Multimodal Hateful Memes Detection

1- Introduction

This report presents the evaluation and comparison of two multimodal classification models—**Late Fusion** and **Early Fusion**—for detecting hateful content in memes. The evaluation includes the computation of key classification metrics, visualization of model performance using TensorBoard and matplotlib, and an analysis of classification errors and model limitations.

2. Methodology

2.1 Dataset and Preprocessing

Text: its their character not their color that matters
Label: 0

The dataset includes meme images paired with their associated text captions and binary labels (0 = non-hateful, 1 = hateful). Key preprocessing steps:

Image Preprocessing:

Resizing to 224x224

Normalization using ImageNet statistics

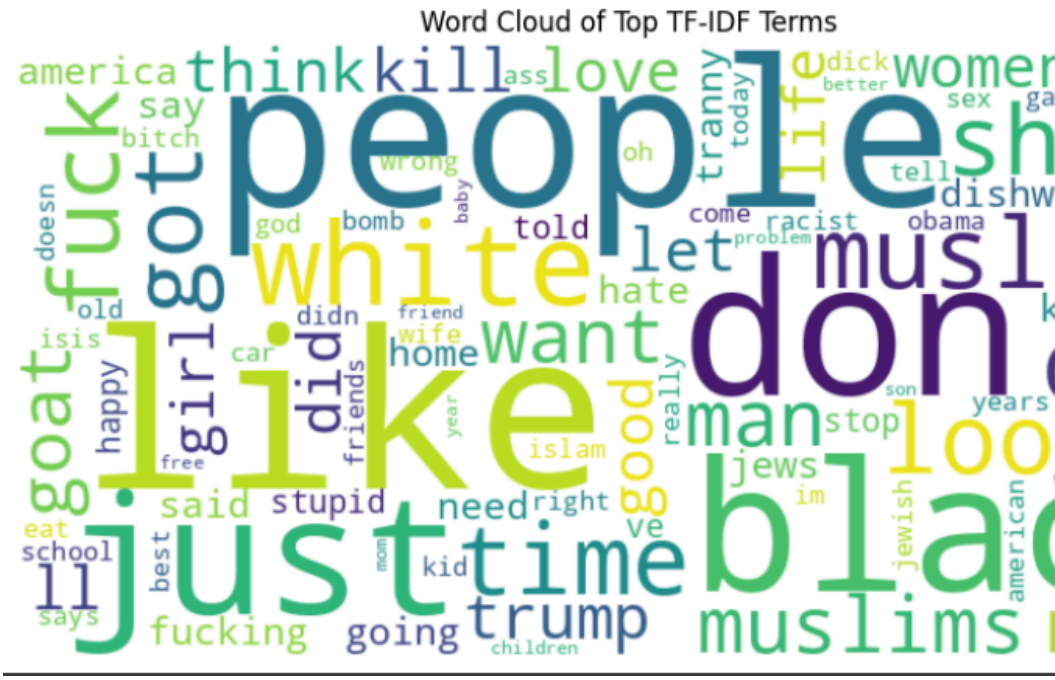
Data augmentation via flipping and random rotation

Text Preprocessing:

