Comparative Analysis of BERT Variants for Sentiment Classification

1. Introduction

Transformer-based language models have revolutionized Natural Language Processing (NLP) tasks, particularly text classification. This project evaluates five popular **BERT variants**—**BERT Base**, **RoBERTa**, **DistilBERT**, **ALBERT**, and **XLNet**—on the IMDb movie reviews dataset for binary sentiment classification. The objective is to identify the most effective model for classifying reviews as positive or negative and analyze the architectural differences influencing performance.

2. Model Descriptions

2.1 BERT (Bidirectional Encoder Representations from Transformers)

- Standard transformer-based model pre-trained on masked language modeling and next sentence prediction.
- Uses bidirectional attention, capturing context from both left and right.
- Strong baseline for text classification tasks.

2.2 RoBERTa (Robustly Optimized BERT)

- Improves BERT by removing next-sentence prediction and training with:
    - Larger mini-batches
    - More data
    - Dynamic masking
- Allows more robust contextual representations.

2.3 DistilBERT

- A distilled version of BERT that reduces parameters by ~40% while retaining ~97% of performance.
- Faster training and inference.
- Suitable for lightweight applications.

2.4 ALBERT (A Lite BERT)

- Reduces model size significantly using:
    - Parameter sharing across layers
    - Factorized embedding parameterization
- Designed for scalability and efficiency.

2.5 XLNet

- Uses permutation-based language modeling, capturing bidirectional context without masking tokens.
- Excels in capturing long-range dependencies and sentence-level coherence.

3. Experimental Setup

- **Dataset:** IMDb (50,000 labeled reviews)
- **Splits:** Training, Validation, Testing with balanced classes
- **Optimizer:** AdamW
- **Loss Function:** CrossEntropy
- **Metrics:** Accuracy, Precision, Recall, F1-Score, MCC, Cohen's Kappa, MAE, RMSE, AUC-ROC, CSI
- **Evaluation:** Early stopping with patience to prevent overfitting

4. Results

4.1 Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score	MCC	Kappa	MAE	RMSE	AUC-ROC	CSI
BERT	86.4%	83.7%	90.4%	86.9%	0.730	0.728	0.136	0.369	0.938	0.769
RoBERTa	**89.5%**	86.6%	93.5%	89.9%	**0.793**	**0.790**	**0.105**	**0.324**	**0.961**	**0.817**
DistilBERT	86.1%	86.6%	85.3%	85.9%	0.721	0.721	0.140	0.373	0.940	0.754
ALBERT	86.7%	86.1%	87.6%	86.8%	0.734	0.734	0.133	0.365	0.941	0.767
XLNet	**89.5%**	**89.0%**	90.1%	89.6%	0.790	0.790	0.105	0.324	**0.961**	0.811

