DEEP LEARNING-BASED TEXT CLASSIFICATION FOR HATE SPEECH IN ONLINE SOCIAL NETWORKS

MAJOR PROJECT

CEN-891

Supervisor :

Dr. Faiyaz Ahmad

Presented By:

Faizan Ahamad 21BCS047

Mantasha Firdous 21BCS049

DEPARTMENT OF COMPUTER ENGINEERING

FACULTY OF ENGINEERING & TECHNOLOGY

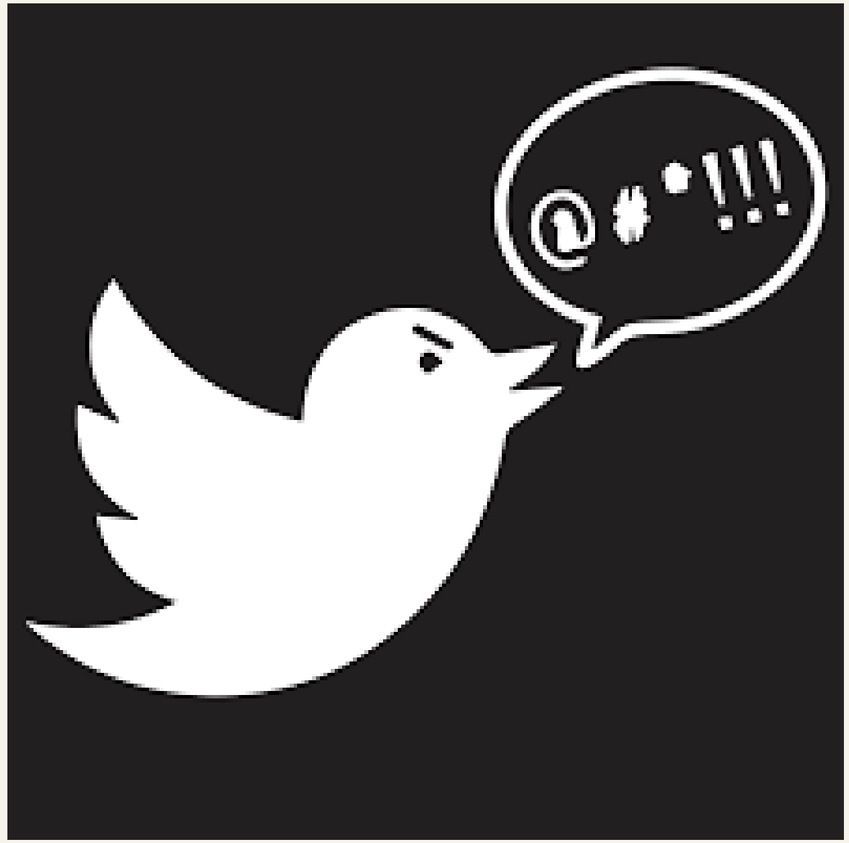
Jamia MIllia Islamia| 2025

CONTENT

- Introduction
- Problem Statement
- How X (Twitter) Detects and Bans Hate Speech
- Literary Review
- Research Gap
- Workflow
- Preprocessing
- Dataset Overview
- Model Architecture
- Experimental Setup
- Experimental Results
- Conclusion
- Future Works
- References

INTRODUCTION

- **The Rise of Social Media:**
  - With affordable smartphones and internet services, platforms have revolutionized global communication.
- **The Dark Side of Connectivity:**
  - The same platforms enabling connection and expression also amplify hate speech and offensive content.
- **The Consequences:**
  - It leads to mental health issues, social polarization, and in extreme cases, cyberbullying-induced suicides.



PROBLEM STATEMENT

🧩 **Problem Statement**

To develop robust automated systems capable of classifying tweets into hate-related categories, using two distinct datasets:

Dataset 1 (D1): Binary Classification

- Objective: Classify tweets as Hate Speech or Non-Hate Speech

Dataset 2 (D2): Multi-Class Classification

- Objective: Categorize tweets into:
  - Hate Speech
  - Offensive Speech
  - Neutral Speech

HOW X (TWITTER) DETECTS AND BANS HATE SPEECH

- Policy Enforcement:
  - X has a comprehensive Hateful Conduct Policy that prohibits behavior targeting individuals or groups based on attributes such as race, ethnicity, gender, religion, and more.
- Automated Detection Systems:
  - The platform utilizes machine learning algorithms to identify and flag potentially hateful content.
- Human Moderation:
  - Human moderators review reported content to assess context and intent, ensuring accurate enforcement of policies.

HOW X (TWITTER) DETECTS AND BANS HATE SPEECH

- User Reporting Mechanism:
  - User's reports are reviewed by the moderation team, and appropriate actions are taken based on the findings.
- Suspension Warnings:
  - Research indicates that issuing warnings to users about potential account suspensions can effectively reduce the use of hateful language.
- Account Suspensions:
  - For severe or repeated violations, X suspends accounts to prevent further dissemination of hate speech. In the first half of 2023, the platform suspended nearly 5.3 million accounts for policy violations, including hateful conduct.

Literary review

Authors	Paper Title & Publisher	Model/ Architecture	Datasets	Remarks
Pradeep Kumar Roy, Asis Kumar Tripathy, Tapan Kumar Das, Xiao-Zhi Gao	A Framework for Hate Speech Detection Using DCNN IEEE Access (2020)	Deep Convolutional Neural Network (DCNN)	Kaggle Twitter Hate Speech Dataset (31,962 tweets: 92.98% NHS, 7.02% HS)	- Outperformed traditional ML models - Achieved F1-score of 0.92 using 10-fold cross-validation - LSTM and CNN-LSTM models underperformed

Literary review

Authors	Paper Title & Publisher	Model/ Architecture	Datasets	Remarks
Jitendra Singh Malik, Hezhe Qiao, Guansong Pang, Anton van den Hengel	Deep Learning for Hate Speech Detection: A Comparative Study (International Journal of Data Science and Analytics) (Springer-2024)	Shallow Models: • SVM • XGBoost • MLP Deep Learning Models: • CNN • LSTM	• **HateXplain** • **Hate Speech Offensive (Davidson dataset)** • **OLID (Offensive Language Identification Dataset)**	• The study provides a comprehensive empirical comparison across 14 different hate speech detection models. • Transformer-based models (especially BERT and RoBERTa) showed superior performance in terms of accuracy and generalization. • Deep models significantly outperform shallow models but are computationally more expensive.

Literary review

Authors	Paper Title & Publisher	Model/ Architecture	Datasets	Remarks
Suraj Futane, Twinkal Bandwal, Dnyaneshwari Dhonde, Sakshi Gudmewar, Aishwarya Kadam	Hate Speech Detection Using Machine Learning, IJRASET (2023)	Support Vector Machine (SVM), Logistic Regression, Random Forest	**Twitter dataset from Kaggle**	Achieved 0.9668 accuracy with SVM and TF-IDF after preprocessing. Focused on detecting hate speech using machine learning algorithms. Best performance with SVM and TF-IDF after preprocessing and grid search.

RESEARCH GAP

1

Feature Extraction Limitations:

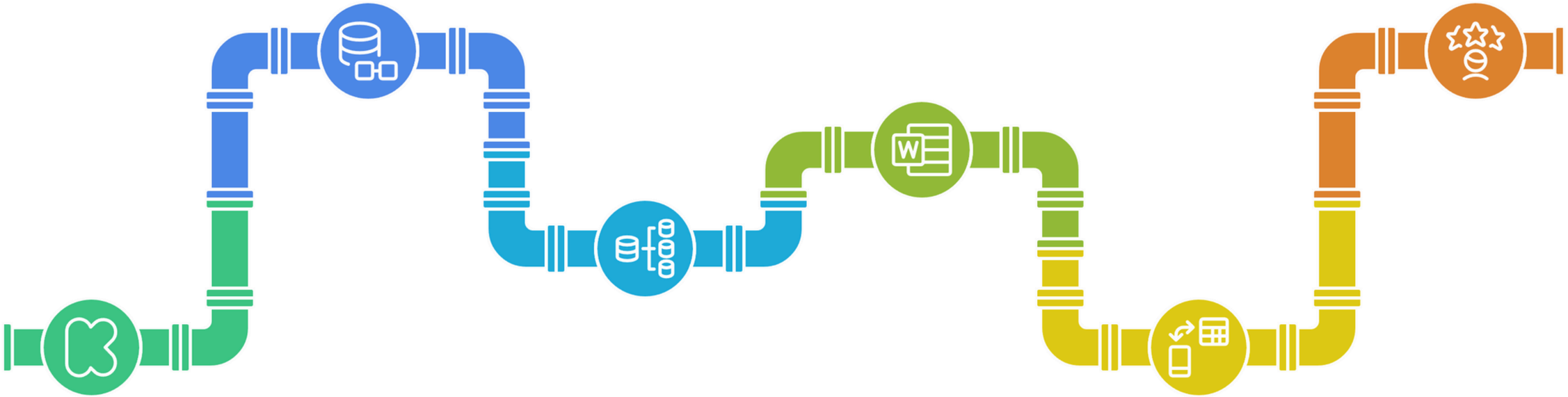
The study uses TF-IDF for feature extraction, which, while effective, may not capture the full semantic meaning of the text. More advanced techniques like word embeddings BERT could potentially improve performance but are not explored in this study.

2

Data Imbalance Handling:

The study does not address class imbalance, which can hinder model performance. Techniques like SMOTE and data augmentation could improve minority class representation and boost classification metrics, but are not explored.

WORKFLOW



01

Start

Initiate the project workflow

02

Load Datasets

Import binary and multi-class datasets

03

Preprocessing Pipeline

Apply preprocessing steps to datasets

04

Generate Word Embeddings

Create embeddings using GloVe or BERT

05

Train Deep Learning Models

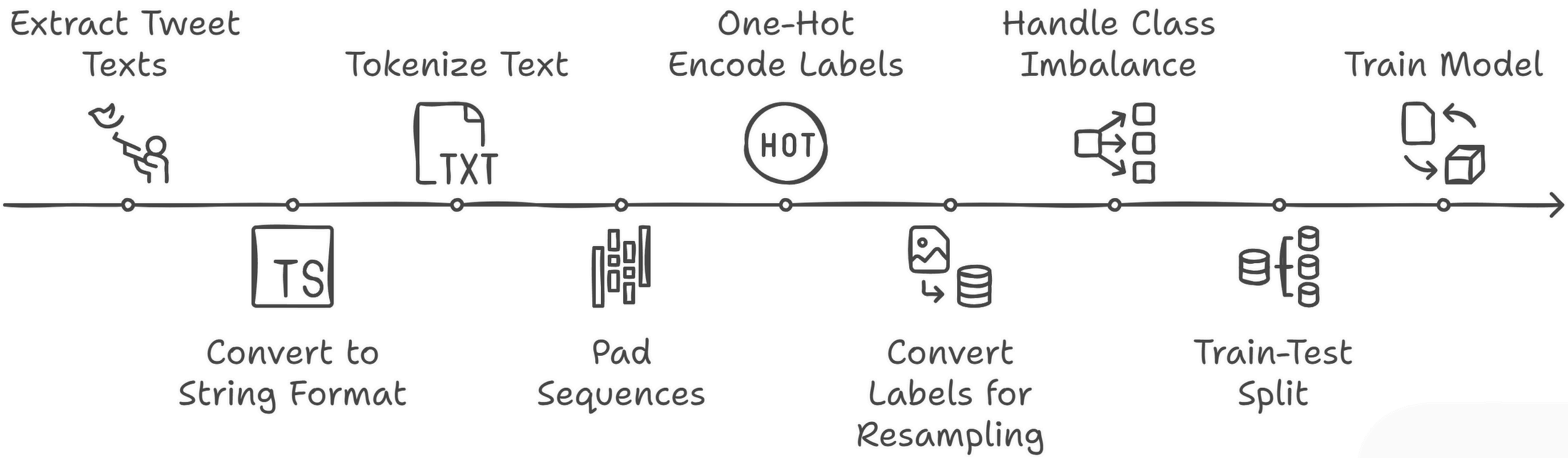
Train CNN and other models

06

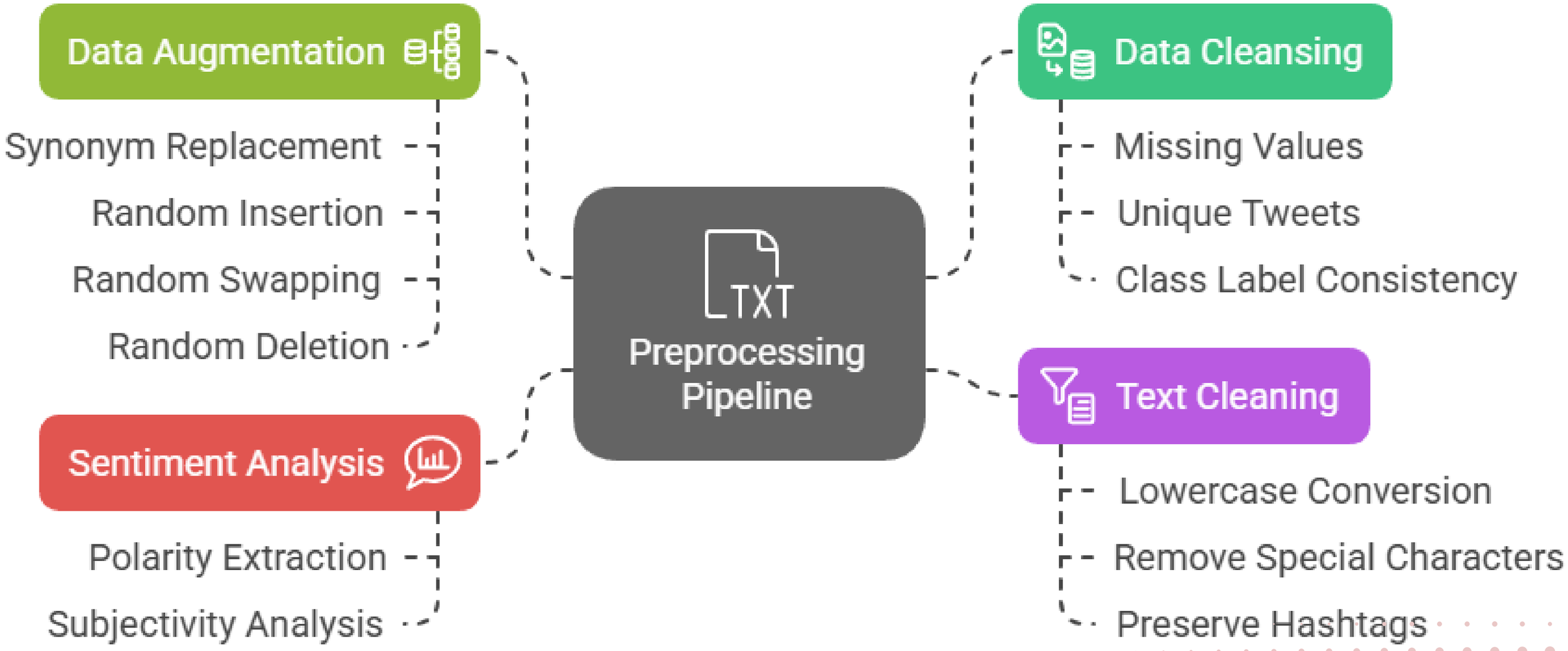
Evaluate Models

Assess model performance with metrics

PREPROCESSING FOR DATASET 1



PREPROCESSING FOR DATASET II

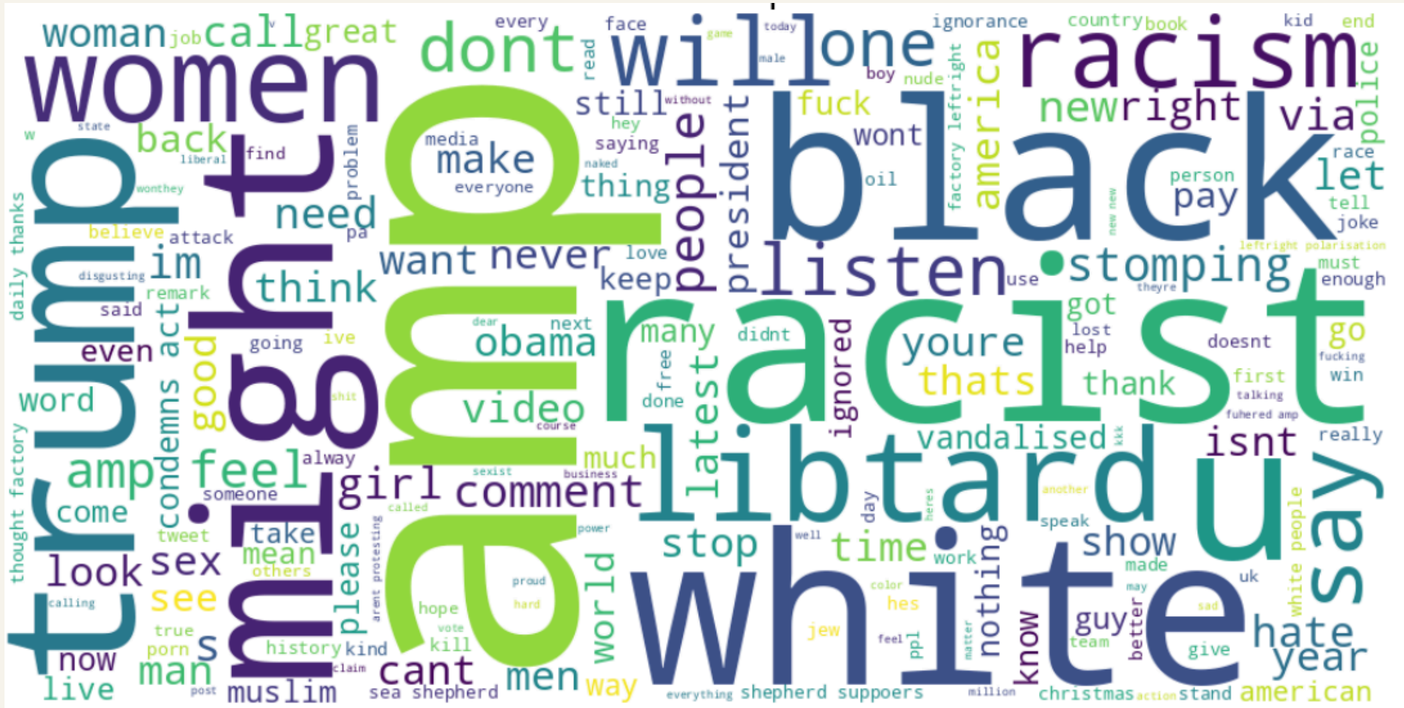


Made with Napkin

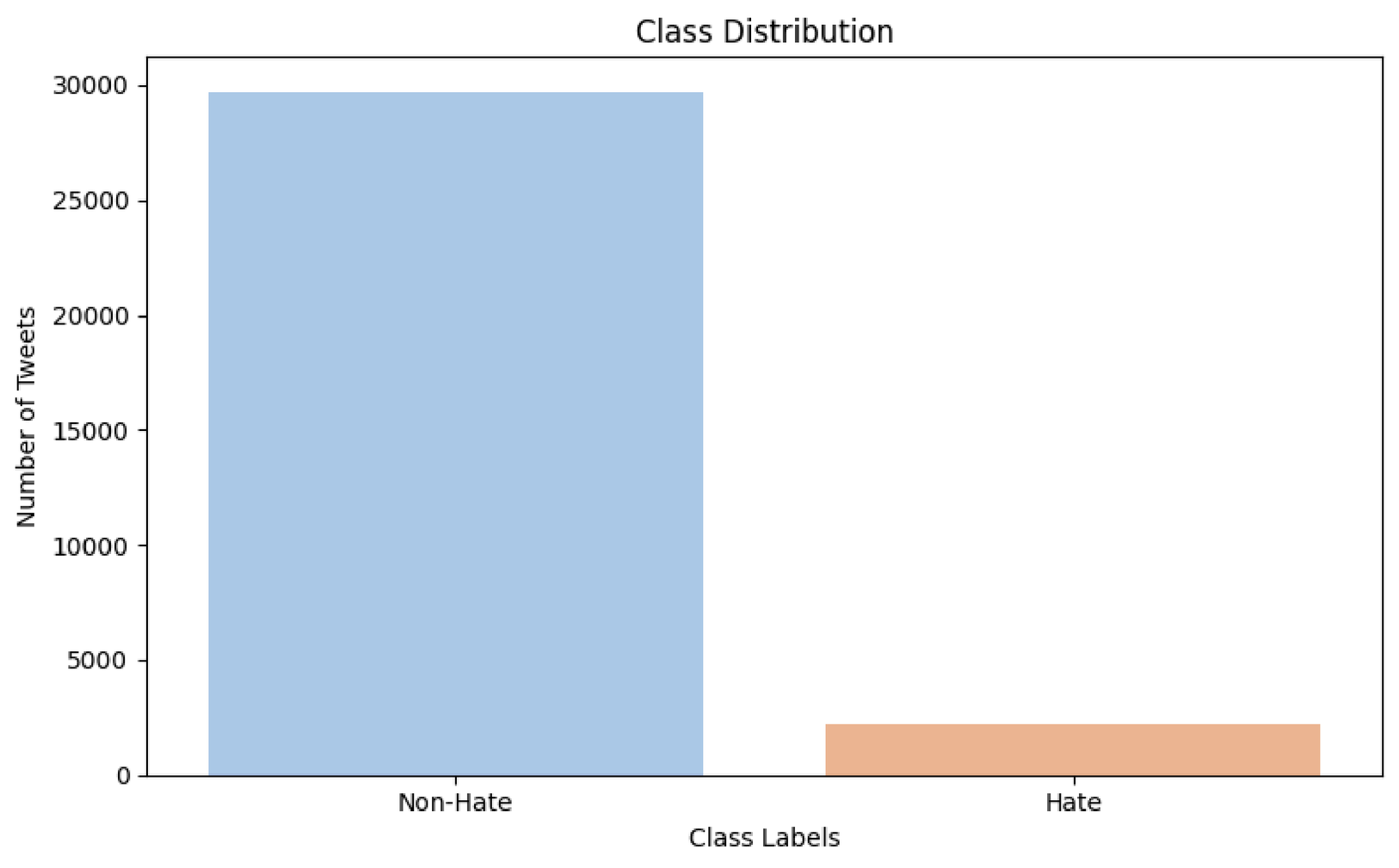
DATASET - I

- **Twitter Hate Speech**

- **Total Text Samples:** 31,962 English tweets
- **Classification Categories:** Two distinct labels
- **Hate Speech (HS)** – 2,242 tweets (7.02%)
- **Non-Hate Speech (NHS)** – 29,720 tweets (92.98%)
- **Data Source:** Collected from Twitter
- **Annotations:** Each tweet is manually labeled as Hate or Non-Hate to ensure accurate binary classification
- **Data Format:** CSV file with fields for tweet ID, tweet text, and binary label