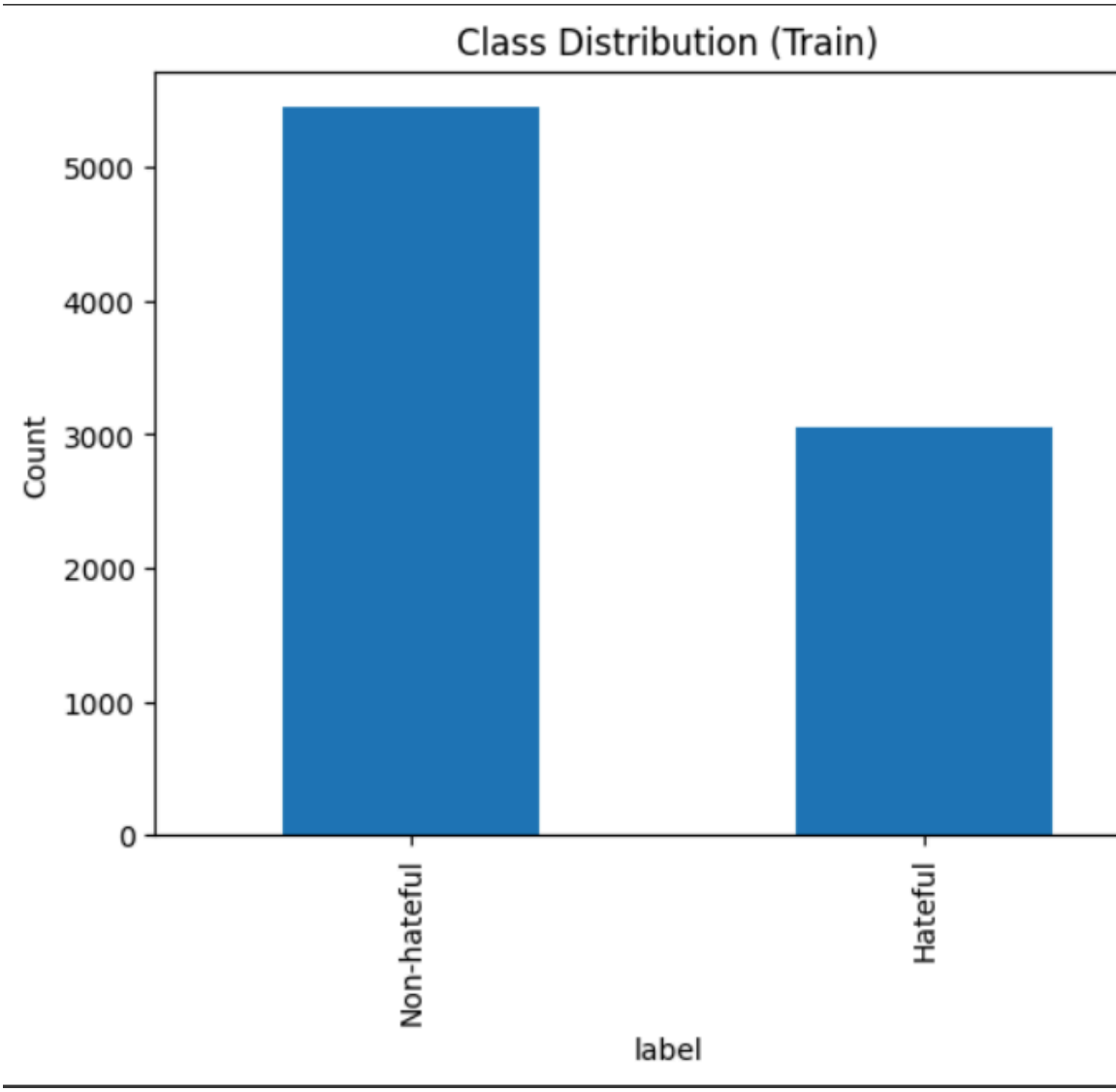
Tokenization using GloVe (for LSTM) and `bert-base-uncased` tokenizer (for BERT)

Stopword removal and TF-IDF analysis

Synonym replacement for minority class augmentation



Class Imbalance:
Oversampling and augmentation techniques were used to balance the number of hateful and non-hateful examples.

