DEEP LEARNING-BASED TOXIC COMMENT CLASSIFICATION

FINAL PROJECT — DTSA-5511
INTRODUCTION TO DEEP LEARNING

JINNAJATE ACHALAPONG

INTRODUCTION & PROBLEM DEFINITION

Introduction

- Toxic comments pose serious challenges to online communities.
- Manual moderation is time-consuming and not scalable.
- Deep learning offers automated, scalable solutions for text classification.

Problem Definition

- **Objective**: To build deep learning models that detect toxic comments in online discussions.
- **Task Type**: Multi-label text classification
- **Target Classes**: toxic, severe_toxic, obscene, threat, insult, identity_hate
- **Dataset**: Wikipedia talk page comments (Jigsaw/Kaggle)
- **Approach**: Use pre-trained word embeddings with various neural network architectures

DATASET DESCRIPTION

JIGSAW/CONVERSATION AI · FEATURED PREDICTION COMPETITION · 7 YEARS AGO

Toxic Comment Classification Challenge

Identify and classify toxic online comments

Dataset Overview

- **Source**: Jigsaw Toxic Comment Classification Challenge (Kaggle, 2018)
- **Domain**: User comments from Wikipedia talk pages
- **Total records**: 159,571 comments
- **Task**: Multi-label classification – each comment may belong to multiple toxic categories

Dataset Characteristics

- **Multi-label**: One comment can have multiple toxic tags
- **Unstructured text**: Requires preprocessing for deep learning models
- **Real-world noise**: Includes slang, misspellings, and informal language

SUMMARY OF EXPLORATORY DATA ANALYSIS (EDA)

Class Distribution

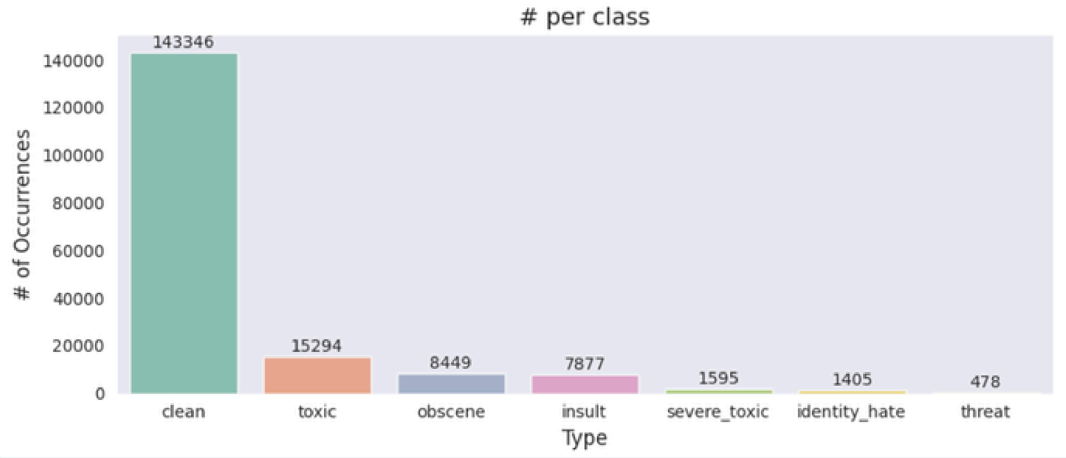
- Strong **class imbalance** observed
- Most comments are **clean**
- Frequent toxic types: toxic, obscene, insult
- Rare types: threat, severe_toxic, identity_hate