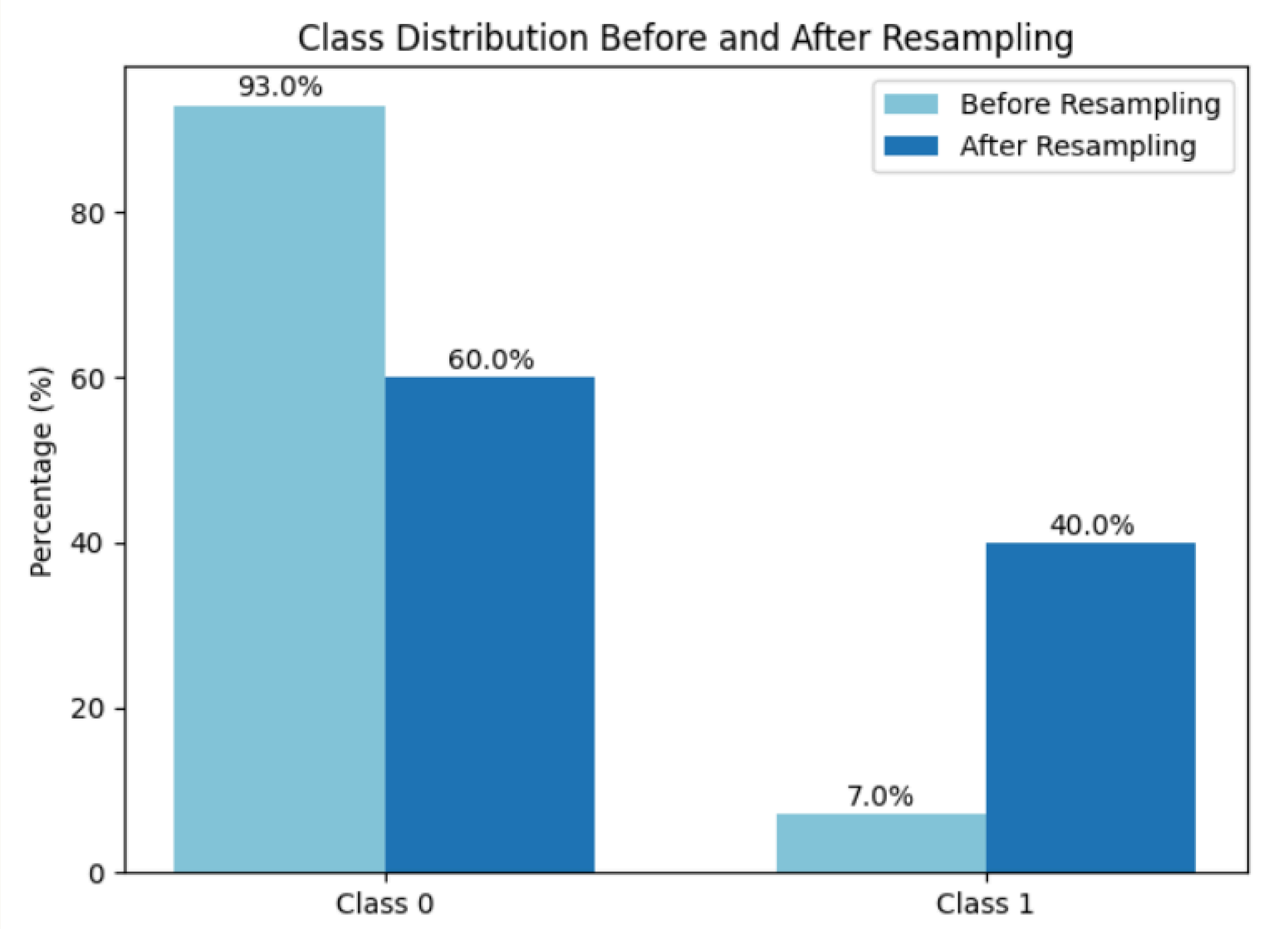

women
trump
might
amp
white
black
racist
libtard
u
racism
will
one
dont
call
great
woman
back
need
make
people
president
listen
stomping
want
never
think
obama
video
feel
girl
comment
stop
time
show
say
hate
year
american
christmas
shepherd
suppoers
sea shepherd
muslim
cant
men
way
world
latest
ignored
vandalised
isnt
youre
thats
thank
newright
via
let
pay
fuck
wont
america
still
thing
keep
many
much
look
sex
see
man
now
live
come
word
amp
even
im
condemns
please
take
mean
nothing
know
guy
hate
give
stand
go
doesnt
joke
tell
enough
must
police
race
person
people
white
better
made
uk
speak
work
hes
jew
kill
history
kind
hope
porn
true
someone
tweet
good
act
said
remark
disgusting
believe
attack
find
problem
media
everyone
saying
hey
without
read
every
face
ignorance
country
book
kid
end
boy
nude
oil
factory
leftright
polarisation
daily thanks
thought factory
fucking
win
really
first
talking
fuhered amp
another
made
may
sad
color
hard
proud
power
business
course
sexist
called
claim
post
everything
million
action
theyre
use
lost
help
didnt
done
free
ive
going
next
dear
shit
kkk
today
male
game
state
liberal
worthey
naked
calling
arent protesting
day
well
theres
matter
feel
ppl
team
hard
vote
sexist
alway
shit
claim
kill
hope
kind
history
porn
true
calling
post
sea shepherd
everything
shepherd
suppoers
million
christmas
action
stand
american
year
hate
give
better
made
uk
white people
speak
work
hes
jew
know
nothing
time
stop
show
comment
girl
latest
ignored
vandalised
isnt
really
go
win
fucking
first
talking
fuhered amp
another
doesnt
joke
tell
enough
must
police
race
person
people
white
better
made
uk

DATASET -1 OVERVIEW



class	count
0	29720
1	2242

DATASET -1 OVERVIEW



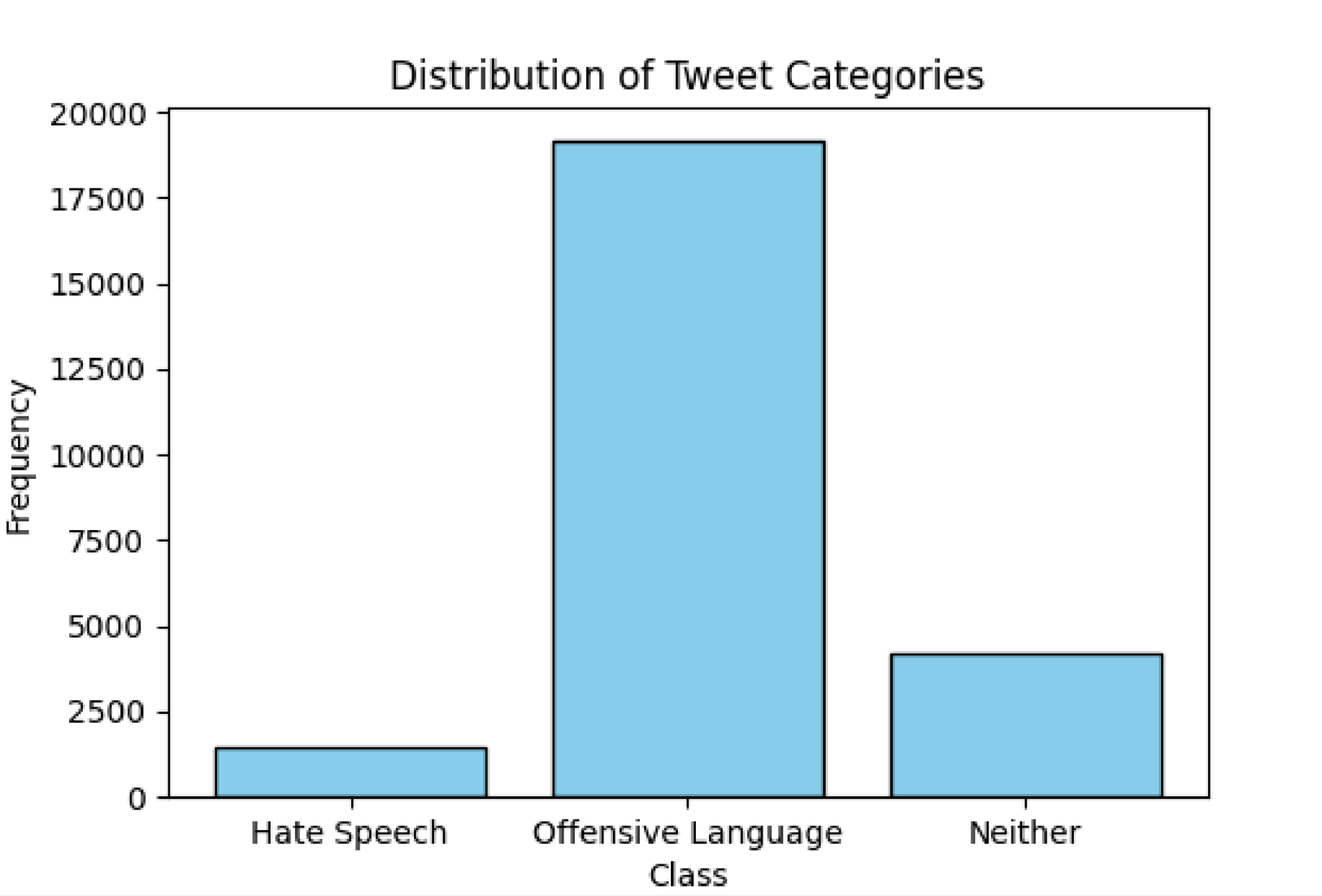
DATASET - II

- **Hate Speech and Offensive Language**

  - Total Text Samples: 24,783 tweets
  - Classification Categories: Three distinct labels:
    - Hate Speech
    - Offensive Language
    - Neither
  - Data Source: Collected from Twitter
  - Annotations: Each tweet is manually labeled to facilitate accurate classification
  - Data Format: Provided in CSV format, including fields for tweet ID, tweet text, and corresponding label

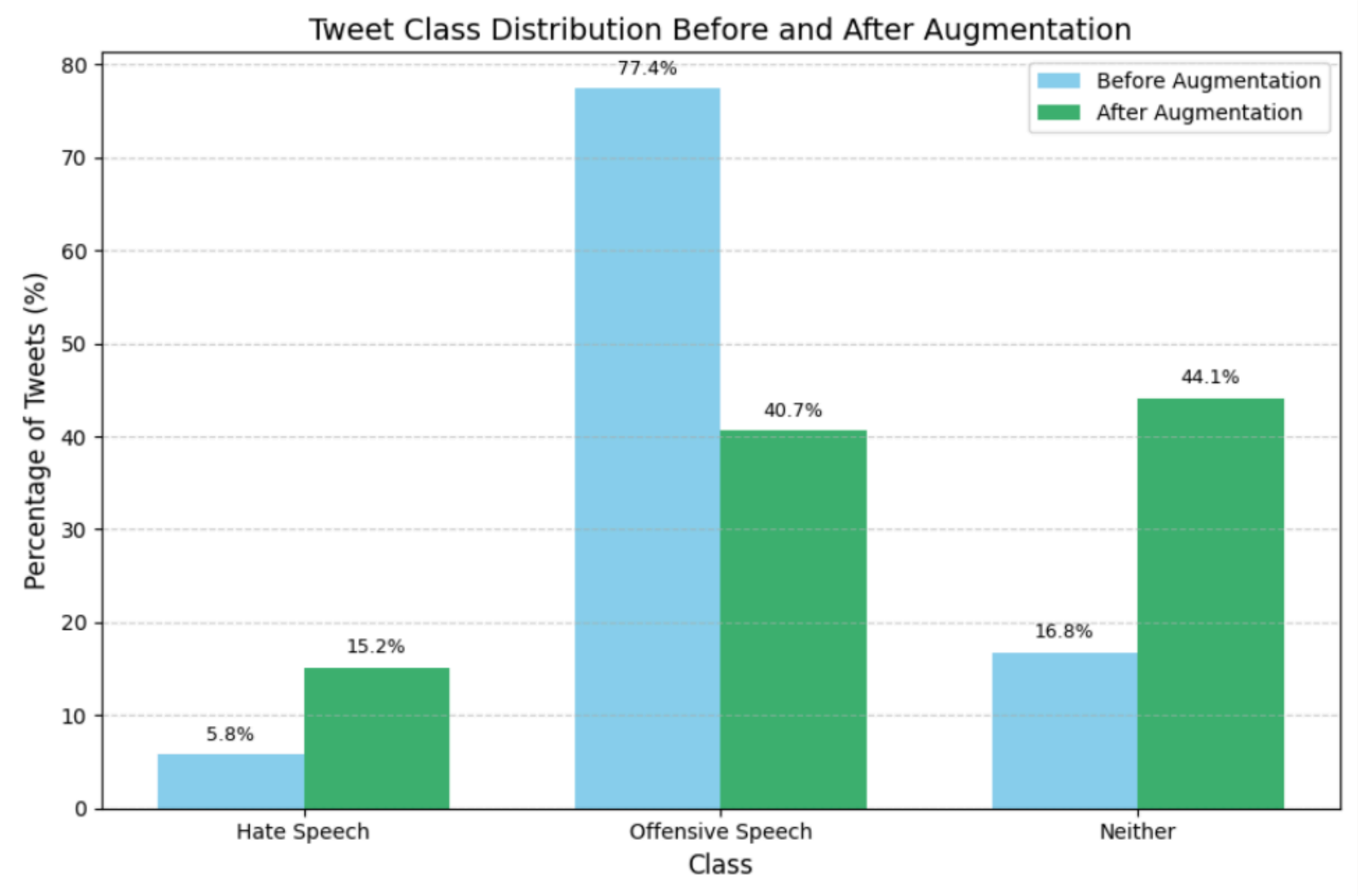
WORDCLOUD OF THE DATASET

DATASET -II OVERVIEW

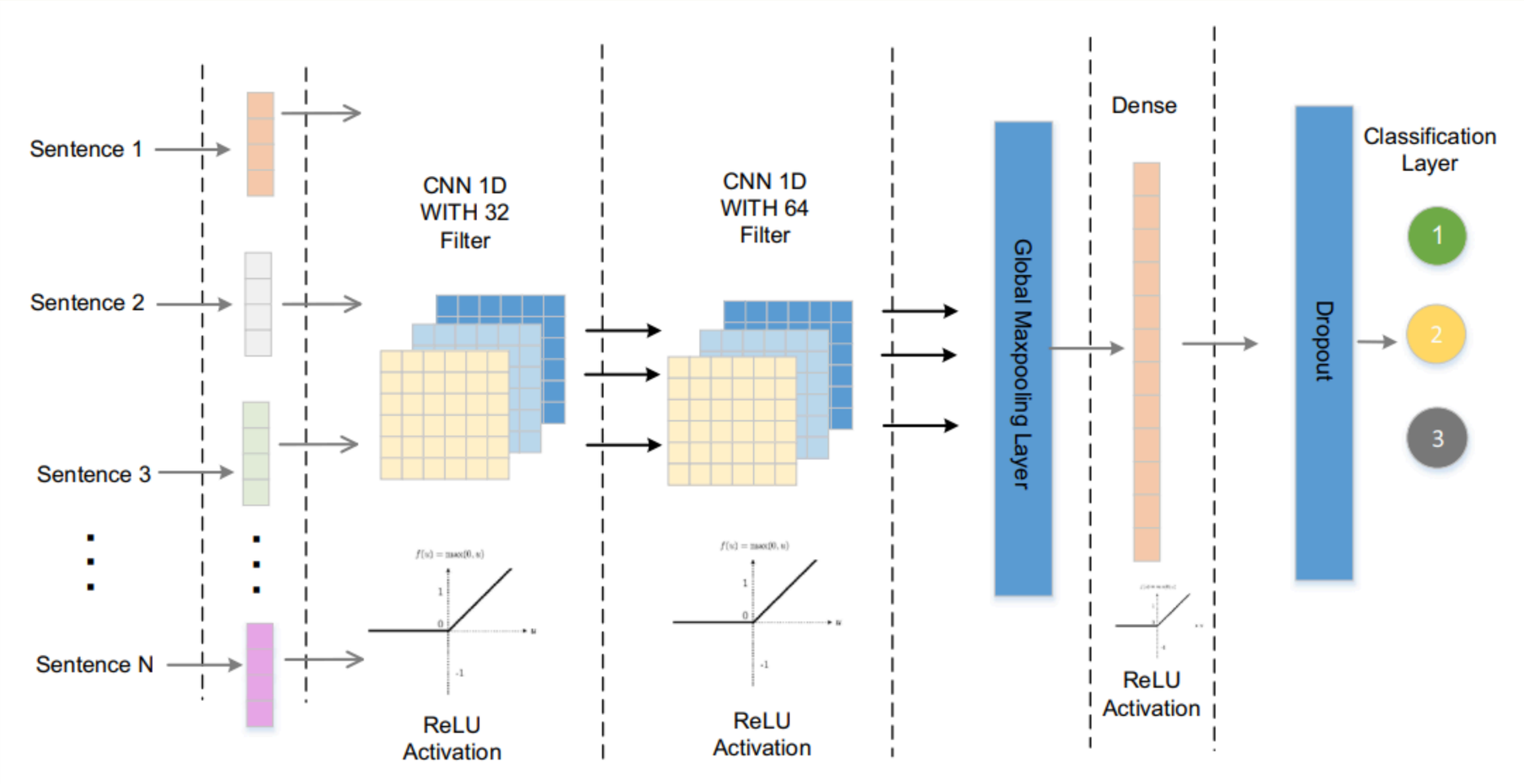


class	count
1	19190
2	4163
0	1430

DATASET -II AFTER AUGMENTATION

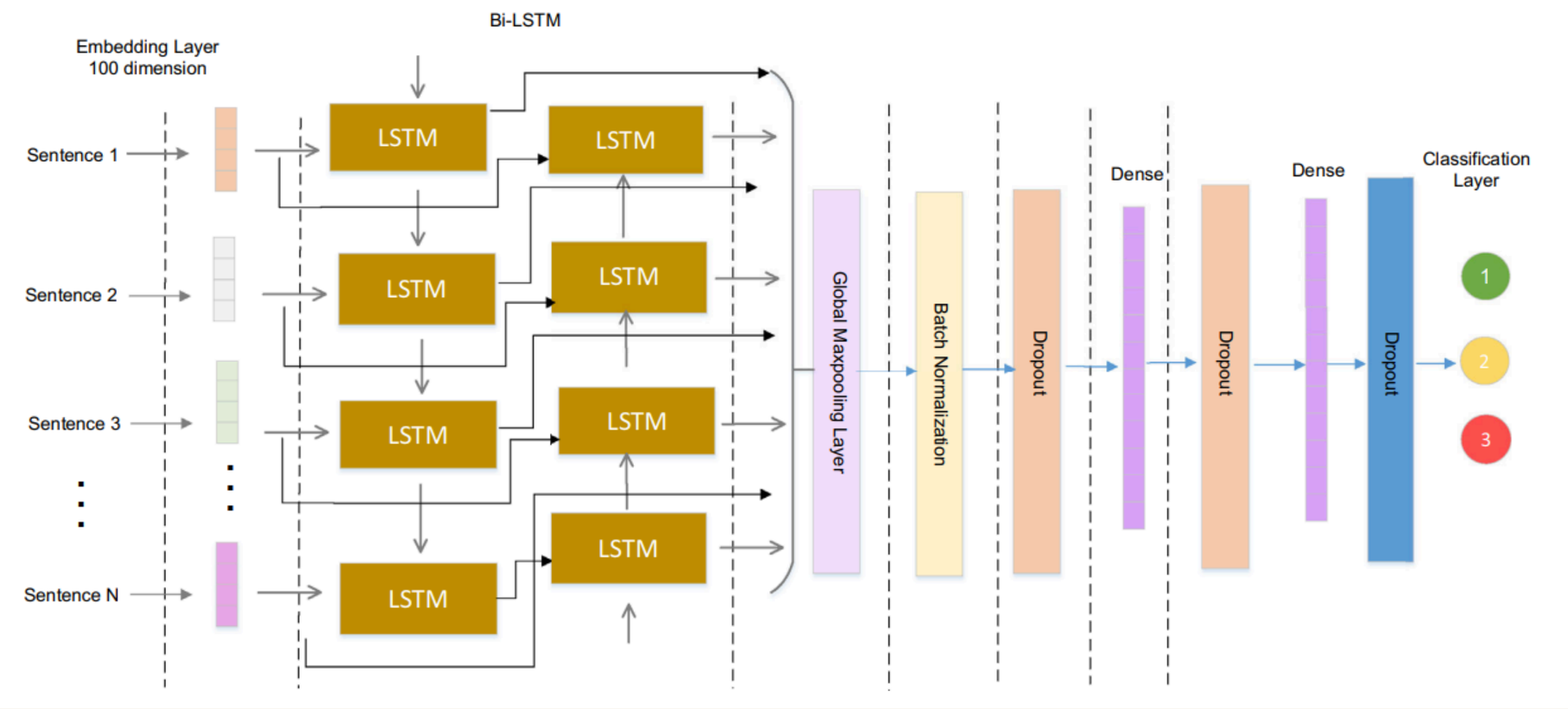


MODEL ARCHITECTURE FOR DATASET II



CNN ARCHITECTURE USED IN [2]

MODEL ARCHITECTURE FOR DATASET II



Bi-LSTM ARCHITECTURE USED IN [2]

EXPERIMENTAL SETUP

Aspect	Dataset 1 (D1)	Dataset 2 (D2)
Classification Type	**Binary Classification** (Hate / Non-Hate)	**Multi-Class Classification** (Hate, Offensive, Neutral)
Text Preprocessing	- Tokenization (Keras) - Padding (max_length = 30)	Same as D1
Embeddings Used	- **GloVe (glove.6B.100d.txt)** - Embedding dim: 100 - Non-trainable	- **GloVe (Twitter 100d)** - **TF-IDF (5000 features)** - **Transformer-based embeddings only**: BERT, Small BERT, AL-BERT, ELECTRA
Models Evaluated	- CNN - LSTM - cLSTM	- CNN, MLP, Bi-LSTM - SVM, XGBoost
Train-Test Split	80/20	80/20
Validation Split	20% of training data	Same as D1
Loss Function	Categorical Crossentropy	Categorical Crossentropy
Optimizer	RMSprop	Adam
Epochs	50	50
Batch Size	50	50

EVALUATION METRICES

- **PRECISION (PI):** THE PROPORTION OF TRUE POSITIVE PREDICTIONS AMONG ALL PREDICTED POSITIVES FOR A CLASS.

  PRECISION = TP / (TP + FP)

- **RECALL (RI):** THE PROPORTION OF TRUE POSITIVE PREDICTIONS AMONG ALL ACTUAL POSITIVE INSTANCES FOR A CLASS.

  RECALL = TP / (TP + FN)

- **F1 SCORE:** THE HARMONIC MEAN OF PRECISION AND RECALL, BALANCING BOTH FALSE POSITIVES AND FALSE NEGATIVES.

  F1 = 2 × (PI × RI) / (PI + RI)

- **MACRO F1 SCORE** : AVERAGES THE F1 SCORE EQUALLY ACROSS ALL CLASSES, REGARDLESS OF THEIR FREQUENCY.

$$\text{Macro F1} = \frac{1}{C}\sum_{i=1}^{C} \text{F1}_i$$

- **WEIGHTED F1 SCORE :**AVERAGES THE F1 SCORE WEIGHTED BY CLASS SIZE, GIVING MORE IMPORTANCE TO FREQUENT CLASSES.