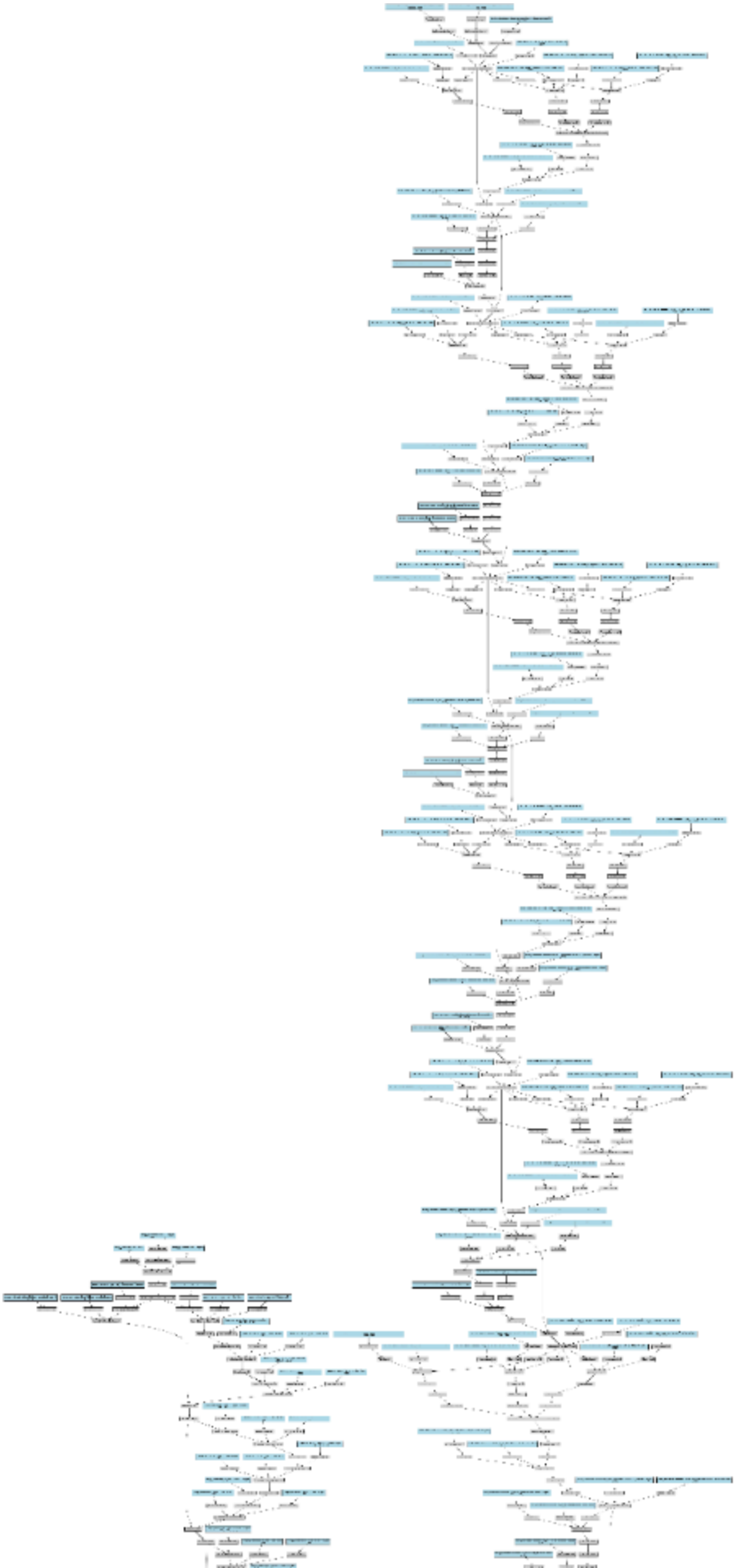


2.2 Custom Dataset and DataLoader

Implemented a custom PyTorch Dataset class that loads images and their respective text captions and