Model Performance Analysis Report

Logistic Regression Evaluation:

TOXIC
Accuracy: 0.9441959531416401
Precision: 0.9004739336492891
Recall : 0.4408352668213457
F1 Score : 0.5919003115264797

ABUSIVE
Accuracy : 0.9929712460063898
Precision: 0.4444444444444444
Recall : 0.125
F1 Score : 0.1951219512195122

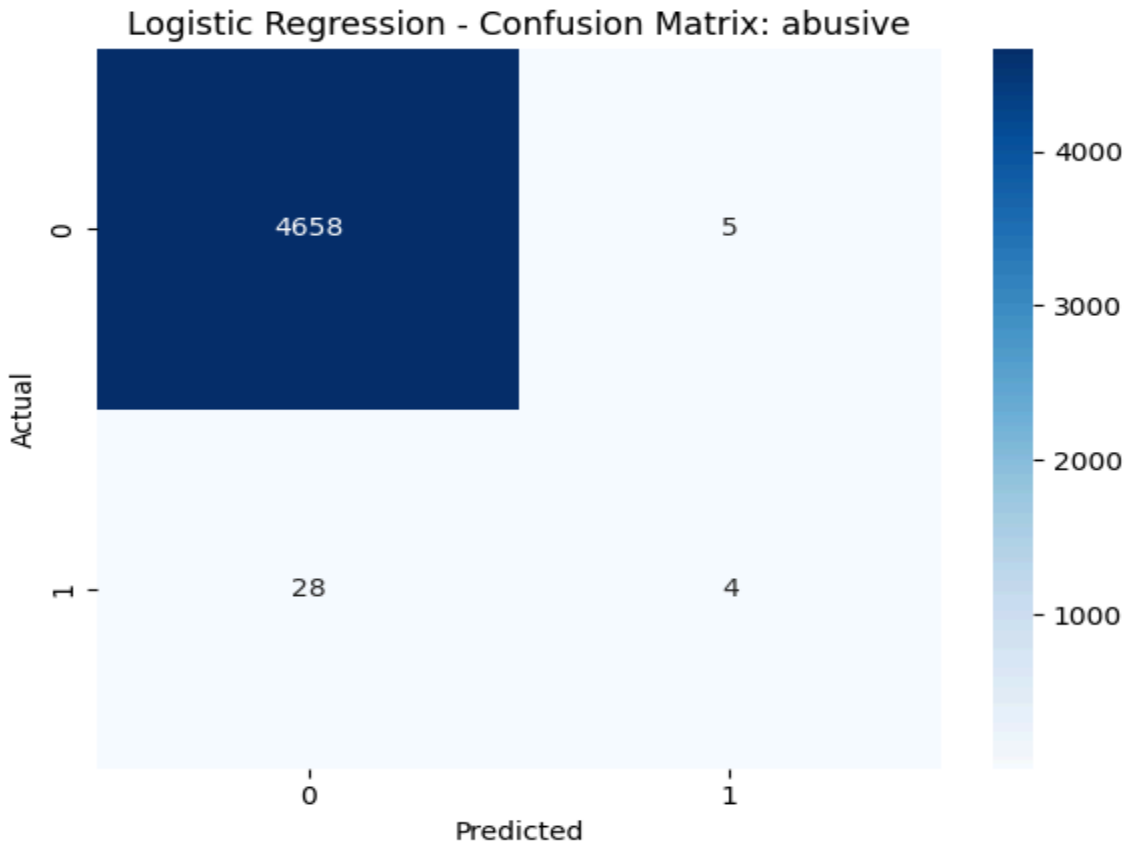
VULGAR
Accuracy : 0.9684771033013845
Precision: 0.8976377952755905
Recall : 0.4578313253012048
F1 Score : 0.6063829787234043

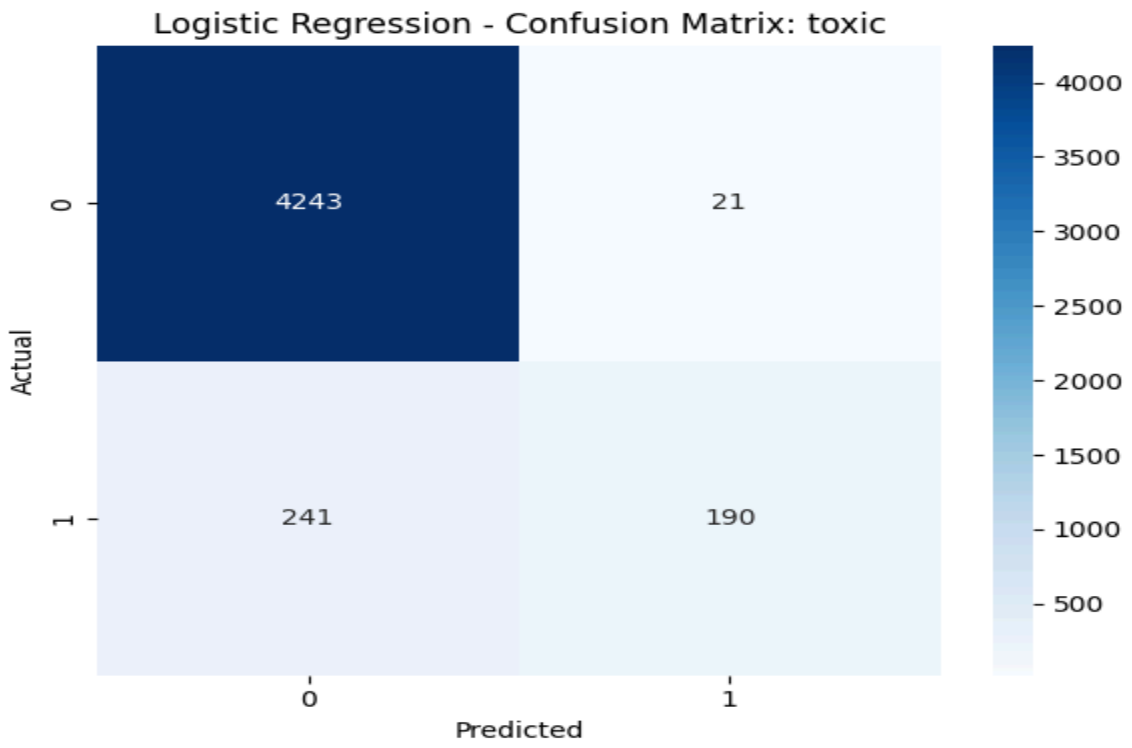
MENACE
Accuracy : 0.9957401490947817
Precision: 0.0
Recall : 0.0
F1 Score : 0.0

