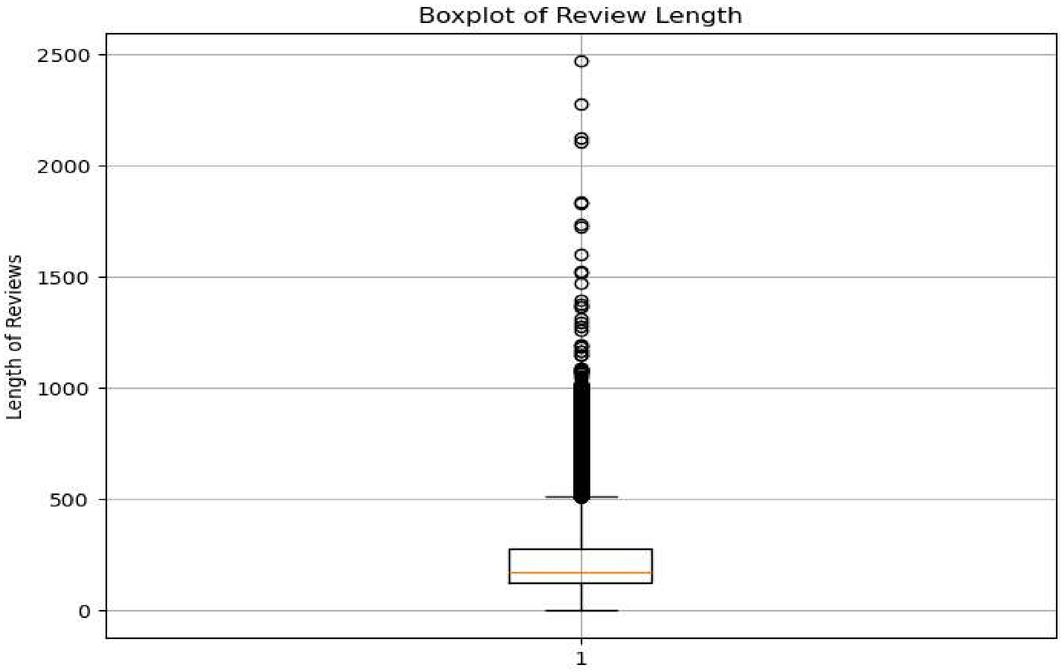


- Checking for missing values and data types of both the columns and converting them to str data type.

- Checking the distribution of number of words in the review column using a histogram.



- Finding outliers by plotting a box plot of number of words in review column per document. Then removing the outliers by fixing the maximum value to quartile 3.



3. Data Preprocessing

To prepare the textual data for modeling, several NLP preprocessing steps were applied:

- Tokenization using nltk.word_tokenize
- Lowercasing all words
- Removal of punctuation and stopwords
- Lemmatization using WordNetLemmatizer

This resulted in cleaned and standardized textual data suitable for model training.

4. Feature Engineering

- Calculated new columns for word count, character count and average word length.
- Vectorizing the review column using TF-IDF (TfidfVectorizer) with max_features as 5000.

5. Model Building and Selection

Train-test split was performed to evaluate model performance
Several classification models were tested:

- Logistic Regression- Accuracy score-88.31%
- Naive Bayes-Accuracy Score- 85.33%
- Support Vector Machines (SVM)- Accuracy Score- 87.61%
- Random Forest Model- Accuracy Score- 83.72%

, and accuracy scores were computed.
At last Logistic regression was the best performing model with accuracy score of 88.31%.

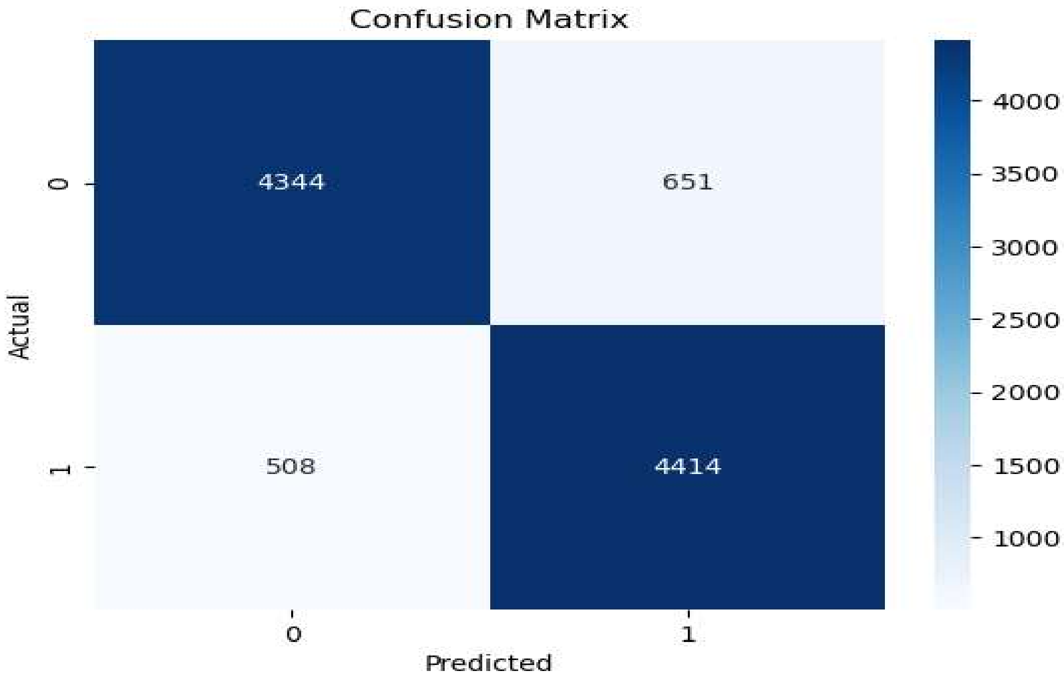
6. Model Evaluation

Logistic Regression model's performance was assessed using:

- Accuracy Score – 88.31%
- Classification Report (Precision, Recall, F1 Score)

	Precision	**Recall**	**F1-Score**	**Support**
Negative	0.90	0.87	0.88	4995
Positive	0.87	0.90	0.88	4922
Accuracy			0.88	9917
Macro avg	0.88	0.88	0.88	9917
Weighted avg	0.88	0.88	0.88	9917

- Confusion Matrix-



Confusion matrix detailed review: High True Positive (4411) and True Negative (4347) values indicate strong model performance on both classes. The majority of predictions are correct, suggesting good overall accuracy.

The logistic classification model showed the highest accuracy, followed closely by SVM.

7. Conclusion

This project successfully demonstrated the process of sentiment classification using IMDb reviews. NLP preprocessing was essential for transforming raw text into useful features, and machine learning classifiers performed well in distinguishing between positive and negative sentiments. SVM emerged as the best-performing model.