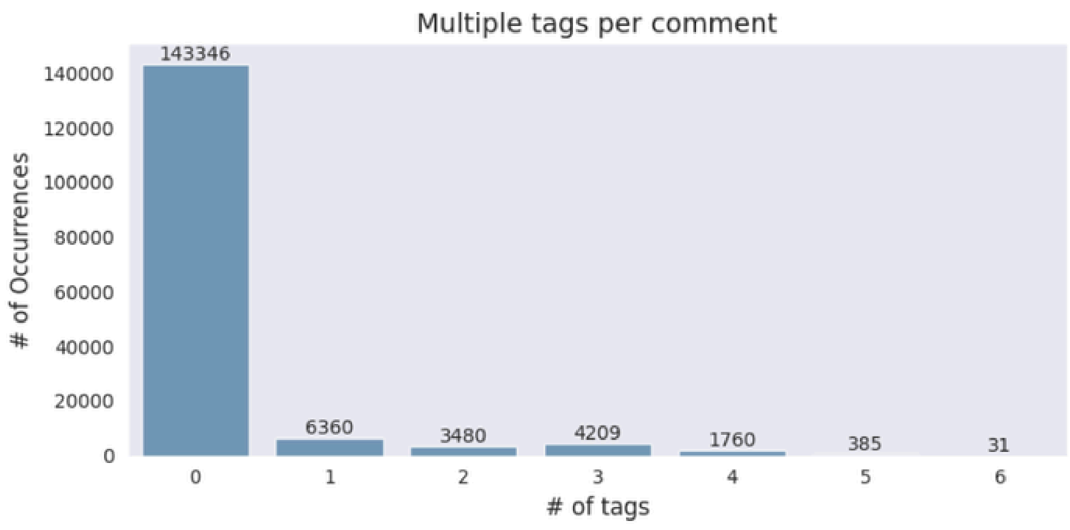
Multi-Label Comments

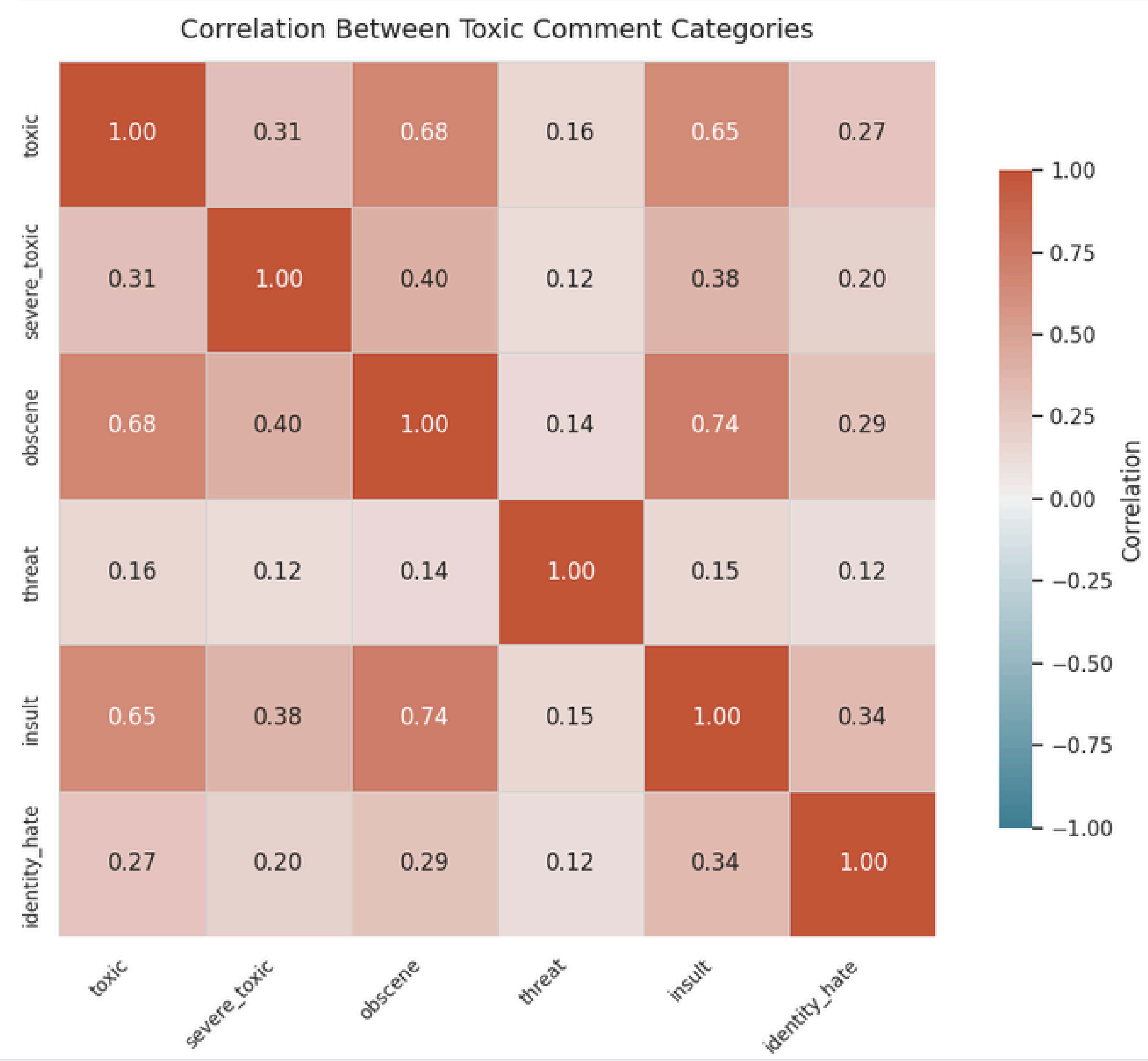
- ~10% of comments have **one or more toxic tags**
- 31 comments have **all six toxic categories**
- Multi-label modeling is required