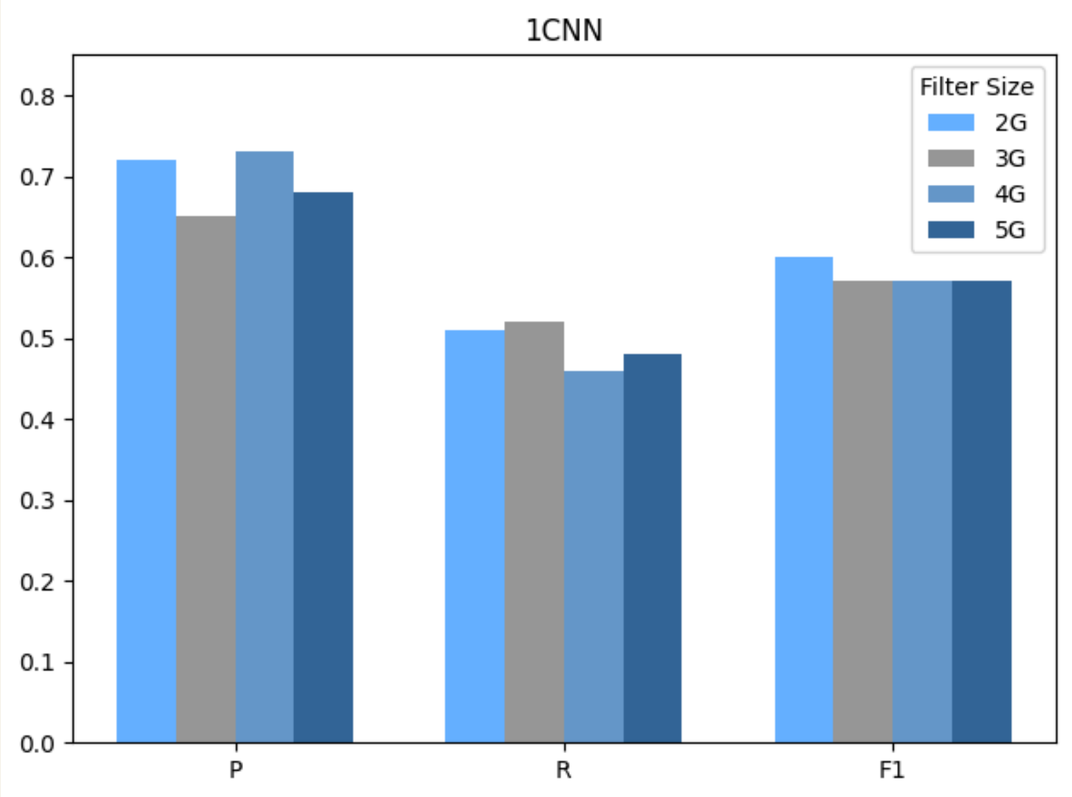
$$\text{Weighted F1} = \frac{1}{N}\sum_{i=1}^{C} \text{F1}_i \times N_i$$

EXPERIMENTAL RESULTS – 1

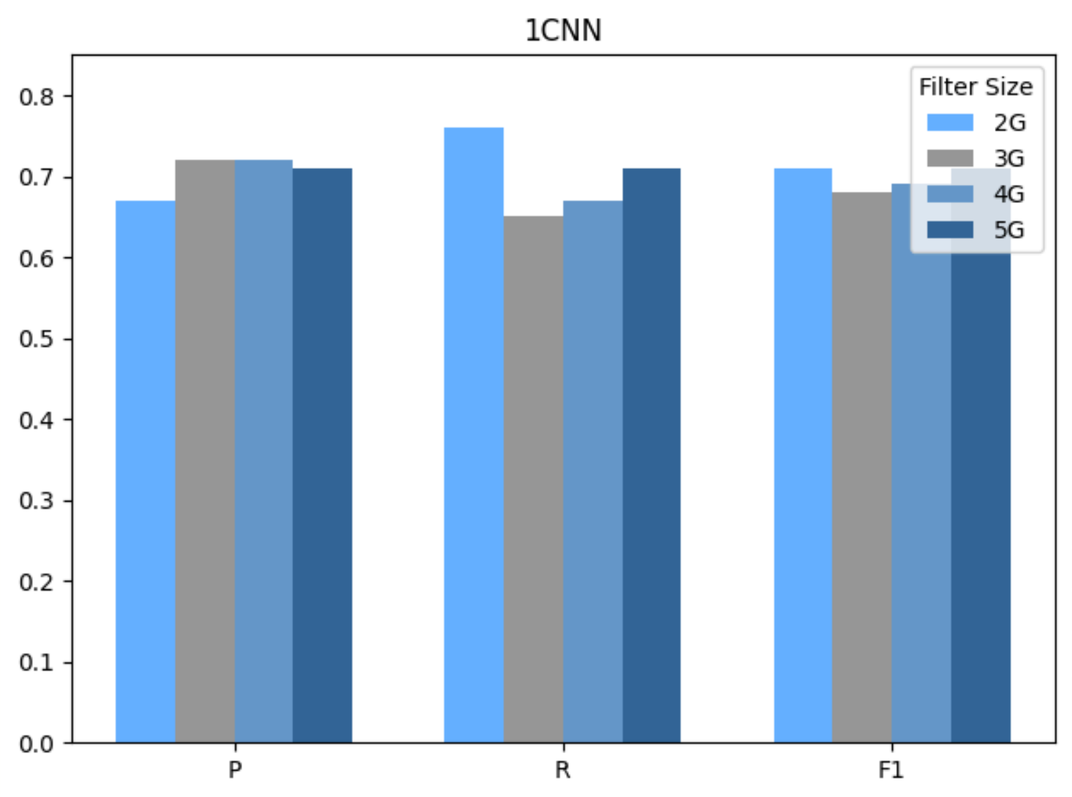
RESULTS OF 1-CNN MODEL WITH (2-5)G FILTERS

1-CNN Model	Reported Results from research paper[1]			Reproduced Results		
Filter Size	**P**	**R**	**F1**	**P**	**R**	**F1**
2G	0.72	0.51	0.60	0.67	**0.76**	0.71
3G	0.65	**0.52**	0.57	0.72	0.65	0.68
4G	0.73	0.46	0.57	0.72	0.67	0.69
5G	0.68	0.48	0.57	0.71	0.71	0.71

RESULTS OF 1-CNN MODEL WITH (2-5)G FILTERS

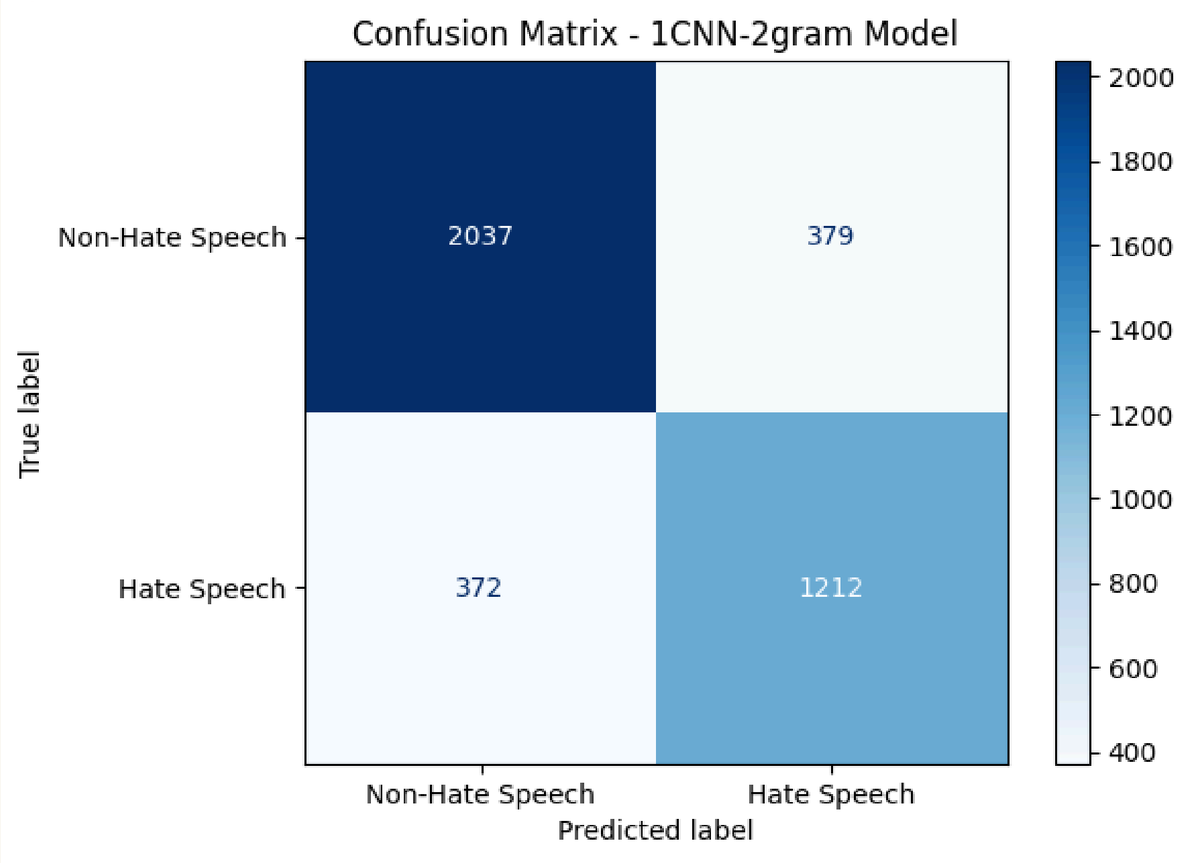


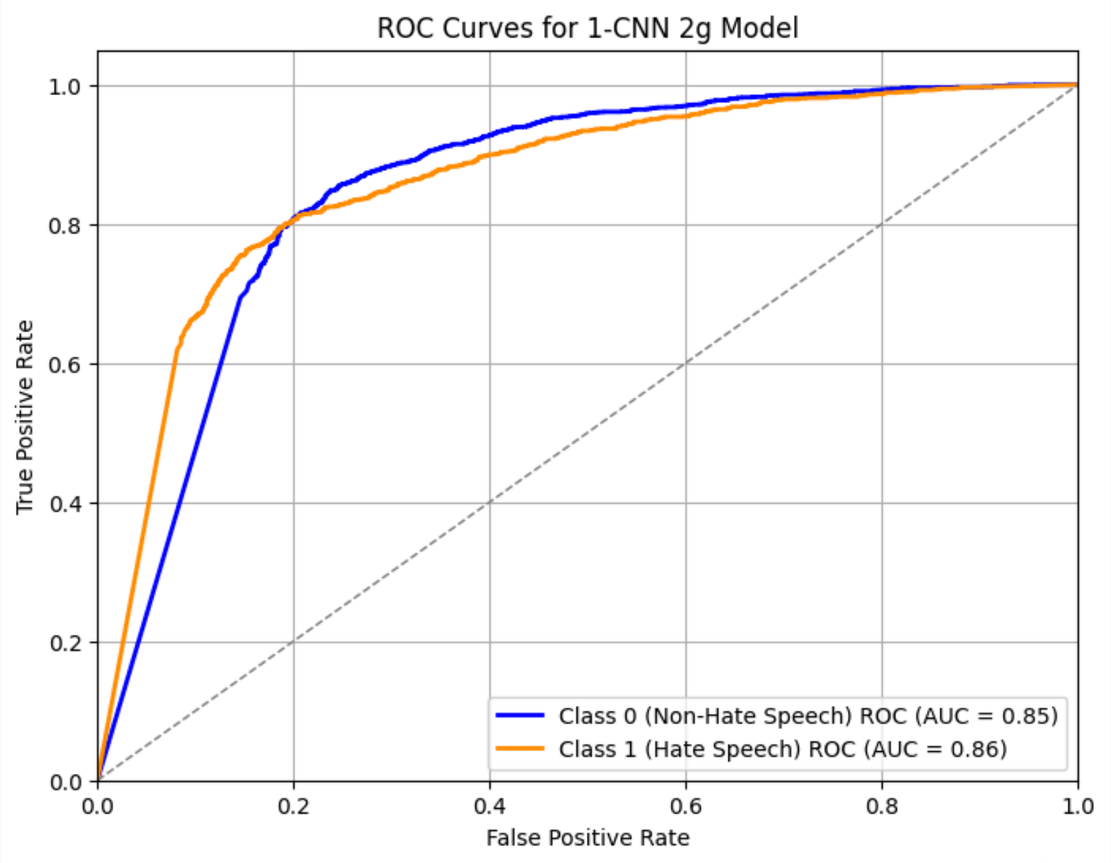
BEFORE RESAMPLING



AFTER RESAMPLING

RESULTS OF 1-CNN MODEL (2G)

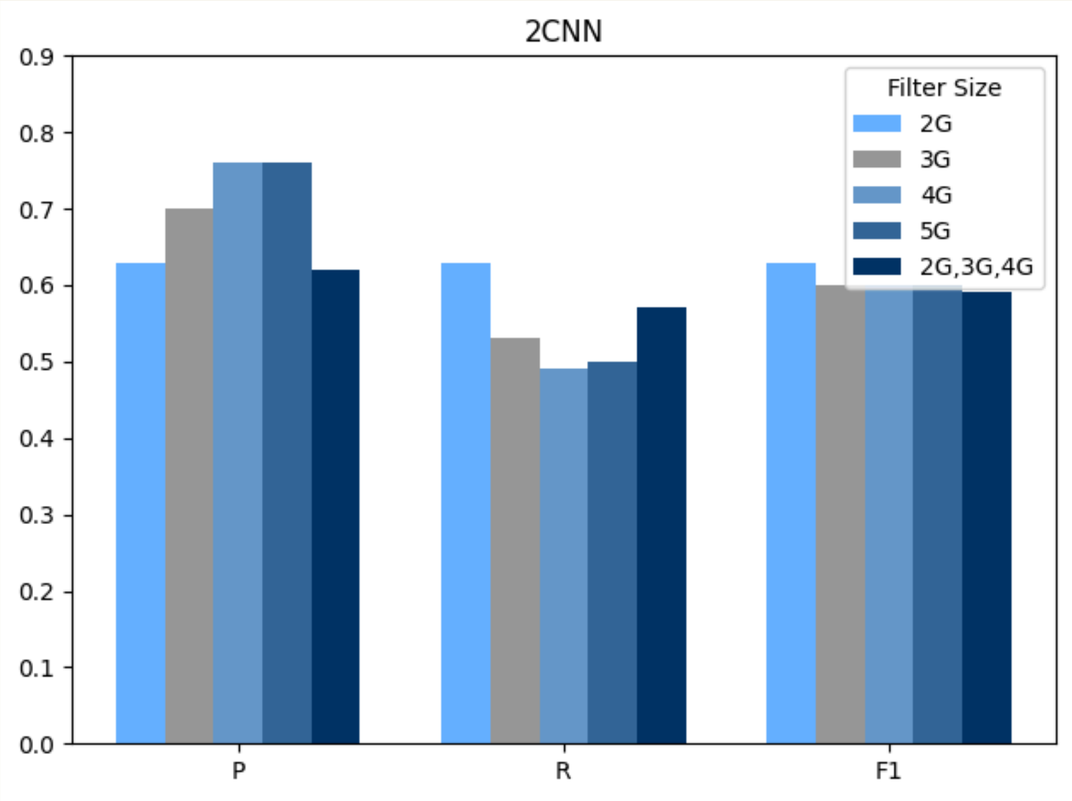




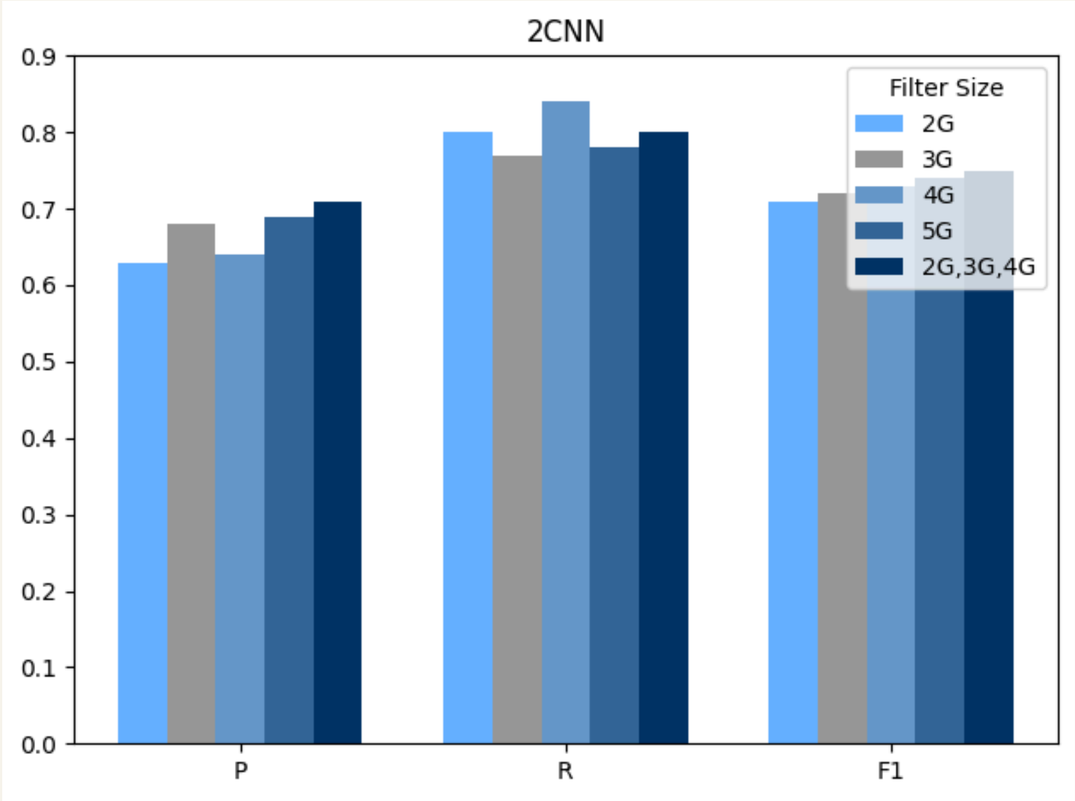
RESULT OF 2-CNN MODEL WITH (2-5)G FILTERS

2-CNN Model	Reported Results from research paper[1]			Reproduced Results		
Filter Size	**P**	**R**	**F1**	**P**	**R**	**F1**
2G	0.63	**0.63**	0.63	0.63	0.80	0.71
3G	0.70	0.53	0.60	0.68	0.77	0.72
4G	0.76	0.49	0.60	0.64	**0.84**	0.73
5G	0.76	0.50	0.60	0.69	0.78	0.74
2G,3G,4G	0.62	0.57	0.59	0.71	0.80	0.75