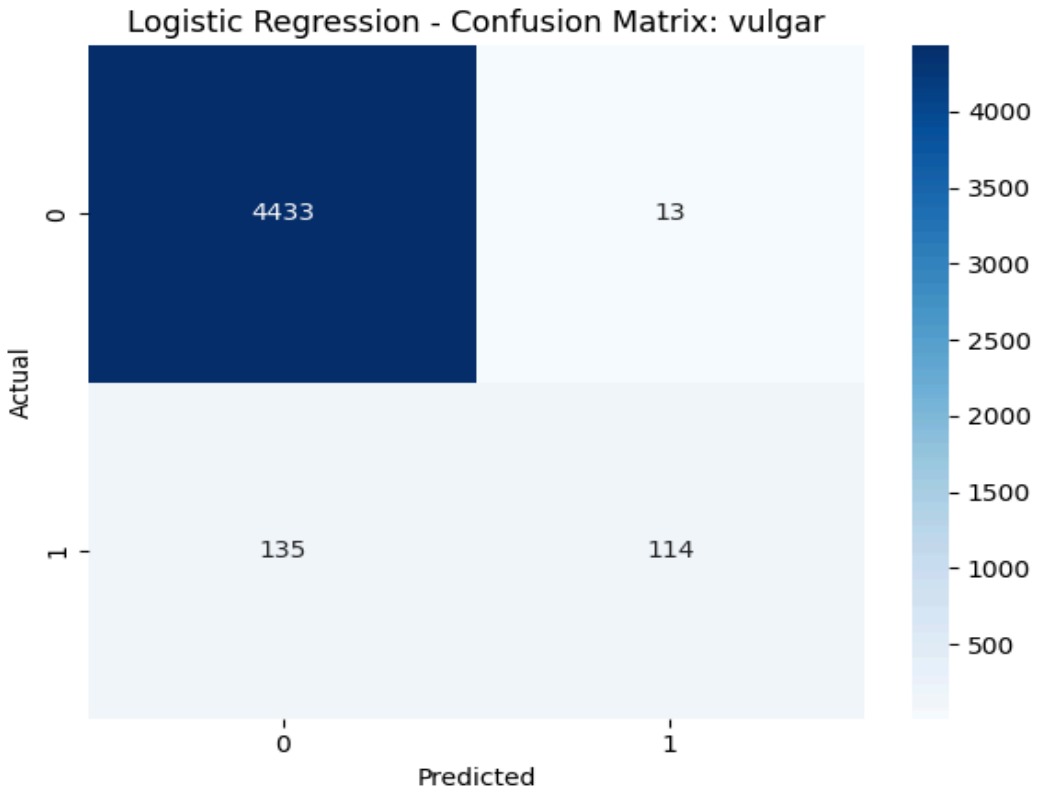
OFFENSE
Accuracy : 0.967199148029819
Precision: 0.9222222222222223
Recall : 0.36086956521739133
F1 Score : 0.51875

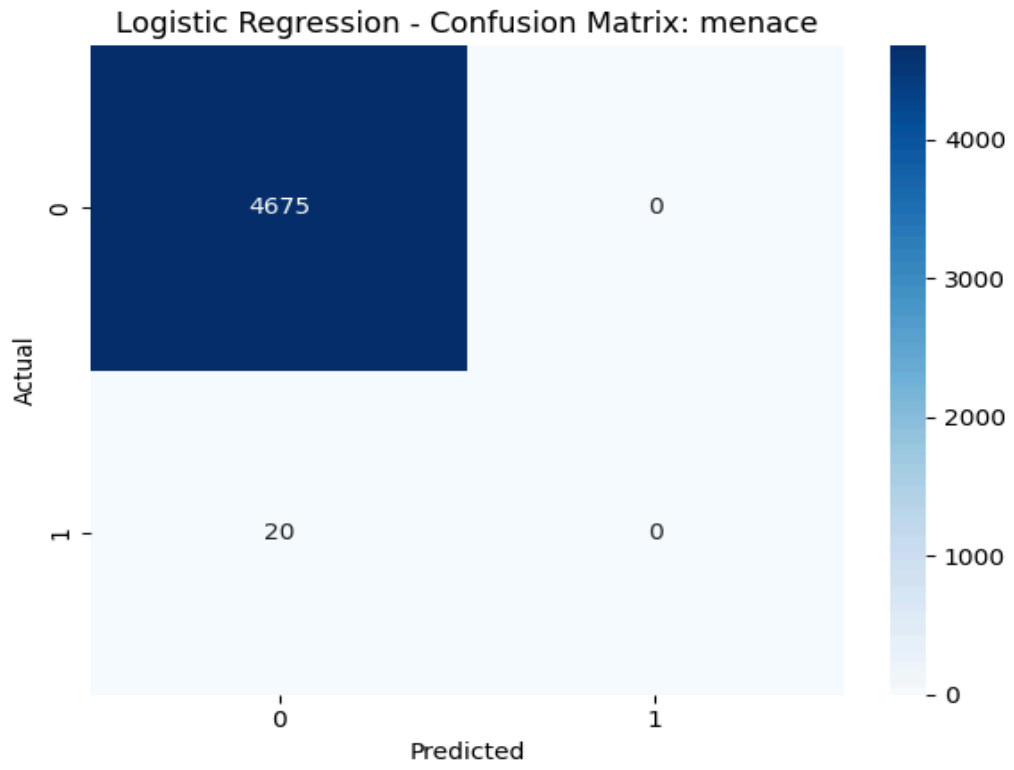
BIGOTRY
Accuracy : 0.9919062832800852
Precision: 0.5
Recall : 0.05263157894736842
F1 Score : 0.09523809523809523

