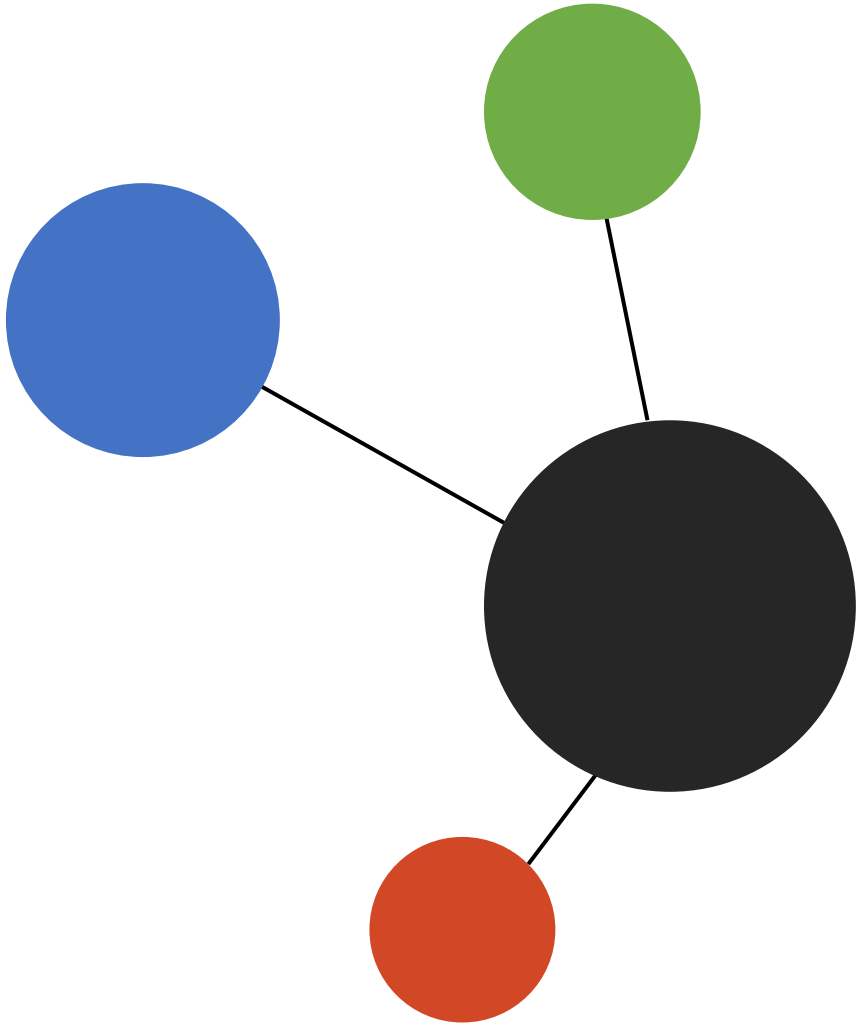
Hate Speech Detection in Movie Reviews



Presented By :
Daniel Suresh (Group 1)

The Problem

- Hate speech in online movie reviews created a toxic environment.
- Discouraged participation and harmed individuals.
- Identifying and moderating hate speech is crucial for a healthy online community.