RESULT OF 2-CNN MODEL WITH (2-5)G FILTERS

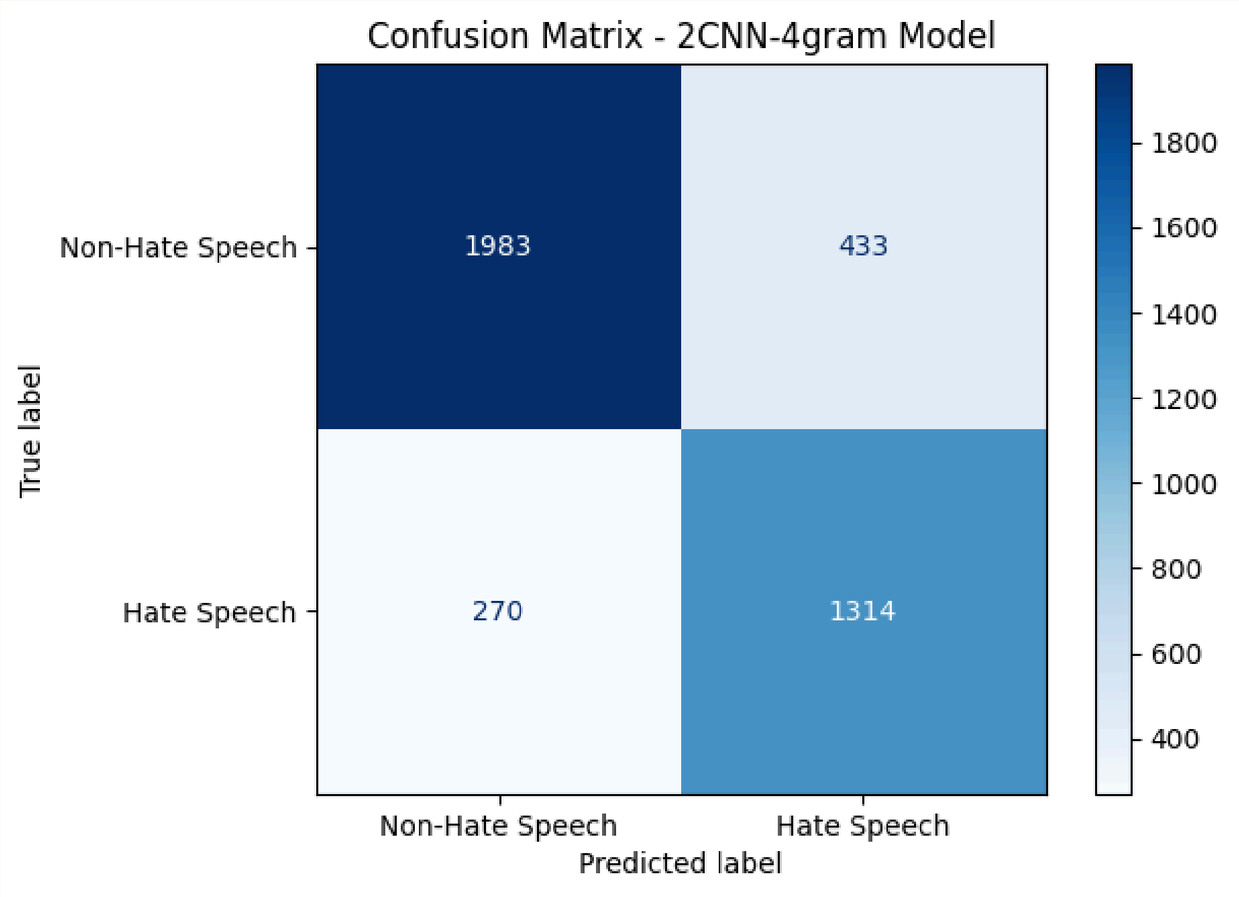


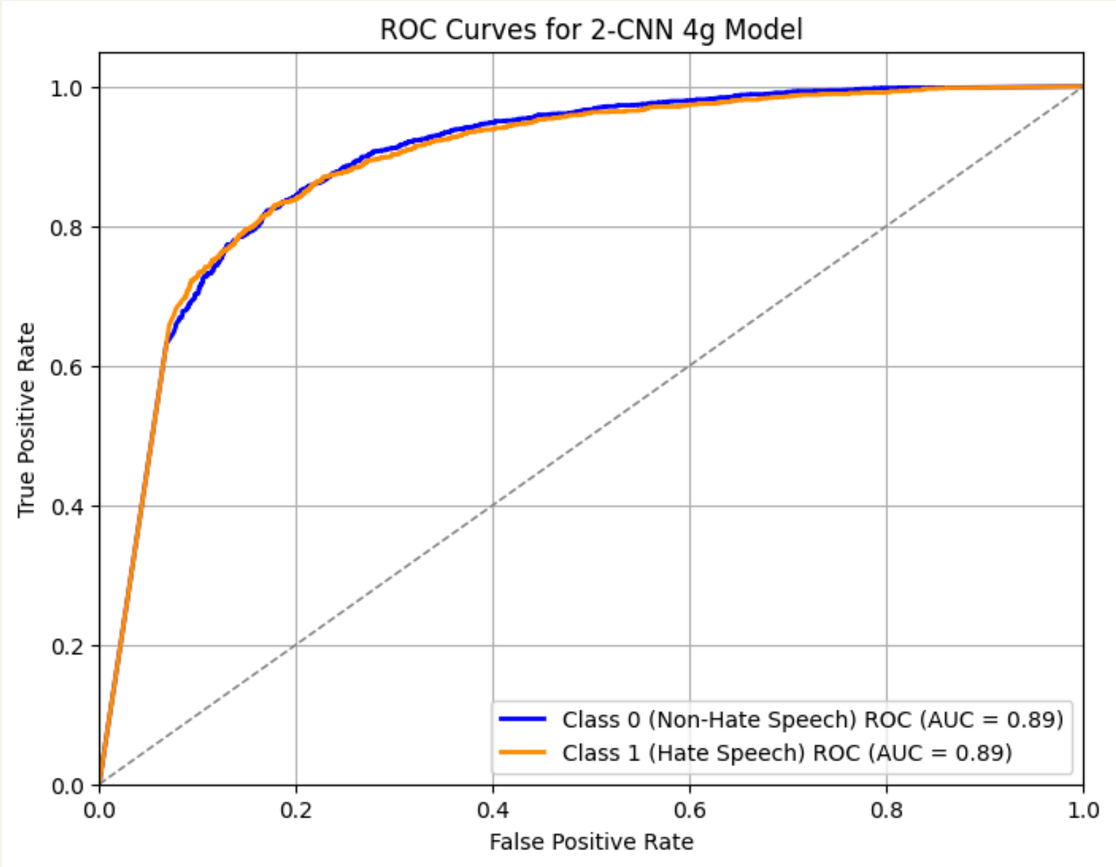
BEFORE RESAMPLING



AFTER RESAMPLING

RESULT OF 2-CNN MODEL WITH 4G FILTER

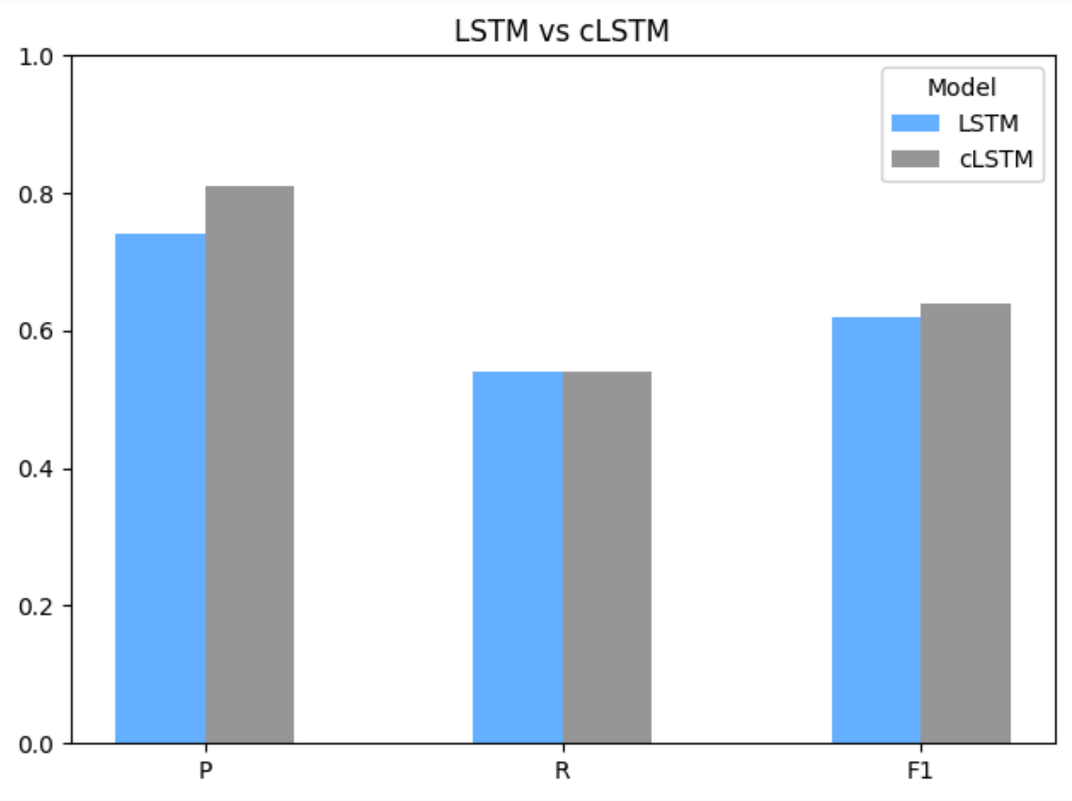




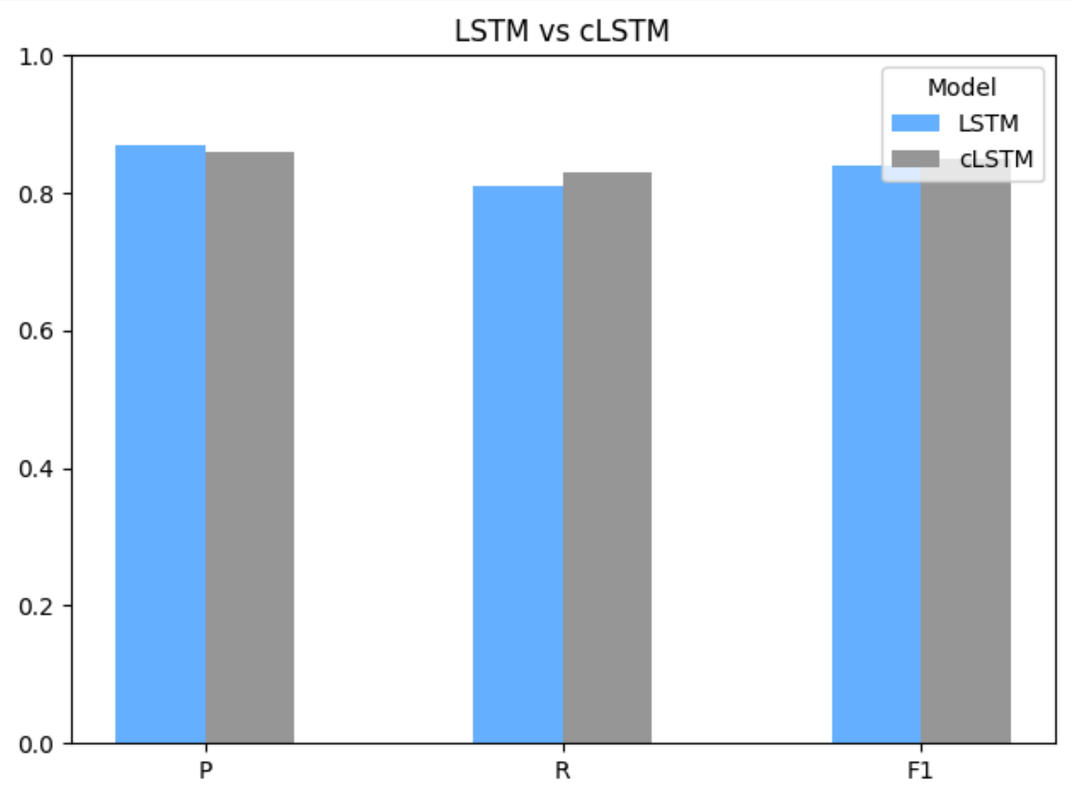
RESULTS OF LSTM AND C-LSTM MODEL

Model	Reported Results from research paper[1]			Reproduced Results		
	P	**R**	**F1**	**P**	**R**	**F1**
LSTM	0.74	**0.54**	0.62	0.87	0.81	0.84
C-LSTM	0.81	0.54	0.64	0.86	**0.83**	0.85

RESULTS OF LSTM AND C-LSTM MODEL

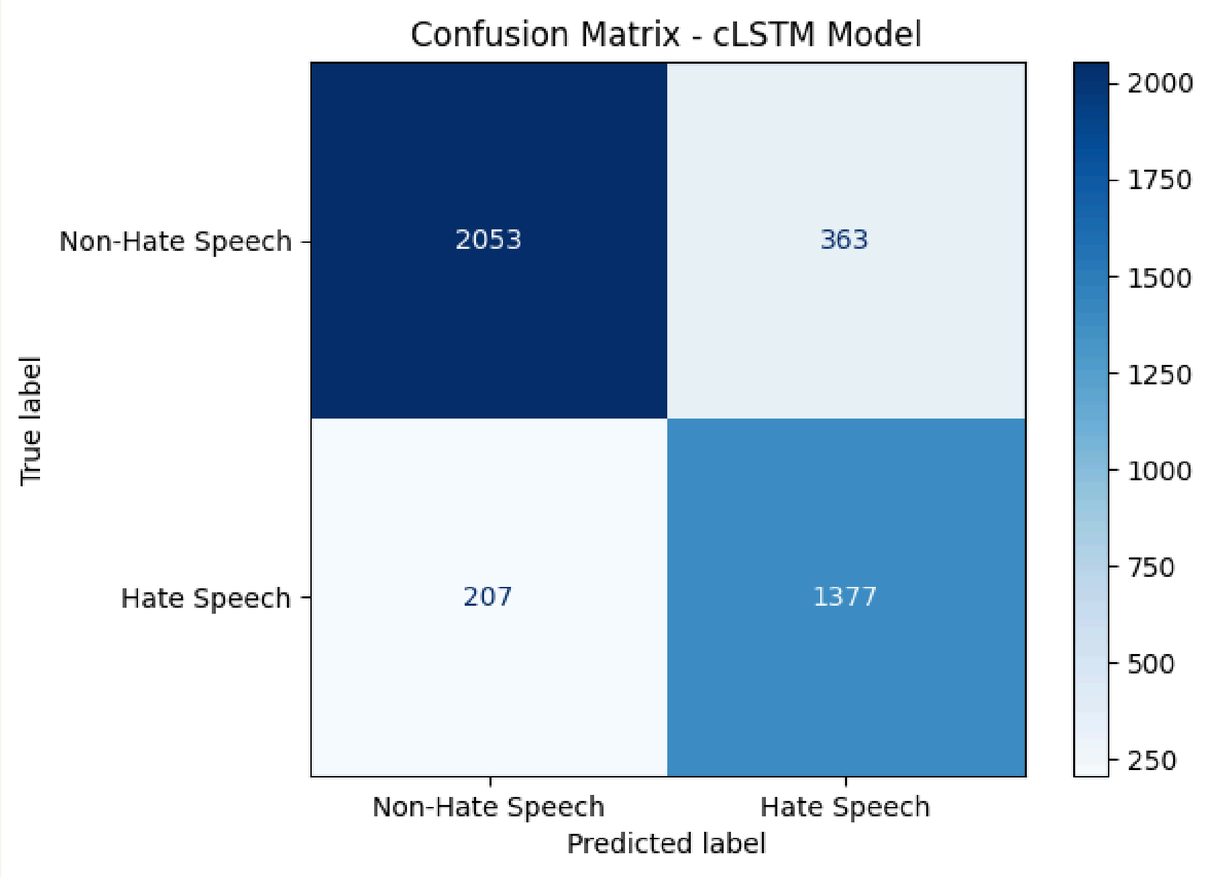


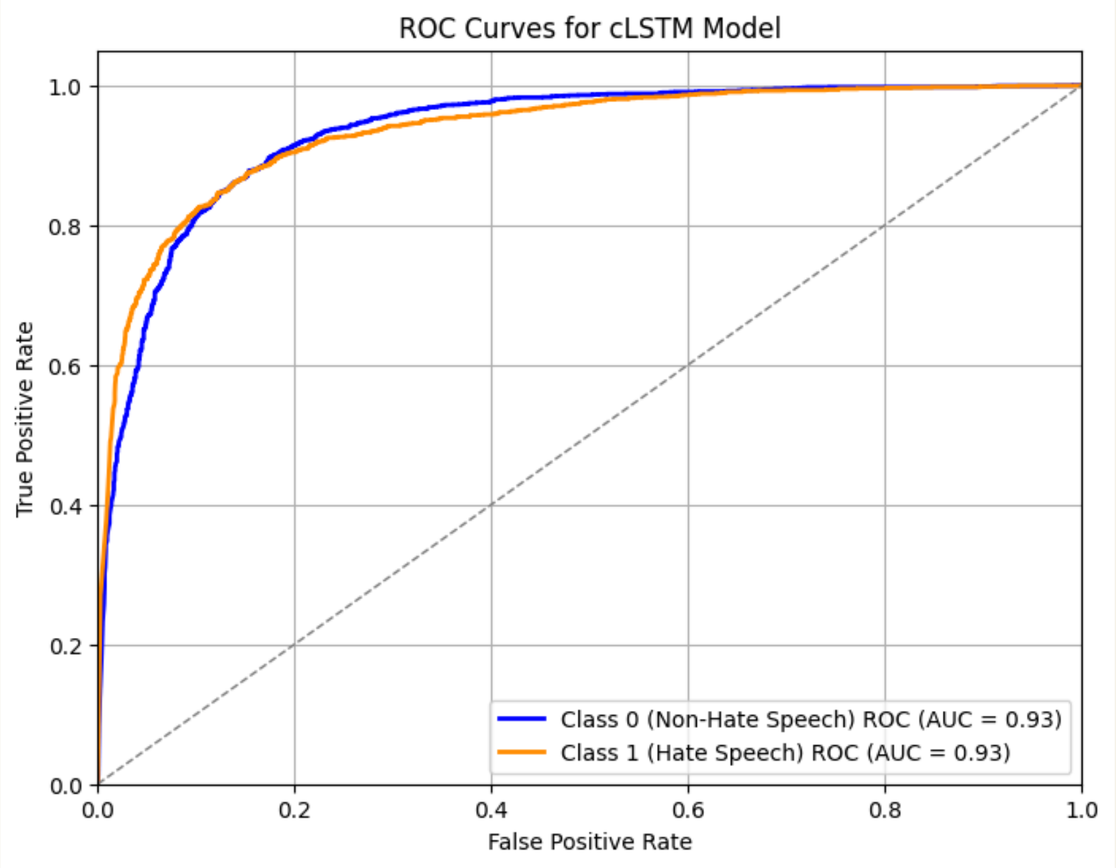
BEFORE RESAMPLING



AFTER RESAMPLING

RESULTS OF C-LSTM MODEL





EFFECTIVENESS OF DIFFERENT MODELS

From the experimental results, a clear progression in performance is observed across the different model groups: from shallow CNNs to deeper CNNs and eventually to LSTM-based architectures. This trend demonstrates how increasing the model's capacity to capture richer features leads to improved hate speech detection.

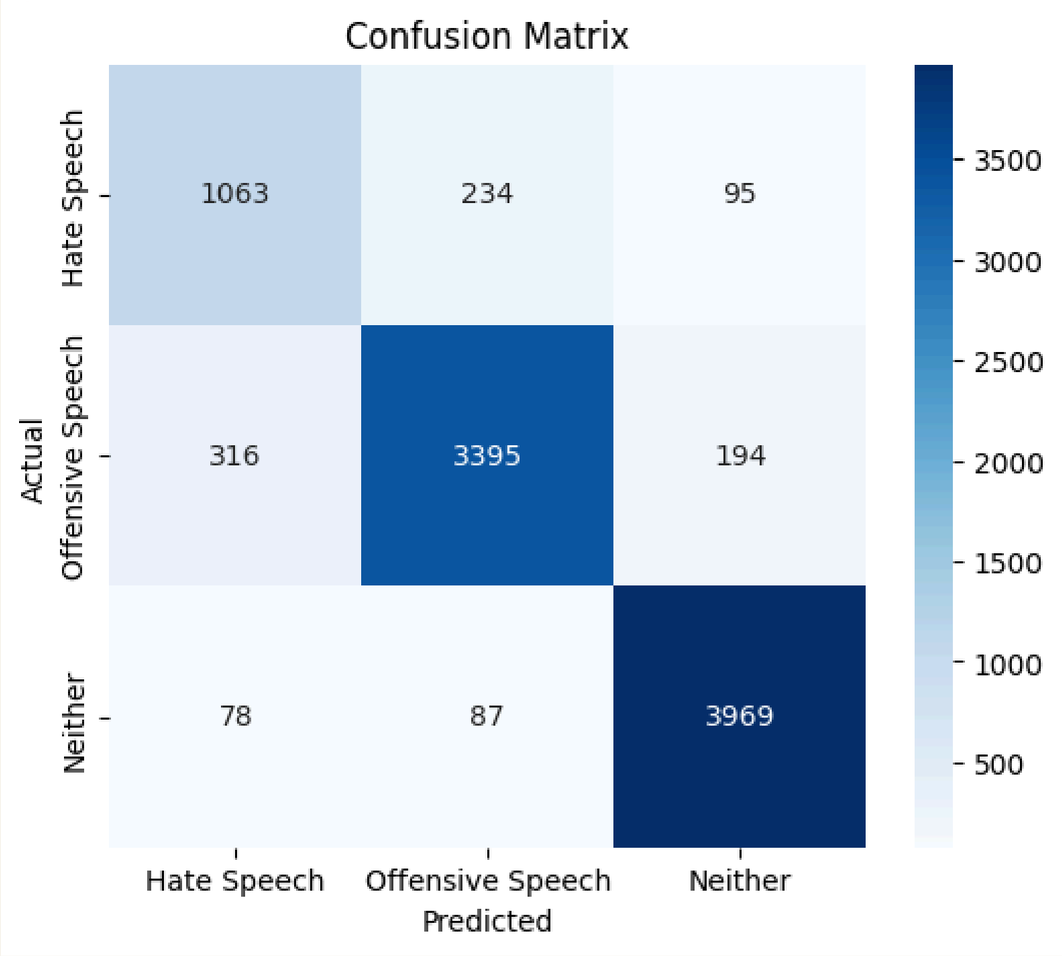
- Starting with 1-CNN, the use of a single 2-gram convolution filter yields an F1-score of **0.71** and ROC-AUC values of **0.85** and **0.86**. This serves as a solid baseline, effectively capturing short-range word dependencies.
- Extending to 2-CNN, which utilizes multiple filter sizes (2G, 3G, 4G), leads to a noticeable improvement. The best configuration achieves an F1-score of 0.73 and ROC-AUC scores of 0.89 for both classes. This improvement mirrors the shift from shallow to deeper convolutional feature extraction, enhancing the model's ability to capture diverse n-gram patterns.
- A similar leap in performance is observed when moving from LSTM to cLSTM. The basic LSTM achieves a strong F1-score of 0.84, while the cLSTM pushes this further to 0.85, with both ROC-AUC scores reaching 0.93. This progression is analogous to the CNN group, where adding complexity (via convolutional preprocessing in cLSTM) refines the model's sequence understanding and improves classification accuracy.

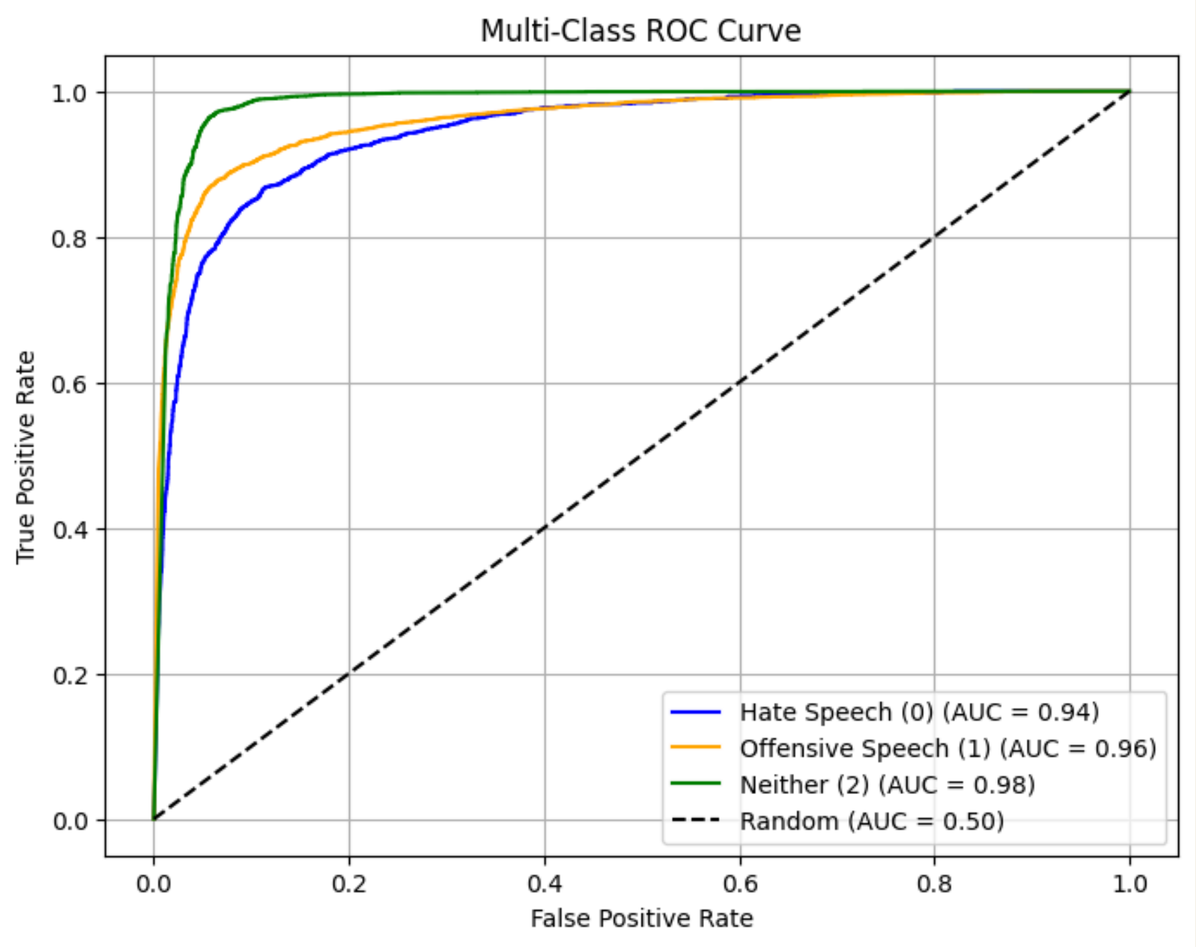
EXPERIMENTAL RESULTS – II

USING GLOVE EMBEDDING

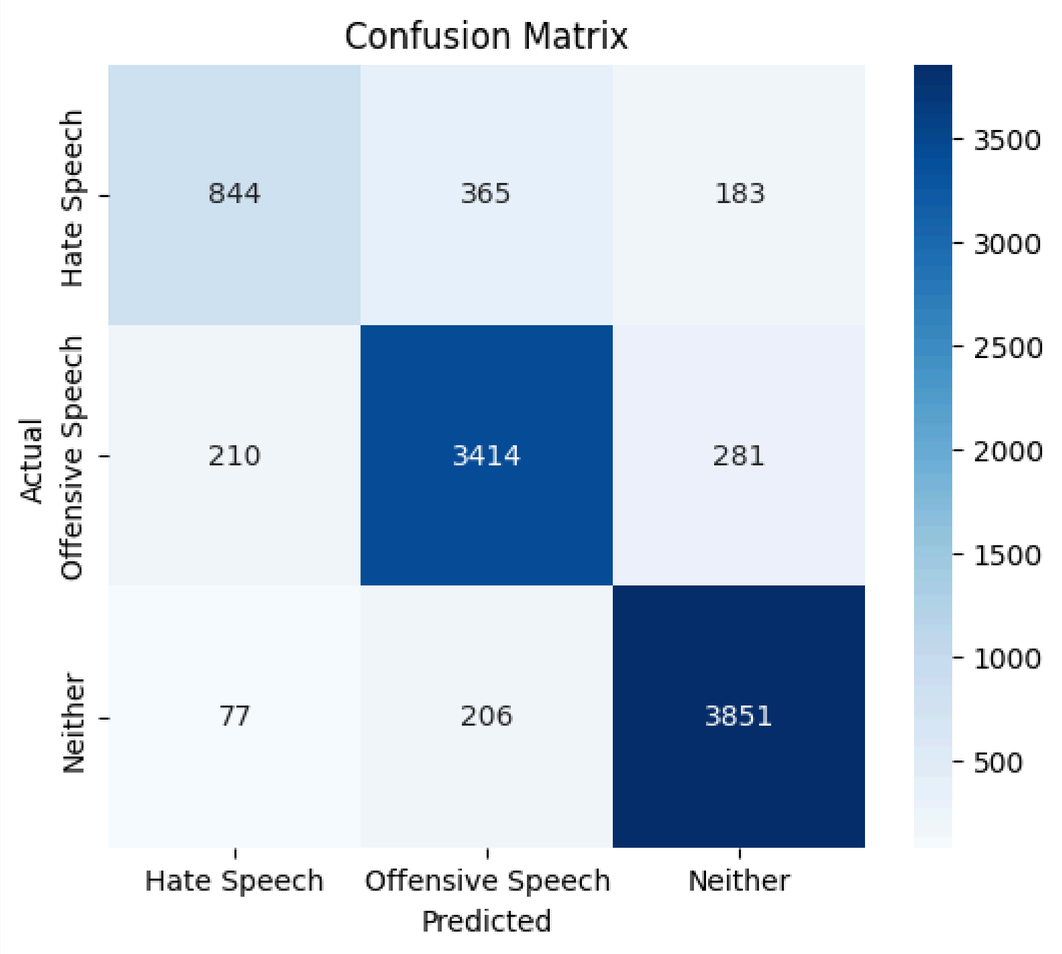
		Results from research paper[2]		Reproduced Results	
Model	**Metric**	**Macro**	**Weighted Avg.**	**Macro**	**Weighted Avg.**
Glove + CNN	**P**	0.66	0.88	0.86	0.89
	R	0.73	0.85	0.86	0.89
	F1	0.69	0.86	0.86	0.89
Glove + MLP	**P**	0.72	0.85	0.83	0.86
	R	0.61	**0.86**	0.80	0.86
	F1	0.65	0.85	0.82	0.86
Glove + Bi-LSTM	**P**	0.67	0.88	0.88	0.92
	R	**0.75**	0.84	**0.91**	**0.91**
	F1	0.69	0.86	0.89	**0.91**