Label Correlation

- **Strong correlation**:
  - toxic and obscene (0.68)
  - toxic and insult (0.65)
- **Weak correlation**:
  - threat with other categories







	toxic	severe_toxic	obscene	threat	insult	identity_hate
toxic	1.00	0.31	0.68	0.16	0.65	0.27
severe_toxic	0.31	1.00	0.40	0.12	0.38	0.20
obscene	0.68	0.40	1.00	0.14	0.74	0.29
threat	0.16	0.12	0.14	1.00	0.15	0.12
insult	0.65	0.38	0.74	0.15	1.00	0.34
identity_hate	0.27	0.20	0.29	0.12	0.34	1.00

WORDCLOUD ANALYSIS

Clean Comments WordCloud

- Focused on article editing:
- article, edit, wikipedia, thank, source, page
- Language is collaborative, neutral, and constructive
- Reflects typical Wikipedia community interaction

⚠️ Interpretation & Caution

- Clear lexical contrast between clean and toxic comments
- Highlights need for robust preprocessing
- (e.g., lowercasing, profanity masking, embedding strategies)
- Offensive content is presented for analytical purposes only

Toxic Categories WordClouds

- Frequent appearance of profanities, slurs, threatening verbs
- fuck, die, kill, nigger, moron, faggot, ass appear prominently
- Tone is hostile, aggressive, or discriminatory

TEXT PREPROCESSING WITH GLOVE

Objective

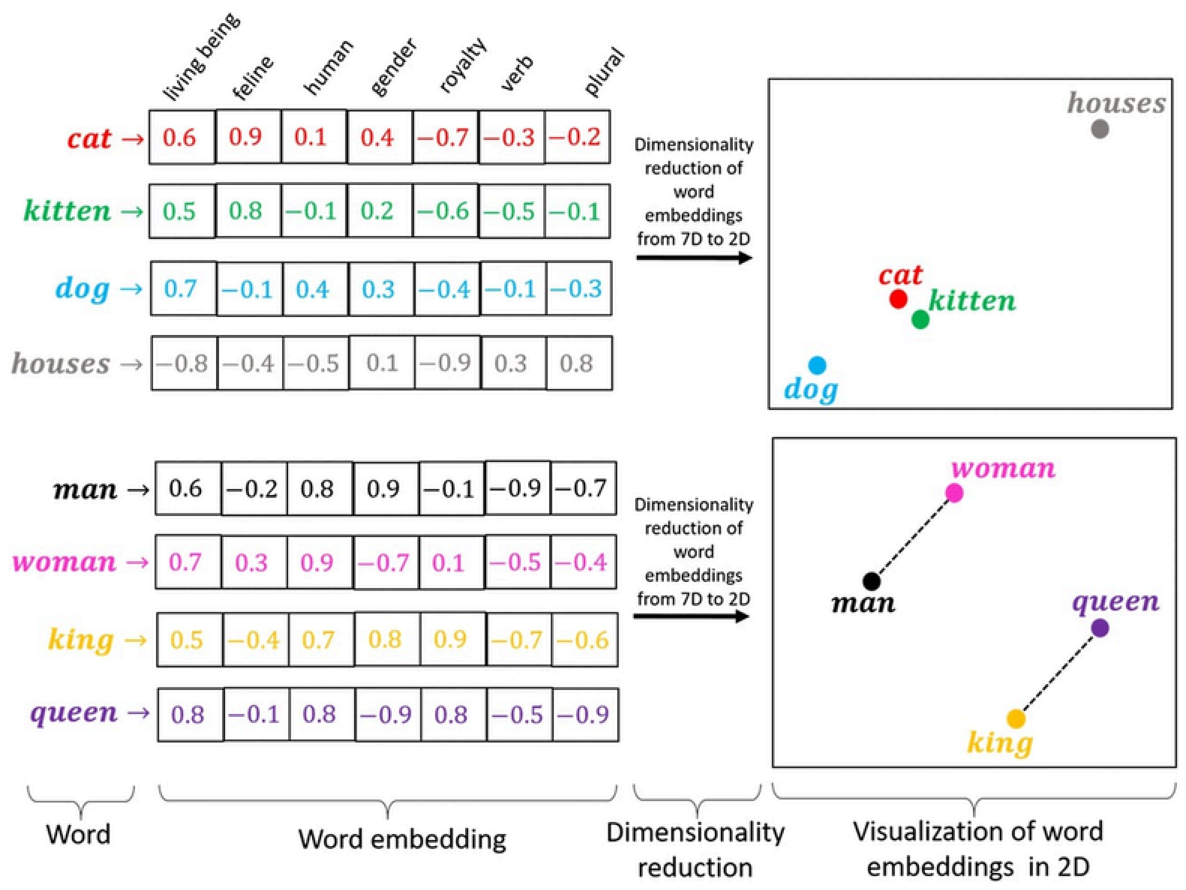
- Prepare text data for deep learning using pre-trained word embeddings (GloVe)

Key Preprocessing Steps

- **Text Cleaning**
  - Lowercasing, removing punctuation/special characters
- **Tokenization & Padding**
  - Convert words to integer sequences
  - Pad sequences to uniform length (e.g., 150 tokens)
- **Load GloVe Vectors**
  - 100-dimensional GloVe embeddings (glove.6B.100d.txt)
- **Build Embedding Matrix**
  - Map each word in our vocabulary to its GloVe vector
  - Words not found in GloVe are initialized as zero vectors
- **Embedding Layer Initialization**
  - Create a non-trainable embedding layer using the matrix

Why Use GloVe?

