Automated Cyberbullying Detection System

This project aims to create an NLP-based system to identify offensive and hateful comments on social media platforms. The goal is to reduce user exposure to cyberbullying and promote a safer online environment.

Project Objectives

1. **Detect Harmful Content**

   Develop an NLP system to identify offensive and hateful comments on social media.

2. **Improve User Experience**

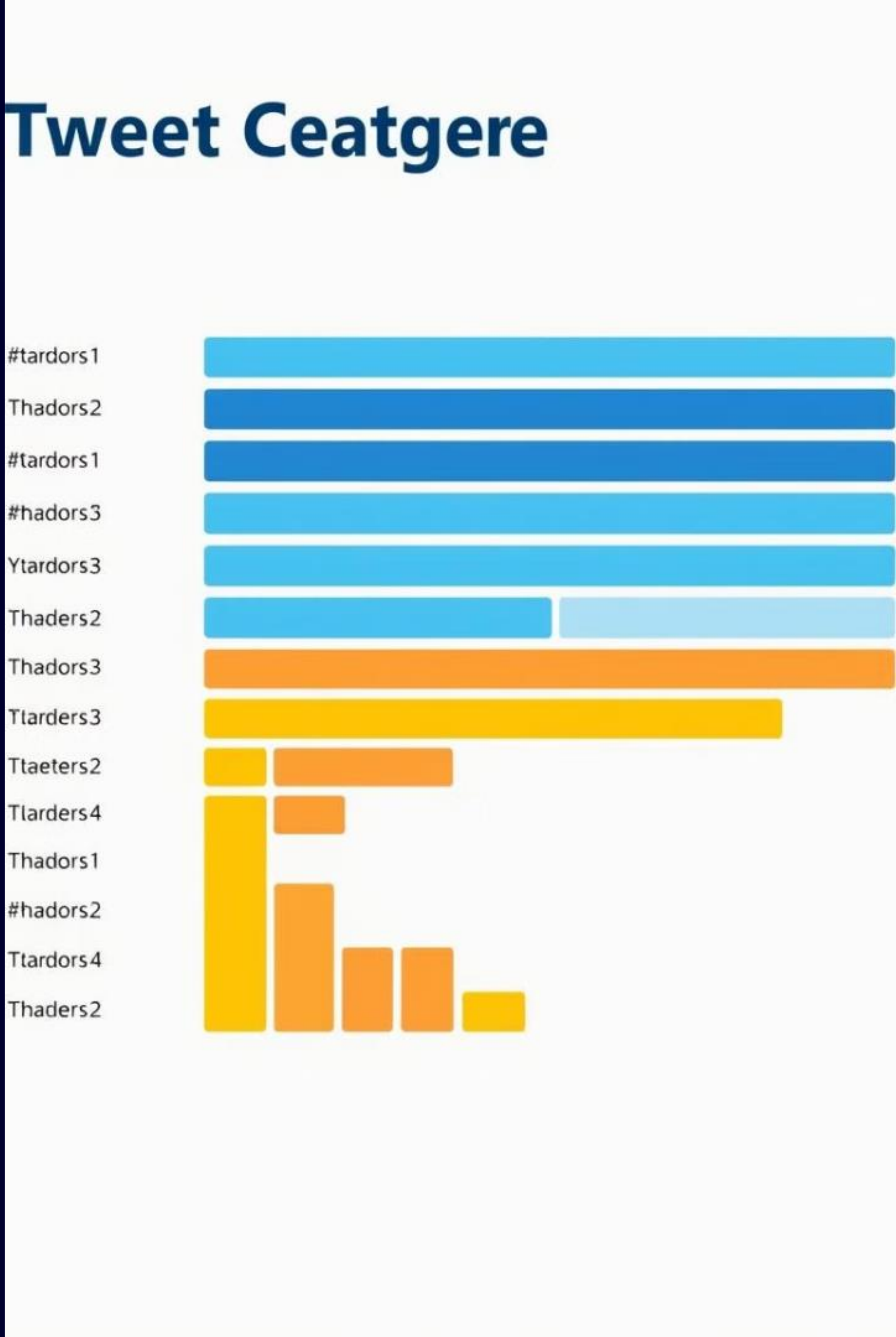
   Reduce exposure to cyberbullying and promote a safer online environment.

