Performance Analysis Report

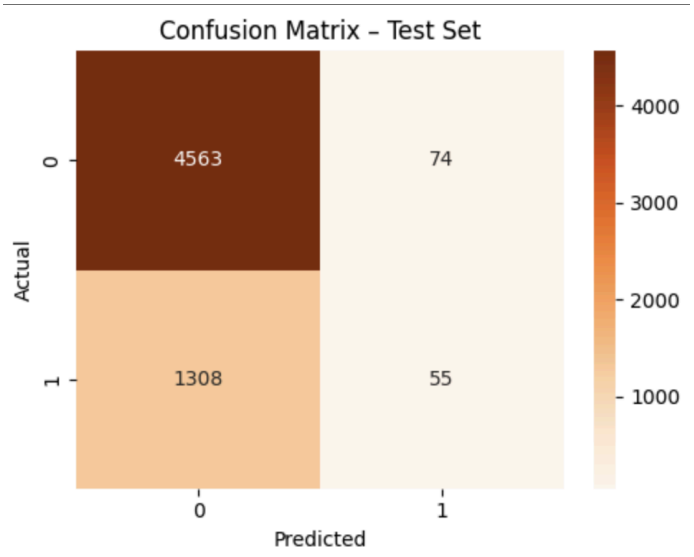
Project : Offensive Language Classification

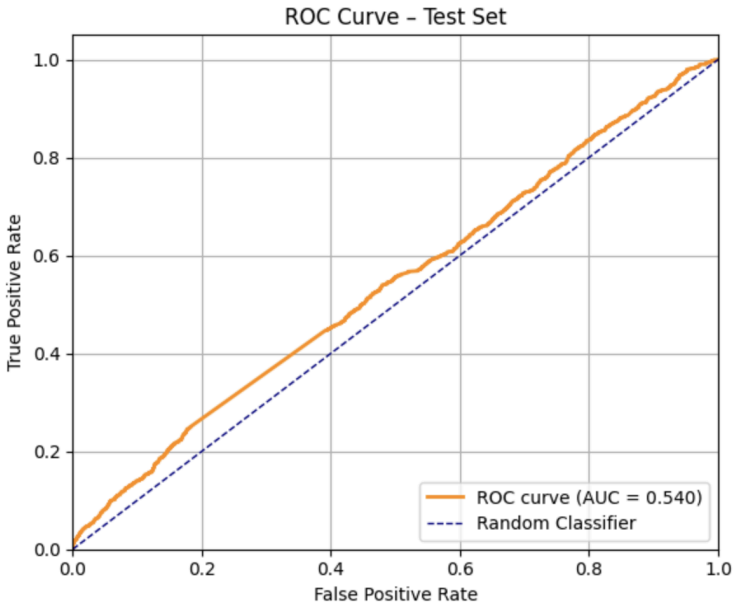
Baseline Model (Logistic Regression)

Performance metrics :

Label	Precision	Recall	F1-score	Support
0	0.78	0.98	0.87	4637
1	0.43	0.04	0.07	1363
Accuracy	**0.770**			
Macro Avg	0.60	0.51	0.47	
Weighted Avg	0.70	0.77	0.69	

AUC-ROC Score: 0.540





Key observations

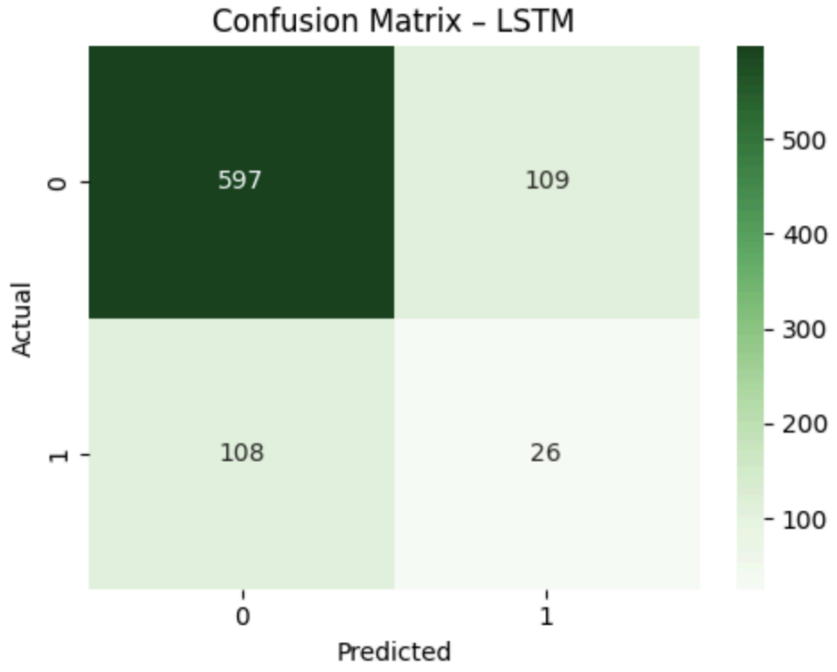
- High accuracy is primarily due to the model predicting the majority class (label 0) correctly.
- Very low recall for label 1 suggests the model fails to detect offensive content effectively.
- AUC is low due to poor classification performance on minority class.