4.2 Training and Validation Curves

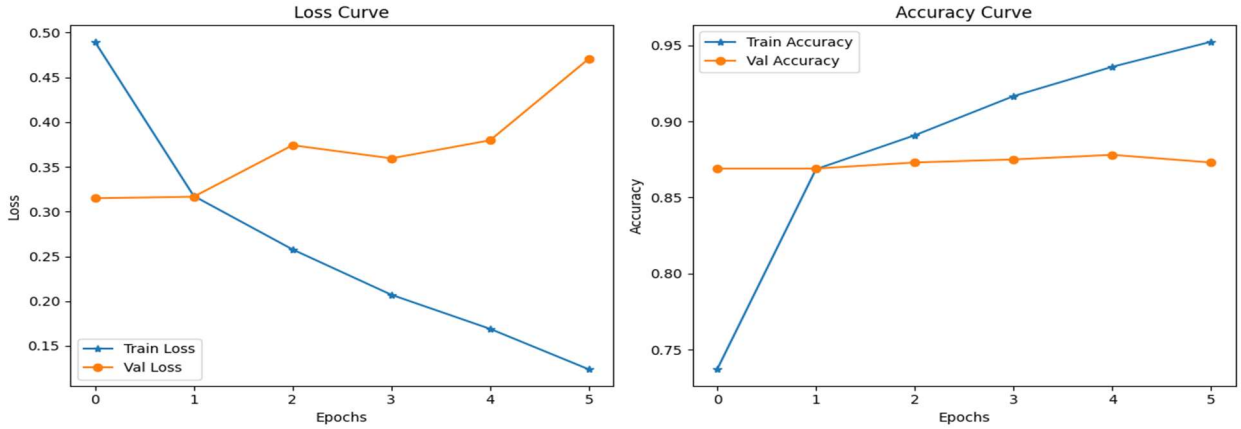

Figure 1: Loss and Accuracy Curve of BERT

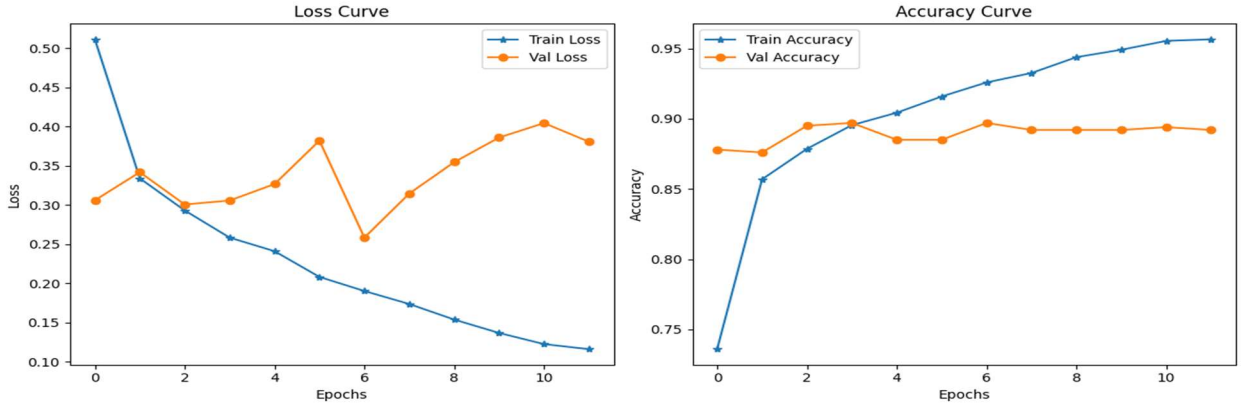

Figure 2: Loss and Accuracy Curve for RoBETRa

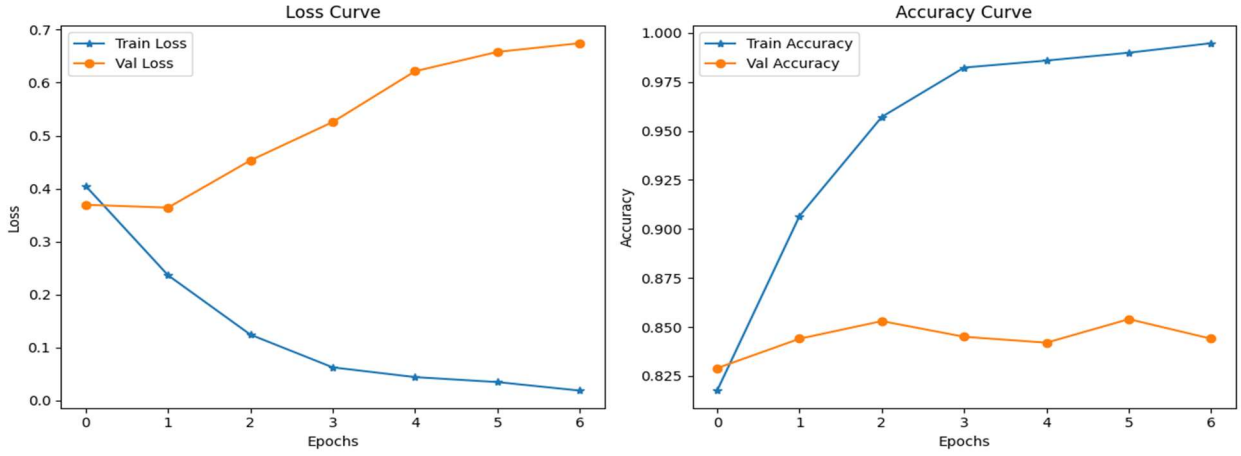

Figure 3: Loss and Accuracy Curve of DistilBERT

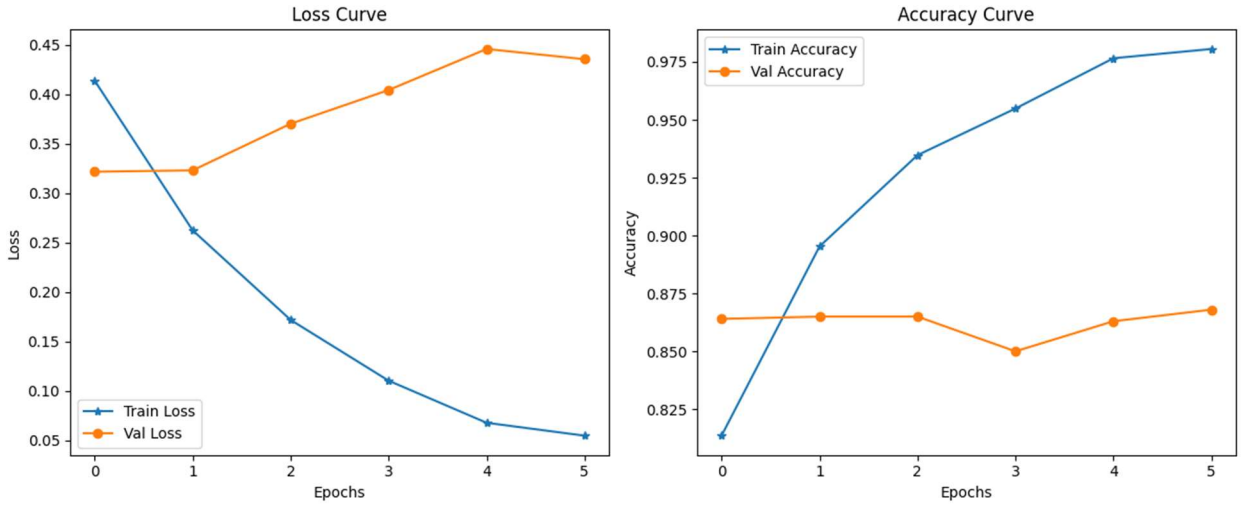

Figure 4: Loss and Accuracy Cruve of ALBERT

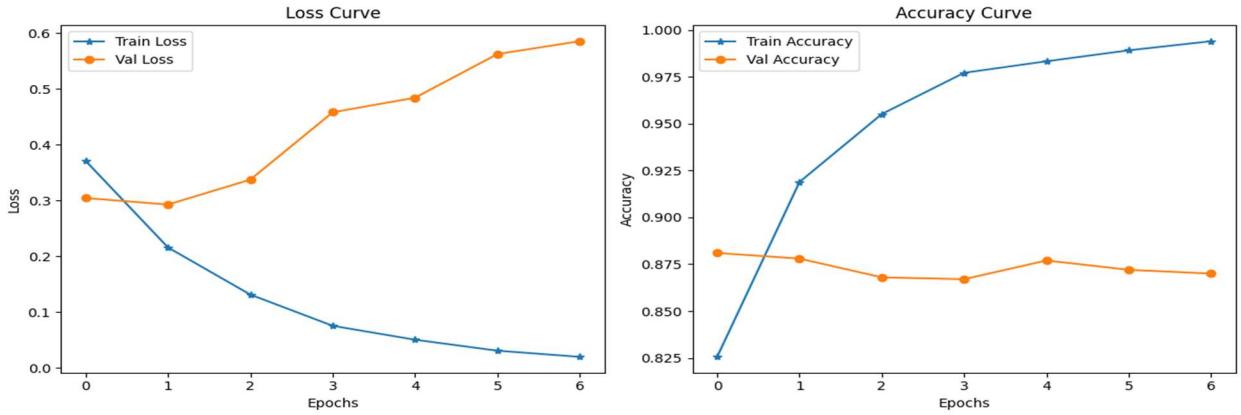

Figure 5: Loss and Accuracy Curve of XLNET

5. Analysis

5.1 Accuracy and F1-Score

- RoBERTa and XLNet achieved the highest accuracy (89.5%) and F1-scores (~89.9% and 89.6%), outperforming the baseline BERT model by ~3%.
- DistilBERT and ALBERT had slightly lower accuracy but provided efficiency advantages.

5.2 Recall and AUC-ROC

- RoBERTa recorded the highest recall (93.5%), meaning it identified more positive sentiments correctly.
- Both RoBERTa and XLNet achieved the highest AUC-ROC (0.961), indicating strong discriminatory power.

5.3 MCC, Kappa, and Error Metrics

- RoBERTa led in MCC (0.793) and Cohen's Kappa (0.790), showing better balanced classification.
- It also achieved the lowest MAE and RMSE, reinforcing its robustness.

5.4 Computational Efficiency

- DistilBERT trained ~2x faster due to fewer parameters.
- ALBERT reduced model size significantly but slightly compromised accuracy.
- XLNet required more time due to its permutation mechanism.

6. Reasoning for Best Model

- **RoBERTa is the best-performing model**, offering:
    - Higher accuracy and F1-score than BERT.
    - Better recall and MCC, reducing false negatives.
    - Lower error metrics and superior AUC-ROC.
- **Architectural advantage:**
    - Trained with **more data and larger batches**.
    - Uses **dynamic masking** instead of static masking.
    - **Removes next sentence prediction**, simplifying pretraining.
    - Learns richer contextual embeddings, improving classification.

Why Not XLNet?

- XLNet achieved similar accuracy and AUC-ROC, but:
    - Requires significantly more computation.
    - Shows slightly lower recall and MCC than RoBERTa.
- Thus, RoBERTa offers a better trade-off between performance and efficiency

7. Conclusion

- RoBERTa outperforms all other BERT variants for IMDb sentiment classification.
- XLNet provides competitive performance but with higher computational cost.
- DistilBERT and ALBERT are suitable for **resource-constrained environments**.
- Future work may involve:
    - Hyperparameter tuning
    - Ensemble methods
    - Exploring larger RoBERTa and XLNet versions