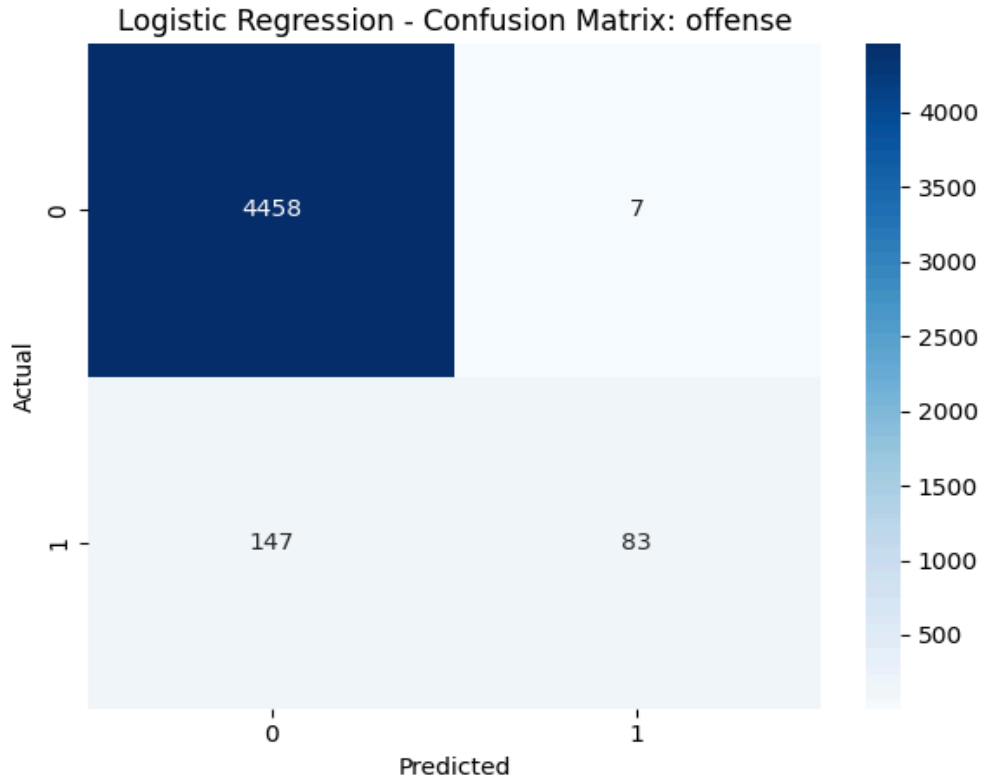
Confusion Metrix:

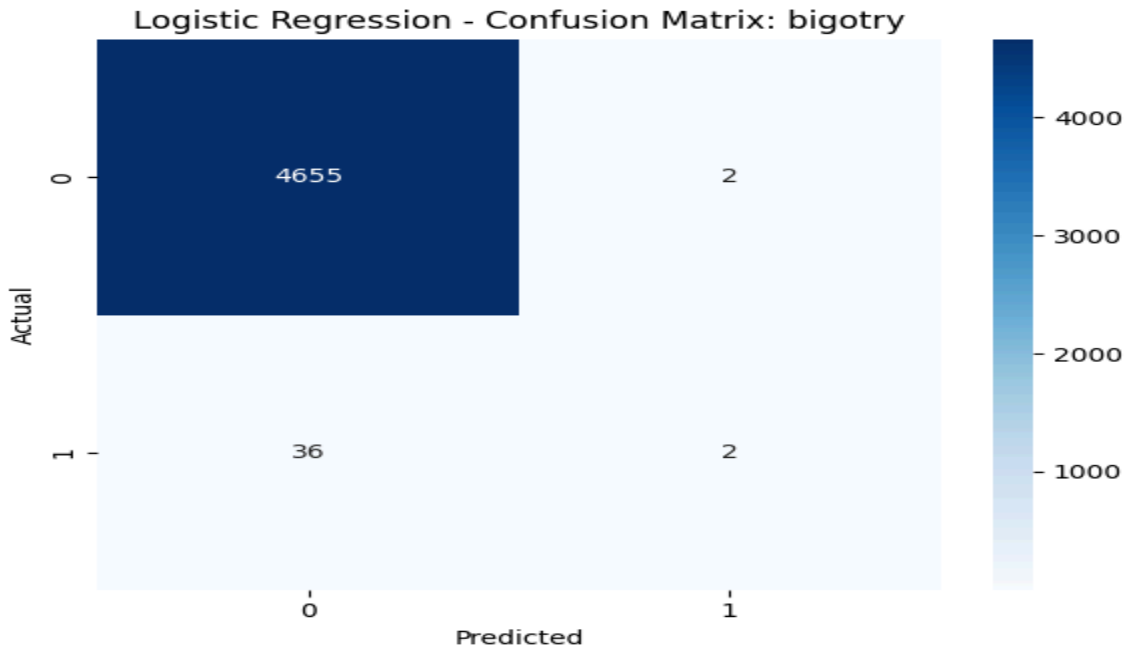

Logistic Regression - Confusion Matrix: abusive
4658
5
28
4
0
1
Actual
0
1
Predicted
4000
3000
2000
1000