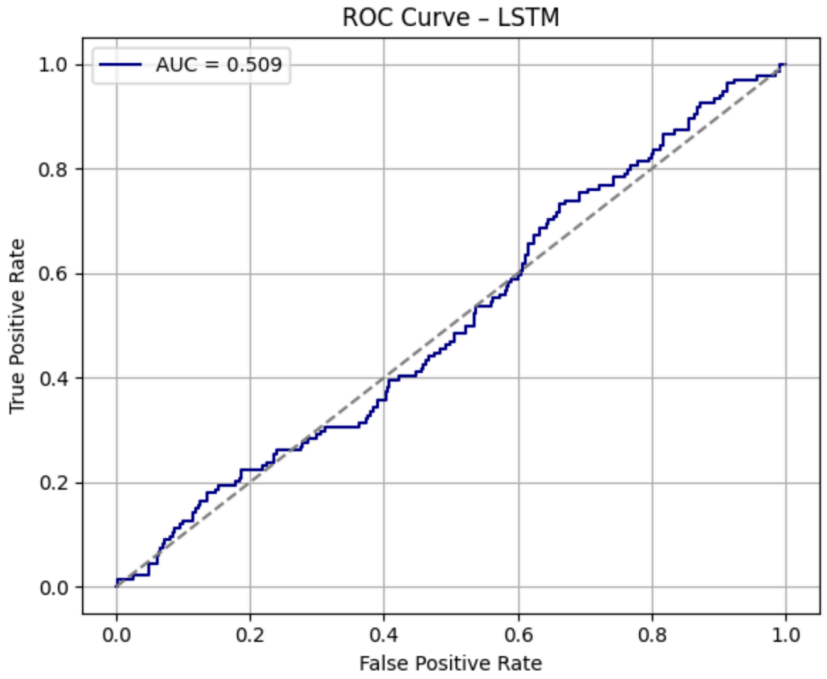
Advanced Model (LSTM)

Validation Set Performance

Label	Precision	Recall	F1-score	Support
0	0.85	0.85	0.85	706
1	0.19	0.19	0.19	134
Accuracy	**0.742**			
Macro Avg	0.52	0.52	0.52	
Weighted Avg	0.74	0.74	0.74	

AUC-ROC Score: 0.509





Due to low performance, The model is proceeded to be tuned.