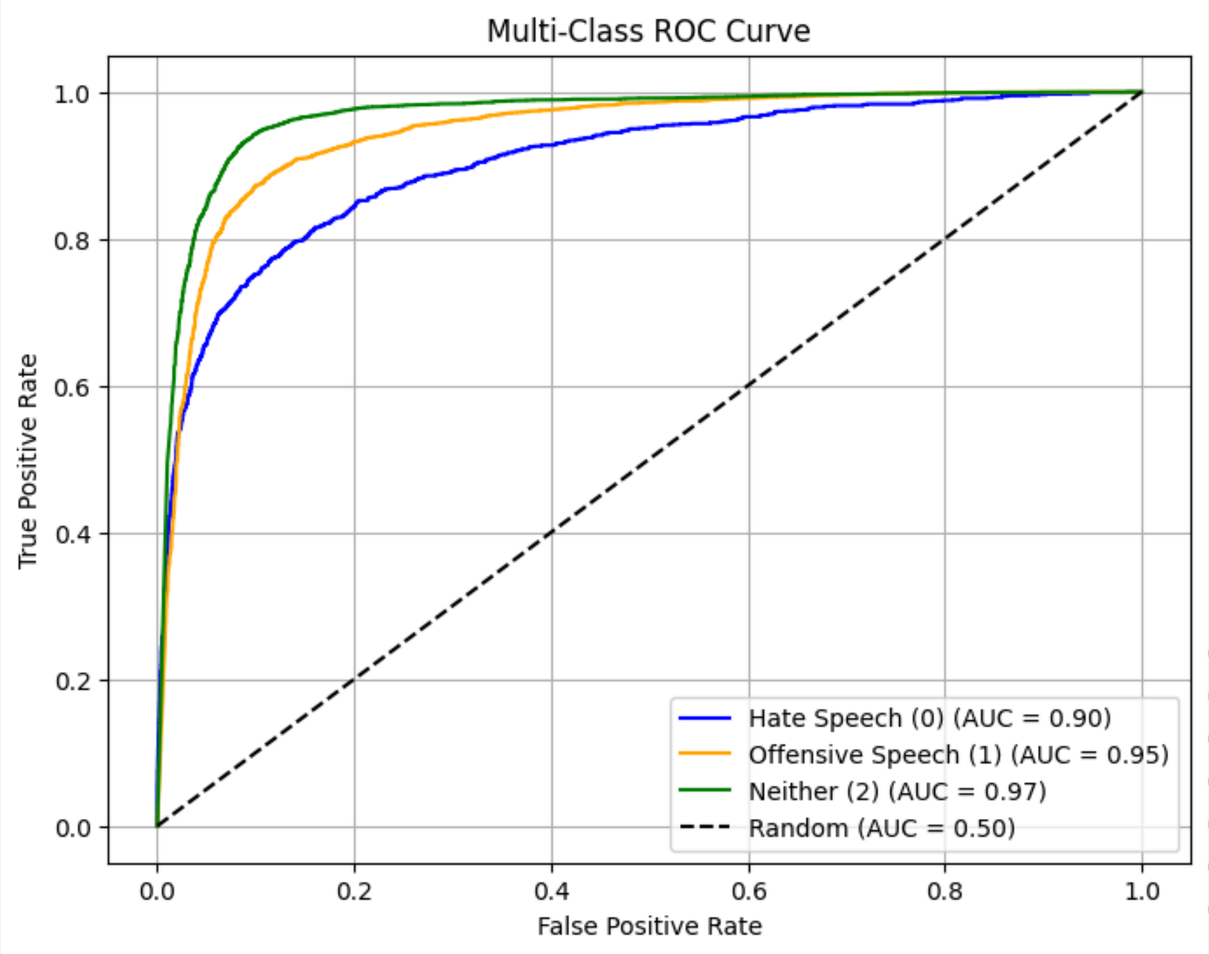
GLOVE + CNN


Confusion Matrix
Actual
Hate Speech
Offensive Speech
Neither
1063
234
95
316
3395
194
78
87
3969
Hate Speech
Offensive Speech
Neither
Predicted
3500
3000
2500
2000
1500
1000
500

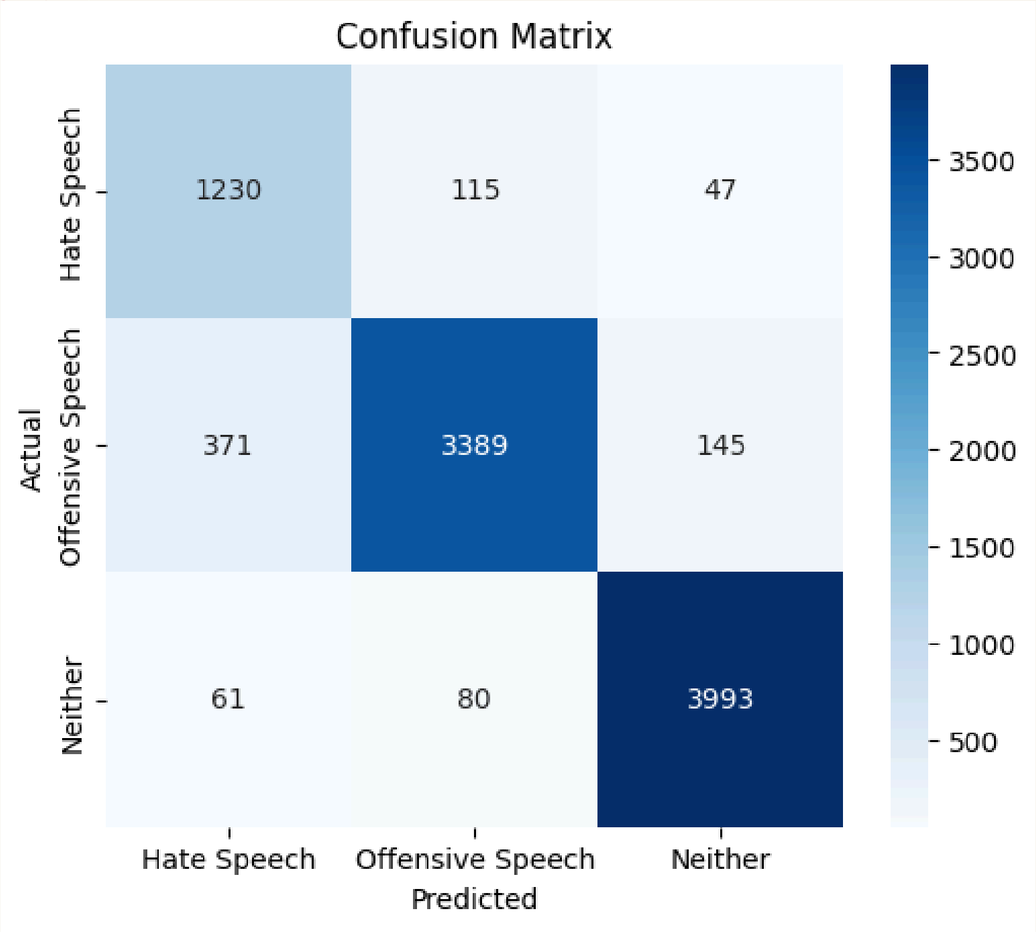

Multi-Class ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Hate Speech (0) (AUC = 0.94)
Offensive Speech (1) (AUC = 0.96)
Neither (2) (AUC = 0.98)
Random (AUC = 0.50)

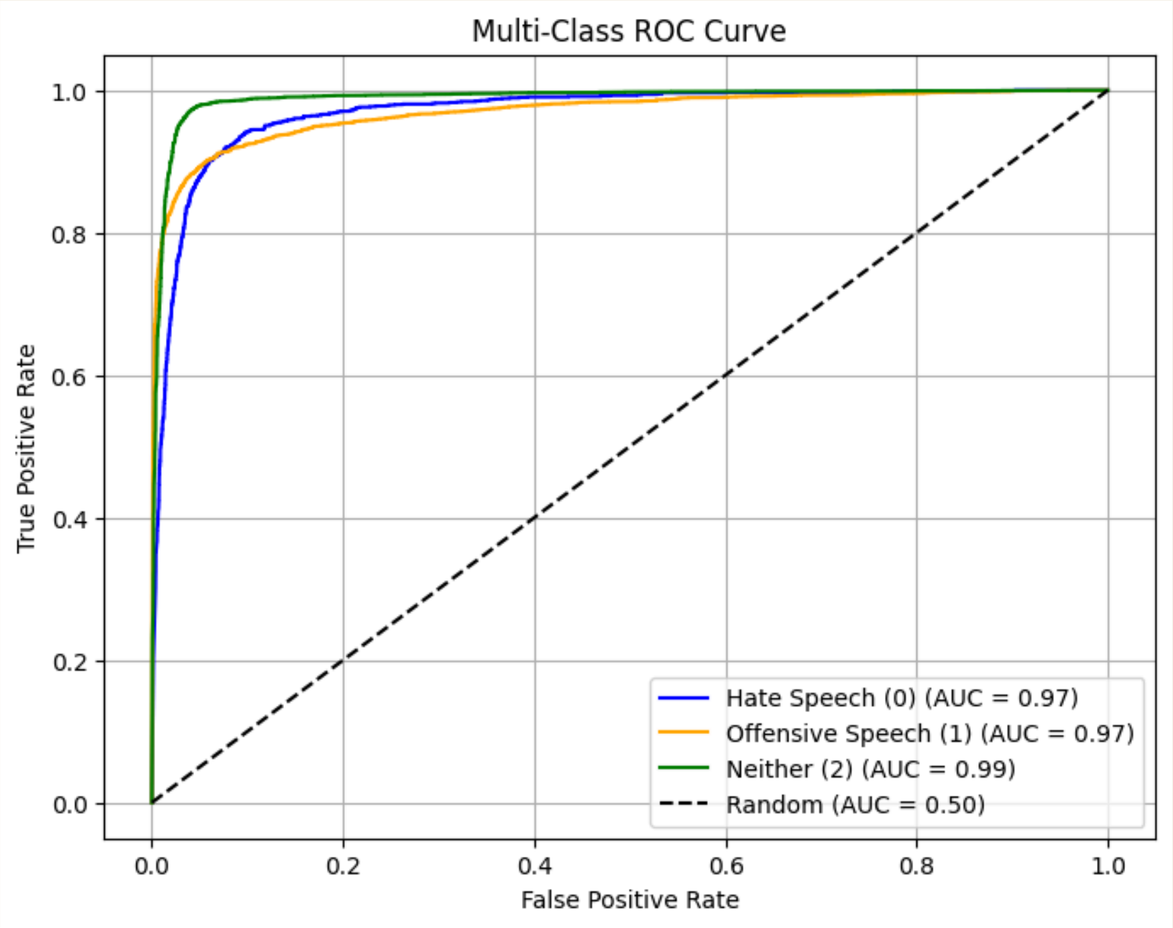
GLOVE + MLP


Confusion Matrix
Actual
Hate Speech
Offensive Speech
Neither
844
365
183
210
3414
281
77
206
3851
Hate Speech
Offensive Speech
Neither
Predicted
3500
3000
2500
2000
1500
1000
500


Multi-Class ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Hate Speech (0) (AUC = 0.90)
Offensive Speech (1) (AUC = 0.95)
Neither (2) (AUC = 0.97)
Random (AUC = 0.50)

GLOVE + BI-LSTM

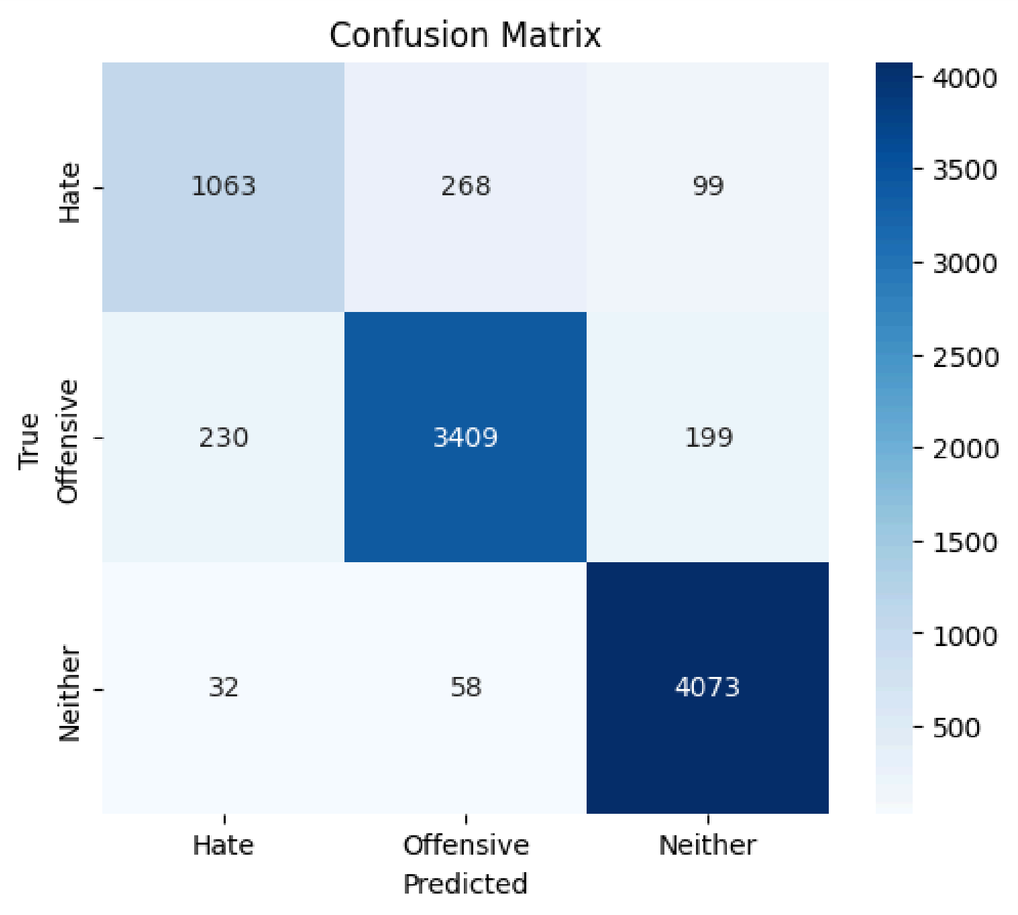


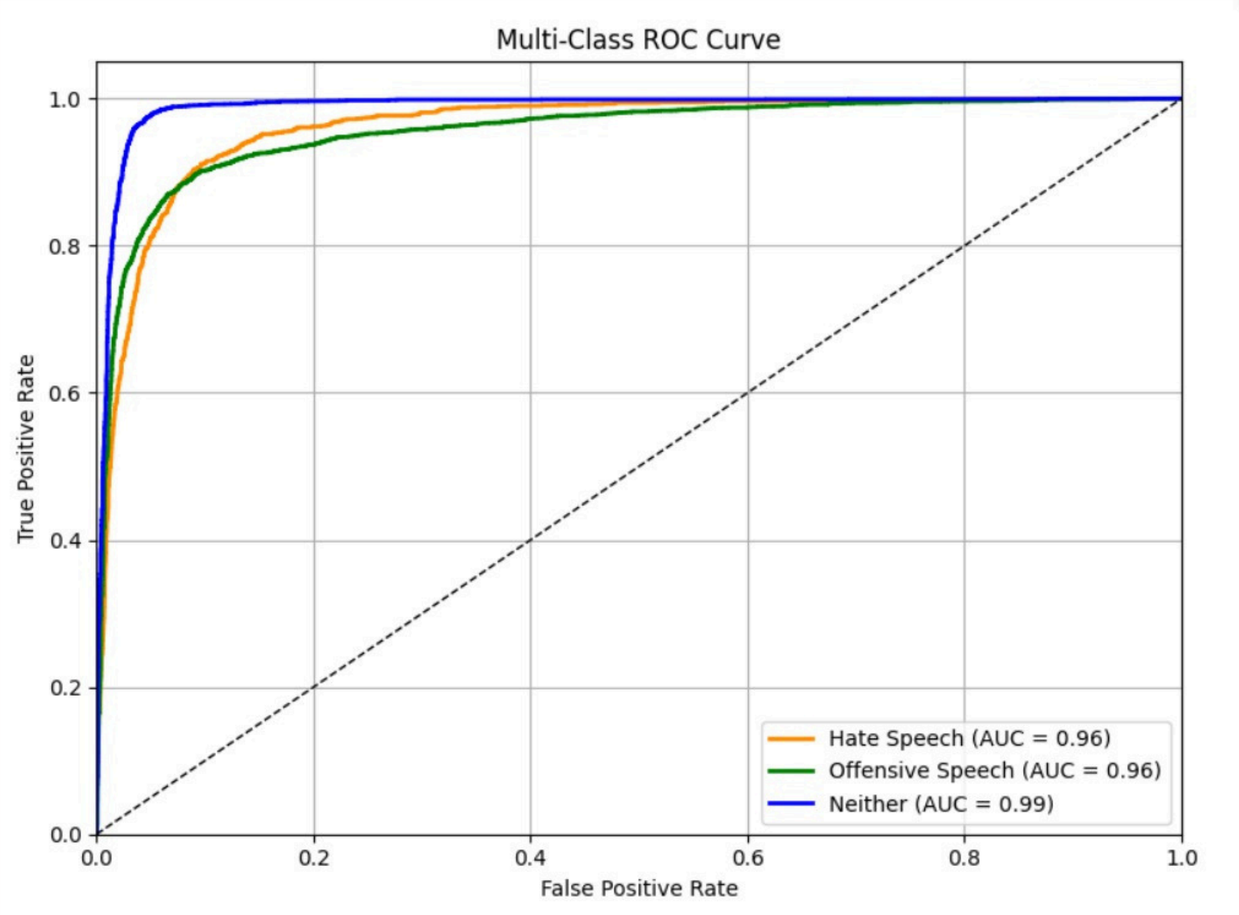


USING TF-IDF

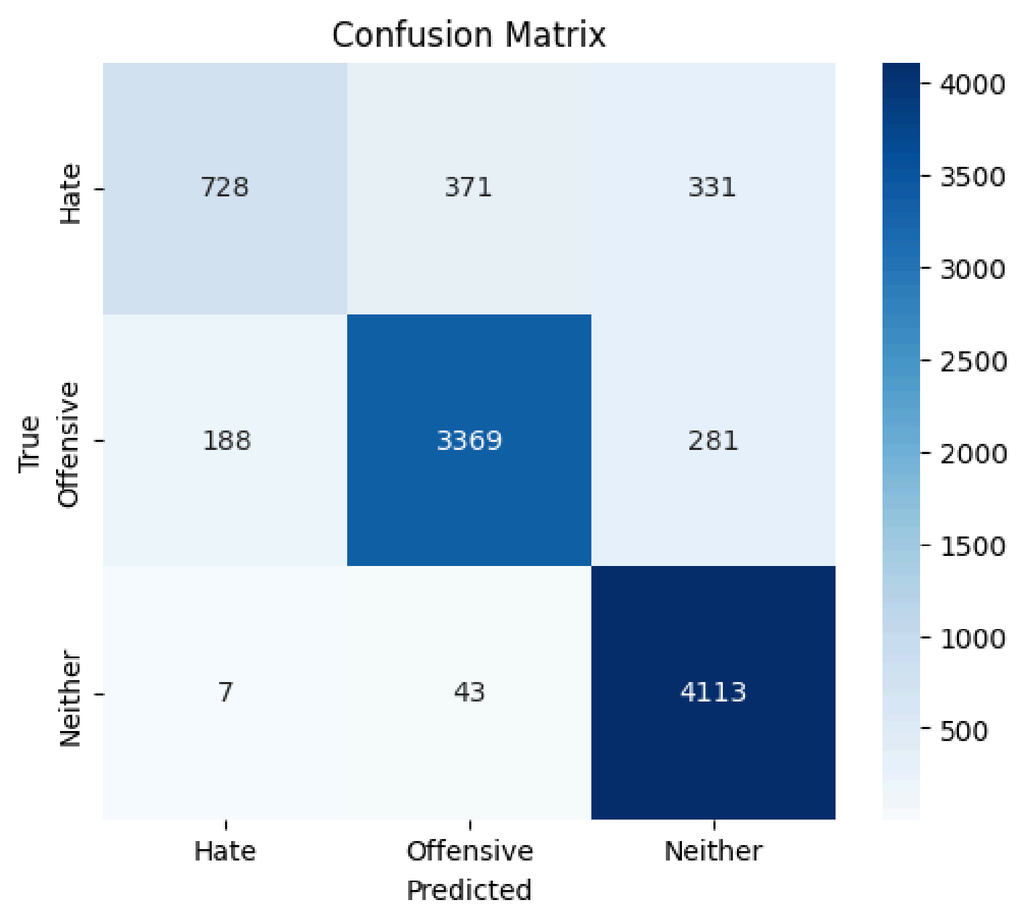
		Results from research paper[2]		Reproduced Results	
Model	**Metric**	**Macro**	**Weighted Avg.**	**Macro**	**Weighted Avg.**
Tf-Idf + SVM	**P**	0.69	0.86	0.88	0.90
	R	0.64	0.87	0.87	0.91
	F1	0.66	0.86	0.88	0.90
Tf-Idf + XGB	**P**	0.74	0.89	0.85	0.87
	R	**0.69**	**0.90**	0.79	0.87
	F1	0.70	0.90	0.81	0.86
Tf-Idf + MLP	**P**	0.68	0.85	0.93	0.95
	R	0.63	0.86	**0.95**	**0.95**
	F1	0.66	0.85	0.94	**0.95**

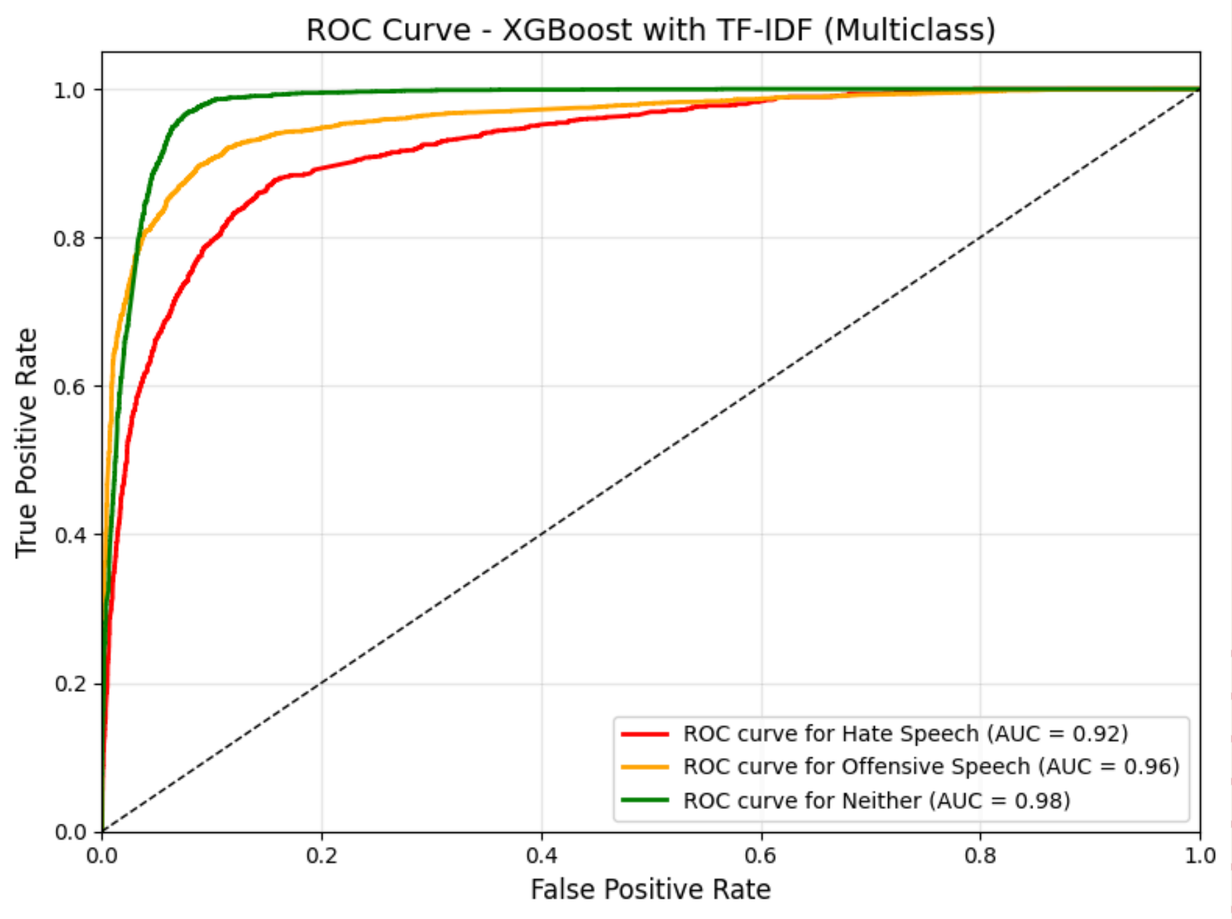
TF-IDF + SVM


Confusion Matrix
True
Hate
Offensive
Neither
1063
268
99
230
3409
199
32
58
4073
Hate
Offensive
Neither
Predicted
4000
3500
3000
2500
2000
1500
1000
500

