Punjab University College of Information Technology

Project Title:

Sentiment Analysis on Movie Reviews

Submitted By:

Group No 3

Sajeela Safdar	**BCSF22M001**
Areeba Abdullah	**BCSF22M004**
Omama Arshad	**BCSF22M035**
Hiba Noor	**BCSF22M036**
Ayesha Yaqoob	**BCSF22M037**

Abstract:

This project focuses on performing **sentiment analysis** on movie reviews using a transformer-based deep learning model. The objective was to classify reviews as either positive or negative.

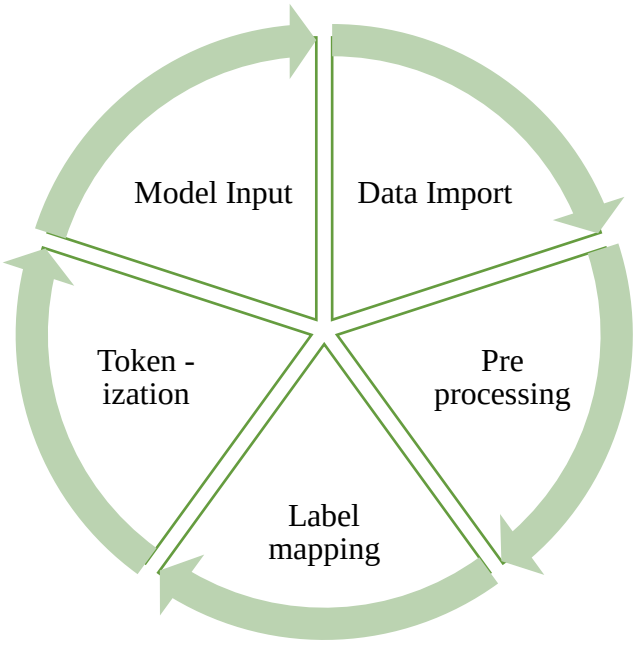
A labelled dataset of movie reviews was pre-processed to retain relevant columns and convert textual sentiments into binary labels. The data was then split into training and validation sets.

To leverage the power of state-of-the-art language models, the DistilBERT model—a lighter and faster version of BERT—was employed.

Tokenization was handled using the corresponding tokenizer, and model fine-tuning was performed using the Hugging Face Trainer API with appropriate training arguments.

Evaluation metrics such as accuracy and classification reports were used to assess model performance.

The **results** demonstrated that the fine-tuned DistilBERT model achieved strong accuracy, indicating its effectiveness in understanding and classifying human sentiment in text data.



Problem Statement:

"To develop an automated system that accurately classifies the sentiment of movie reviews as positive or negative".

Project's Significance:

- Sentiment analysis plays a vital role in areas such as customer service, market research, product improvement, and social media monitoring.

- By leveraging modern natural language processing (NLP) techniques, this project demonstrates how deep learning and pre-trained transformer models can efficiently interpret human emotions from text.
- The results not only validate the effectiveness of transformer models in sentiment classification but also offer a scalable and practical solution for real-world applications.

Algorithm/Models used:

- This project utilizes the DistilBERT model, a transformer-based architecture that is a smaller, faster, and more efficient version of BERT (Bidirectional Encoder Representations from Transformers). DistilBERT maintains over 95% of BERT's language understanding capabilities while being 40% smaller and 60% faster, making it suitable for tasks like sentiment classification.

Justification of chosen method:

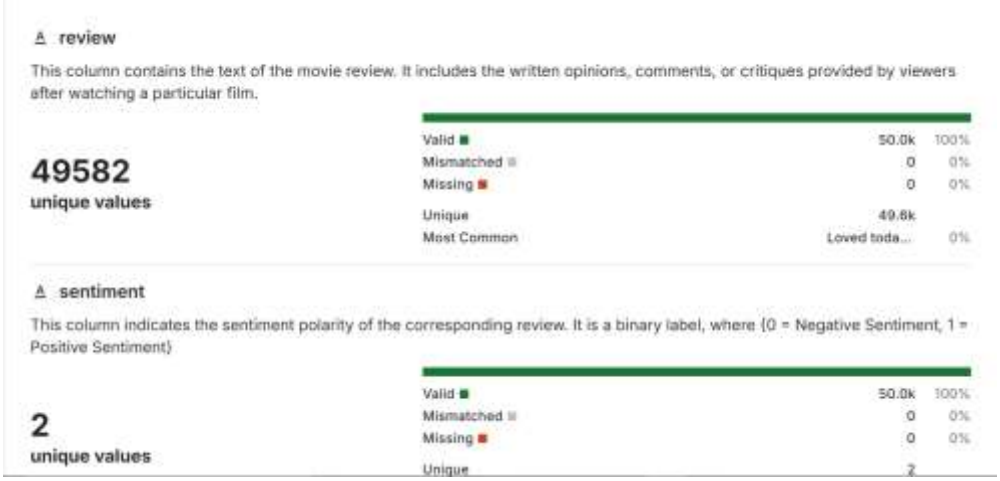
- DistilBERT was selected due to its balance of speed and accuracy, especially for sentiment analysis tasks involving natural language understanding.
- Unlike traditional machine learning models or shallow neural networks, transformer-based models like DistilBERT capture contextual relationships between words, which significantly improve performance on text classification problems.

Data set description:

The dataset used is a **CSV file containing 5,000 labelled movie reviews**, downloaded from **Kaggle**. Each entry consists of a review and its corresponding sentiment label (positive or negative). Preprocessing steps included:

- Removing missing values
- Renaming columns for consistency
- Mapping sentiment labels to binary values (positive = 1, negative = 0)
- Tokenizing text using DistilBERT's tokenizer

This prepared data was then split into training and validation sets for fine-tuning the model.

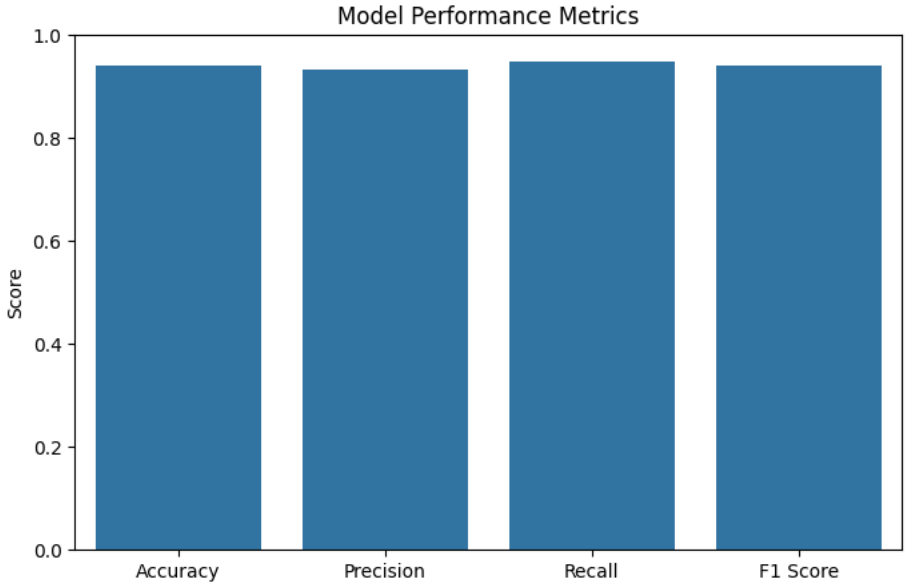


Results:

Performance Metrics:

The performance of the fine-tuned DistilBERT model was evaluated using standard classification metrics:

- **Accuracy**: Measures the overall correctness of the model.
- **Precision**: Indicates how many of the predicted positive reviews were actually positive.
- **Recall**: Measures how many actual positive reviews were correctly identified.
- **F1-Score**: Harmonic mean of precision and recall, providing a balanced measure.