- Captures semantic relationships between words (e.g., king–queen, hate–love)
- Reduces the need for large labeled datasets
- Improves model generalization, especially for rare/complex terms



MODEL 1: ARTIFICIAL NEURAL NETWORK (ANN)

Model Architecture

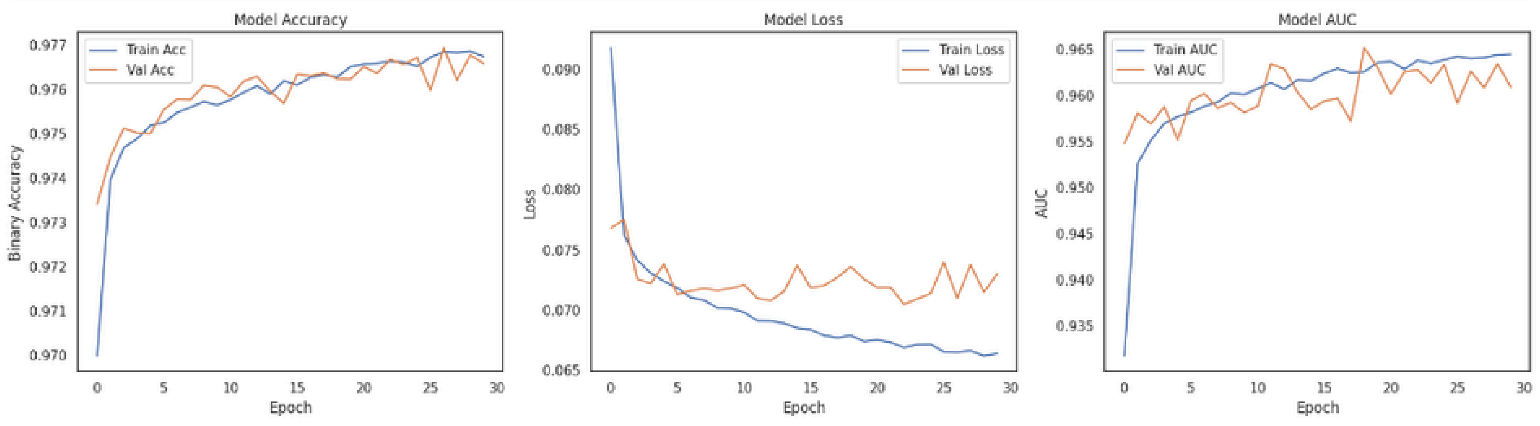
- **Input**: Pre-trained GloVe embeddings (100D)
- **Flatten**: Converts embedding output into 1D vector
- **Dense** Layer: ReLU activation
- **Dropout**: To reduce overfitting
- **Output Layer**: 6 sigmoid units (multi-label classification)

Hyperparameter Tuning with KerasTuner

Hyperparameter	Description	Best Value Found
`units`	Number of neurons in the Dense hidden layer	128
`dropout`	Dropout rate after the hidden layer	0.2
`learning_rate`	Learning rate for the Adam optimizer	0.01

Training Performance

- High training and validation **accuracy (~97.5%)**
- Some **overfitting** observed in validation loss
- **Validation AUC ~0.96**, stable performance across 30 epochs



Classification Results (Per Class)

- **Macro F1 Score**: 0.38
- **Macro ROC AUC**: 0.95
- **Precision** generally higher than recall
- Weak recall for rare classes (threat, severe_toxic)

```
1496/1496 ━━━━━━━━━━━━━━━━━━━━ 2s 1ms/step
📊 Classification Report (per class):
               precision    recall  f1-score   support

        toxic       0.82      0.56      0.66      4582
 severe_toxic       0.65      0.19      0.30       486
      obscene       0.80      0.54      0.65      2556
       threat       0.50      0.03      0.06       136
       insult       0.78      0.46      0.58      2389
identity_hate       0.68      0.03      0.07       432

    micro avg       0.80      0.49      0.61     10581
    macro avg       0.70      0.30      0.38     10581
 weighted avg       0.79      0.49      0.59     10581
  samples avg       0.05      0.04      0.04     10581

🧠 ROC AUC Score (macro): 0.9541
⚖️ F1 Score (macro): 0.3847
```