Logistic Regression - Confusion Matrix: toxic
4243
21
241
190
Actual
0
1
Predicted
0
1
4000
3500
3000
2500
2000
1500
1000
500


Logistic Regression - Confusion Matrix: vulgar
4433
13
135
114
Actual
0
1
Predicted
0
1
4000
3500
3000
2500
2000
1500
1000
500


Logistic Regression - Confusion Matrix: menace
4675
0
20
0
Actual
Predicted
0
1
4000
3000
2000
1000
0


Logistic Regression - Confusion Matrix: offense
4458
7
147
83
Actual
Predicted
0
1
4000
3500
3000
2500
2000
1500
1000
500



Key Observations:

1. **Severe Class Imbalance Issues**:
    - Extremely high accuracy scores (mostly >90%) contrast sharply with poor recall and F1 scores
    - This pattern indicates the model is biased toward the majority class (non-toxic examples)
2. **Performance by Category**:
    - **Best Performing**: Toxic and Vulgar categories show relatively better F1 scores (~0.6)
    - **Worst Performing**: Menace (complete failure with 0 scores) and Bigotry (F1=0.095)
    - **Precision-Recall Tradeoff**: High precision but low recall across most categories suggests the model is conservative - only predicting positive when very confident
3. **Critical Failures**:
    - The model completely fails to detect Menace comments
    - Bigotry detection only catches 5% of actual cases (recall=0.05)
    - Abusive language detection is particularly weak (recall=0.125)

LSTM Model Evalutaion

TOXIC
Accuracy : 0.9456869009584664
Precision: 0.7514285714285714
Recall : 0.6102088167053364
F1 Score : 0.6734955185659411

ABUSIVE
Accuracy : 0.9931842385516507
Precision: 0.0
Recall : 0.0
F1 Score : 0.0

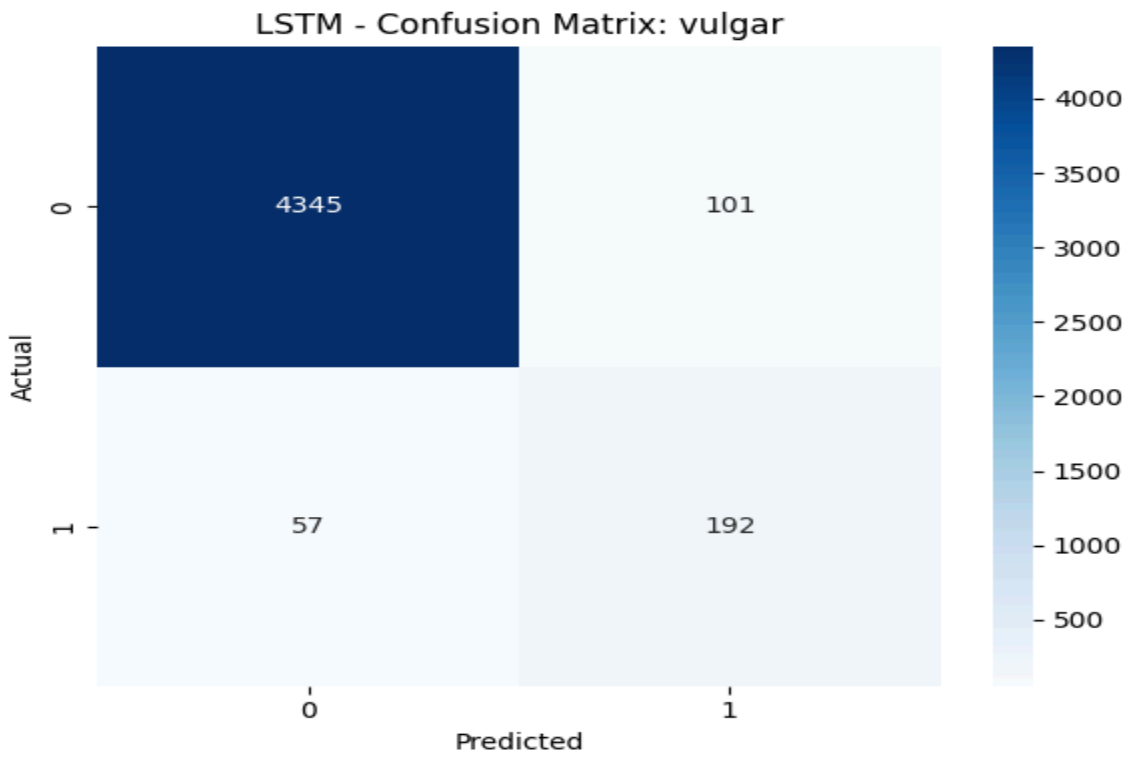
VULGAR
Accuracy : 0.9663471778487753
Precision: 0.6552901023890785
Recall : 0.7710843373493976
F1 Score : 0.7084870848708487