3. **Support Mental Health**

   Mitigate psychological impacts like anxiety, depression, and low self-esteem.

Data Source and Structure

Source	Hugging Face's "tdavidson/hate_speech_offensive"
Content	Tweets labeled for hate speech, offensive language, or neither
Columns	count, hate_speech_count, offensive_language_count, neither_count, class, tweet



Key Stakeholders

Social Media Users

Benefit from reduced exposure to offensive language.

Moderation Teams

Need tools to effectively monitor and control comment sections.

Data Scientists

Interested in building and refining cyberbullying detection models.

Executive Leadership

Focused on user retention, satisfaction, and brand reputation.

Project Methodology

1. **Business Understanding**
   Define problem, objectives, and stakeholders for cyberbullying detection strategy.

2. **Data Understanding**
   Collect and explore datasets with labeled comments categorized by severity.

3. **Data Preparation**
   Clean and preprocess text data, handling misspellings, slang, and emojis.

4. **Modeling**
   Use NLP techniques and train neural network models for text classification.

5. **Evaluation**
   Assess model performance using metrics like accuracy for text classification.

Data Analysis Insights

Class Distribution

Class 1 (offensive language) most common with 19,190 tweets. Class 2 (hate speech) has 4,163 tweets. Class 0 (neutral) least frequent with 1,430 tweets.

Tweet Lengths

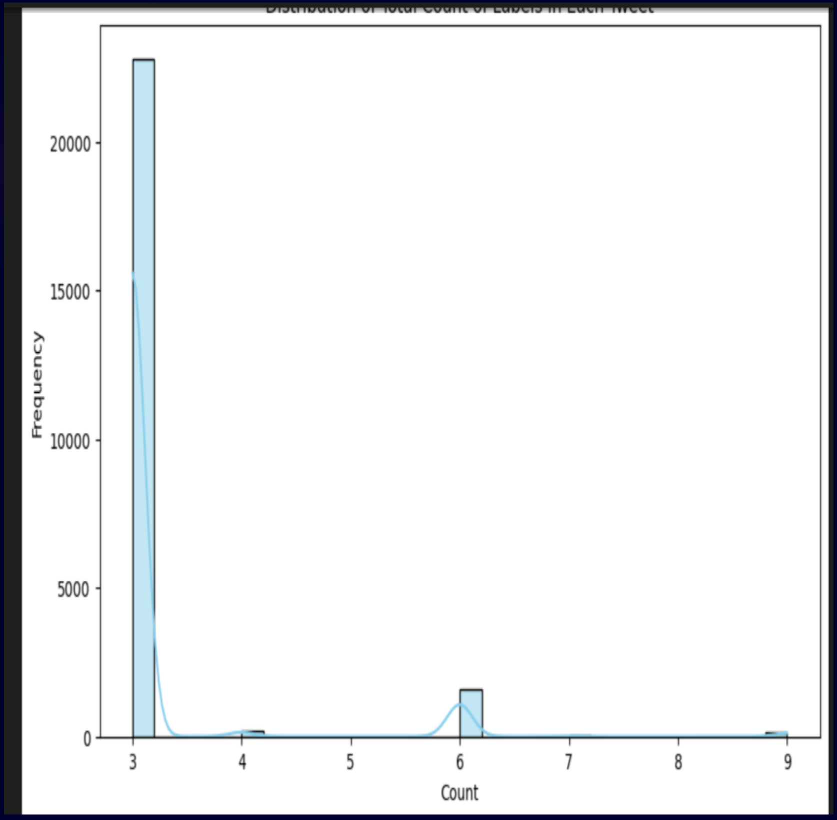
Hate speech tweets generally longer, averaging 95 characters. Offensive language tweets show most diverse range of lengths.