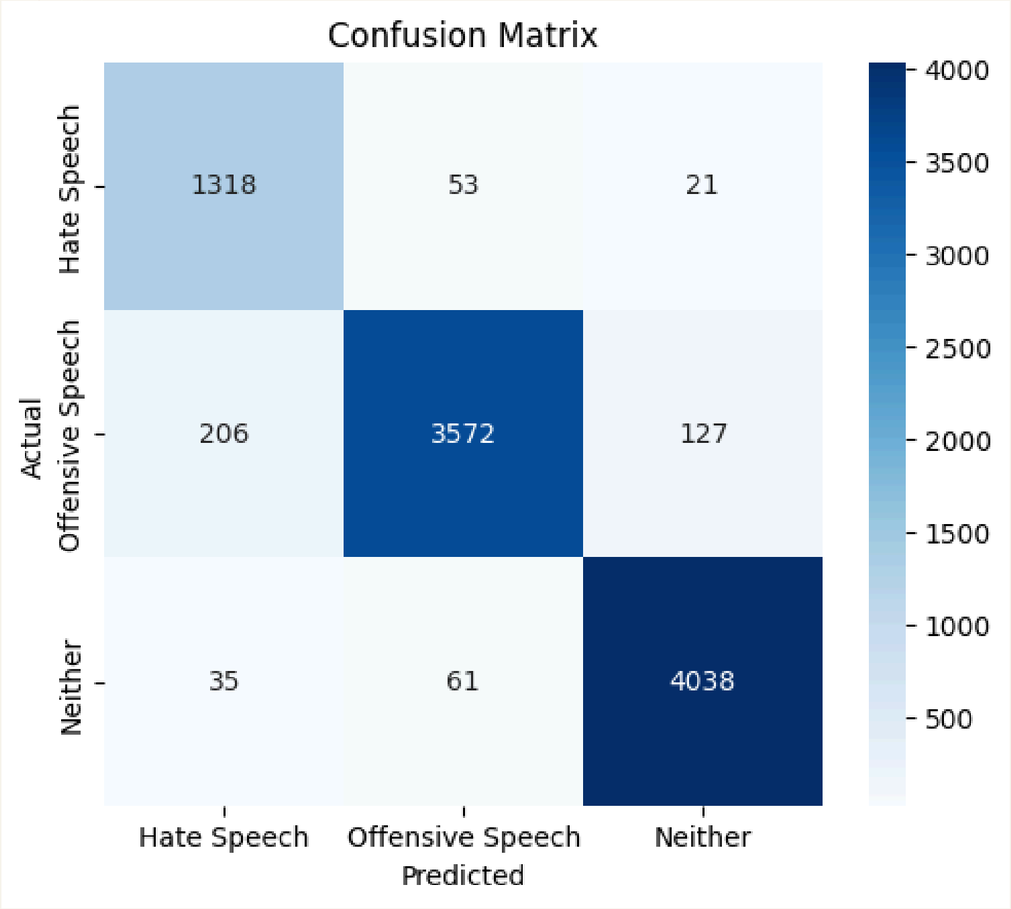

Multi-Class ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Hate Speech (AUC = 0.96)
Offensive Speech (AUC = 0.96)
Neither (AUC = 0.99)

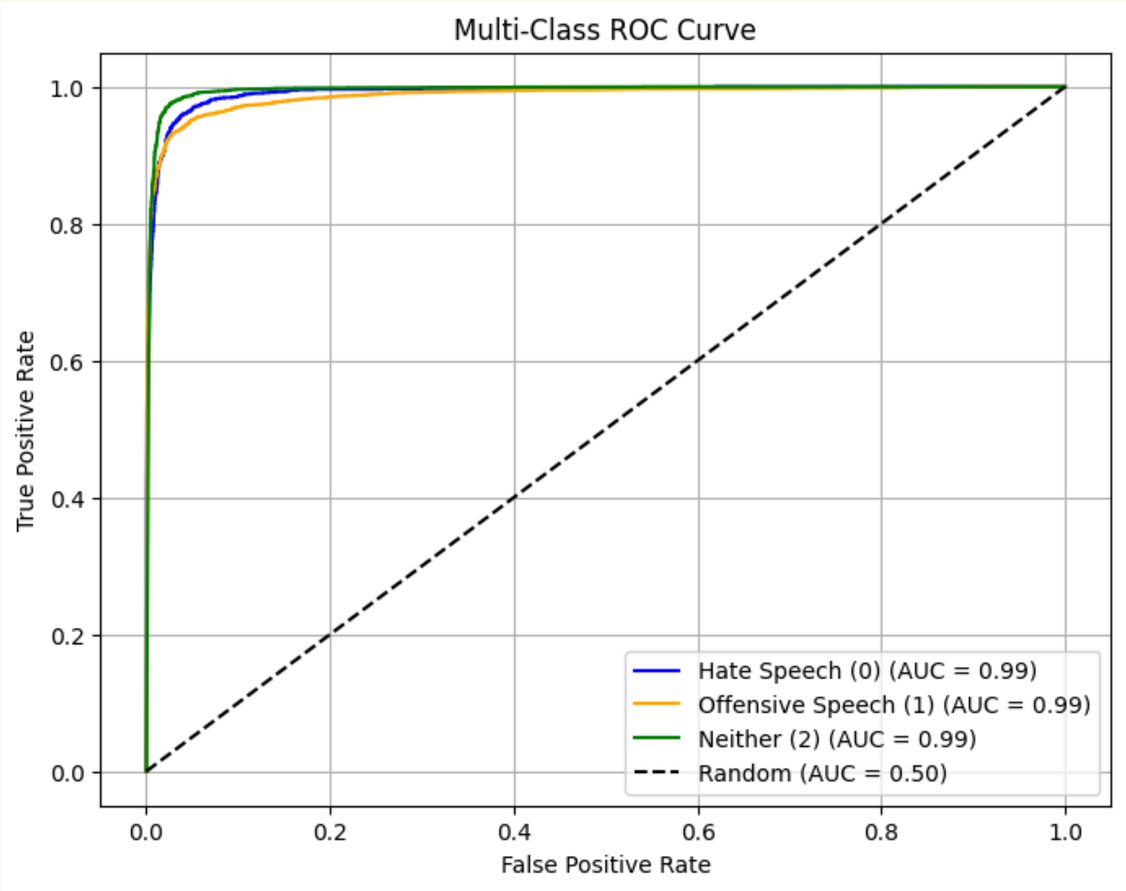
TF-IDF + XGB


Confusion Matrix
Hate
Offensive
Neither
True
728
371
331
188
3369
281
7
43
4113
Hate
Offensive
Neither
Predicted
4000
3500
3000
2500
2000
1500
1000
500


ROC Curve - XGBoost with TF-IDF (Multiclass)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve for Hate Speech (AUC = 0.92)
ROC curve for Offensive Speech (AUC = 0.96)
ROC curve for Neither (AUC = 0.98)

TF-IDF + MLP

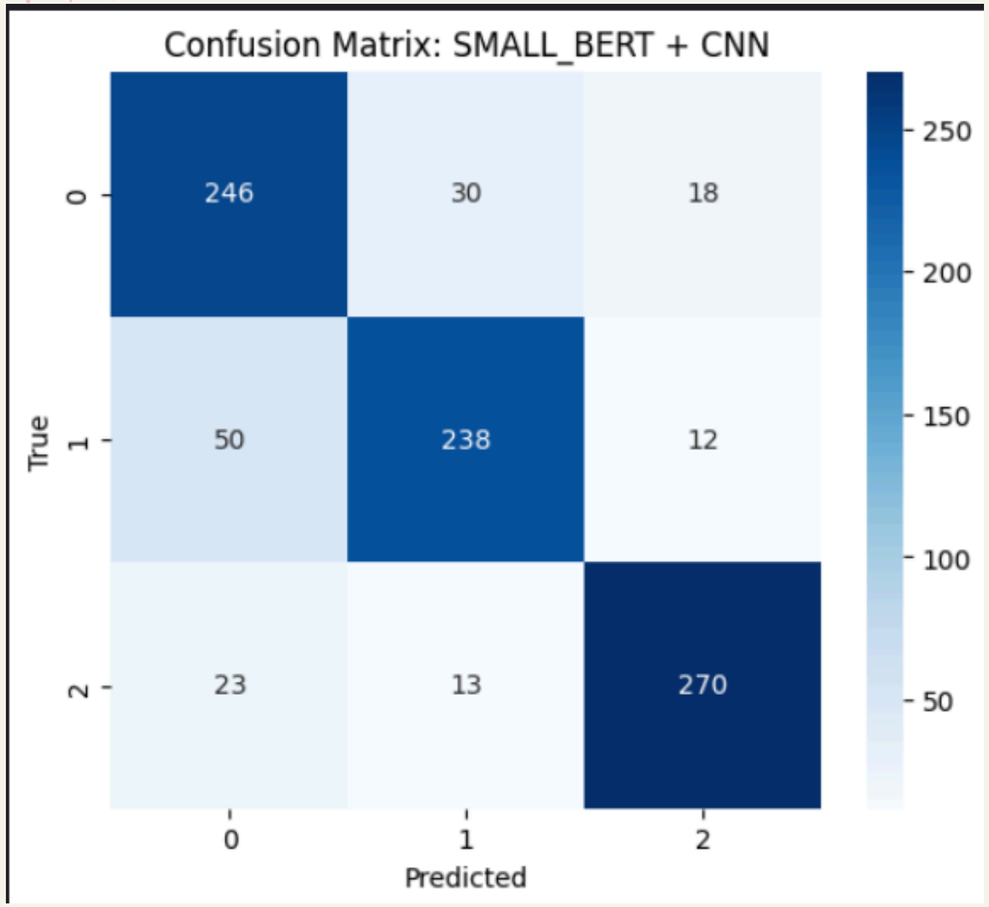

Confusion Matrix
Actual
Hate Speech
Offensive Speech
Neither
1318
53
21
206
3572
127
35
61
4038
Hate Speech
Offensive Speech
Neither
Predicted
4000
3500
3000
2500
2000
1500
1000
500

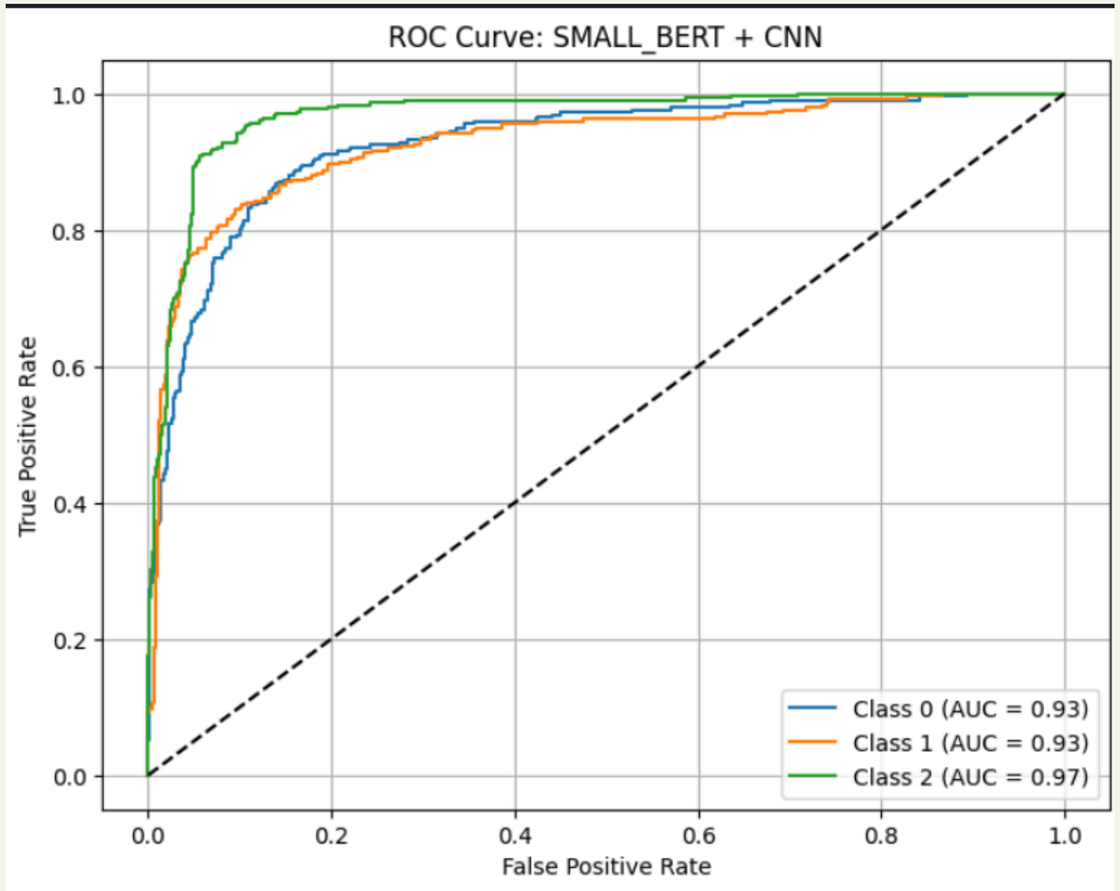

Multi-Class ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Hate Speech (0) (AUC = 0.99)
Offensive Speech (1) (AUC = 0.99)
Neither (2) (AUC = 0.99)
Random (AUC = 0.50)

USING TRANSFORMERS

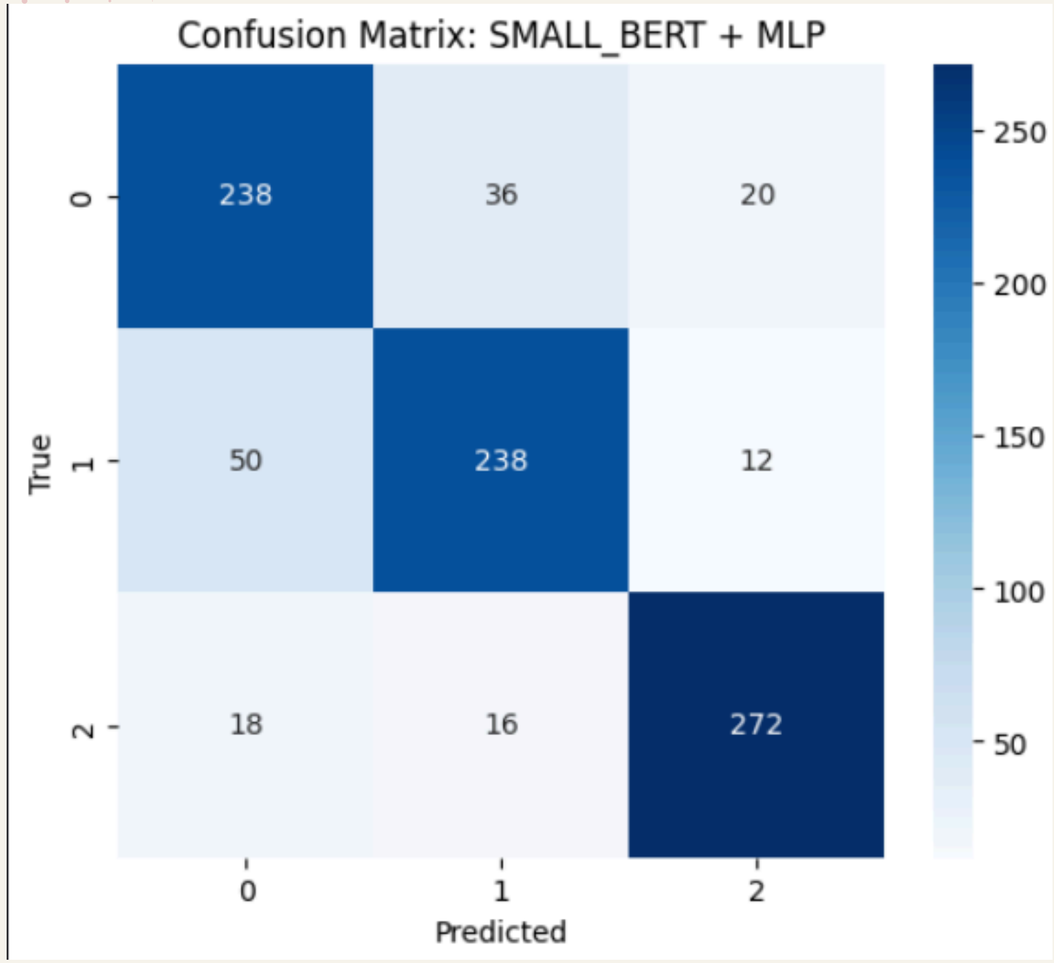
		Results from research paper[2]		Reproduced Results	
Model	**Metric**	**Macro**	**Weighted Avg.**	**Macro**	**Weighted Avg.**
Small Bert + CNN	**P**	0.72	0.91	0.84	0.84
	R	**0.82**	0.85	0.84	0.84
	F1	0.75	0.87	0.84	0.84
Small Bert + MLP	**P**	0.71	0.91	0.83	0.83
	R	0.81	0.85	0.83	0.83
	F1	0.74	0.87	0.83	0.83
Bert + CNN	**P**	0.78	0.91	0.89	0.89
	R	0.75	**0.91**	0.89	0.88
	F1	0.76	0.91	0.76	0.89
Bert + MLP	**P**	0.75	0.90	0.90	0.91
	R	0.72	0.90	**0.90**	**0.90**
	F1	0.74	0.90	0.91	0.90