MODEL 2: SIMPLE RECURRENT NEURAL NETWORK (RNN)

Model Architecture

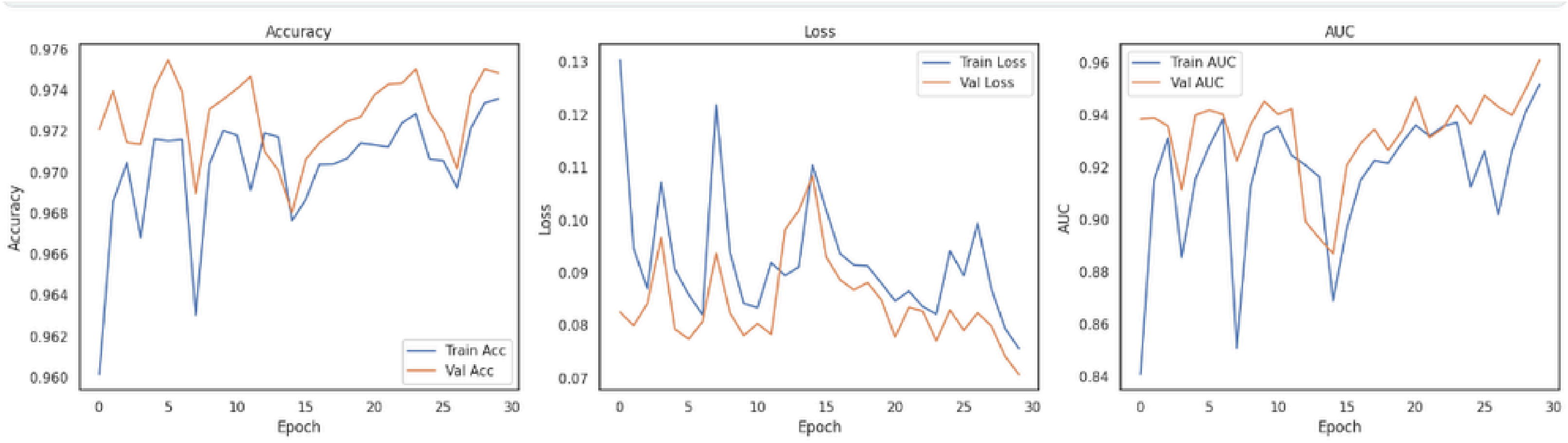
- **Input**: Pre-trained GloVe embeddings (100D)
- **Recurrent Layer**: SimpleRNN
- **Dropout**: Applied after RNN
- **Output**: 6 sigmoid units for multi-label prediction

Hyperparameter Tuning with KerasTuner

Hyperparameter	Description	Best Value Found
`units`	Number of neurons in the Dense hidden layer	64
`dropout`	Dropout rate after the hidden layer	0.5
`learning_rate`	Learning rate for the Adam optimizer	0.0005

Training Performance

- Validation accuracy ~97.5%
- AUC shows gradual improvement, ends at ~0.96
- High **variance in loss**, suggests instability or sensitivity



Classification Results (Per Class)

- **Macro F1 Score**: 0.34
- **Macro ROC AUC**: 0.946
- F1 score remains **low for minority classes**, e.g. threat, identity_hate
- toxic, obscene, and insult perform relatively well

```
1496/1496 ━━━━━━━━━━━━━━━━━━━━ 6s 4ms/step
Classification Report:
               precision    recall  f1-score   support

        toxic       0.72      0.61      0.66      4582
 severe_toxic       0.41      0.06      0.11       486
      obscene       0.75      0.58      0.66      2556
       threat       0.00      0.00      0.00       136
       insult       0.70      0.52      0.59      2389
identity_hate       0.19      0.01      0.02       432

    micro avg       0.72      0.52      0.61     10581
    macro avg       0.46      0.30      0.34     10581
 weighted avg       0.68      0.52      0.58     10581
  samples avg       0.05      0.05      0.05     10581

ROC AUC Score (macro): 0.9461
F1 Score (macro): 0.3391
```

MODEL 3: LONG SHORT-TERM MEMORY (LSTM)