Correlations

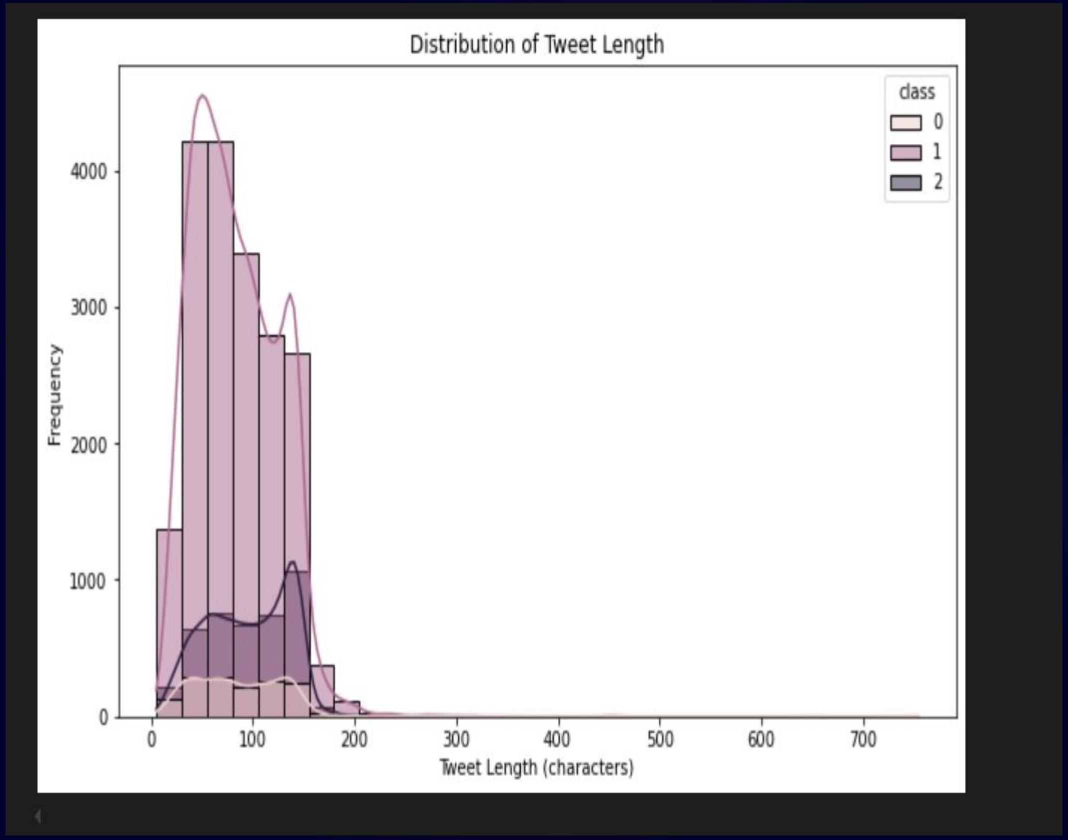
Moderate negative correlation between hate speech and offensive language counts. Strong positive correlation between neutral count and higher class labels.