Solution

1. To develop, a **AI** model to detect hate speech in movie reviews.

2. The model receives **text** as input, **pre-processes** it, and **predicts** whether it is hate speech or not

Dataset Description

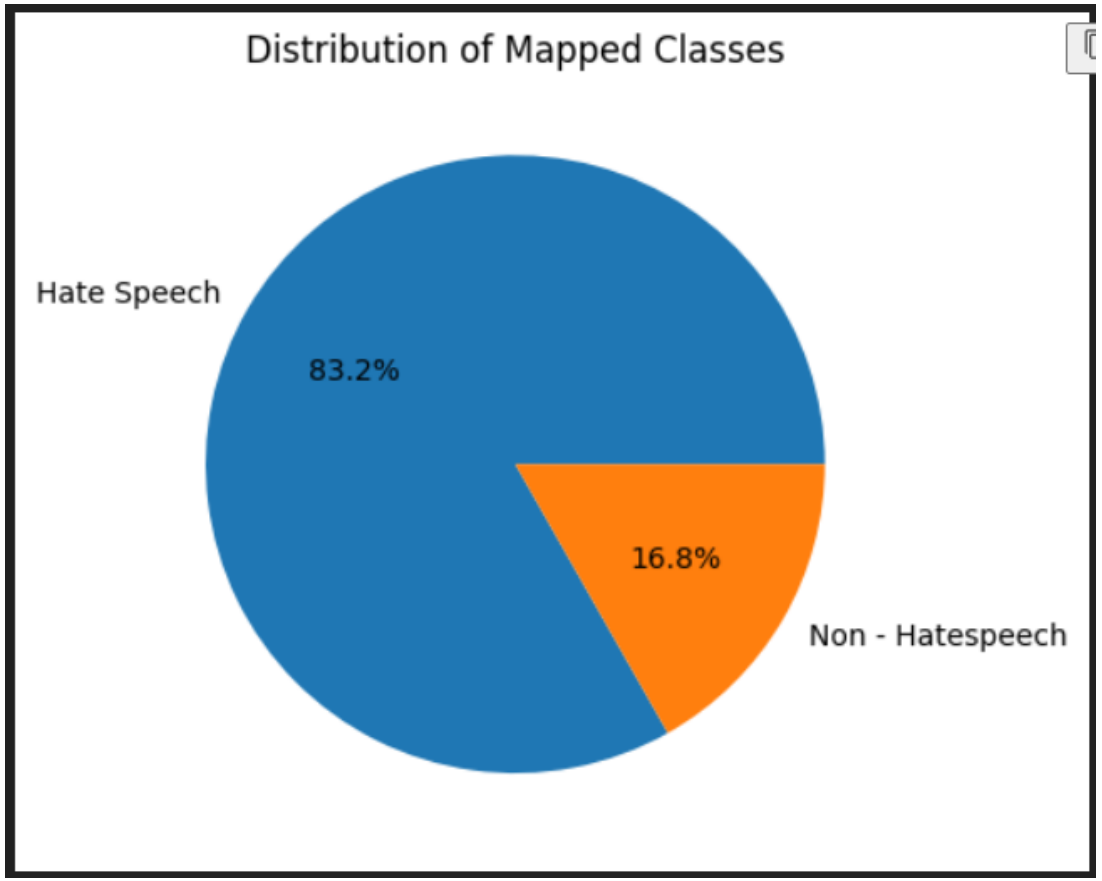
- Used Davidson dataset
- Contains 24k + tweets which is classified as
  - 0 – Hate speech
  - 1 – Offensive language
  - 2 – Neither
- Train and Test Data Split
  - Train Data Size: 19,826 samples
  - Test Data Size: 4,957 samples
  - Split Ratio: 80% train, 20% test

Data Visualization

- To focus more on Hate speech detection, we **mapped** offensive and hate speech to the **same class**:

- Hate Speech (Class 0): 83.2%
- Non- Hate Speech (Class 1): 16.8%

- **Class weights** were used during model training to address the imbalance in the dataset



Data Preprocessing

- Removed usernames, sequences, URLs.
- Converted to lowercase.
- Removed punctuations.
- Replaced words (abbreviations, misspellings).
- Removed stopwords.
- Lemmatization.

Tokenization and Embedding Techniques

- **Tokenization Techniques**
  - Used Keras Tokenizer to convert text into sequences.
- **Embedding Techniques Considered**
  - One-Hot Encoding: Rejected (didn't capture semantic relationships).
  - TF-IDF: Explored but less effective.
  - **GloVe**: Pre-trained embeddings on a large corpus, **Selected** for capturing semantic meanings and relationships.

Modeling

Machine Learning Models

- Logistic Regression: Explored but not selected.
- Random Forest Classifier (RFC): Explored but not selected.
- Support Vector Machine (SVM): Explored but not selected (selected for ml script)

Deep Learning Models

- Convolutional Neural Network (CNN): Explored but not selected.
- Bi-Directional LSTM: Explored but not selected.
- **LSTM with Attention**: Selected for final model.
  - Reason **:**
    - **Best balance of accuracy and efficiency,**
    - **Focuses on important parts of the text.**
    - **Had best f1 score among other models**

Evaluation Metrics

Accuracy: Overall correctness of the model.