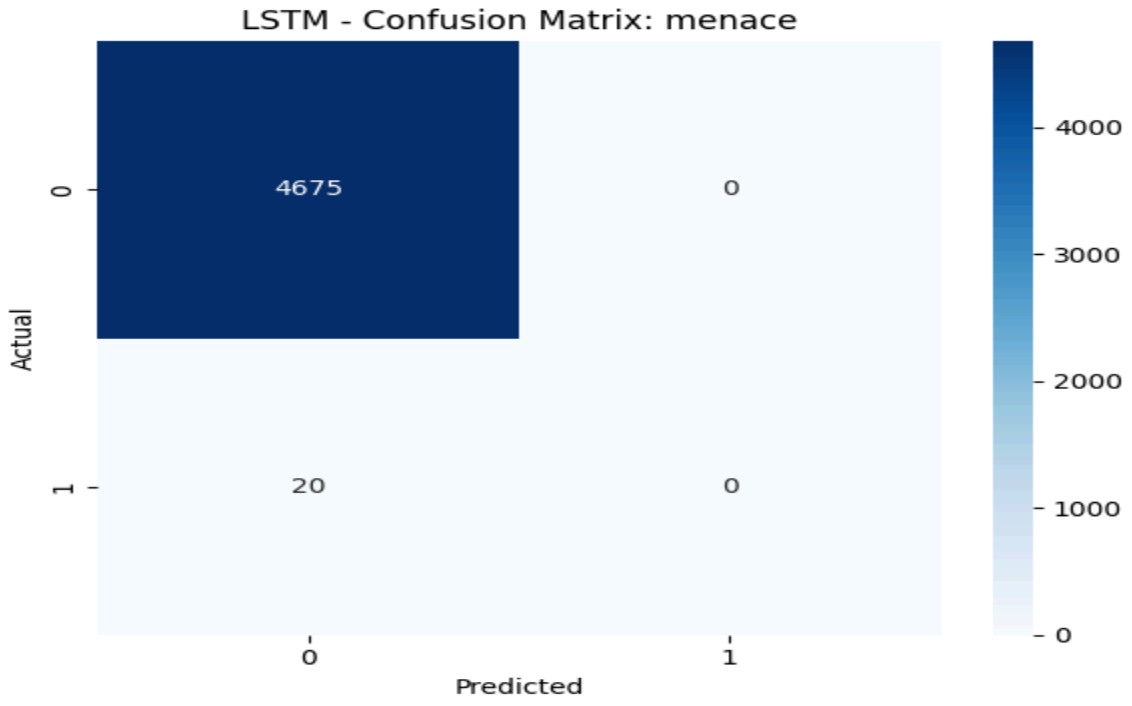
MENACE
Accuracy : 0.9957401490947817
Precision: 0.0
Recall : 0.0
F1 Score : 0.0

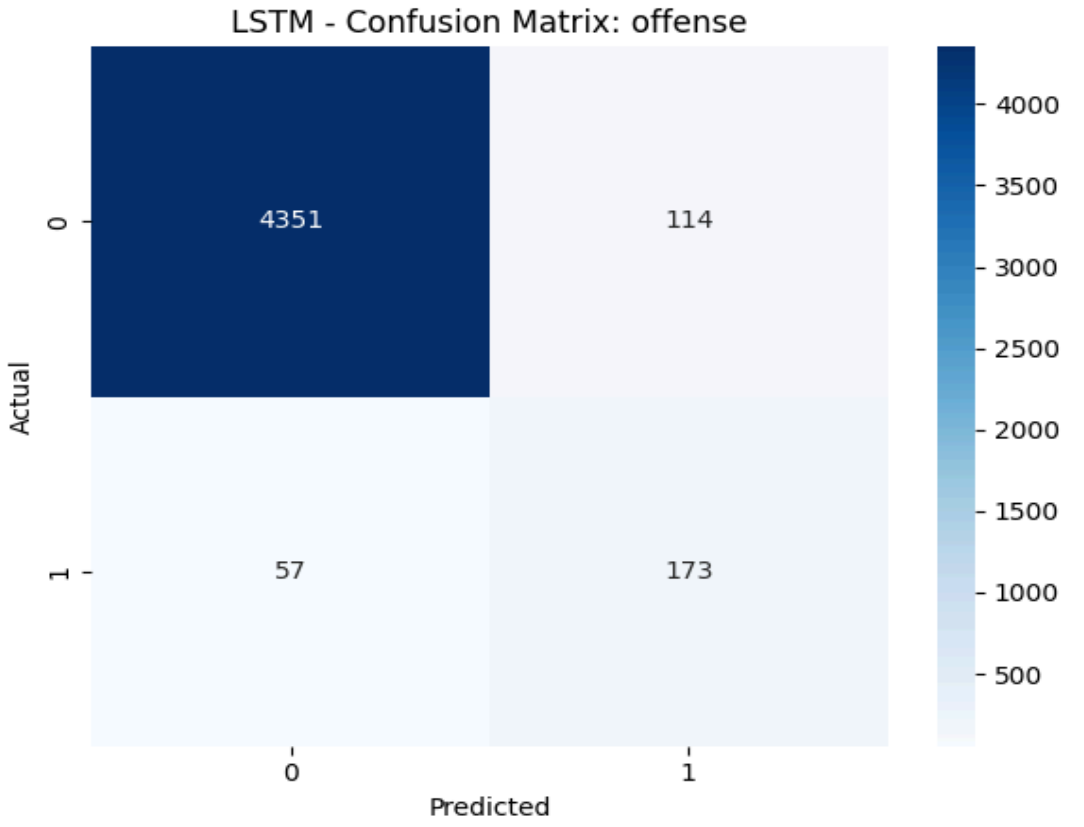
OFFENSE
Accuracy : 0.9635782747603834
Precision: 0.6027874564459931
Recall : 0.7521739130434782
F1 Score : 0.6692456479690522