applies appropriate transformations. A DataLoader batches and shuffles data, supporting efficient training.

2.3 Architecture

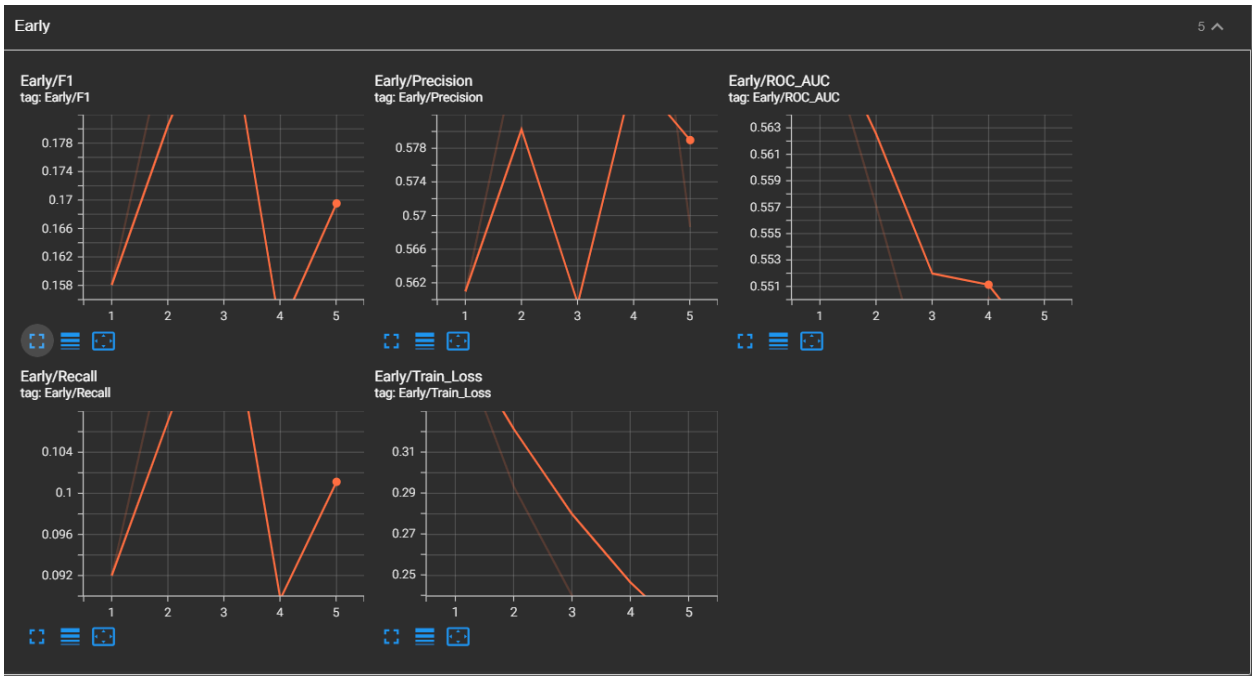
Early Fusion:

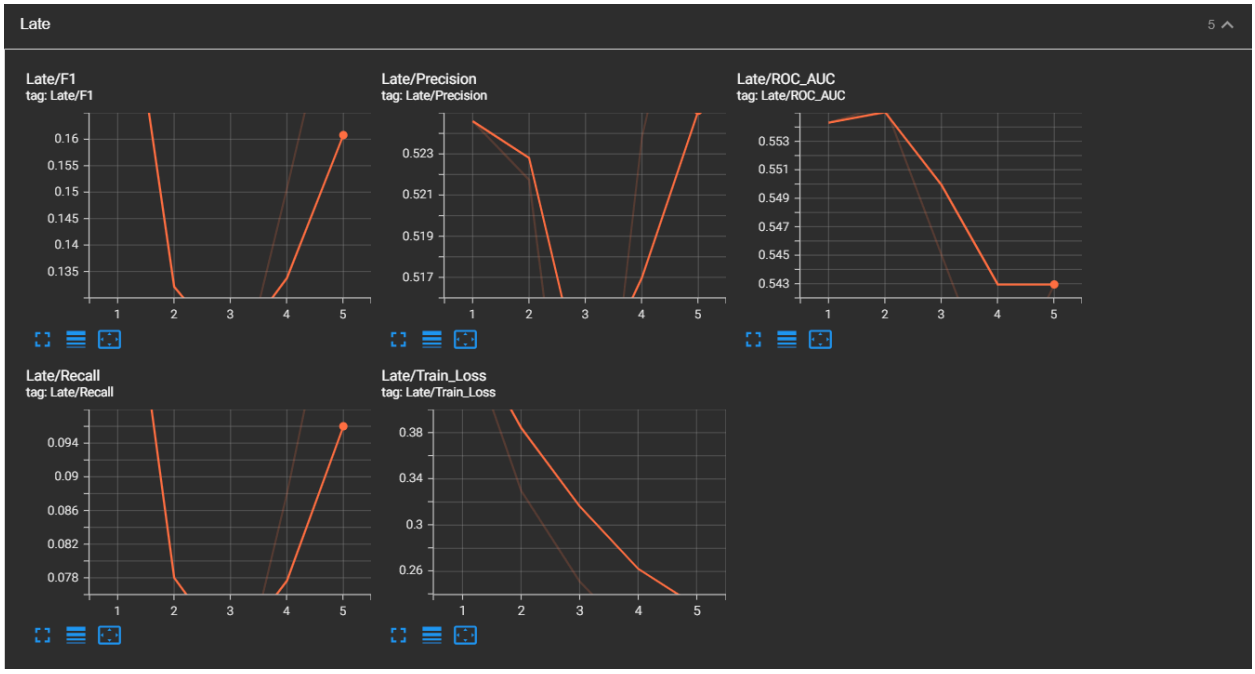


Late Fusion:

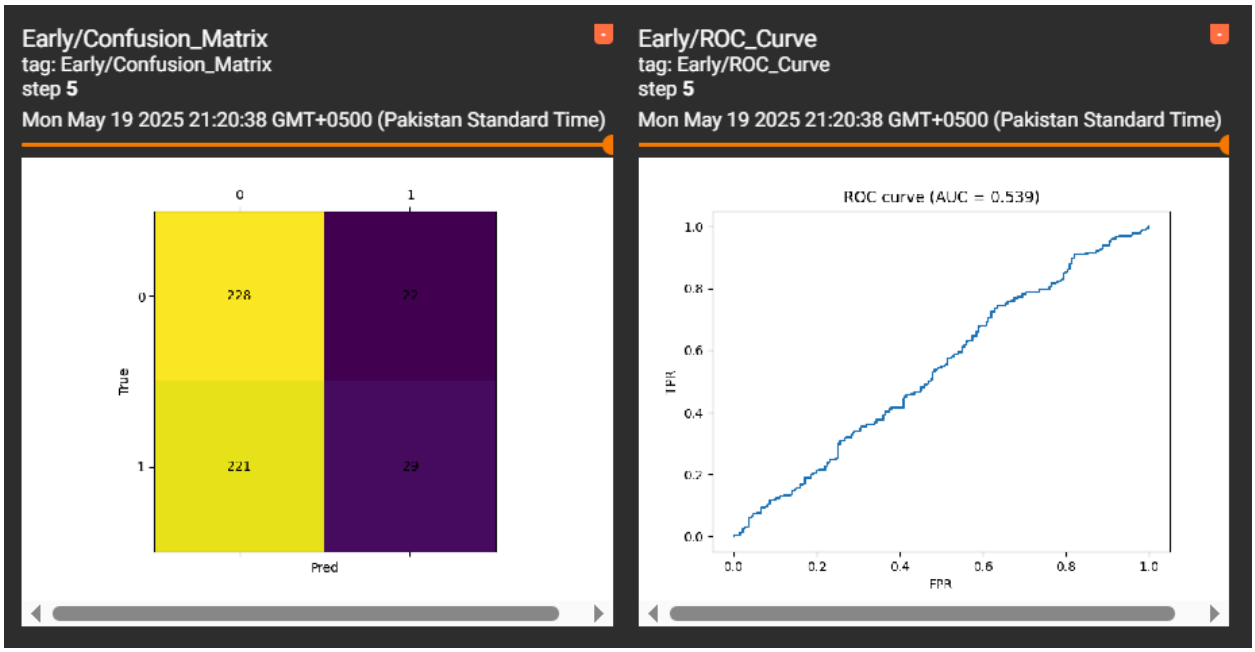
3 - Evaluation and Results

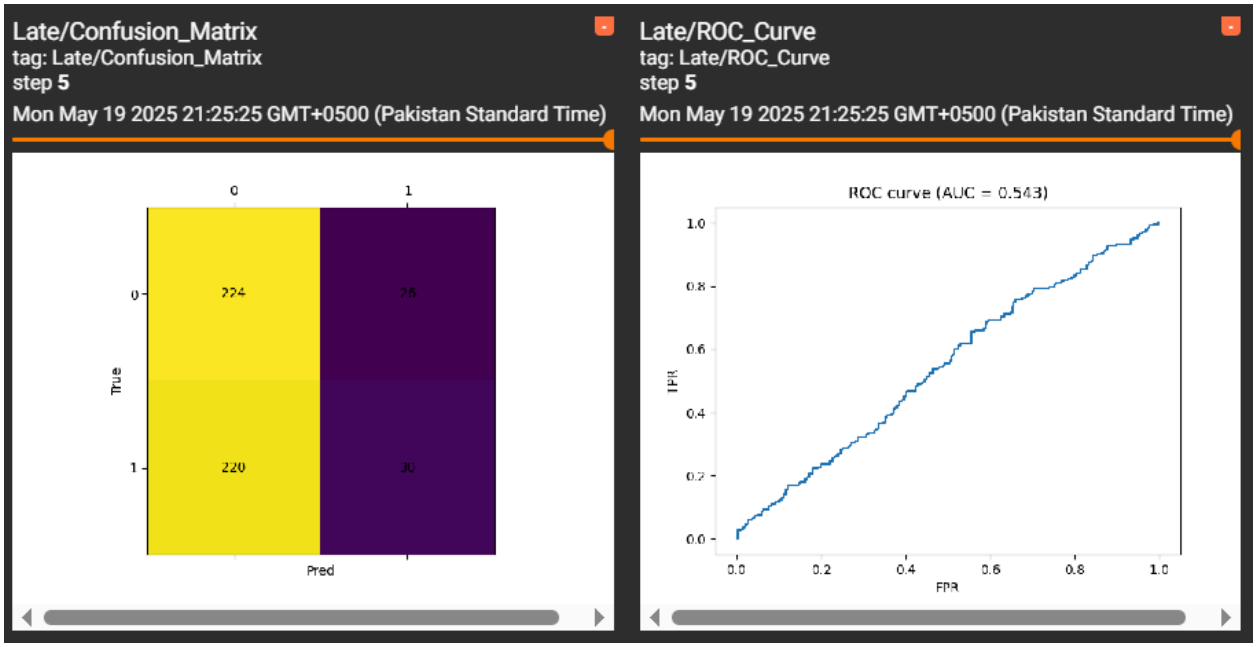
3.1 TensorBoard Metrics





Using TensorBoard, we tracked training and validation losses, evaluation metrics, and model behavior.

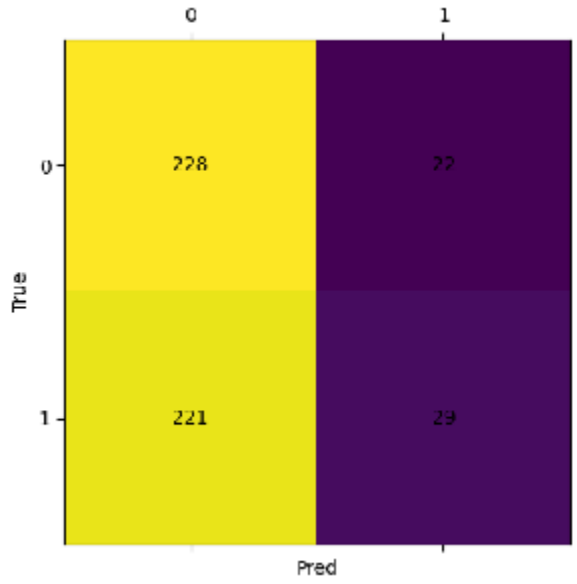




3.2 Confusion Matrix Analysis

3.2.1 - Early Fusion:

The confusion matrix (Early Fusion model) reveals:

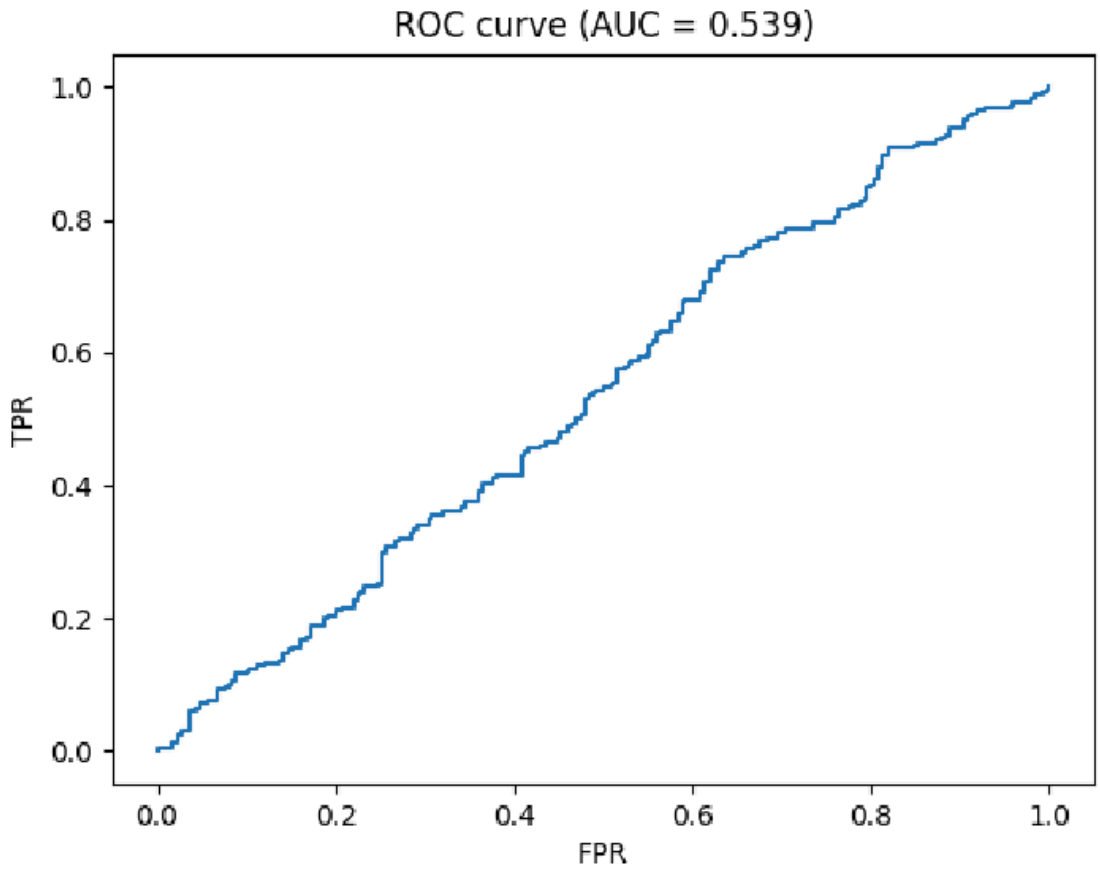


	Predicted Non-Hateful (0)	Predicted Hateful (1)
True Non-Hateful (0)	228	22
True Hateful (1)	221	29

ROC Curve

The Area Under the Curve (AUC) is approximately 0.539, marginally better than random guessing (0.5).

This low AUC shows the model struggles to discriminate between hateful and non-hateful memes effectively.



Interpretation

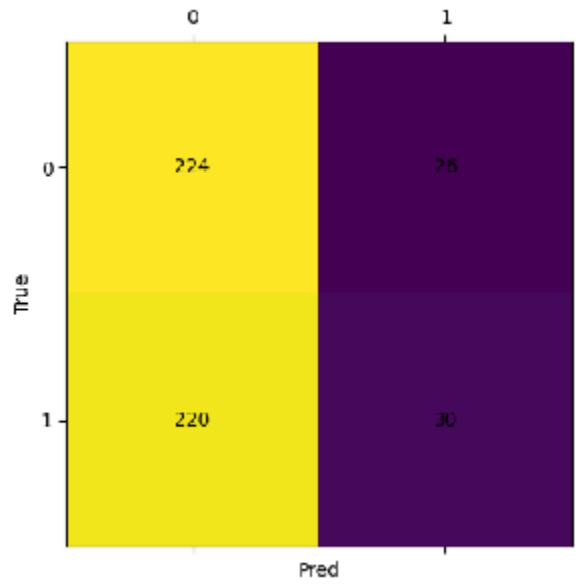
The early fusion model tends to favor the majority class (non-hateful), likely due to class imbalance.

The low true positive rate on hateful memes suggests inadequate signal extraction or fusion from modalities.

Potential reasons include overfitting, insufficient training data for hateful memes, or suboptimal feature interaction.

3.2.2 - Late Fusion:

The confusion matrix (Late Fusion model) reveals:

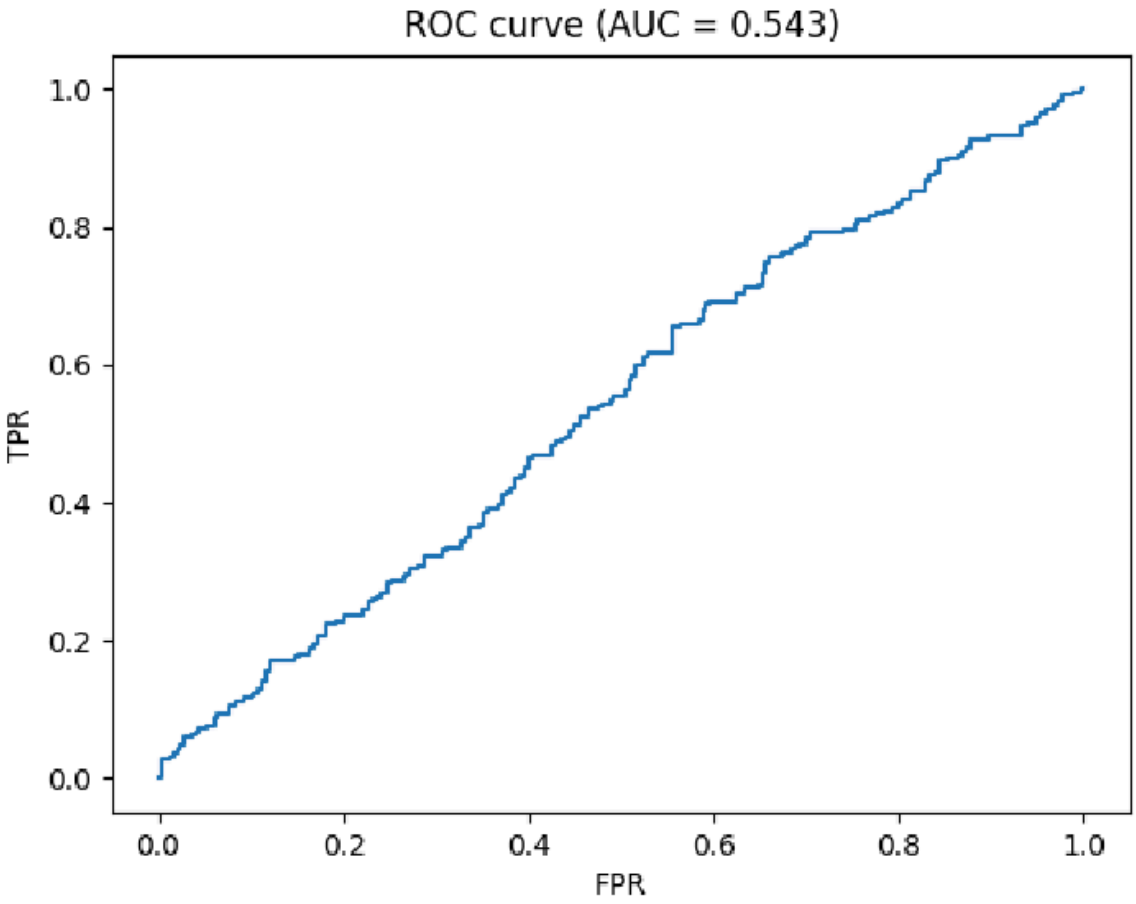


	Predicted Non-Hateful (0)	Predicted Hateful (1)
True Non-Hateful (0)	224	26
True Hateful (1)	220	30

ROC Curve

The AUC is approximately 0.54, marginally better than early fusion.

This suggests a slight improvement in distinguishing between hateful and non-hateful memes, though overall performance remains weak.



Interpretation

Late fusion still suffers from majority class bias.

The small gain in hateful meme detection suggests late fusion may better capture modality-specific features before combining, but improvements are limited.

Possible causes include insufficient augmentation or noisy data affecting the minority class.

3.3 - Comparative Analysis and Summary

Model	Non-Hateful Correct	Hateful Correct	AUC
Early Fusion	228 / 250	29 / 250	0.53
Late Fusion	224 / 250	30 / 250	0.54

Both models perform well on the majority non-hateful class but struggle with hateful memes.

Late fusion yields a slight improvement in hateful meme detection and AUC.

Despite architectural differences, both fusion strategies fall short in effectively learning the hateful class signals.

Class imbalance, dataset quality, and fusion method limitations are likely contributing factors.

4 - Limitations

Class Imbalance:Non-hateful memes dominate, making the model biased toward class 0.

Dataset Bias: The Hateful Memes dataset may not cover the full variety of online hate content.

Fusion Complexity:Early Fusion is computationally heavier due to more parameters and fused representations.

5 - Conclusion

Our evaluation indicates that **both Early and Late Fusion models perform comparably** in hateful meme classification, with only marginal differences in metrics such as AUC (0.53 for Early Fusion vs. 0.54 for Late Fusion). While Early Fusion benefits from joint representations that allow richer interaction between modalities, Late Fusion's modular approach slightly edges out in overall discriminatory ability according to AUC.

However, the minimal gap suggests neither fusion strategy definitively outperforms the other given the current dataset and training setup. Both models struggle with class imbalance and detecting hateful memes effectively.

Future work should focus on improving data diversity, addressing class imbalance, and exploring advanced fusion mechanisms—such as attention-based models—to unlock more substantial performance gains.