Data Analysis Insights

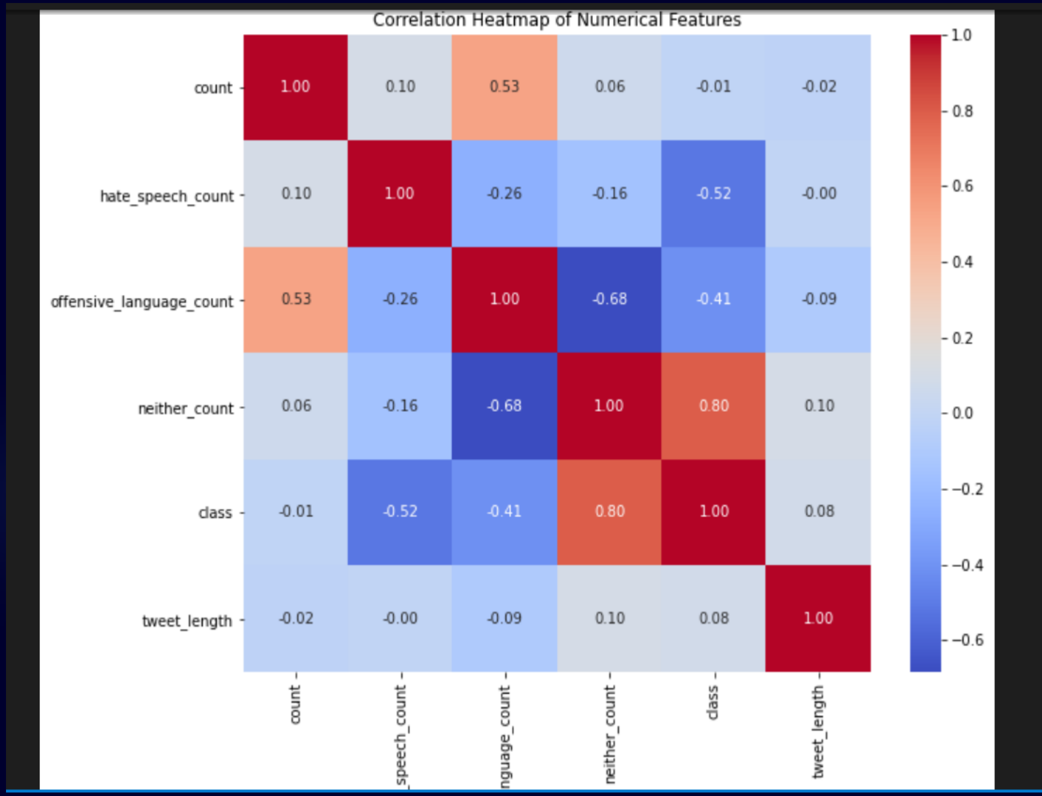
Class Distribution



Tweet Lengths



Correlations



Text Processing

Removal of stop words

Removal of stop words for easier classification of texts as stop words may not have significant meaning to texts. It is also useful to reduce dimensionality.

Handling imbalance

Class_weight is a dictionary where the keys are the class indices and the values are weights assigned to each class. This helps adjust the model's learning process to give more importance to the underrepresented classes in the dataset.

Tokenization & Padding

1. Tokenize the tweet columns to individual words.
2. Convert texts to sequences.
3. Define a max length for truncating the sequences before padding.
4. Padding the train data

