HATE SPEECH DETECTION ACROSS CODE MIXED INDO-ARYAN LANGUAGES

SOURUDRA NAG, SOURIKTA NAG, POULOMI BANERJEE

Problem Statement

Languages are a key part of communication and culture, especially in multilingual regions like South Asia. Hindi, spoken by over 609 million people, and Bengali, spoken by 272 million, are among the most widely used languages in the world. With the rise of online platforms, code-mixed forms like Hinglish (Hindi-English) and Banglish (Bengali-English) have become popular, especially among younger users. However, this has also led to a rise in hate speech online. Detecting hate speech in these mixed languages is challenging due to their informal and evolving nature. This project focuses on creating methods to detect hate speech in Hinglish and Banglish, aiming to make online spaces safer and support multilingual NLP research.

Importance and Applications

1.Enhancing Digital Safety: The project helps reduce hate speech online, creating safer and more inclusive digital spaces, especially in multilingual regions like South Asia.

2.Advancing Multilingual NLP: By addressing the challenges of code-mixed languages, it contributes to improving natural language processing models for under-resourced and mixed linguistic data.

3.Supporting Social Media Platforms: The models can assist platforms like Facebook and YouTube in moderating hate speech, ensuring compliance with community guidelines and promoting positive interactions.

4.Policy and Law Enforcement: Governments and organizations can use this work to enforce hate speech regulations, contributing to societal harmony and digital accountability.

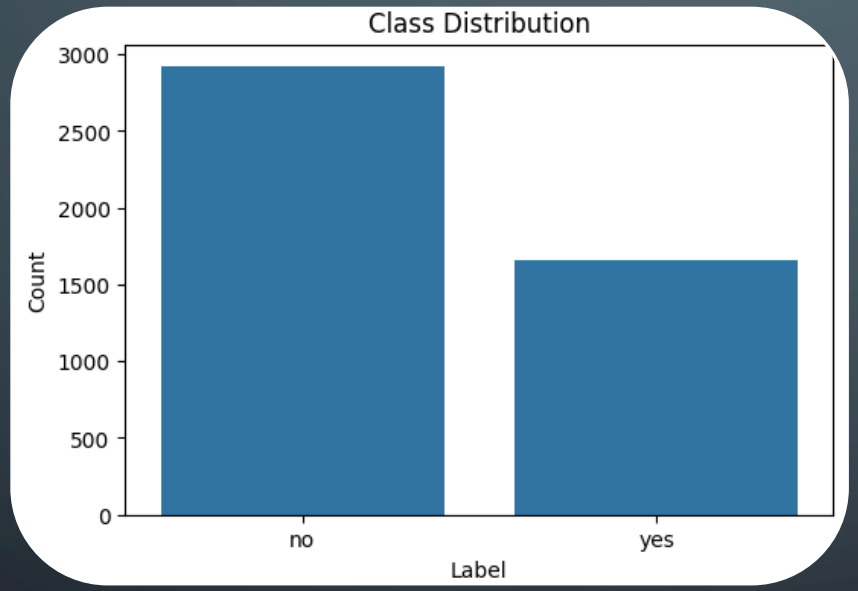
5.Expanding Research Opportunities: The findings lay the groundwork for future studies in hate speech detection across other code-mixed or multilingual settings, advancing global NLP capabilities.

