1. Introduction

For this assignment sentiment analysis will be conducted, a Natural Language Processing task (NLP) which aims to identify and analyse any given texts. It withholds various applications and it has a range from monitoring social media trends to public sentiment and understanding customer feedback. Machine learning models in general, particularly transformer-based, have demonstrated outstanding performance in text classification.

In this assignments experiment a fine-tuned DistilBert model is revealed to analyse sentiment in the IMDb movie's dataset, using a pre-trained transformer to avoid the long and extensive training requirements and resource consumption but while also showing and maintaining a high efficiency. The objective of this sentiment analysis task is to evaluate the performance of the model through metrics such as accuracy, precision, recall and F1 score, making sure that the performance adheres to the previous benchmarks also or if not, where they stand.

2. Research Evaluation

Behind all the research and looking over all the models and their performance, the chosen one is fine-tuned DistilBERT model for this project due to its efficiency and massive performance compared to the larger models. As said before, DistilBERT is a distilled version of BERT which achieves an impressive 97% of BERT's accuracy while operating at a whopping 60% faster speed and using approximately 40% less resources. The choice was inevitably not that hard, DistilBERT is well suited for a scenario like this which requires low computational overhead without sacrificing power.

Also, DistilBERT is also fine-tuned, what does that mean? Fine-tuning enhances the base model to enable it generalise well to data that it has never seen before, it also enhances the model by further adjusting its weight to suit specific tasks which in this case is binary classification of sentiment. In cases like these, where the analysis relies heavily on contextual understanding, the bidirectional attention mechanism of transformers is critical for understanding patterns within textual data.

The dataset used is IMDb dataset, this dataset was sourced from the HugginFace datasets library. This dataset was selected previously as the benchmark dataset due to its wide usage for this type of analysis regarding research and availability as a structured, labelled corpus. This dataset has 50K movie reviews evenly split into training and testing subsets with the binary sentiment labels as follows: Positive (1) and Negative (0).

Even though the dataset is evenly distributed, in practice it often introduces class imbalance and particularly in smaller subsets. To bypass this issue, a dynamic balancing approach is implemented which ensures proportional representation of classes without manually tampering with the equality. The resulting balanced dataset maintains real-world diversity while minimising performance biased during evaluation.

In this case, the model we have here mitigates these challenges by dynamic data balancing, ensuring fairness in evaluation by selecting random subsets while keeping it in proportions of both positive and negative reviews, transfer learning and utilising the pre-trained model to avoid training processes which are resource consuming while after all this, still retaining generalisation capabilities and evaluation.

3. Implementation Design

The implementation was made possible by using PyCharm and using the Python programming language, with libraries such as HuggingFace Transformers, PyTorch, Scikit-learn and Colorama. These tools made it possible to streamline model integration, making the evaluations and visual enhancements. All of these tools ensure a scalable and ease of experimentation possibilities.

All these dependencies were installed using pip ensuring consistency across the system. The visual enhancement makes the program more eye pleasing with additions like coloured-code and progress bars which improve the overall look and usability.

The preprocessing pipeline began with tokenization of textual inputs using the DistilBERT tokenizer, this converts the text in numerical sequences which are suitable for processing by transformer models. Then, dynamic balance is applied to extract a 250 sample subset, which makes 120 positive and 130 negative reviews in this case of run which ensures proportional representation of both classes without enforcing strict equality.

Lastly, after tokenization and dynamic balancing the dataset gets shuffled using a random seed for reproducibility and padded sequences are generated to standardise the input lengths. These are preprocessing steps which ensure compatibility with the model's requirements while preserving randomness and fairness.

4. Evaluation and Results

Evaluation process involves direct predictions using the fine-tuned DistilBERT model. The outputs are divided in positive and negative categories and key metrics are computed to assess the performance. The shown results below might give an idea:

```
Step 1: Loading IMDb dataset...
Step 2: Tokenizing the dataset...
Tokenizing Data: 100%|██████████| 25000/25000 [00:03<00:00, 7244.74 examples/s]
Tokenizing Data: 100%|██████████| 25000/25000 [00:03<00:00, 7424.75 examples/s]
Tokenizing Data: 100%|██████████| 50000/50000 [00:06<00:00, 7313.83 examples/s]
Step 3: Shuffling and dynamically balancing the dataset...
Balanced dataset size: 250
Step 4: Loading pre-trained fine-tuned DistilBERT model...
Step 5: Making predictions on the full dataset...
Positive Reviews: 120
Negative Reviews: 130
Step 7: Evaluating performance metrics...

Evaluation Results:
Metric      Value
----------------------
Accuracy    0.88
Precision   0.89
Recall      0.86
F1 Score    0.87
```

The results show the fine-tuned DistilBERT model exceeds 85% across all the metrics and its performance is high, meeting all the theoretical expectations for transformer-based classifiers. After more observation, the datasets has balance with the dynamic sampling approach which results in a dataset with 120 positives and 130 negative reviews, these reflect real-world variability without sacrificing fairness and efficiency where processing speeds average 7300 examples/second which showcases DistilBERT's computational efficiency, even on consumer-grade hardware.

5. Conclusion

The implementation highlights the strengths of fine-tuned transformer models in handling sentiment analysis tasks effectively. The balanced dataset approach minimizes biases while maintaining randomness, and the use of pre-trained weights avoids computational overheads associated with model training. Despite its strong performance, the classifier exhibits slight limitations due to reliance on binary labels. Future work can extend this approach to multi-class sentiment classification or explore context-aware sentiment detection using hierarchical transformers.

To conclude, this project successfully implemented a fine-tuned DistilBERT classifier with IMDb reviews dataset. The results validate DistilBERT's efficiency and performance, demonstrating its ability and these findings align with the theoretical benchmarks and suggest promising directions for further improvements which include fine-tuning on domain-specific-data and multi-class tasks and deployment-ready pipelines.