Modeling Approach

CNN Base Model

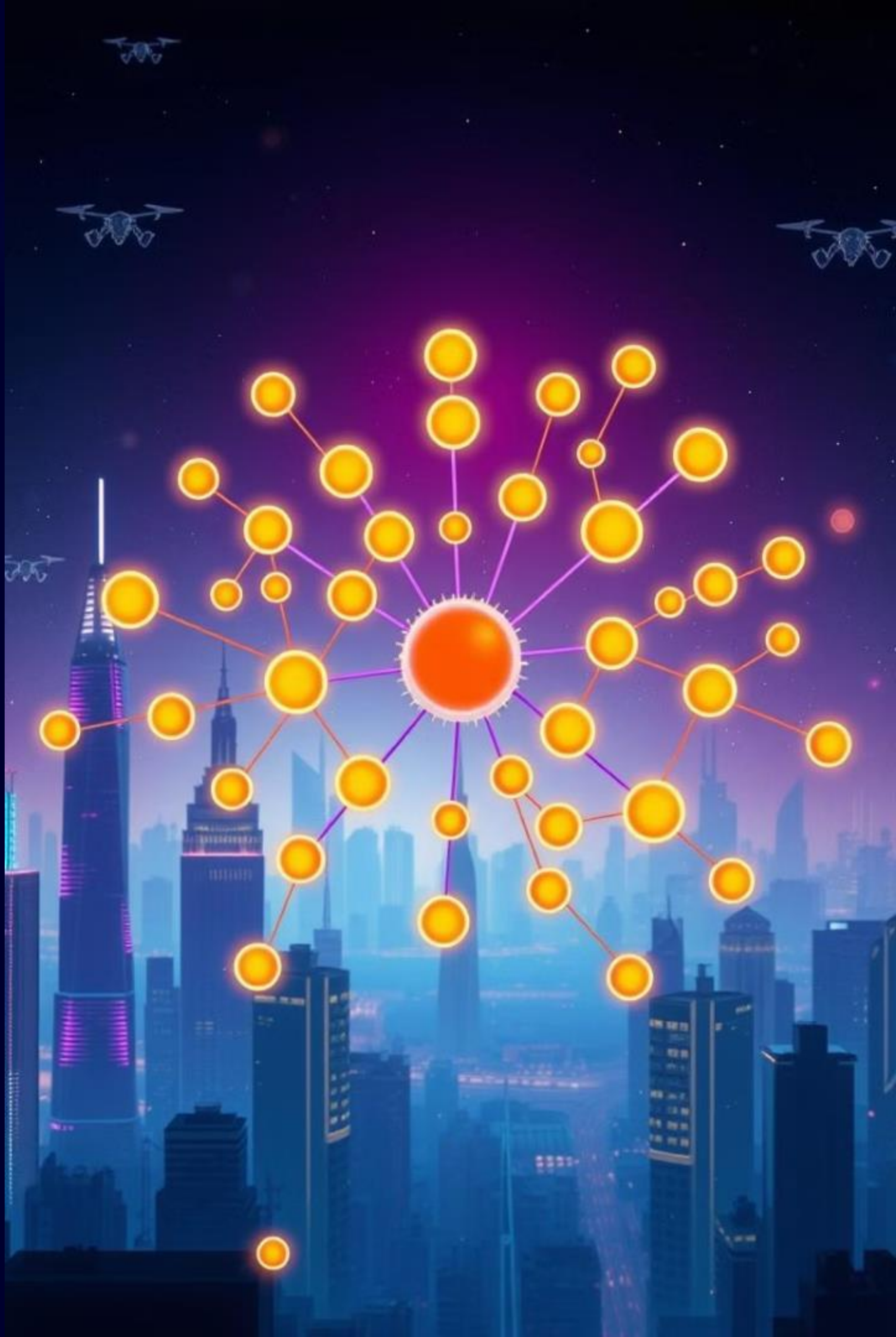
Simple model to test performance on training data. Achieved 86% accuracy.

Global Max Pooling Model

Used to reduce dimensionality and focus on important features. Achieved 68% accuracy.

Bidirectional LSTM Model

Captured sequence context for comprehensive feature representation. Achieved 84% accuracy.



Best Model Selection

1

2

3

CNN base model

High accuracy of 86% compared to the other two models hence selected.

Global max pooling Model

Moderate accuracy but showed signs of underfitting. Not selected.

Bidirectional LSTM Model

Had an accuracy of 84% second highest accuracy but not selected.