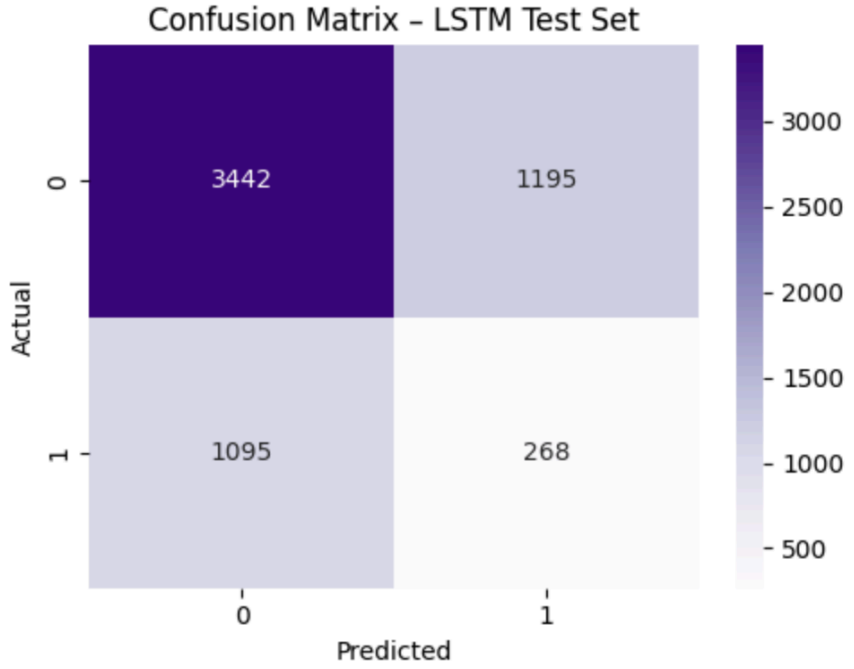
Best Tuned Hyperparameters

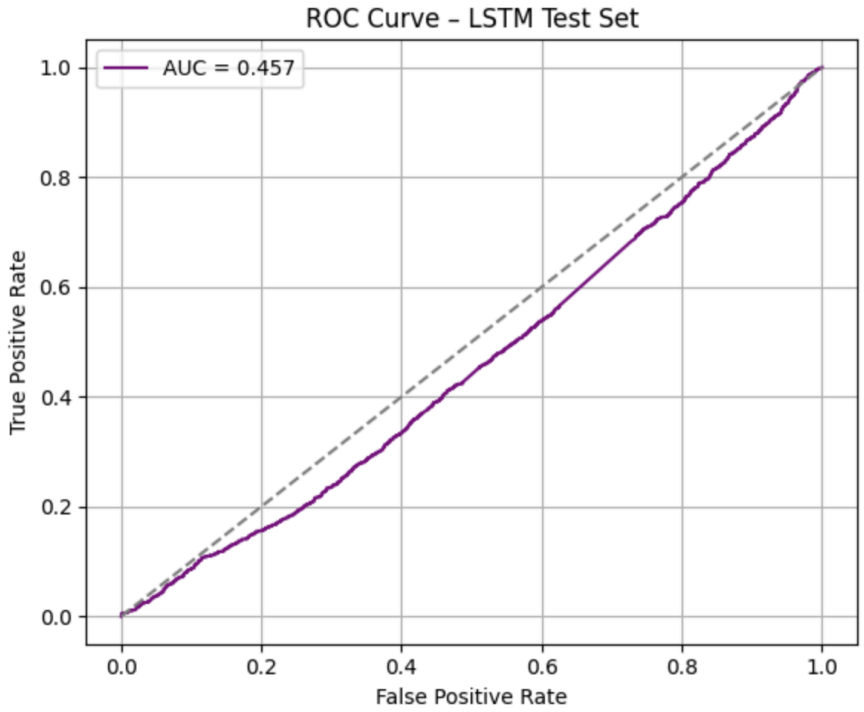
- LSTM Units: 64
- Dropout: 0.5
- Optimizer: Adam
- Dense Units: 32
- Embedding Dimension: 128

Test Set Performance

Label	Precision	Recall	F1-score	Support
0	0.76	0.74	0.75	4637
1	0.18	0.20	0.19	1363
Accuracy	**0.618**			
Macro Avg	0.47	0.47	0.47	
Weighted Avg	0.63	0.62	0.62	

AUC-ROC Score: 0.457





Key observations

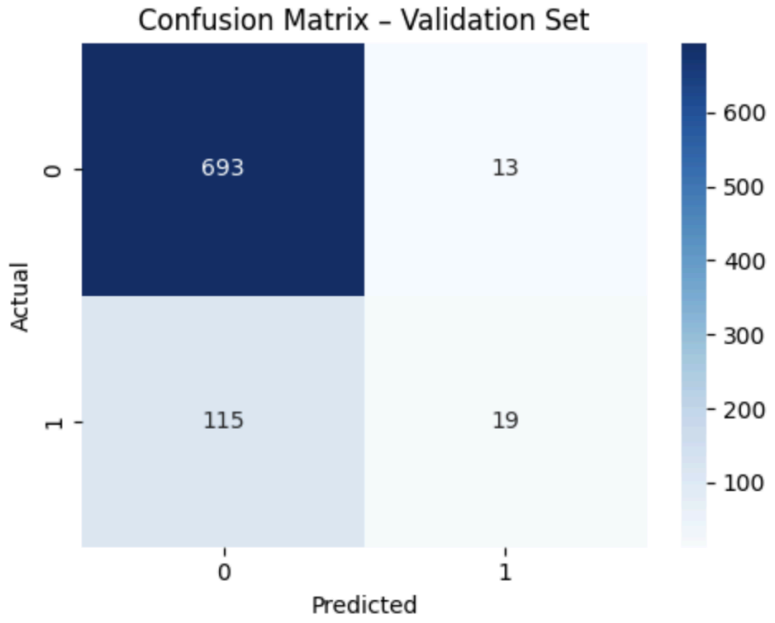
- Although performance improved with hyperparameter tuning, the model still struggles to identify label 1 accurately.
- ROC remains low, indicating poor discriminatory power, especially with the imbalance in labels.