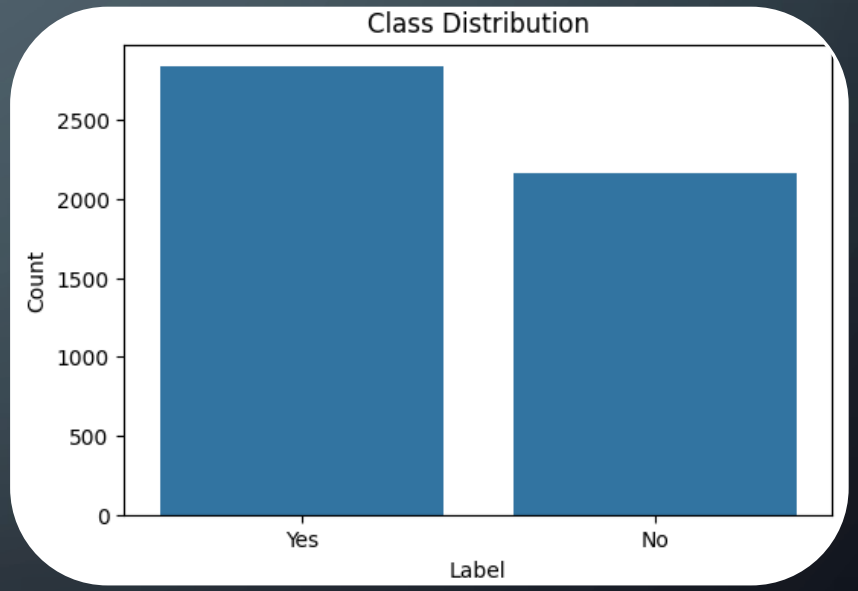
Related Studies

	Name	Year	Author	Approach	Results
1.	Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages using Machine Learning Models	December 2022	Gunjan Kumar, Jyoti Prakash Singh	The study implemented seven ML classifiers and a 1D-CNN model to classify tweets into NOT and HOF classes. Word-level and character-level features were extracted using tf-idf vectorization. The 1D-CNN used pre-trained GloVe embeddings with a multi-layer architecture to extract vital features for classification.	Logistic Regression (LR) achieved the highest accuracy of 0.72. CNN underperformed compared to ML classifiers, achieving accuracies of 0.69 performance using 300-dimensional GloVe embeddings.
2.	Interpretable Multi Labeled Bengali Toxic Comments Classification using Deep Learning	February 2023	Tanveer Ahmed Belal; G. M. Shahariar; Md. Hasanul Kabir	A deep learning pipeline for Bengali toxic comment classification, using LSTM with BERT embeddings for binary toxicity detection and CNN-BiLSTM with attention for categorizing toxic comments into six types. LIME ensures interpretability.	The binary classification model achieved 89.42% accuracy, while the multi-label CNN-BiLSTM classifier obtained 78.92% accuracy and a weighted F1-score of 0.86.
3.	HateDetector: Multilingual technique for the analysis and detection of online hate speech in social network	August 2023	Anjum, Rahul Katarya	Utilizing BERT with a Multi-Layer Perceptron for code conversion and similarity checks to generate vector representations. The approach incorporates the Profanity Check Technique with a ReLU activation function and logistic regression to classify tweets into hate speech.	The proposed technique achieved a classification accuracy of 97.9%, outperforming state-of-the-art models in detecting hate speech from multilingual and complex texts.

Data Collection

1. Hinglish Dataset: The dataset contains 4579 rows and 2 columns: **Comment** and **Hate**. Data Link
2. Banglish Dataset: It contains 5000 rows and 4 columns: **SL.**, **Comment**, **Hate**, and **Type**. Data Link





Hinglish Dataset

Banglish Dataset

Approach

Data Preprocessing

1. Clean the text by converting it to lowercase, removing special characters, and eliminating URLs.
2. Convert labels into binary values (hate or non-hate) for classification.
3. Use BERT's tokenizer to break text into sub-word units effectively.

Tokenization and Attention Mask

1. Convert each token into a numerical ID to represent the input text.
2. Apply an attention mask to identify meaningful tokens (1) and ignore padding tokens (0) during training.

Model Architecture

1. Utilize **BERT-base-multilingual-cased** to extract contextual embeddings from the input text using pre-trained transformer layers.
2. Add a fully connected Feed Forward Neural Network classification layer with a sigmoid activation function to categorize the text into hate speech or non-hate speech.