Evaluation Approach

Accuracy on the test data

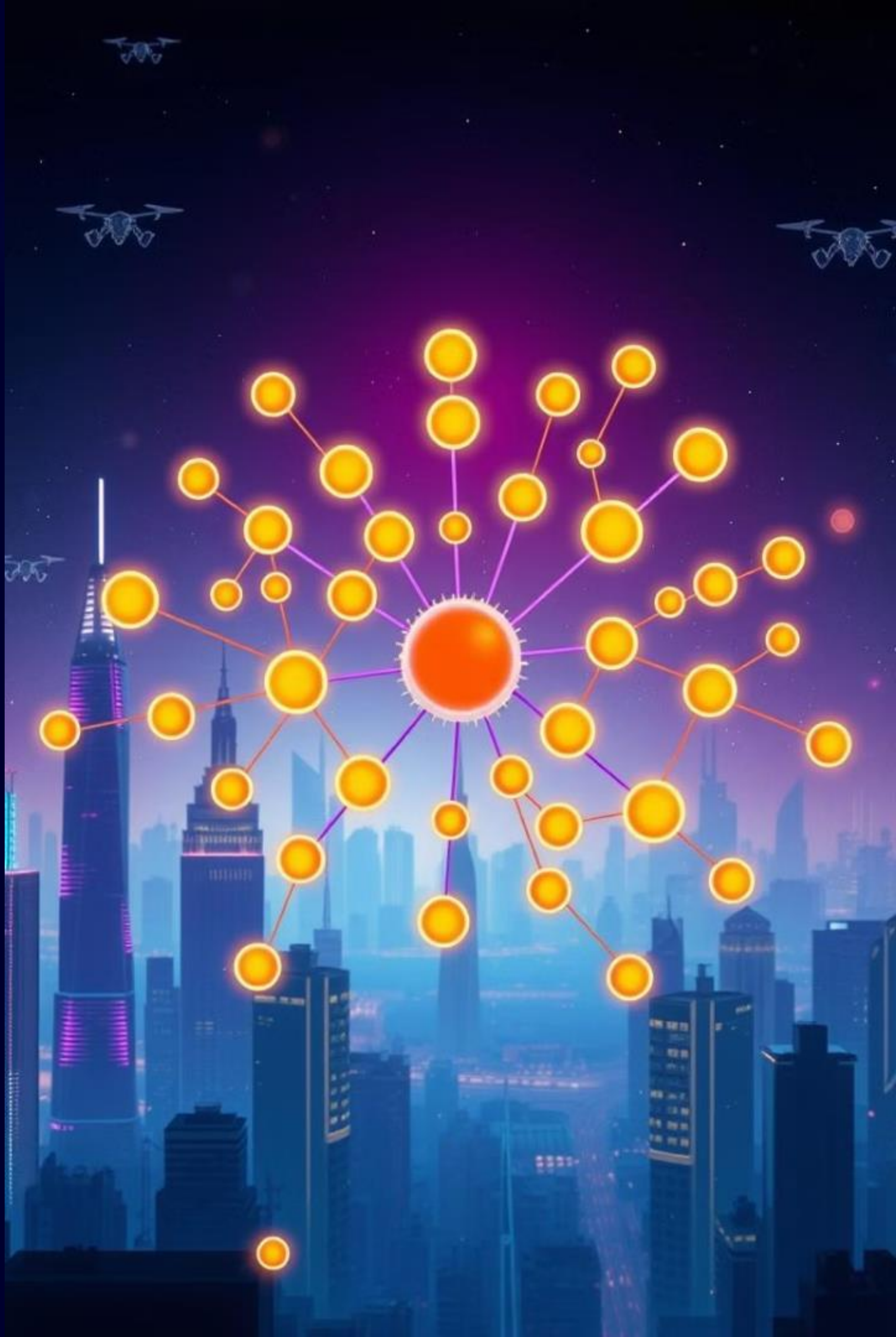
The accuracy obtained from the test data was 72%. This shows that our model is able to predict 72% of test data or unseen data