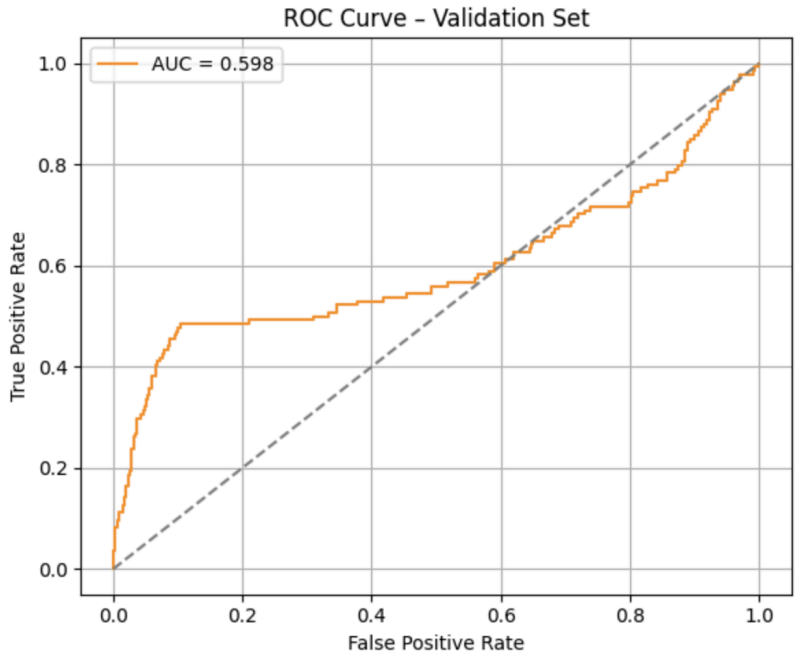
Transformer-based Model (Fine-tuned BERT)

Validation Set Performance

Label	Precision	Recall	F1-score	Support
0	0.86	0.98	0.92	706
1	0.59	0.14	0.23	134
Accuracy	**0.848**			
Macro Avg	0.73	0.56	0.57	
Weighted Avg	0.82	0.85	0.81	