Model Architecture

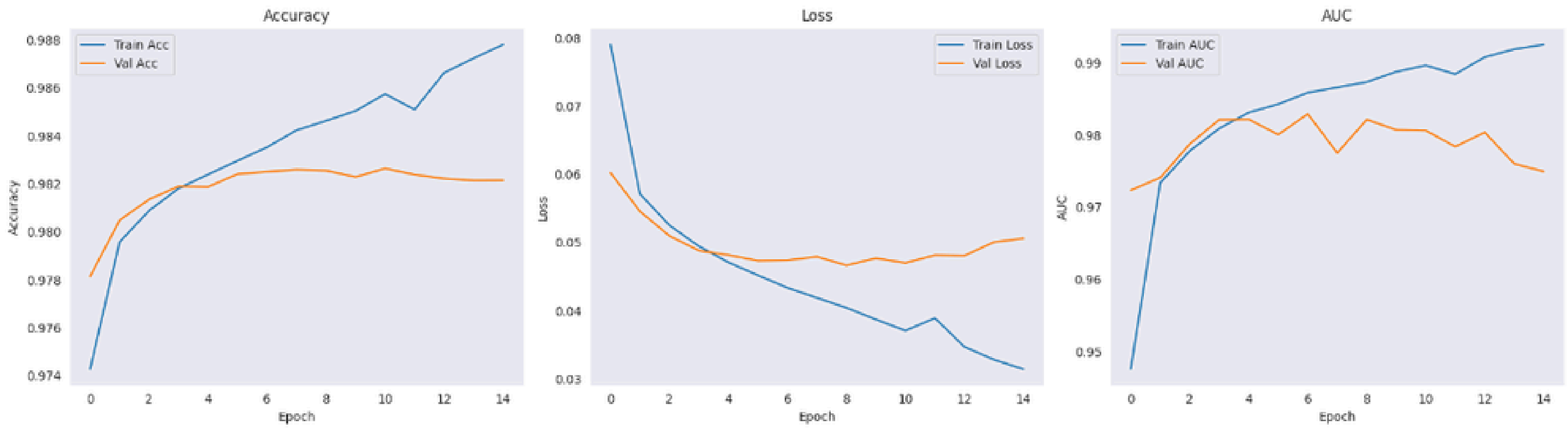
- **Input**: Pre-trained GloVe embeddings (100D)
- **Recurrent Layer**: LSTM (return_sequences=False)
- **Dropout**: Applied after LSTM for regularization
- **Output**: 6 sigmoid units for multi-label classification

Hyperparameter Tuning with KerasTuner

Hyperparameter	Description	Best Value Found
`units`	Number of neurons in the Dense hidden layer	192
`dropout`	Dropout rate after the hidden layer	0.2
`learning_rate`	Learning rate for the Adam optimizer	0.001

Training Performance

- **Validation Accuracy** ~98.2%
- **Train–Val Gap is small** → stable generalization
- AUC ~0.98+ throughout, steady improvement in training



Classification Results (Per Class)

- **Macro F1 Score**: 0.5993
- **Macro ROC AUC**: 0.9810
- Significant improvement across minority classes:
- All classes show balanced precision/recall

```
1496/1496 ━━━━━━━━━━━━━━━━━━━━ 7s 4ms/step
Classification Report:
               precision    recall  f1-score   support

        toxic       0.83      0.75      0.79      4582
 severe_toxic       0.53      0.32      0.40       486
      obscene       0.84      0.75      0.79      2556
       threat       0.52      0.40      0.46       136
       insult       0.74      0.69      0.72      2389
identity_hate       0.62      0.34      0.44       432

    micro avg       0.79      0.70      0.74     10581
    macro avg       0.68      0.54      0.60     10581
 weighted avg       0.79      0.70      0.74     10581
  samples avg       0.07      0.06      0.06     10581

ROC AUC Score (macro): 0.9810
F1 Score (macro): 0.5993
```

MODEL 4: BIDIRECTIONAL LSTM (BI-LSTM)

Model Architecture

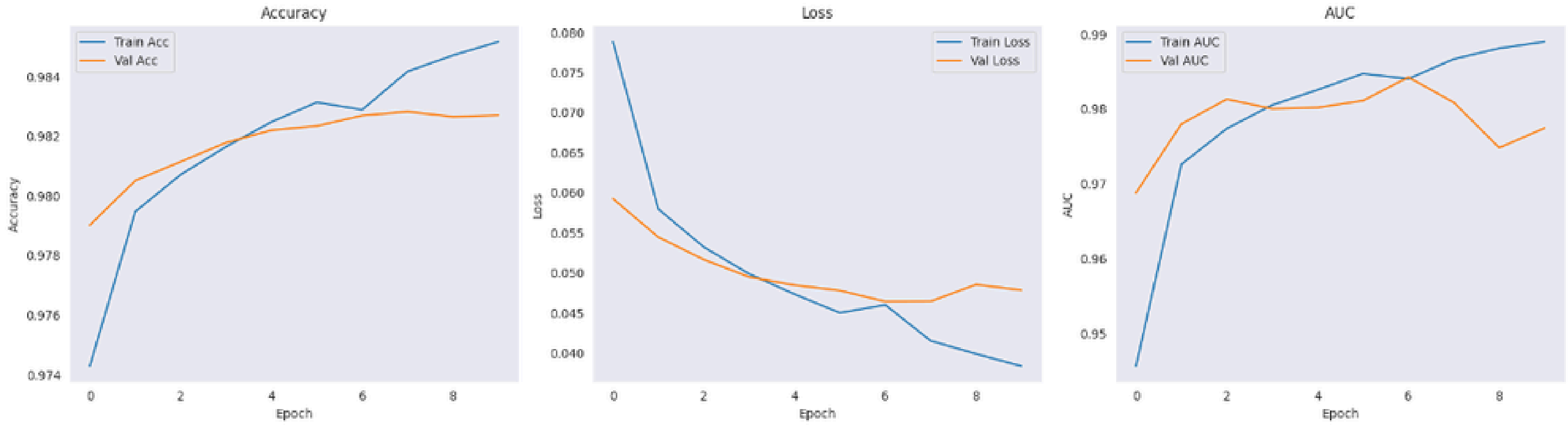
- **Input**: Pre-trained GloVe embeddings (100D)
- **Recurrent Layer**: Bidirectional LSTM
- **Dropout**: Applied after LSTM for regularization
- **Output**: 6 sigmoid units for multi-label classification