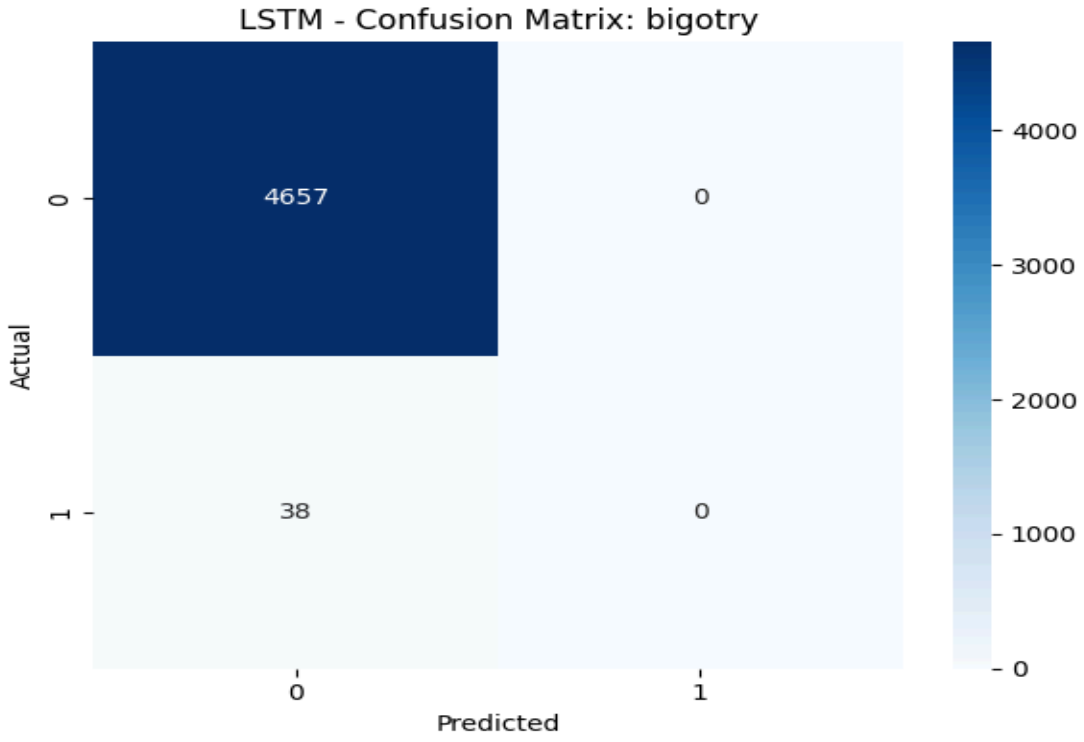
BIGOTRY
Accuracy : 0.9919062832800852
Precision: 0.0
Recall : 0.0
F1 Score : 0.0