AUC-ROC Score: 0.598

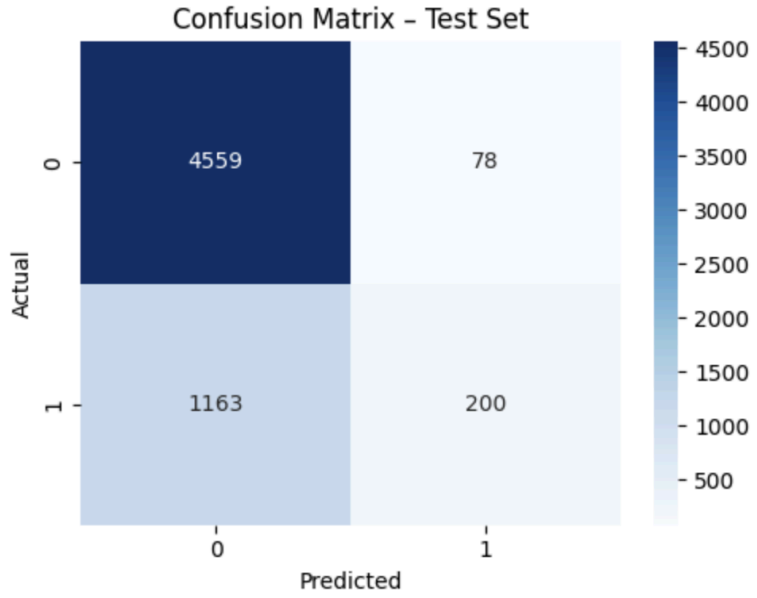


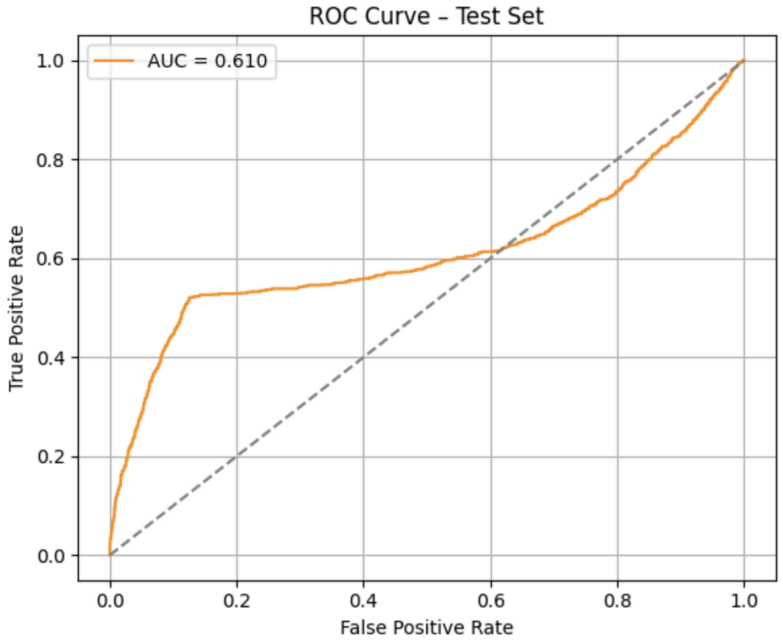


Test Set Performance

Label	Precision	Recall	F1-score	Support
0	0.80	0.98	0.88	4637
1	0.72	0.15	0.24	1363
Accuracy	**0.793**			
Macro Avg	0.76	0.56	0.56	
Weighted Avg	0.78	0.79	0.74	

AUC-ROC Score: 0.610





Key observations

- Best performance among all models in identifying offensive content (label 1).
- Improved precision and recall compared to baseline and LSTM, particularly for the minority class.
- Transformer models handle multilingual context better, but require more resources.

My overal Observations and Analysis

➢ **Accuracy vs. AUC-ROC:**

- Models often show high accuracy by favoring the majority class (label 0).
- AUC-ROC provides a better sense of model's ability to distinguish between classes, especially in imbalanced datasets.
- Low AUC despite high accuracy is a sign of poor minority class (offensive) recognition.

➢ **Impact of Imbalanced Data:**

- All models struggled with label 1 due to its smaller representation in the dataset.
- Techniques like oversampling/undersampling or class-weighted loss could improve performance.

➢ **Visual Insights:**

- Confusion matrices and ROC curves give visual clarity on model behavior.
- They reveal that most false negatives stem from the model ignoring offensive samples.

Conclusion

Let’s look at the model Performance Summary

Model	Accuracy	AUC-ROC	Precision (Label 1)	Recall (Label 1)	F1-Score (Label 1)
Baseline (LogReg)	0.770	0.620	0.43	0.04	0.07
LSTM	0.742	0.509	0.19	0.19	0.19
	0.618	0.457	0.18	0.20	0.19
Transformer (BERT)	0.848	0.598	0.59	0.14	0.23
	0.793	0.610	0.72	0.15	0.24

Best Model: Fine-tuned BERT delivered the most effective performance by striking an optimal balance between overall accuracy and detecting minority class instances.

Unlike traditional models such as Logistic Regression or LSTM, BERT benefits from pre-training on a massive multilingual corpus, allowing it to grasp the nuanced meaning of text across different languages. This was especially important for our task, as the validation and test datasets included multilingual content, while the training data was primarily in English. Even with class imbalance—where toxic comments were fewer than clean ones—BERT outperformed other models in identifying those minority cases. Its self-attention mechanism and contextual embeddings helped it pick up subtle signs of offensiveness that simpler models often failed to catch. Although the LSTM model performed reasonably well in terms of accuracy, it struggled to recognize toxicity in brief or ambiguous comments. In contrast, BERT could capture both local and global relationships within the text, which made it more reliable. After fine-tuning on our dataset, the model demonstrated consistent performance, even on unseen and linguistically diverse feedback. Evaluation metrics like AUC-ROC and F1-score showed clear improvements over baseline models, proving its stronger discriminative ability. Overall, BERT proved to be a more adaptable and robust solution for offensive language detection, making it highly applicable for real-world content moderation systems.