Hyperparameter Tuning with KerasTuner

Hyperparameter	Description	Best Value Found
`units`	Number of neurons in the Dense hidden layer	256
`dropout`	Dropout rate after the hidden layer	0.4
`learning_rate`	Learning rate for the Adam optimizer	0.001

Training Performance

- **Validation Accuracy** ≈ 98.3%
- Strong stability with **low loss variance**
- AUC surpasses 0.98, with smooth upward trend



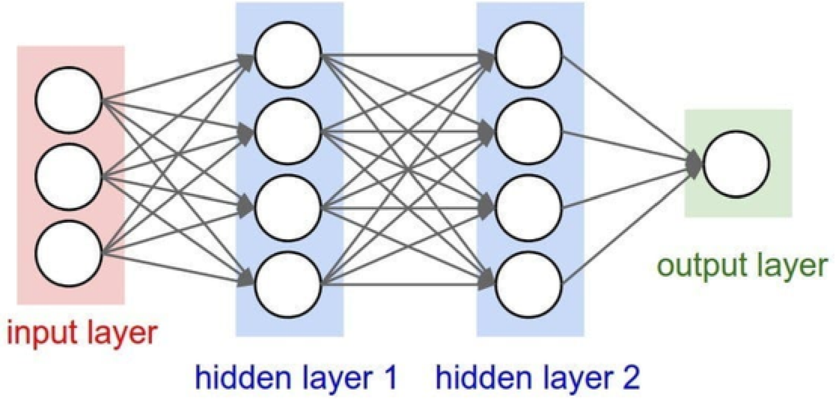
Classification Results (Per Class)

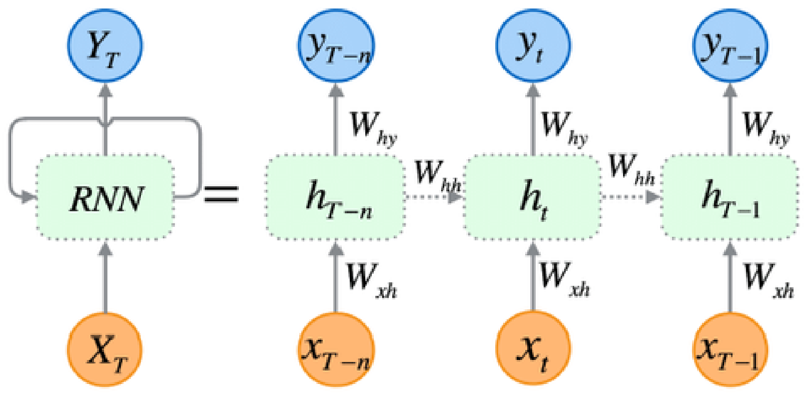
- **Macro F1 Score**: 0.5908
- **Macro ROC AUC**: 0.9826
- Consistently strong across all labels
- toxic, obscene, insult f1-scores ≥ 0.70