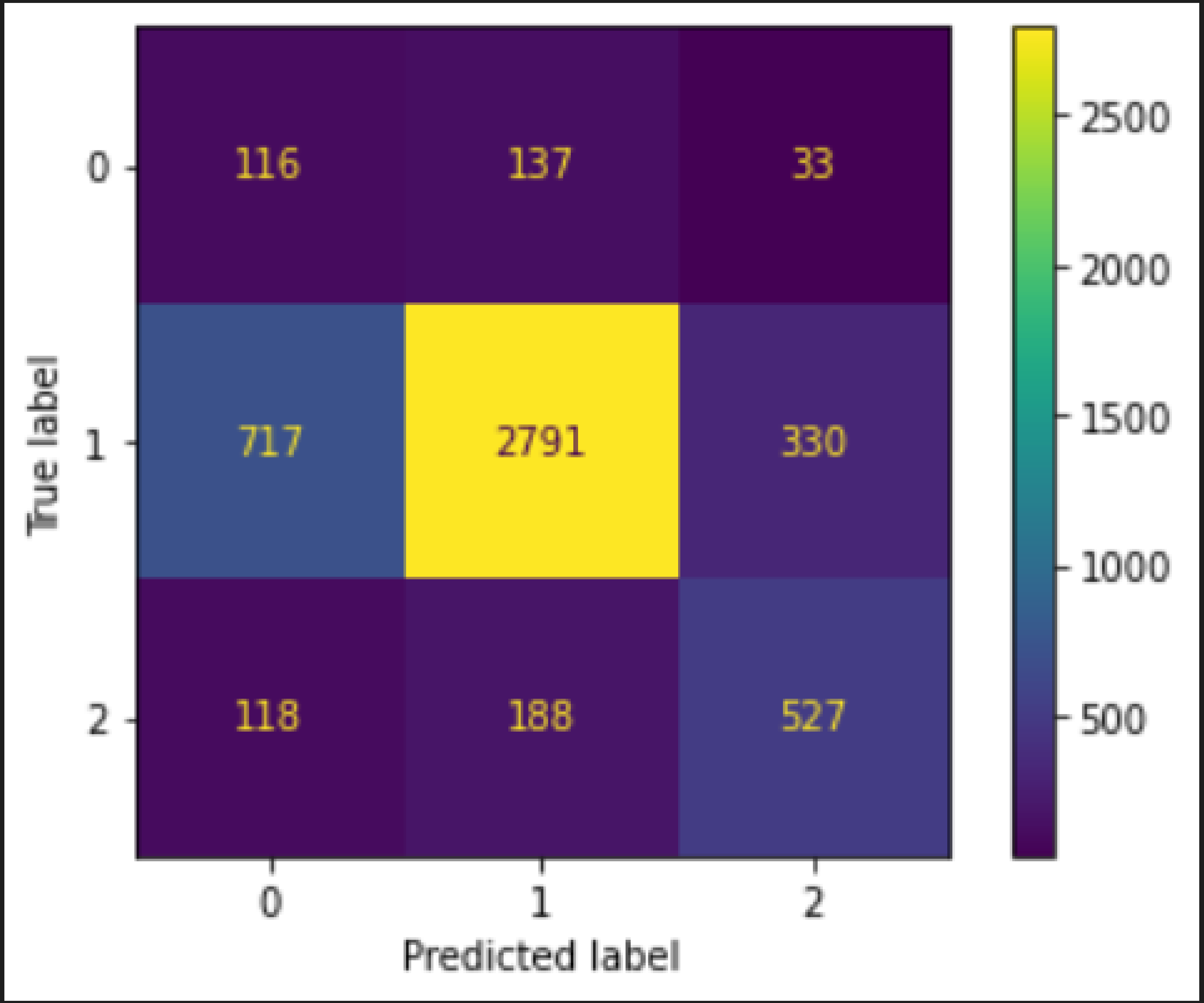
Precision on the test data. data.

The precision obtained from the test data is 54%. This shows that the model can be able to identify instances of positive tweets.

Recall on the test data

The recall obtained was 59%. This shows that the model was able to correctly identify 59% positive instances.




116
137
33
717
2791
330
118
188
527
0
1
2
True label
0
1
2
Predicted label
2500
2000
1500
1000
500

Conclusions

1.

The system achieved high accuracy of 91% being able to detect offensive and hateful language, indicating that the system can reliably identify harmful comments based on the user input in prediction of tweets part. Making it suitable for practical applications.

2.

The use of bidirectional LSTM in the third model significantly improved performance by capturing the context and sequential dependencies of language data.

3.

The system was able to mitigate cyberbullying actions such as offensive language, hate speech which depreciate the mental health of the users which is a growing concern for social media users.

Recommendations

1. Incorporate a user reporting mechanism - To further improve the system, a reporting feature is recommendable to allow users to report the other users who leave negative or hateful comments on their page.

2. Expand the system to other languages - With users coming from different countries and the social platforms been interpreted in different languages, it will be useful if the model is designed further in a way to interpret different languages.