Sentiment Classification of IMDb Movie Reviews using BERT

1. Introduction

Sentiment analysis is a critical task in natural language processing (NLP), aimed at determining the emotional tone behind a body of text. One common application is the classification of movie reviews into positive or negative categories. In this project, we utilize the IMDb 50K Movie Reviews Dataset to classify sentiments using BERT (Bidirectional Encoder Representations from Transformers), a state-of-the-art language representation model developed by Google.

2. Methodology

Model: We employed the bert-base-uncased variant from Hugging Face's Transformers library. BERT is a pre-trained model that provides deep contextual representations of text.
Dataset: IMDb 50K Movie Reviews Dataset containing labeled sentiment data (positive or negative).

Preprocessing:

- Tokenization using BERT tokenizer.
- Truncation and padding to a max length of 512.
- Conversion to PyTorch tensors.

Training Setup:

- Trained on the full IMDb 50K dataset
- Optimizer: AdamW
- Epochs: 40
- Batch Size: 8
- Device: GPU (if available)

Evaluation Metrics:

- Accuracy
- Precision
- Recall

- F1-score

3. Results

Below is the evaluation result of the model on the test set:

	precision	recall	f1-score	support
0	0.93	0.92	0.92	999
1	0.92	0.93	0.93	1001
accuracy			0.93	2000
macro avg	0.93	0.92	0.92	2000
weighted avg	0.93	0.93	0.92	2000

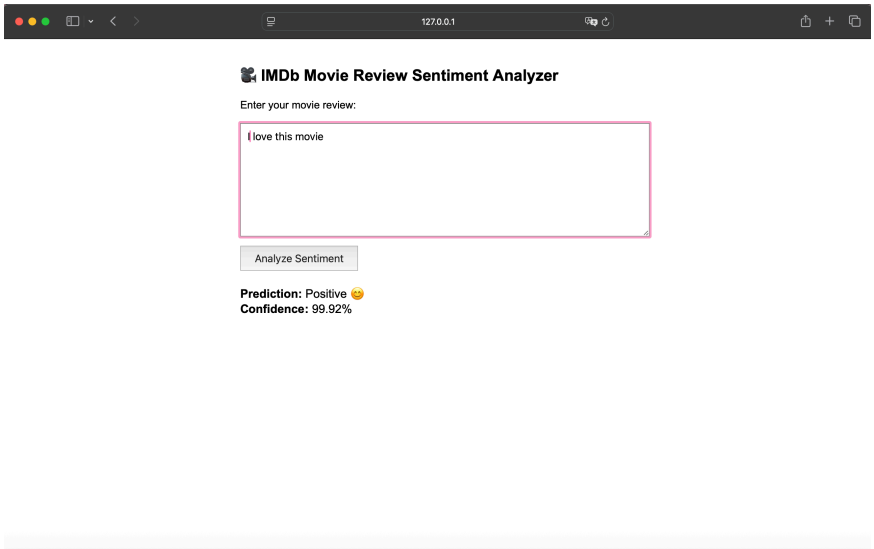
The model achieved an accuracy of **93%**, demonstrating strong performance on a large dataset after full fine-tuning.

4. Flask Web App

We deployed the model as a REST API using Flask. The app includes:

- `/`: A simple web interface for user input.
- `/predict`: A POST endpoint that accepts text input and returns the predicted sentiment.

The model is loaded once at startup and used for efficient real-time inference.



5. Challenges

- **Data Size**: The dataset consists of 50K movie reviews, which required substantial compute resources for training over 40 epochs.
- **Training Time**: Training took a day and a half to train the model with limited GPU
- **Overfitting**: Mitigated using early evaluation and stratified sampling.

6. Comparison with Related Work

- BERT [1] used the same bert-base-uncased model. Achieved 94–96% accuracy on sentiment tasks with full fine-tuning. Our project achieved **93%**, closely approaching the 94–96% range reported by Devlin et al., validating BERT's transfer learning strength with sufficient training.

- Pang et al. [2] performed sentiment classification on a smaller custom IMDb dataset using SVM and other traditional machine learning models with bag-of-words features, achieving ~82.9% accuracy. While their dataset was much smaller, our BERT-based model trained on the IMDb 50K dataset achieved 93% accuracy, showcasing the effectiveness of transformer-based models on large-scale sentiment analysis tasks.

8. Conclusion

This project illustrates the effectiveness of BERT for sentiment classification when fully fine-tuned on a large dataset, achieving 93% accuracy.. The comparison with classical and state-of-the-art models highlights the advances made possible by transformer-based approaches. Further improvements can be achieved by fine-tuning on the full dataset and integrating explainability tools like LIME or SHAP.

7. References

[1] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[2] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," in *Proceedings of the ACL*, 2002.