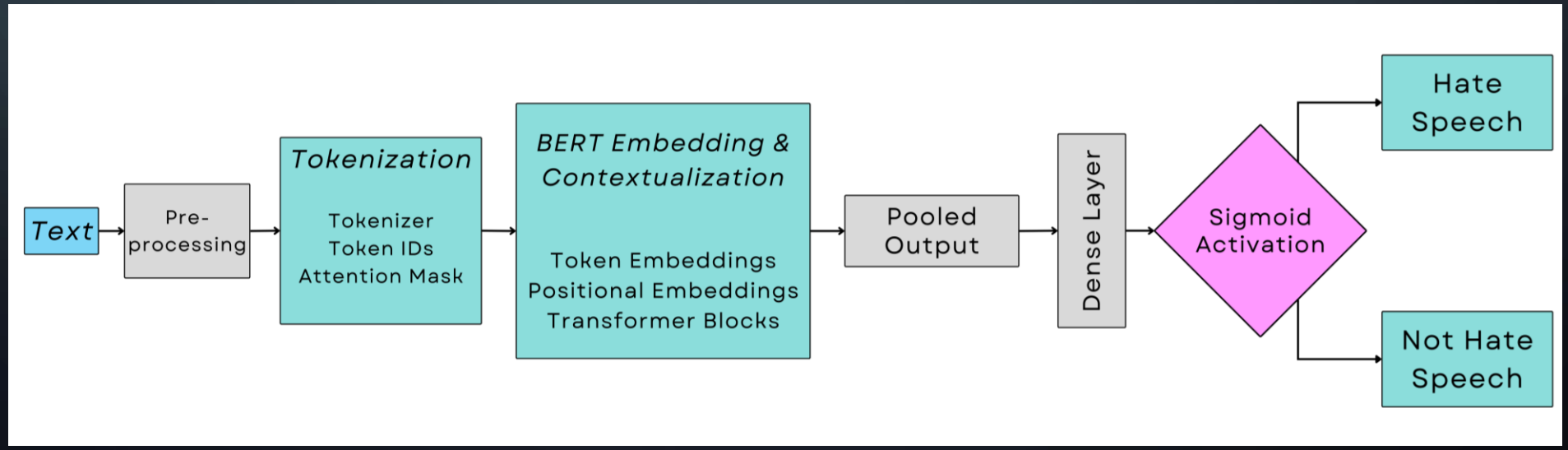
Approach

Training Process

1. Use **binary cross-entropy** as the loss function to measure the difference between predictions and actual labels.
2. Optimize the model using the Adam optimizer with a learning rate of 1×10^{-5}.
3. Monitor the training process with validation metrics such as accuracy and loss.

Evaluation Metrics

1. Measure the model's overall effectiveness using **accuracy** as the primary metric.
2. Generate a classification report detailing precision, recall, and F1-score for both hate and non-hate categories.



Current Progress

To date, Feed Forward Neural Network model showed best result:

- **Hinglish Model**: Achieved an accuracy of **72%**.
- **Banglish Model**: Achieved an accuracy of **75%**.

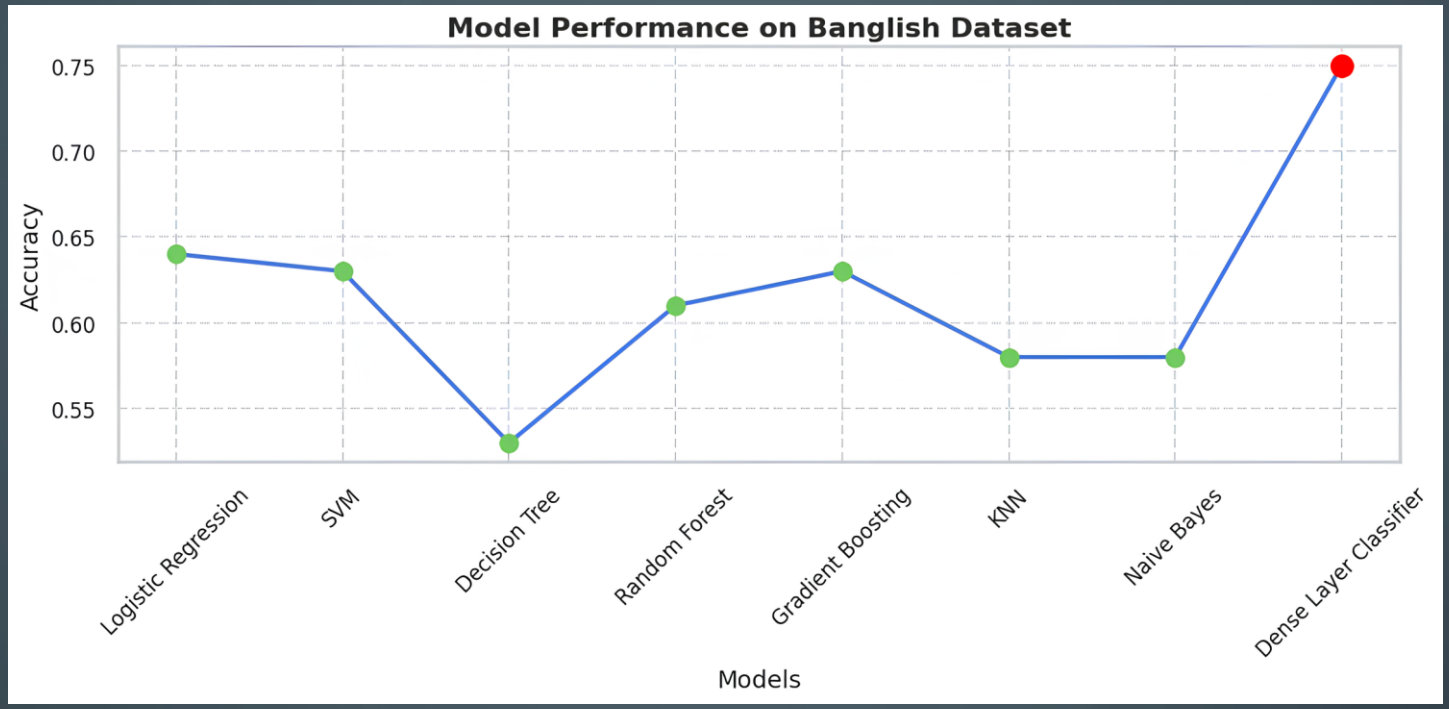
It is built upon the bert-base-multilingual-cased architecture, a pre-trained model that supports over 100 languages. It is well-suited for handling the challenges of code-mixed text.