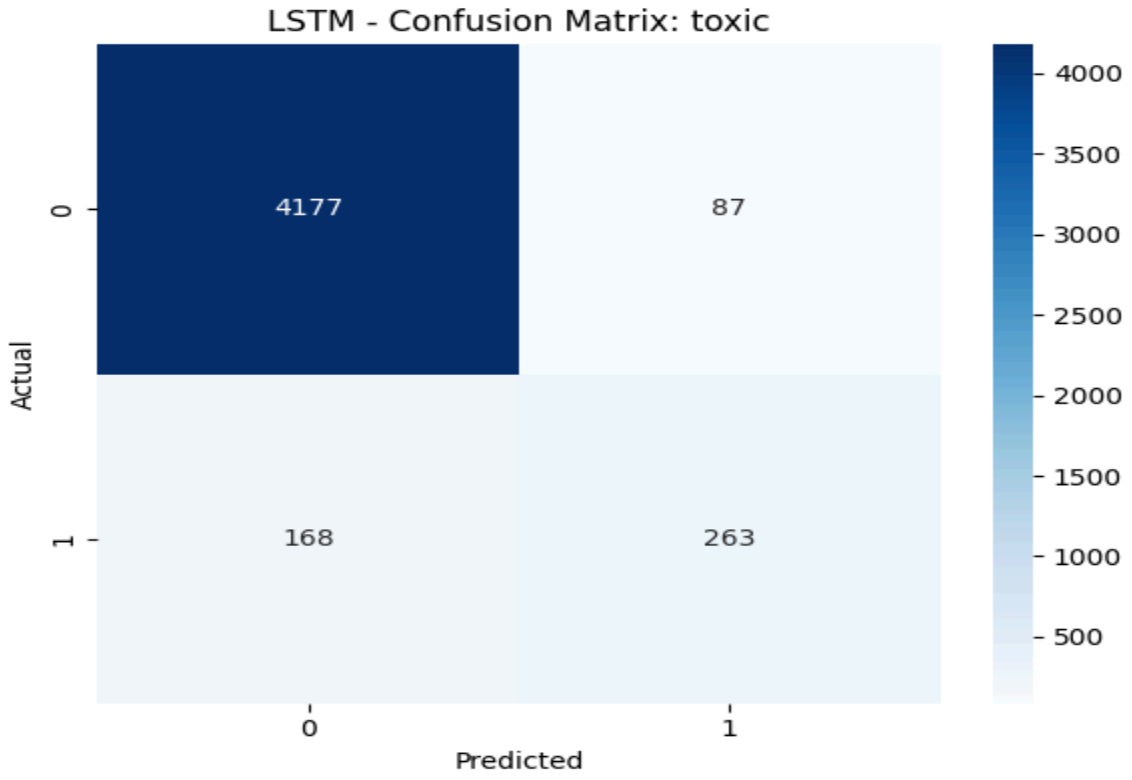
Confusion Metrix

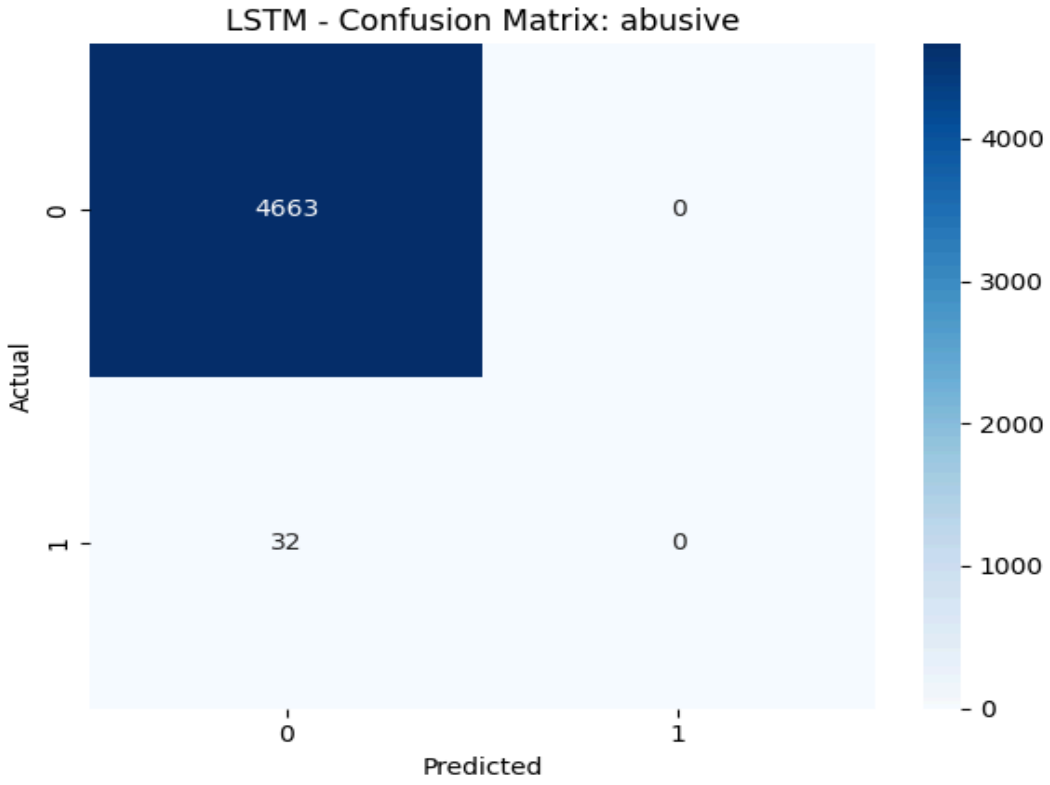

LSTM - Confusion Matrix: vulgar
4345
101
57
192
Actual
Predicted
0
1
4000
3500
3000
2500
2000
1500
1000
500


LSTM - Confusion Matrix: menace
4675
0
20
0
Actual
Predicted
0
1
4000
3000
2000
1000
0


LSTM - Confusion Matrix: offense
4351
114
57
173
Actual
Predicted
0
1
4000
3500
3000
2500
2000
1500
1000
500


LSTM - Confusion Matrix: bigotry
4657
0
38
0
Actual
Predicted
0
1
4000
3000
2000
1000
0


LSTM - Confusion Matrix: toxic
4177
87
168
263
Actual
Predicted
0
1
4000
3500
3000
2500
2000
1500
1000
500


LSTM - Confusion Matrix: abusive
4663
0
32
0
Actual
Predicted
0
1
4000
3000
2000
1000
0

Key Findings by Metric:

1. **Accuracy vs. Recall Disparity**:
    - High accuracy scores (94-99%) mask poor performance on positive cases
    - The model achieves 99.5% accuracy on Menace/Bigotry by simply never predicting these classes
2. **Category-Specific Performance**:
    - **Best Performing**:
        - *Vulgar*: Achieves best balance (F1=0.71, Recall=0.77)
        - *Offense*: Highest recall (0.75) but lower precision (0.60)
    - **Complete Failures**:
        - *Abusive*, *Menace*, *Bigotry*: Zero predictions (Precision=Recall=F1=0)
    - *Toxic*: Moderate performance (F1=0.67) but still misses ~40% of cases
3. **Comparative Improvements**:
    - Better recall than logistic regression for Toxic (+17%), Vulgar (+31%), and Offense (+39%) categories
    - Worse precision for Toxic (-15%) and Vulgar (-24%) compared to logistic regression