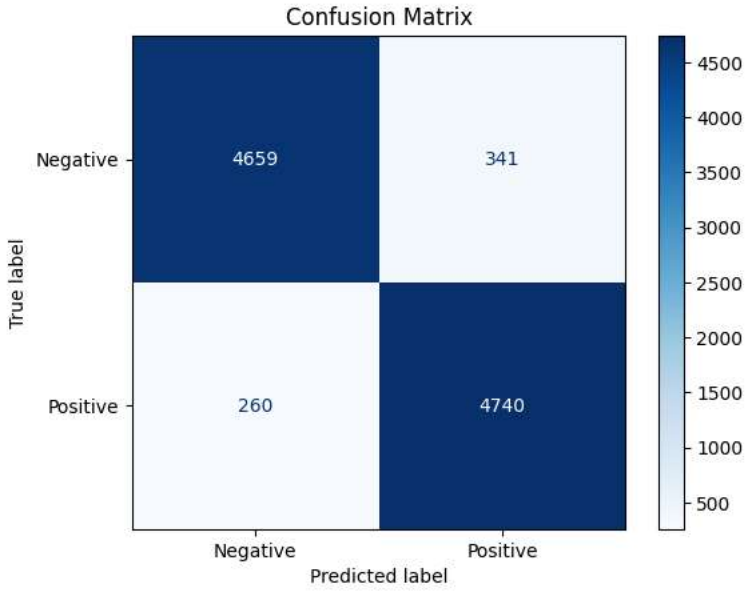


Accuracy: 0.9399

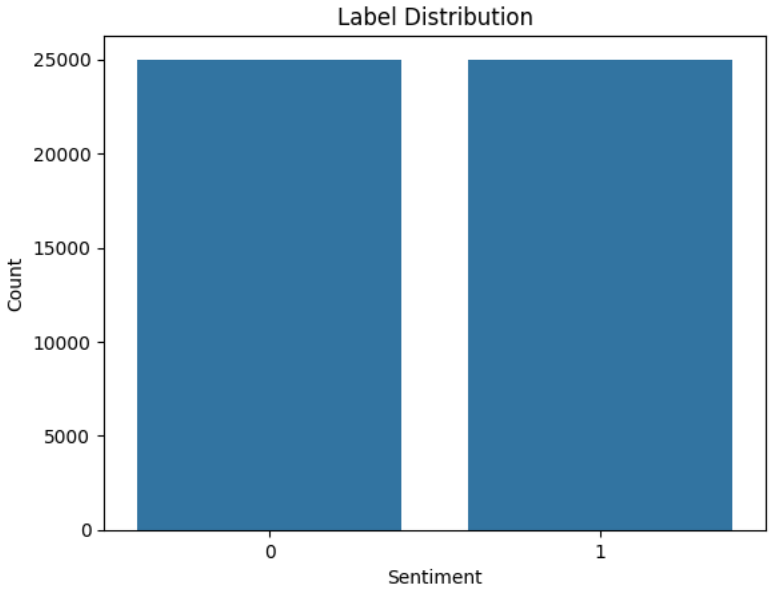
	precision	recall	f1-score	support
0	0.95	0.93	0.94	5000
1	0.93	0.95	0.94	5000
accuracy			0.94	10000
macro avg	0.94	0.94	0.94	10000
weighted avg	0.94	0.94	0.94	10000

Confusion Matrix:

Shows true positives, false positives, true negatives, and false negatives.



Final Results:



Discussion:

Challenges:

- One key challenge encountered during the project was the unexpectedly long training time. Despite having only 5,000 reviews in the dataset, the model took approximately **2 hours** to train.

This is primarily because **transformer models like DistilBERT are computationally intensive**, requiring significant processing power even for relatively small datasets. The model processes each token through multiple attention layers, and tokenized input sequences can be long, especially with detailed movie reviews.

- Additionally, training was done on a CPU, which further slowed down the process. Using a GPU would significantly reduce training time.

Insights Gained:

- Transformer models like DistilBERT are highly effective for sentiment classification, even with limited data.
- Proper pre-processing (e.g., tokenization, label mapping) has a major impact on model performance.
- Evaluation metrics beyond accuracy, such as F1-score, give a more balanced view of model effectiveness.

Comparative Analysis:

Although traditional machine learning models (e.g., SVM, Naive Bayes) could have been used, transformer-based models outperformed them in capturing context and nuance in text. DistilBERT offered a good trade-off between performance and computational efficiency compared to full BERT.

Conclusion:

The project successfully applied the DistilBERT transformer model for sentiment analysis on movie reviews, achieving high accuracy and balanced performance. Despite the training time challenges, the results demonstrate the effectiveness of transformer-based models in understanding textual sentiment. This approach offers a practical solution for automated sentiment classification in real-world applications.

References:

IMDB Movie Reviews (Data set)

https://huggingface.co/distilbert/distilbert-base-uncased

https://discuss.huggingface.co/t/recommendations-for-sentiment-analysis-pre-trained-models/85756 (recommendation)