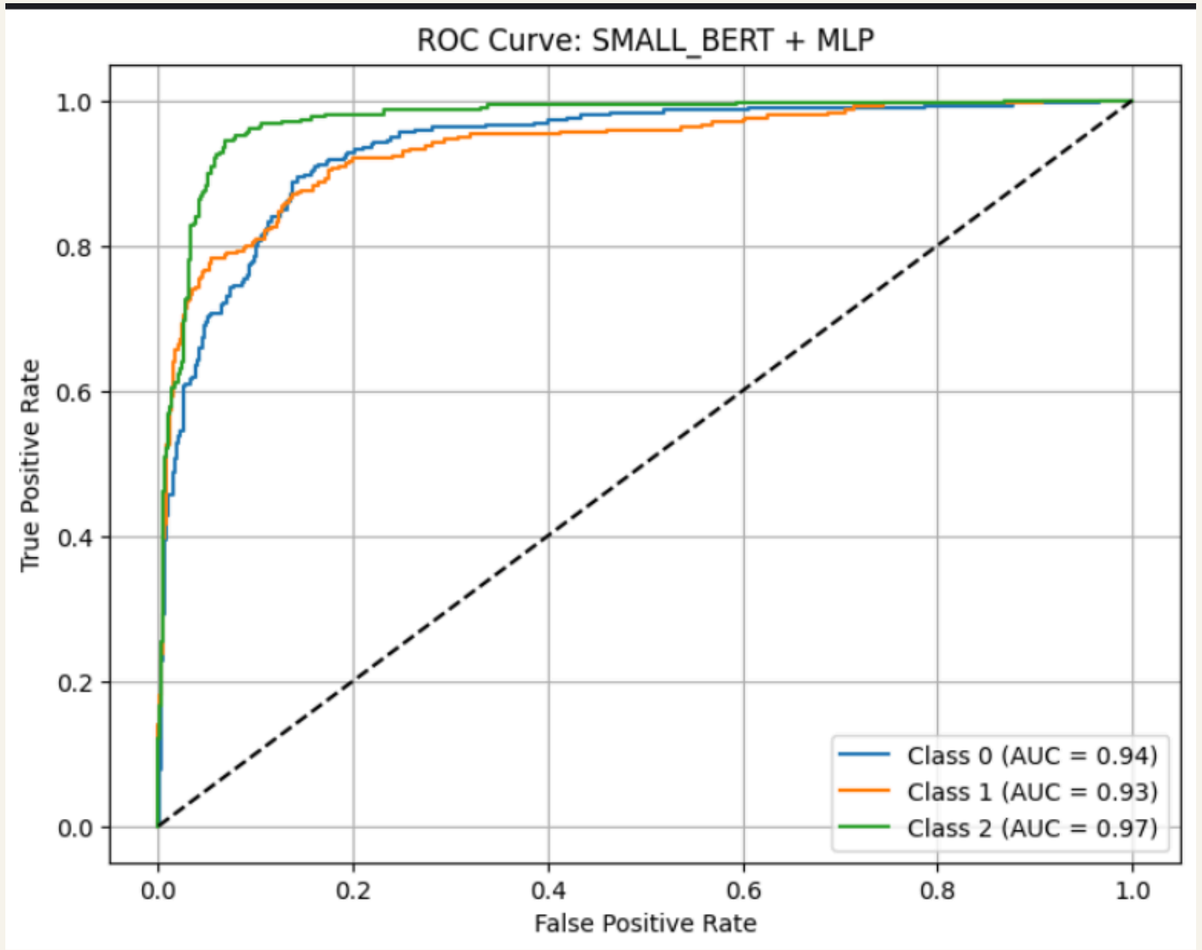
SMALL BERT + CNN



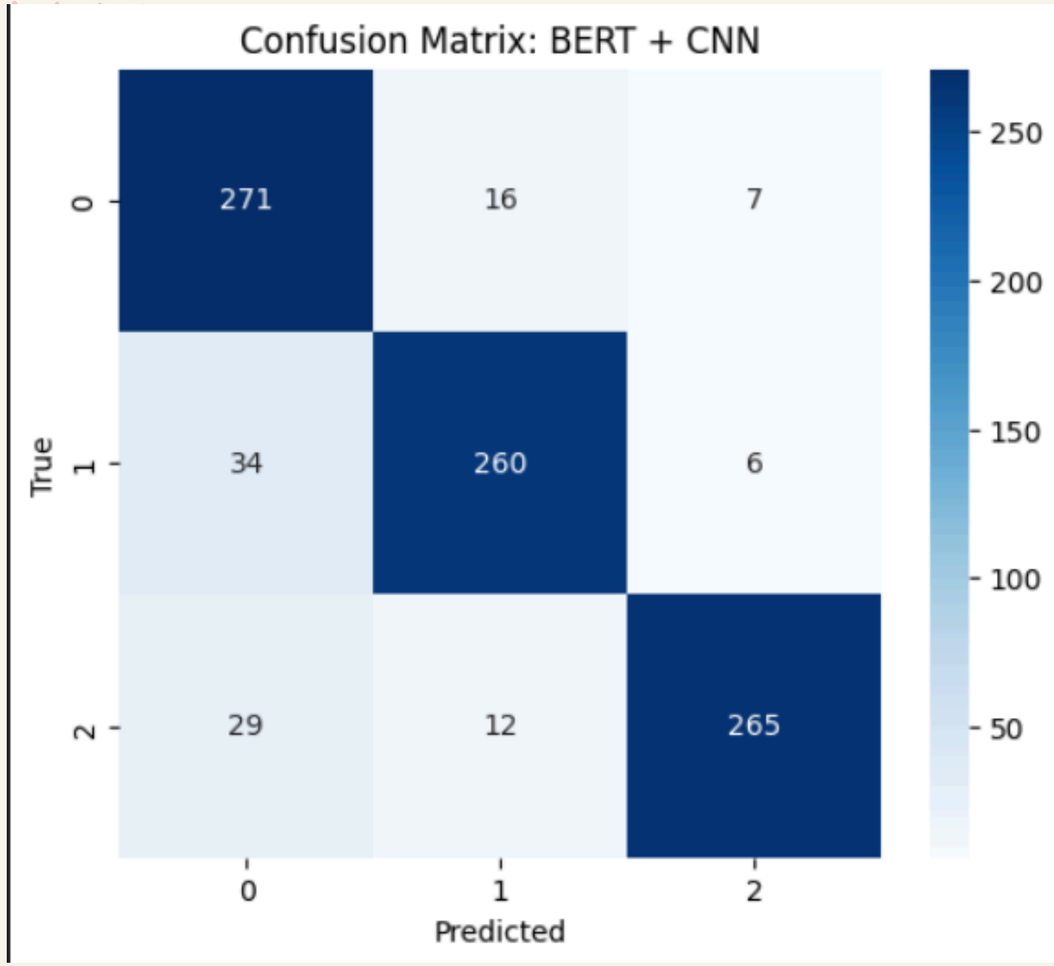


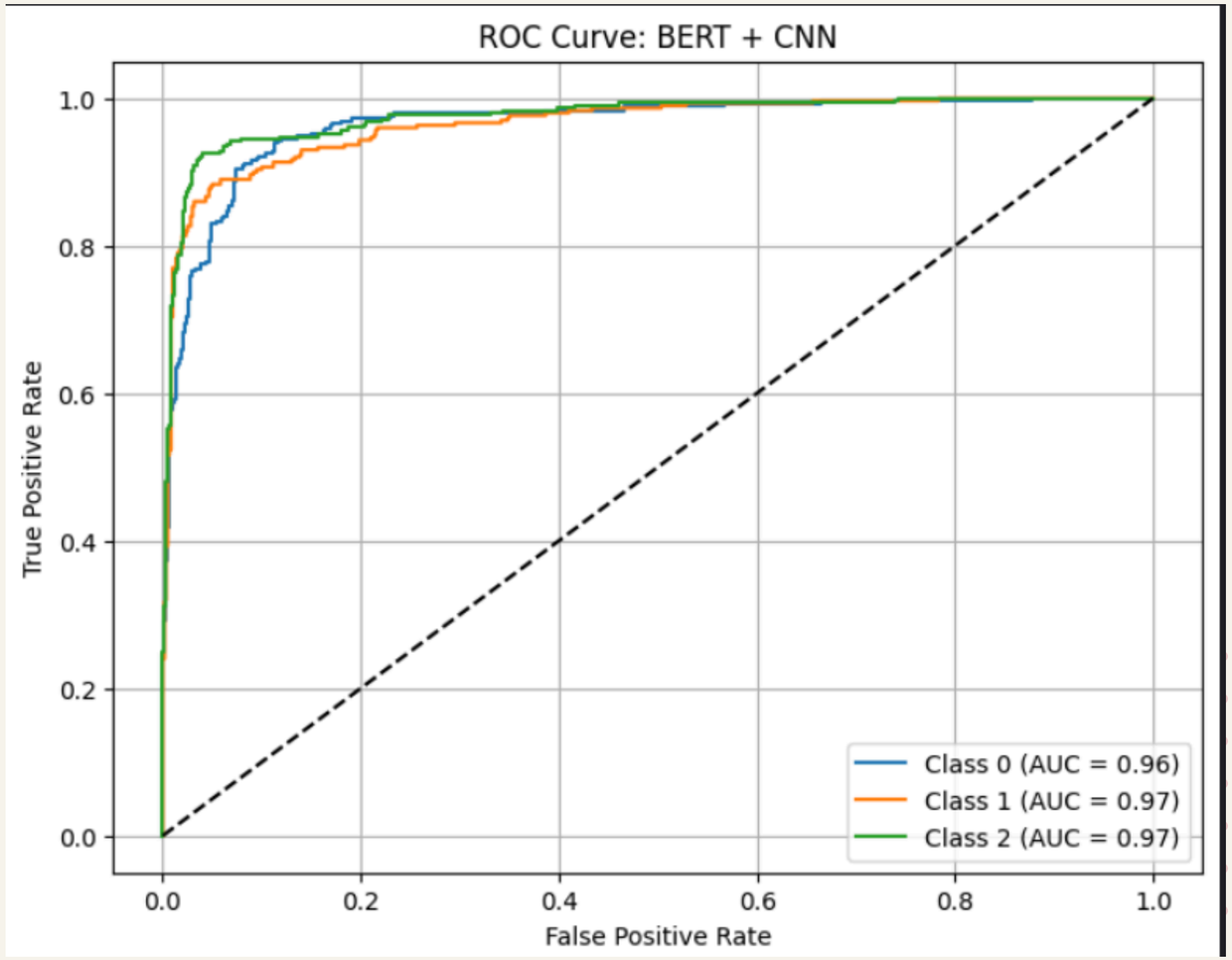
SMALL BERT + MLP



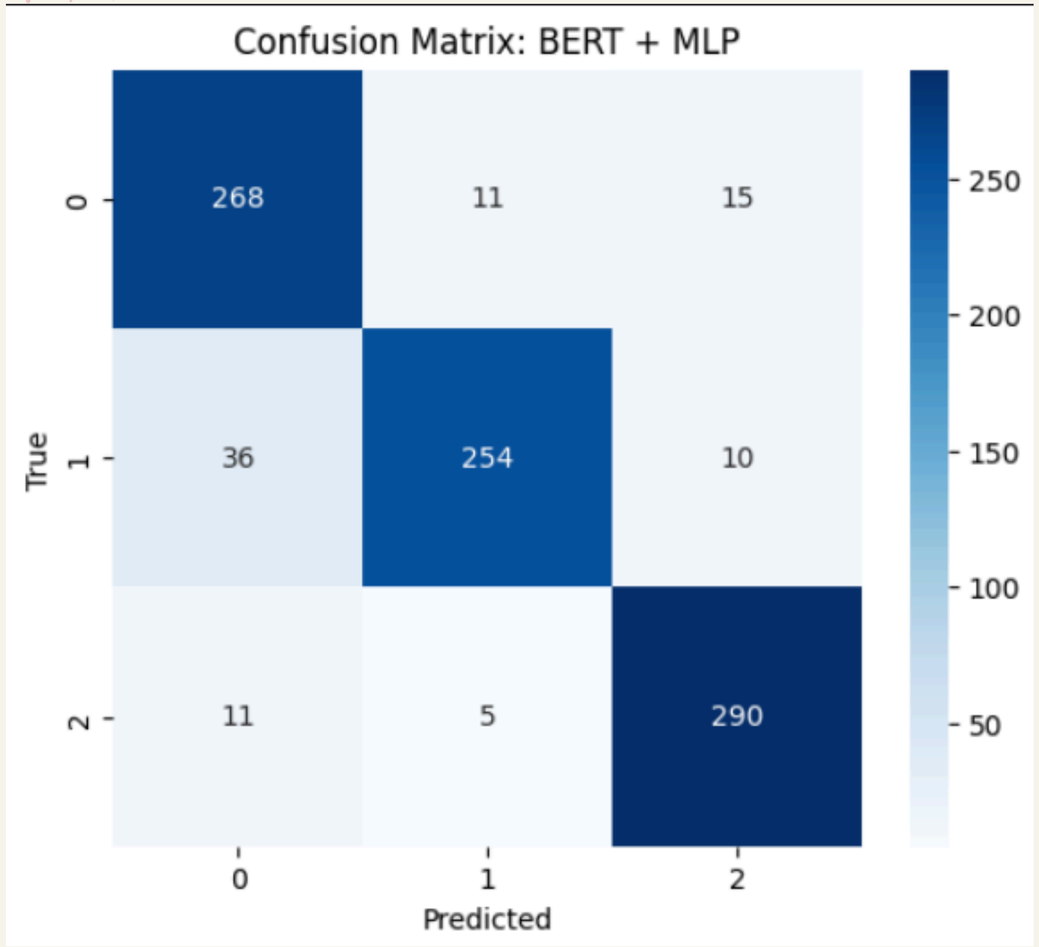


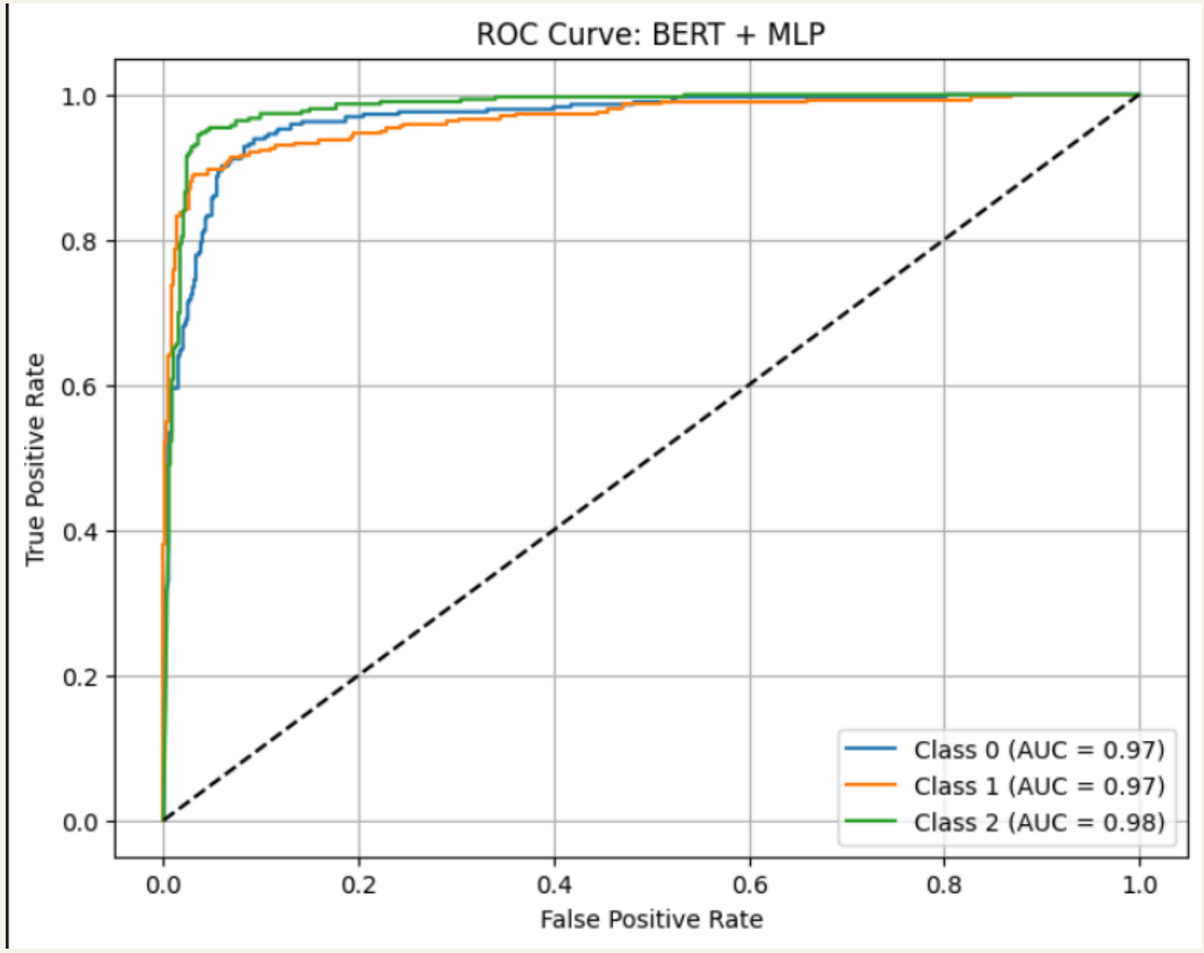
BERT + CNN


Confusion Matrix: BERT + CNN
271
16
7
34
260
6
29
12
265
True
0
1
2
Predicted
0
1
2
250
200
150
100
50


ROC Curve: BERT + CNN
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Class 0 (AUC = 0.96)
Class 1 (AUC = 0.97)
Class 2 (AUC = 0.97)

BERT + MLP

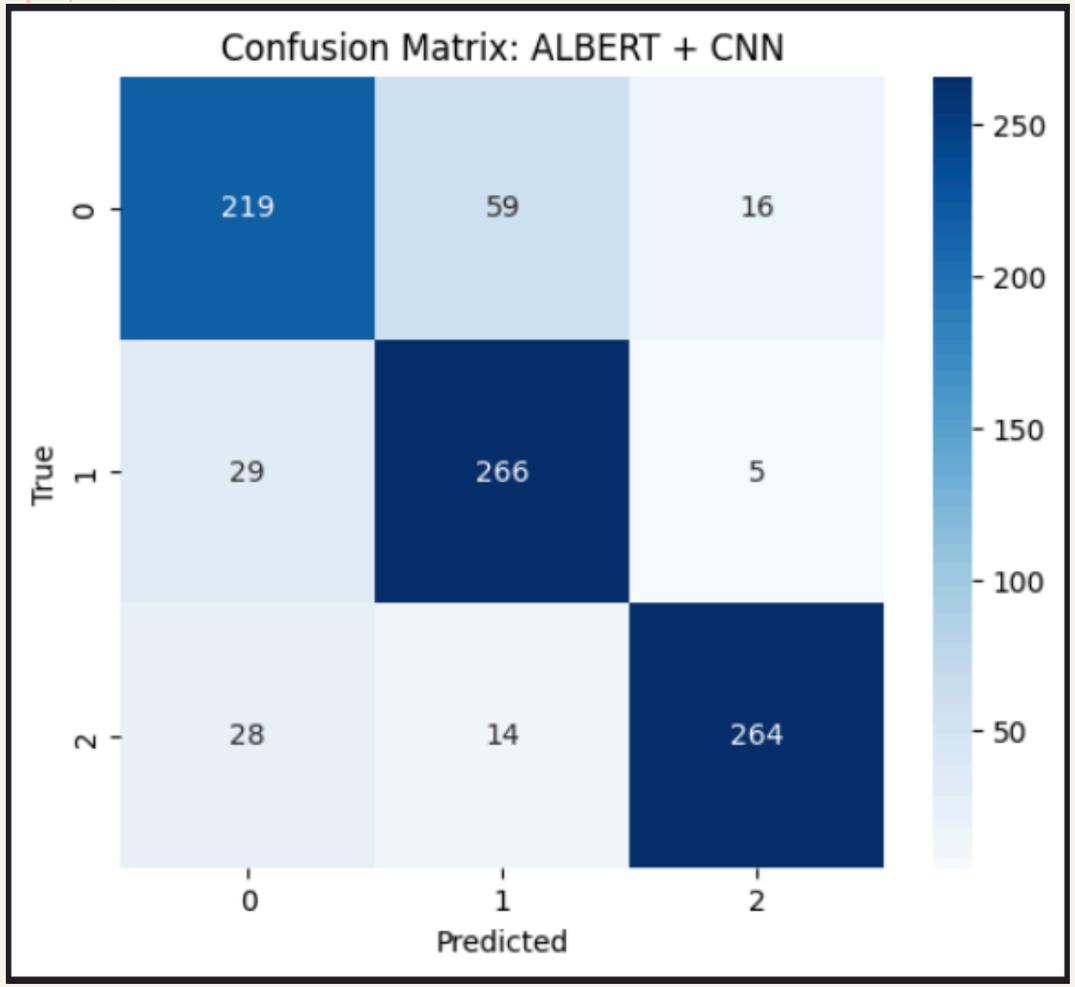

Confusion Matrix: BERT + MLP
True
Predicted
0
1
2
268
11
15
36
254
10
11
5
290
250
200
150
100
50

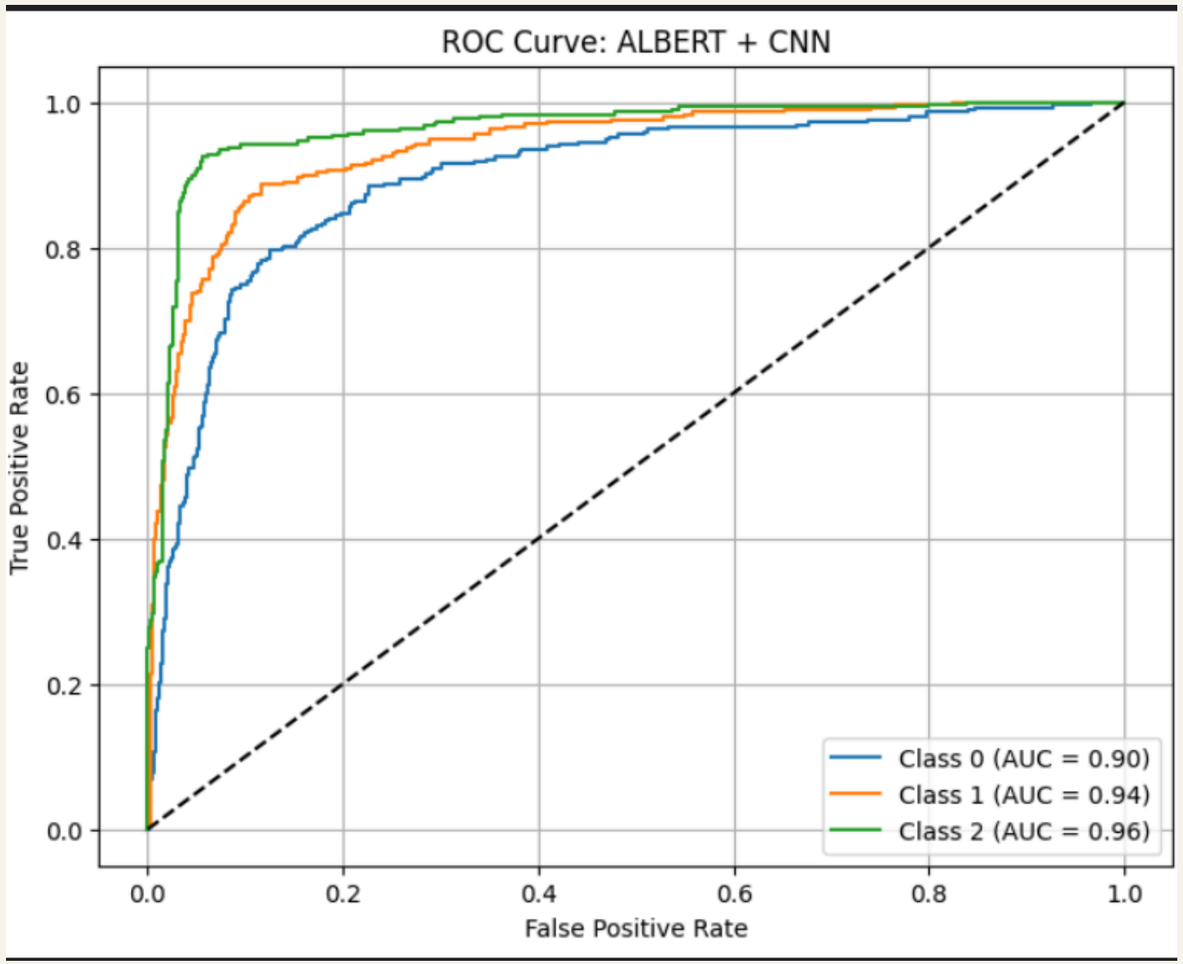

ROC Curve: BERT + MLP
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Class 0 (AUC = 0.97)
Class 1 (AUC = 0.97)
Class 2 (AUC = 0.98)

USING TRANSFORMERS

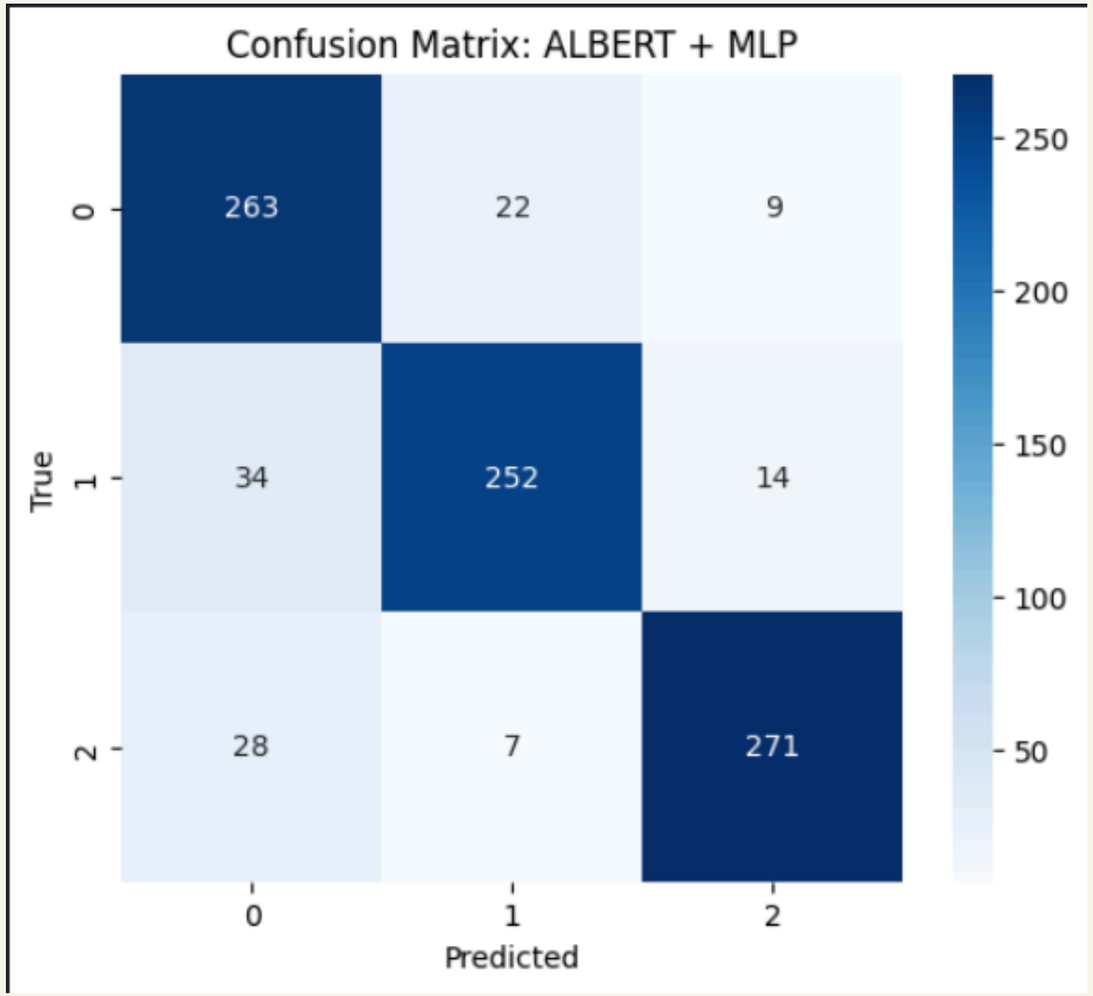
		Results from research paper[2]		Reproduced Results	
Model	**Metric**	**Macro**	**Weighted Avg.**	**Macro**	**Weighted Avg.**
AI-BERT + CNN	**P**	0.76	0.89	0.83	0.84
	R	0.69	0.90	0.83	0.84
	F1	0.72	0.90	0.83	0.84
AI-BERT + MLP	**P**	0.77	0.90	0.87	0.88
	R	0.72	0.91	**0.87**	**0.88**
	F1	0.74	0.90	0.87	0.88
ELECTRA + CNN	**P**	0.75	0.91	0.86	0.87
	R	0.75	**0.91**	0.86	0.87
	F1	0.75	0.91	0.86	0.87
ELECTRA + MLP	**P**	0.75	0.91	0.86	0.87
	R	**0.76**	0.91	0.86	0.87
	F1	0.76	0.91	0.86	0.87