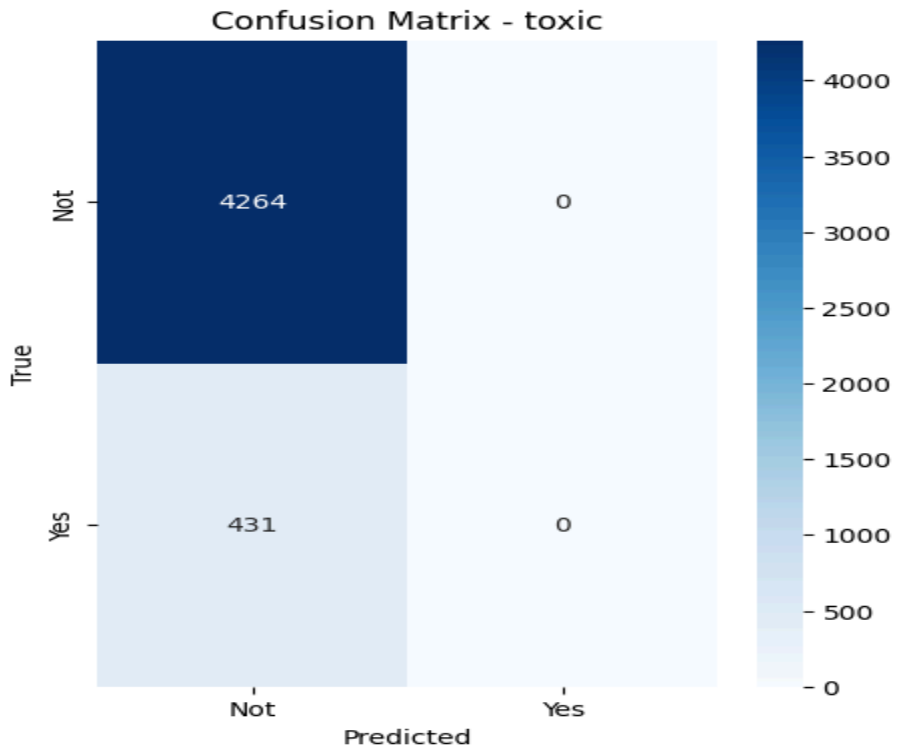
Critical Issues Identified:

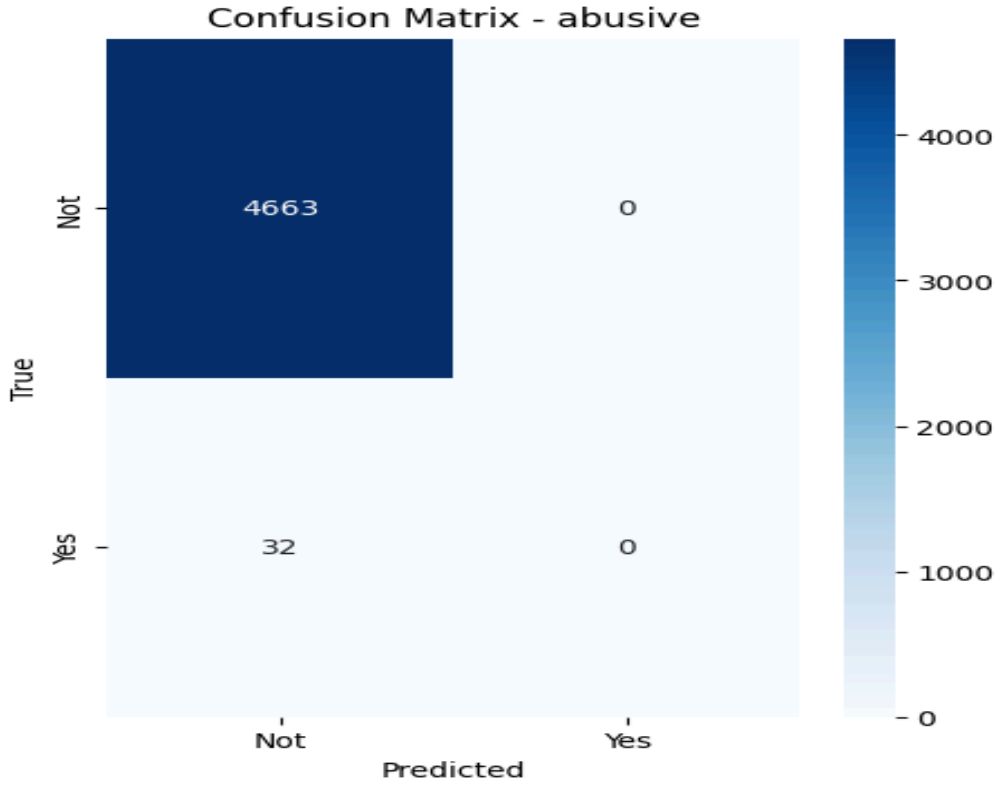
1. **Structural Model Limitations**:
    - LSTM's sequential processing may be inadequate for detecting subtle toxicity patterns
    - Complete failure on 3/6 categories suggests fundamental representation problems
2. **Training Data Problems**:
    - Likely insufficient examples of Abusive, Menace, and Bigotry in training set
    - Possible annotation inconsistencies for these categories
3. **Decision Threshold Problems**:
    - Default 0.5 threshold appears too conservative for rare classes

BERT Model Evaluation

	precision	recall	f1-score	support
toxic	0.00	0.00	0.00	431
abusive	0.00	0.00	0.00	32
vulgar	0.00	0.00	0.00	249
menace	0.00	0.00	0.00	20
offense	0.00	0.00	0.00	230
bigotry	0.00	0.00	0.00	38
micro avg	0.00	0.00	0.00	1000
macro avg	0.00	0.00	0.00	1000
weighted avg	0.00	0.00	0.00	1000
samples avg	0.00	0.00	0.00	1000

Confusion Metrix(Show only for toxic and abusive class here)





Analysis of Zero Performance in BERT Model Evaluation

Root Cause Identification:

1. **Complete Prediction Failure**:
    - Your model is predicting all zeros (non-toxic) for every sample
    - This results in 0 precision, 0 recall, and 0 F1-score across all categories
    - The "support" column shows the actual positive cases being completely missed
2. **Primary Suspects**:
    - **Severe Class Imbalance**: The model learned to always predict negative as the "safe" choice
    - **Incorrect Prediction Threshold**: Using default 0.5 threshold may be too conservative
    - **Training Issues**: Model may not have actually learned meaningful patterns
    - **Implementation Bug**: Possible error in model compilation or evaluation