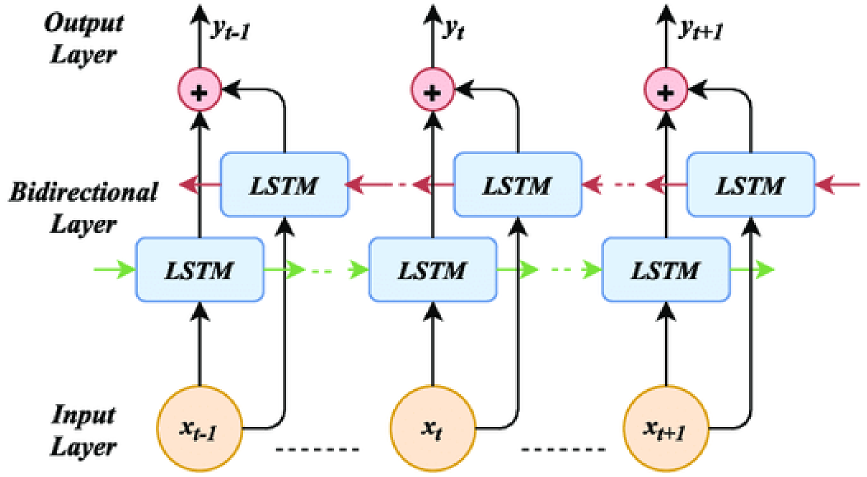
```
1496/1496 ━━━━━━━━━━━━━━━━━━━━ 13s 9ms/step
Classification Report:
               precision    recall  f1-score   support

        toxic       0.85      0.74      0.79      4582
 severe_toxic       0.56      0.29      0.38       486
      obscene       0.83      0.77      0.80      2556
       threat       0.60      0.25      0.35       136
       insult       0.75      0.70      0.73      2389
identity_hate       0.60      0.41      0.49       432

    micro avg       0.80      0.70      0.75     10581
    macro avg       0.70      0.53      0.59     10581
 weighted avg       0.80      0.70      0.74     10581
  samples avg       0.06      0.06      0.06     10581

ROC AUC Score (macro): 0.9826
F1 Score (macro): 0.5908
```

RESULTS & CONCLUSION

Model	Macro F1	ROC AUC (Macro)	Notes
ANN	0.3847	0.9541	Baseline, no sequence modeling
RNN	0.3391	0.9461	Weak sequential capture
LSTM	0.5993	0.9810	Strong overall, best recall
Bi-LSTM	**0.5908**	**0.9826**	Best AUC, bidirectional context



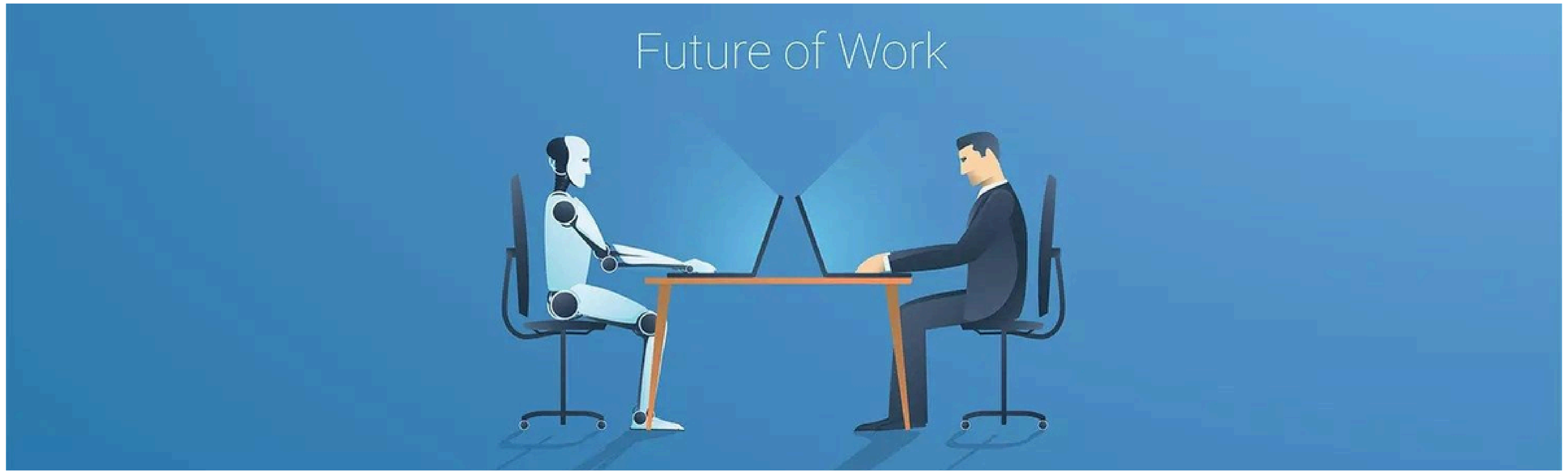




🧠 Conclusion

- **ANN** is insufficient for capturing toxic context
- **Simple RNN** slightly improves recall, but not enough
- **LSTM** delivers substantial performance gain due to memory capabilities
- **Bi-LSTM** achieves best overall performance, especially for hard-to-learn labels like threat and identity_hate

FUTURE WORK



Potential Improvements

1. **Transformers (e.g., BERT)** for deeper semantic modeling and state-of-the-art results
2. **Data Augmentation** to improve learning on low-support classes like threat
3. **Attention Mechanism** focus the model on toxic spans within comments
4. **Model Ensembling** combine ANN + Bi-LSTM + Transformer for robustness
5. **Fine-tuning Embeddings** let GloVe adjust during training (instead of freezing)
6. **Explainability (XAI)** use SHAP/LIME to interpret predictions

THANK YOU