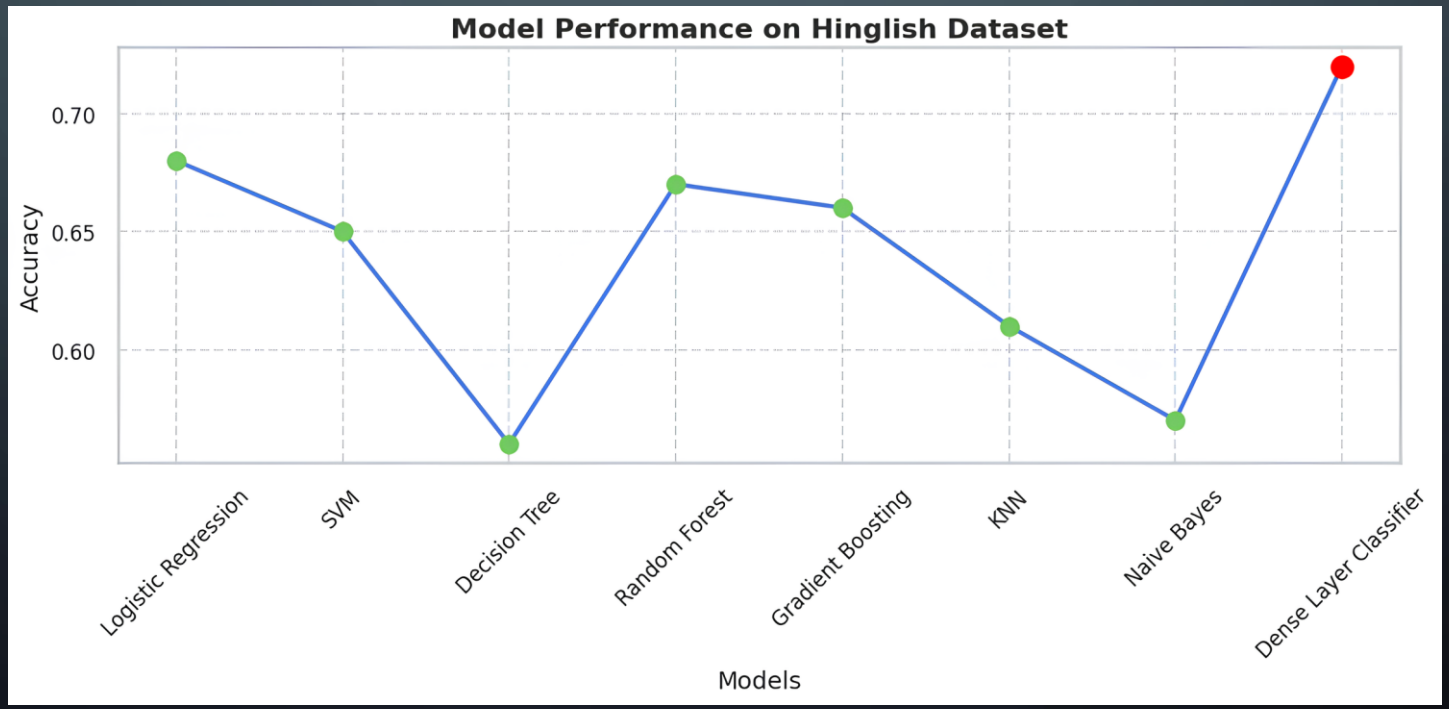
Comparison of Results for Banglish

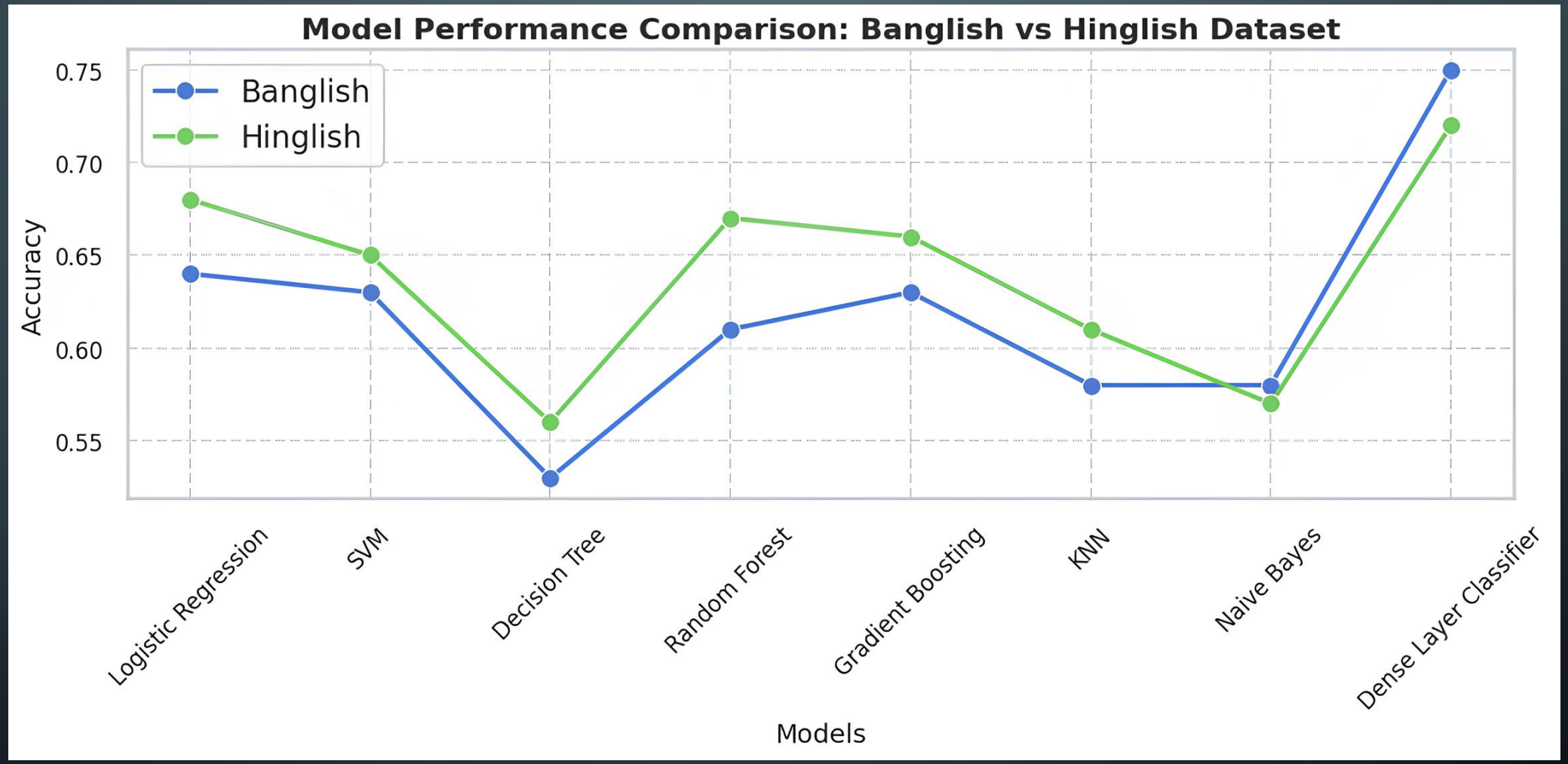
Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.64	0.66	0.72	0.69
Support Vector Machine	0.63	0.65	0.71	0.68
Decision Tree	0.53	0.57	0.58	0.58
Random Forest	0.61	0.62	0.79	0.69
Gradient Boosting	0.63	0.64	0.76	0.69
K-nearest Neighbor	0.58	0.60	0.68	0.64
Naïve Bayes	0.58	0.63	0.58	0.60
Dense Layer Classifier	**0.75**	0.75	0.82	0.78