F1 Score: Balanced precision and recall.

Confusion Matrix: Detailed breakdown of model performance.

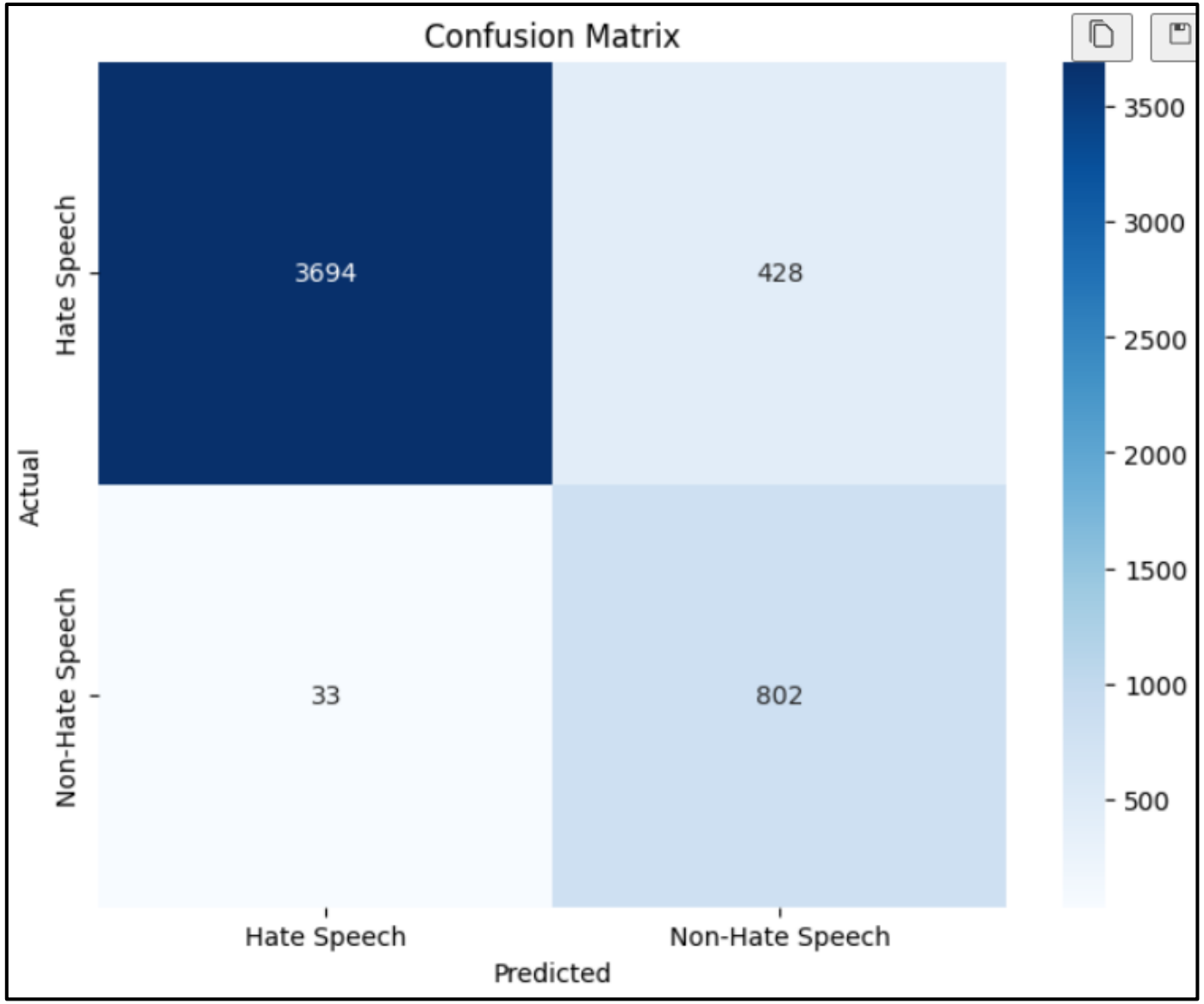
Best Scores Achieved :

Deep Learning (LSTM with Attention)

Accuracy : **90.7 %**

F1 Score : **91.35 %**

Confusion Matrix



Thanks...!