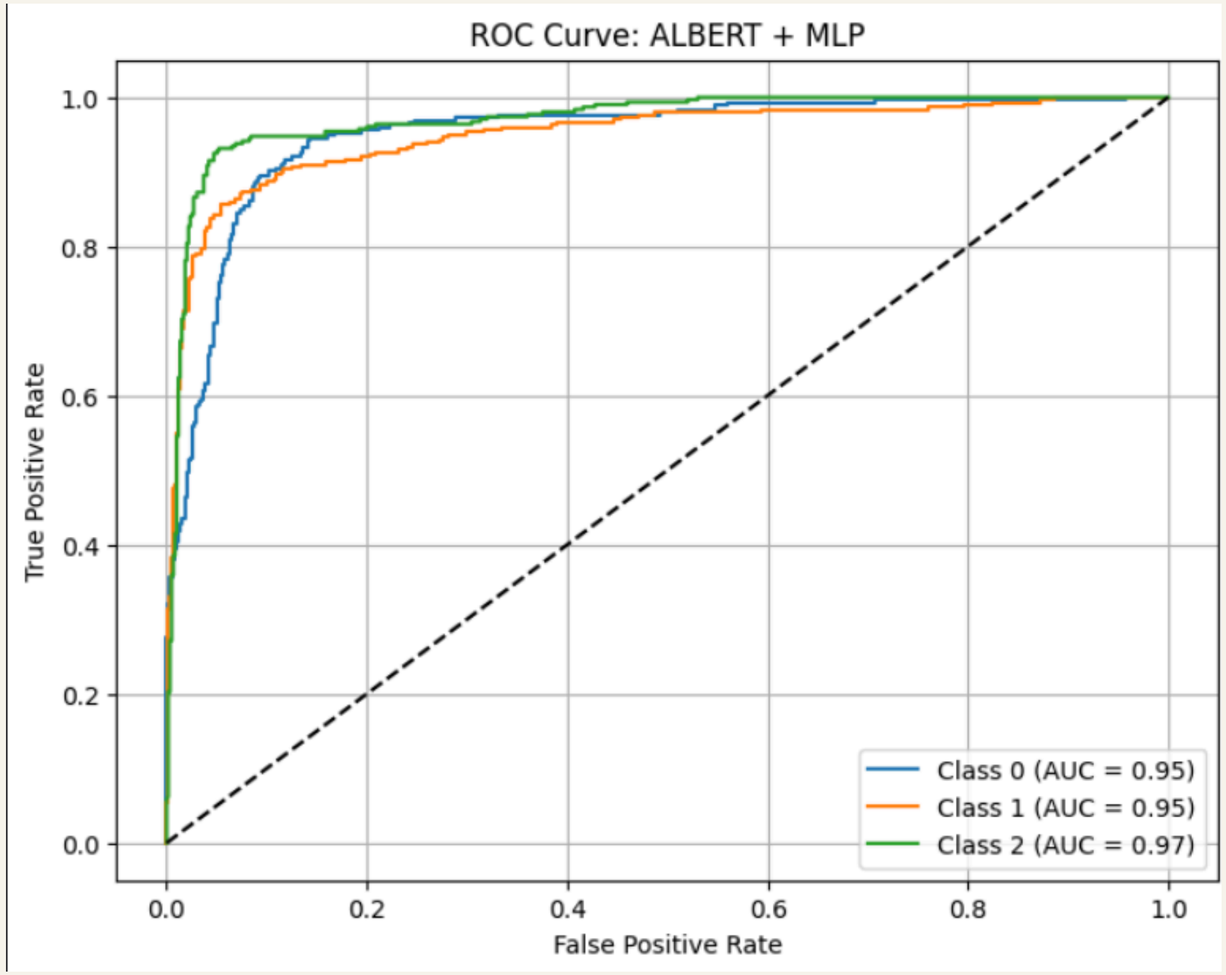
AL-BERT + CNN


Confusion Matrix: ALBERT + CNN
219
59
16
29
266
5
28
14
264
True
0
1
2
Predicted
0
1
2
250
200
150
100
50

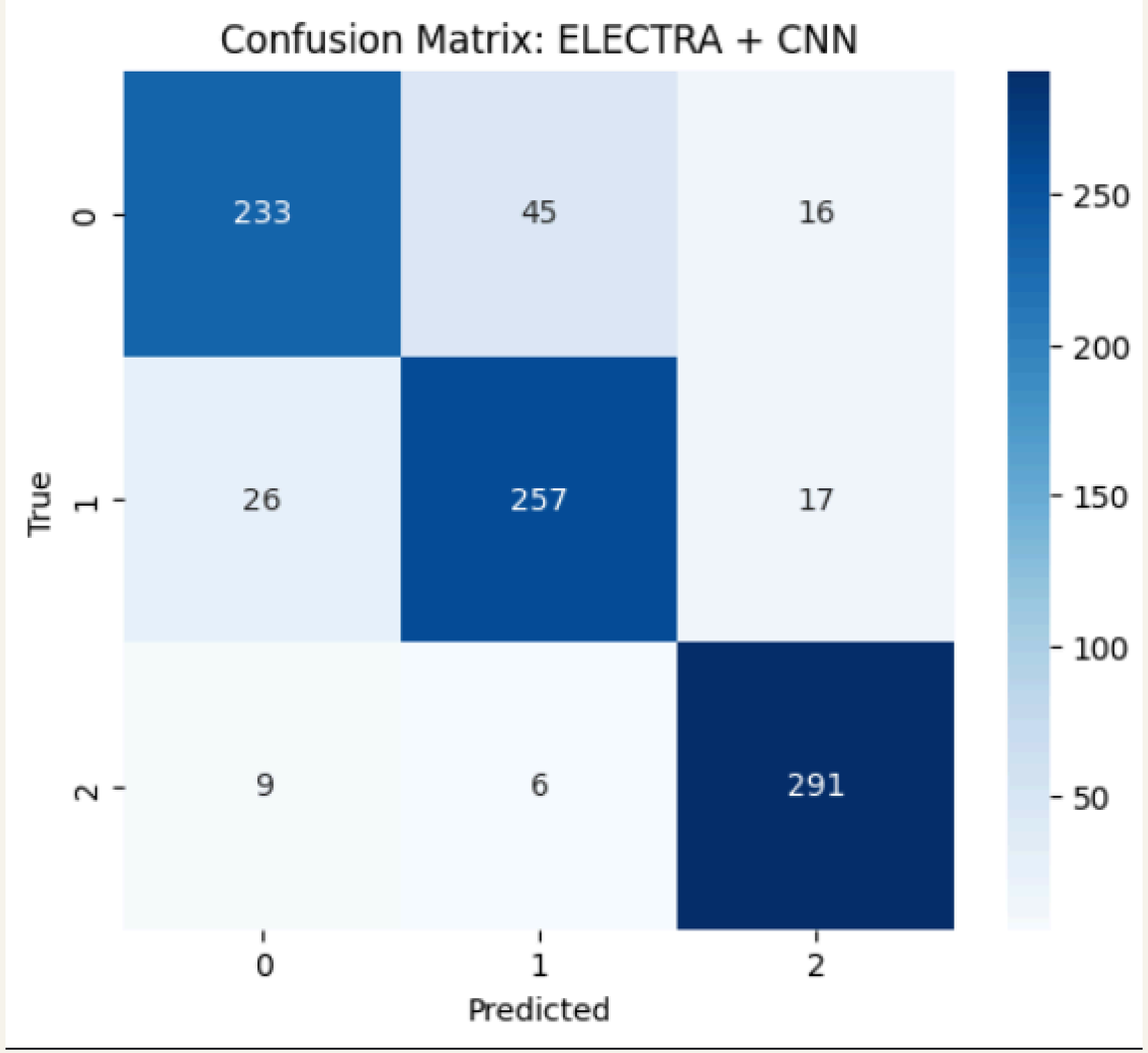

ROC Curve: ALBERT + CNN
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Class 0 (AUC = 0.90)
Class 1 (AUC = 0.94)
Class 2 (AUC = 0.96)

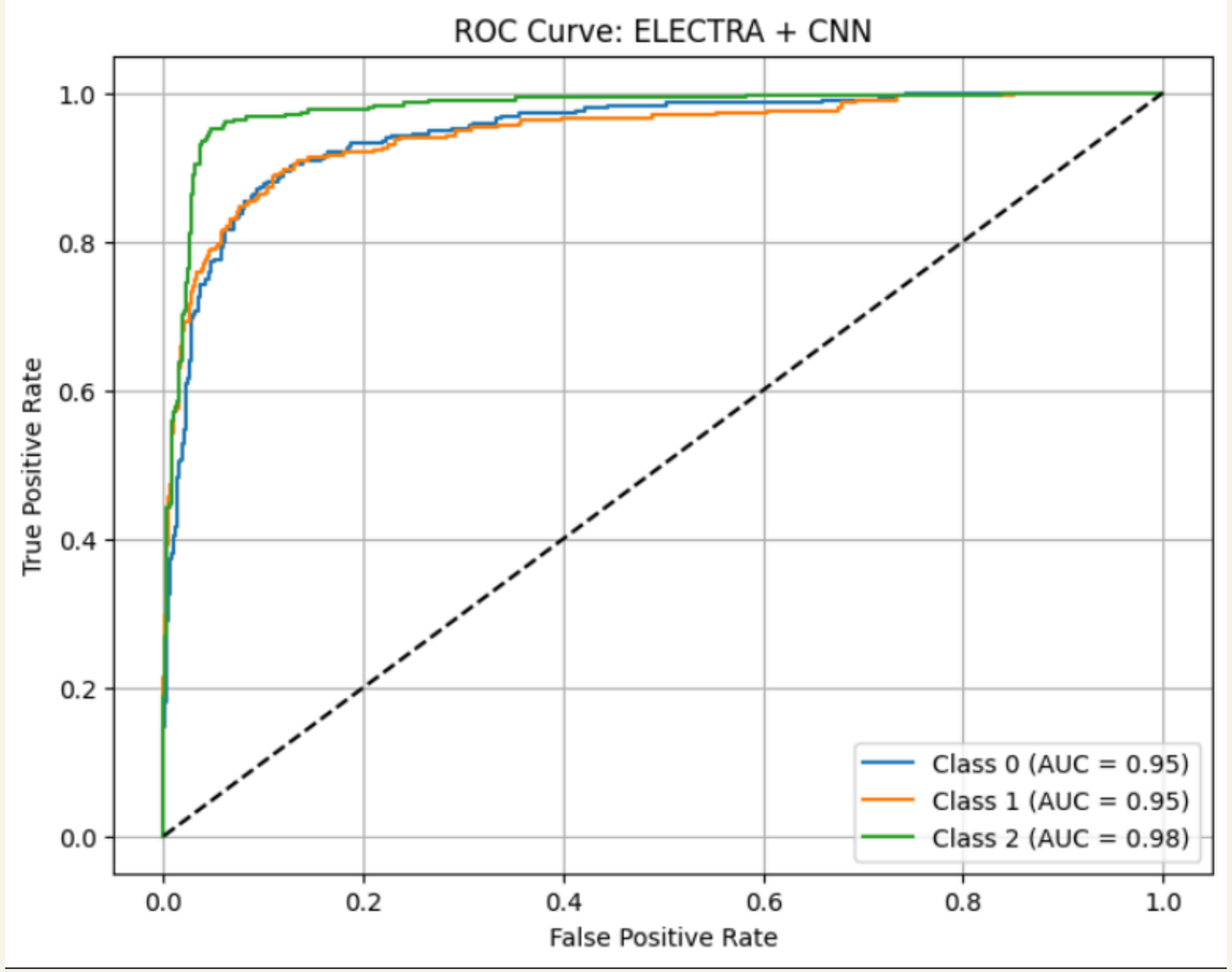
AL-BERT + MLP


Confusion Matrix: ALBERT + MLP
263
22
9
34
252
14
28
7
271
True
0
1
2
Predicted
0
1
2
250
200
150
100
50

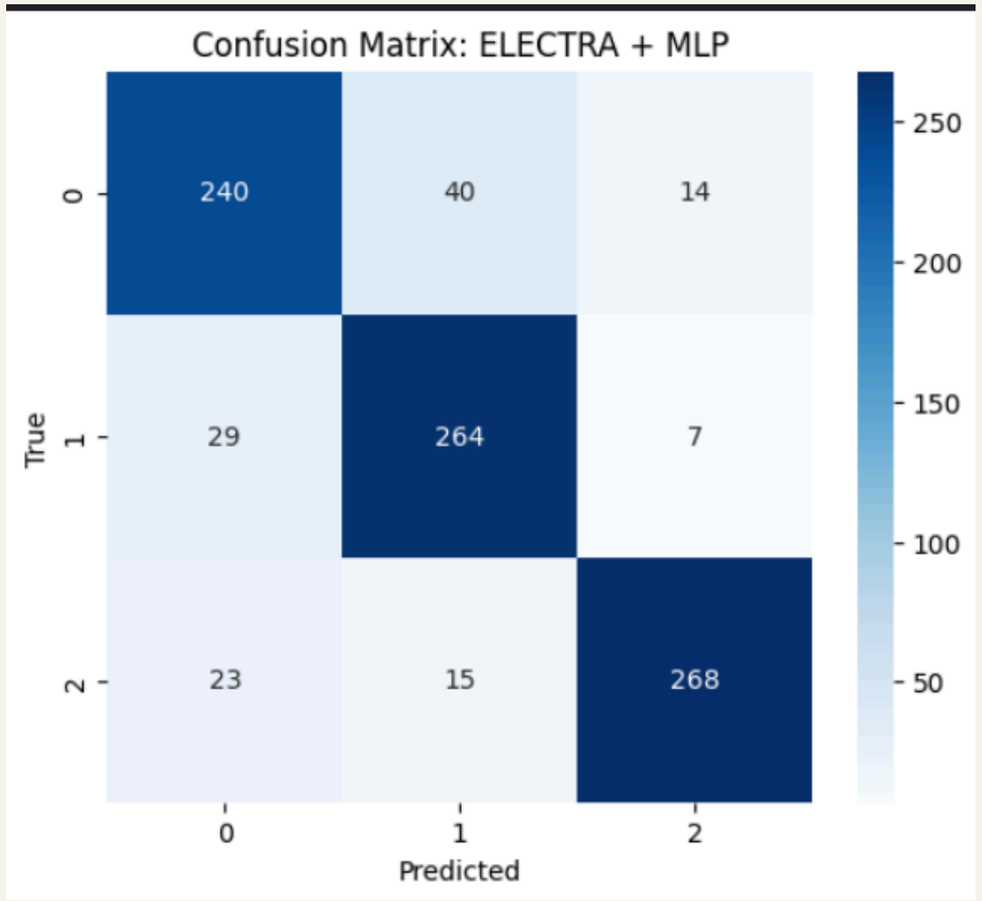

ROC Curve: ALBERT + MLP
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Class 0 (AUC = 0.95)
Class 1 (AUC = 0.95)
Class 2 (AUC = 0.97)

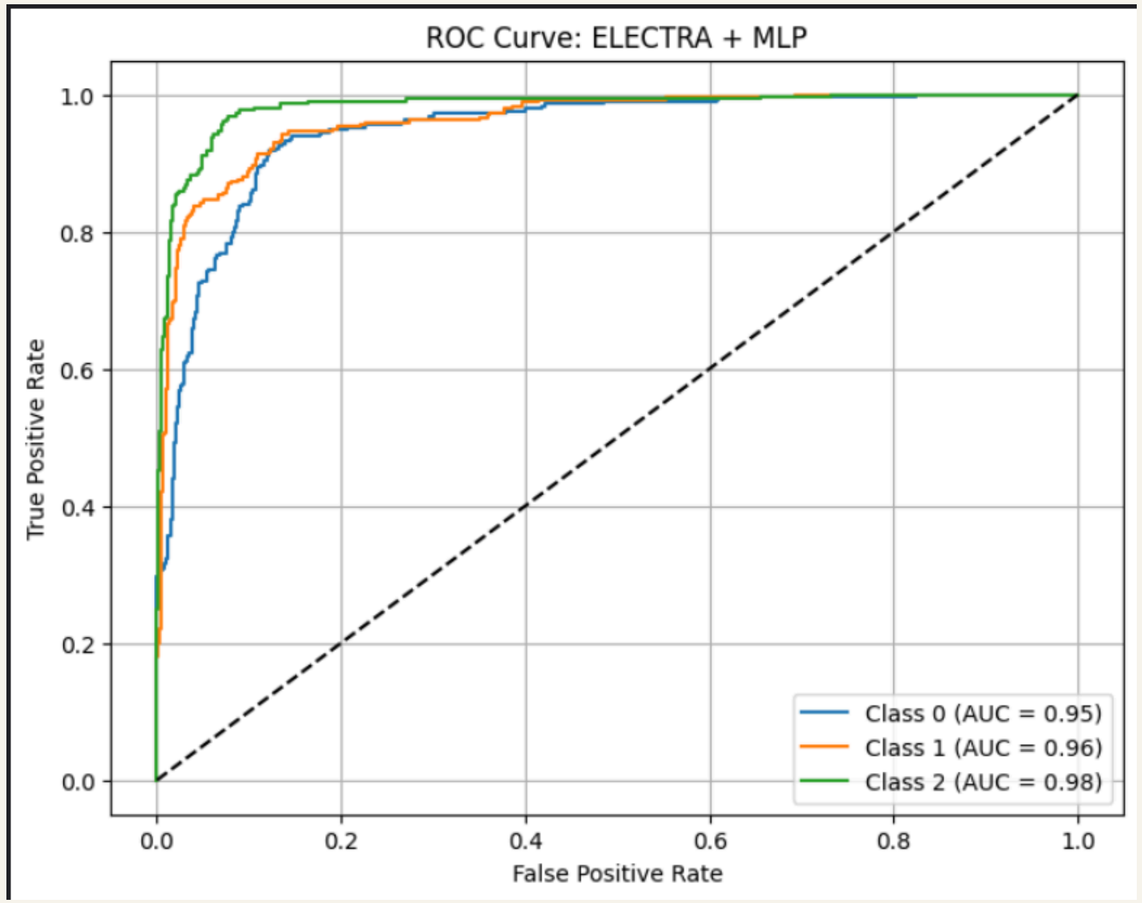
ELECTRA + CNN


Confusion Matrix: ELECTRA + CNN
233
45
16
26
257
17
9
6
291
True
0
1
2
Predicted
250
200
150
100
50


ROC Curve: ELECTRA + CNN
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Class 0 (AUC = 0.95)
Class 1 (AUC = 0.95)
Class 2 (AUC = 0.98)

ELECTRA + MLP


Confusion Matrix: ELECTRA + MLP
240
40
14
29
264
7
23
15
268
True
0
1
2
Predicted
0
1
2
250
200
150
100
50


ROC Curve: ELECTRA + MLP
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Class 0 (AUC = 0.95)
Class 1 (AUC = 0.96)
Class 2 (AUC = 0.98)

CONCLUSION

- In terms of macro F1 score, it can be observed that TF-IDF embeddings, particularly when combined with the MLP model, outperform other methods, achieving the highest macro F1 of 0.94. However, Glove embeddings also show strong performance—especially with the Bi-LSTM architecture, which reaches a macro F1 score of 0.89, close to the best TF-IDF model. This suggests that contextual information captured by Glove combined with sequential modeling (Bi-LSTM) can provide results comparable to top-performing TF-IDF-based models.
- While the Glove + CNN and Glove + MLP models obtain macro F1 scores of 0.86 and 0.82 respectively, they still lag behind TF-IDF + SVM (0.88) and TF-IDF + MLP (0.94). The TF-IDF + XGB model, although slightly weaker in macro recall (0.79), performs similarly to Glove + CNN in terms of F1 (0.81 vs 0.86).
- In terms of weighted average F1 score, the TF-IDF + MLP model again leads with 0.95, followed closely by TF-IDF + SVM (0.90) and Glove + Bi-LSTM (0.91). This indicates that TF-IDF with MLP is not only effective at predicting minority classes but also performs consistently across all classes.

Overall, while Glove embeddings perform reasonably well—especially when combined with deeper architectures like Bi-LSTM—the TF-IDF-based approaches, particularly with MLP, show the most robust and balanced results on this dataset.

CONCLUSION

Small BERT (Lightweight & Efficient)

- Designed to reduce model size and inference time
- Performs competitively with F1-scores ~0.83–0.84
- Achieves ROC-AUC up to 0.97, making it suitable for real-time applications or resource-limited environments
- Best paired with MLP for balanced accuracy and efficiency

ELECTRA (Discriminator-Based Pretraining)

- Pretraining strategy leads to more data-efficient learning
- ELECTRA + MLP delivers ROC-AUC = 0.98
- Strong F1 performance (0.87) across classes
- Ideal for scenarios where early detection and low false negatives are critical

AL-BERT (A Lite BERT)

- Embedding compression makes it faster with fewer parameters
- Offers strong results in combination with MLP:
- F1 = 0.87, ROC-AUC = 0.97
- Effective middle ground between model size and classification power
- Good option where deployment speed and memory are constraints

BERT (Baseline Transformer)

- Achieves the highest overall scores among all models:
- F1-score = 0.91
- ROC-AUC up to 0.98 with BERT + MLP
- Shows strong generalization and superior contextual understanding
- BERT + MLP outperforms all others in both macro and weighted metrics
- Best suited for high-accuracy production systems

- **MLP consistently outperforms CNN when paired with transformer embeddings**
- **BERT + MLP and ELECTRA + MLP are top performers across all metrics**
- **Small BERT and AL-BERT offer strong alternatives with less computational cost**
- **Transformer + MLP combinations provide robust, high-recall systems for hate speech detection**

FUTURE WORKS

Multimodal Detection:

- Future research can incorporate images with embedded text or analyze video-based content to enhance the detection of hate speech in more complex and diverse formats, especially across platforms like Twitter and Instagram.

Multilingual Support:

- The current model is limited to English-only tweets. Expanding the dataset to include other widely used languages such as Hindi, Tamil, Japanese, and more can improve the model's applicability and global relevance.

Improved Recall & Coverage:

- Although the model achieves a recall of 0.91, some hate speech content still goes undetected. Future work should focus on developing models that can minimize false negatives and fully capture hate speech nuances across varying contexts.

Dataset Expansion:

- Building a more generalizable and robust framework requires a larger and more diverse training dataset. Future efforts can include extending the current dataset with more annotated samples to improve accuracy and reduce bias.

REFERENCES

[1] Roy, P. K., Tripathy, A. K., Das, T. K., & Gao, X.-Z. (2020). A Framework for Hate Speech Detection Using Deep Convolutional Neural Network. IEEE Access. Retrieved from https://doi.org/10.1109/ACCESS.2020.3037073

[2] Malik, J. S., Qiao, H., Pang, G., & van den Hengel, A. (2024). Deep learning for hate speech detection: a comparative study. International Journal of Data Science and Analytics. Retrieved from https://doi.org/10.1007/s41060-024-00650-6

[3] K. J., Bhanushree, et al. (2023). Automatic Hate Speech Detection using Ensemble Method and Natural Language Processing Techniques. IEEE NMITCON 2023. Retrieved from https://www.researchgate.net/publication/374786398

[4] Futane, S., Bandwal, T., Dhonde, D., Gudmewar, S., & Kadam, A. (2023). Hate Speech Detection Using Machine Learning. International Journal for Research in Applied Science & Engineering Technology (IJRASET), 11(IV). Retrieved from https://www.ijraset.com

[5] Chhabra, A., & Vishwakarma, D. K. (2022). Fuzzy and Machine Learning Classifiers for Hate Content Detection: A Comparative Analysis. 2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST). IEEE. Retrieved from https://doi.org/10.1109/AIST.2022

REFERENCES

[6] Jemima, D. P., et al. (2022). Hate Speech Detection using Machine Learning. 2022 7th International Conference on Communication and Electronics Systems (ICCES). IEEE. Retrieved from https://doi.org/10.1109/ICCES.2022

[7] Naseem, U., Razzak, I., Musial, K., & Imran, M. (2021). Offensive Language and Hate Speech Detection with Deep Learning and Transfer Learning. Retrieved from https://www.researchgate.net/publication/353729289

[8]Gambäck, B., & Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. In Proceedings of the First Workshop on Abusive Language Online (pp. 85–90).

[10] Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion .

[9] Gala, J., Gandhi, D., Mehta, J., & Talat, Z. (2023). A federated approach for hate speech detection. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL).

[10] Plaza-del Arco, F. M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for Spanish hate speech detection. Expert Systems with Applications, 166, 114120. https://doi.org/10.1016/j.eswa.2020.114120

THANK YOU

Jamia MIllia Islamia| 2025