Comparison of Results for Hinglish

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.68	0.52	0.43	0.47
Support Vector Machine	0.65	0.48	0.43	0.45
Decision Tree	0.56	0.35	0.38	0.36
Random Forest	0.67	0.53	0.13	0.21
Gradient Boosting	0.66	0.49	0.23	0.31
K-nearest Neighbor	0.61	0.41	0.38	0.39
Naïve Bayes	0.57	0.39	0.53	0.45
Dense Layer Classifier	**0.72**	0.63	0.40	0.49


Model Performance on Banglish Dataset
Accuracy
0.75
0.70
0.65
0.60
0.55
Logistic Regression
SVM
Decision Tree
Random Forest
Gradient Boosting
KNN
Naive Bayes
Dense Layer Classifier
Models


Model Performance on Hinglish Dataset
Accuracy
0.70
0.65
0.60
Logistic Regression
SVM
Decision Tree
Random Forest
Gradient Boosting
KNN
Naive Bayes
Dense Layer Classifier
Models

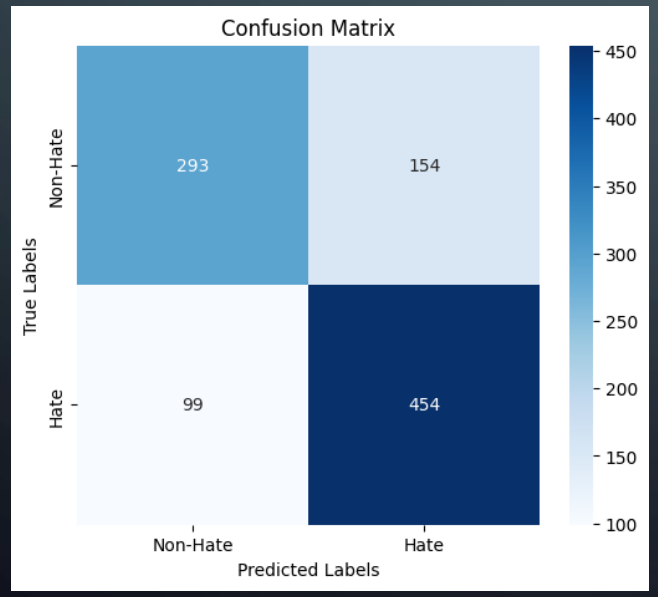

Model Performance Comparison: Banglish vs Hinglish Dataset
Banglish
Hinglish
Accuracy
0.75
0.70
0.65
0.60
0.55
Logistic Regression
SVM
Decision Tree
Random Forest
Gradient Boosting
KNN
Naive Bayes
Dense Layer Classifier
Models

Comparison Of Best Results

BANGLISH

Classification Report:

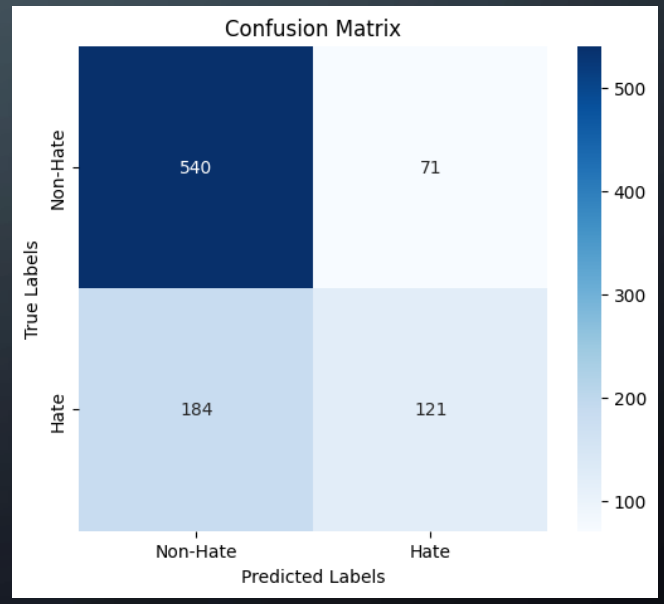
	precision	recall	f1-score	support
Non-Hate	0.75	0.66	0.70	447
Hate	0.75	0.82	0.78	553
accuracy			0.75	1000
macro avg	0.75	0.74	0.74	1000
weighted avg	0.75	0.75	0.74	1000



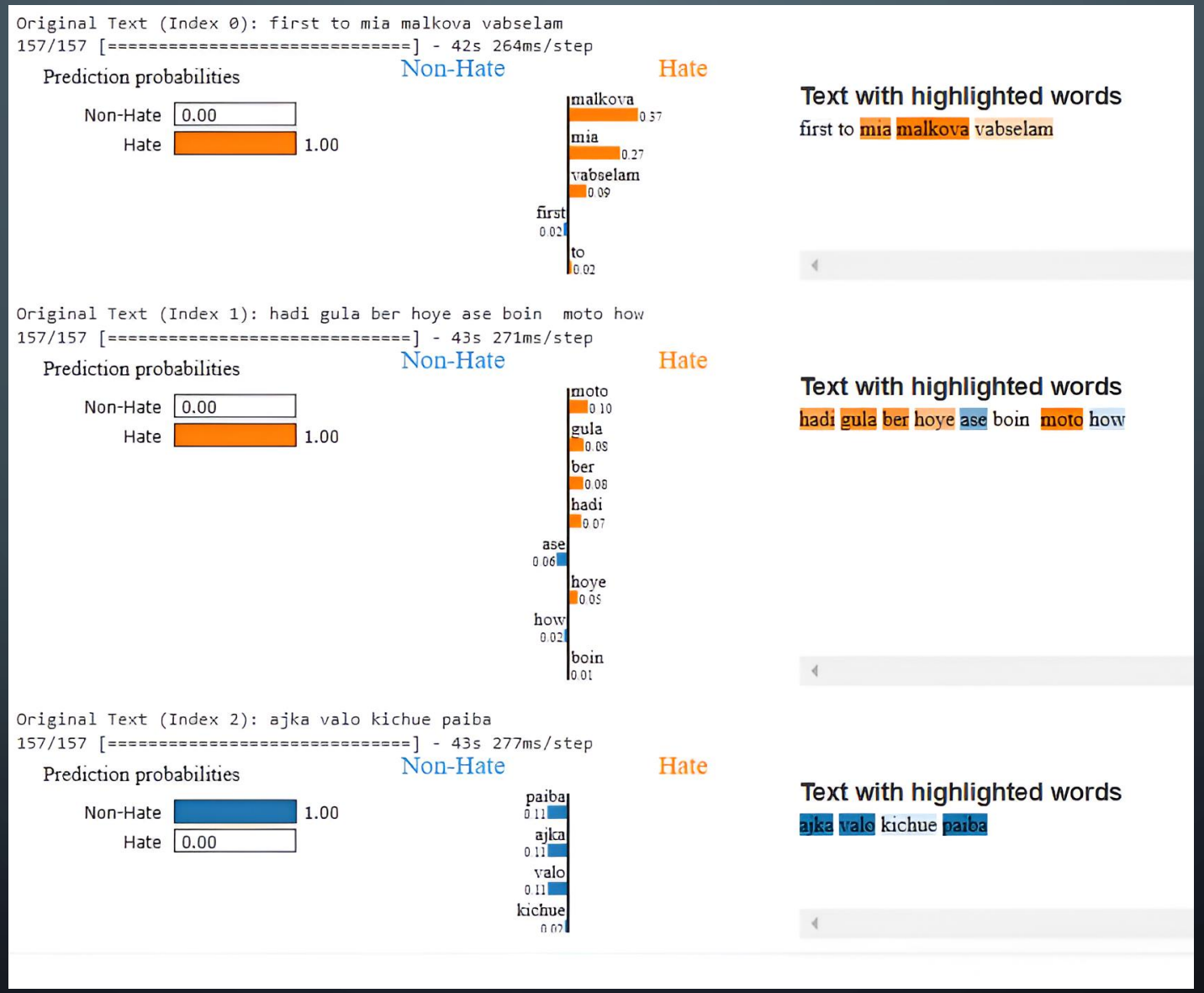
HINGLISH

Classification Report:

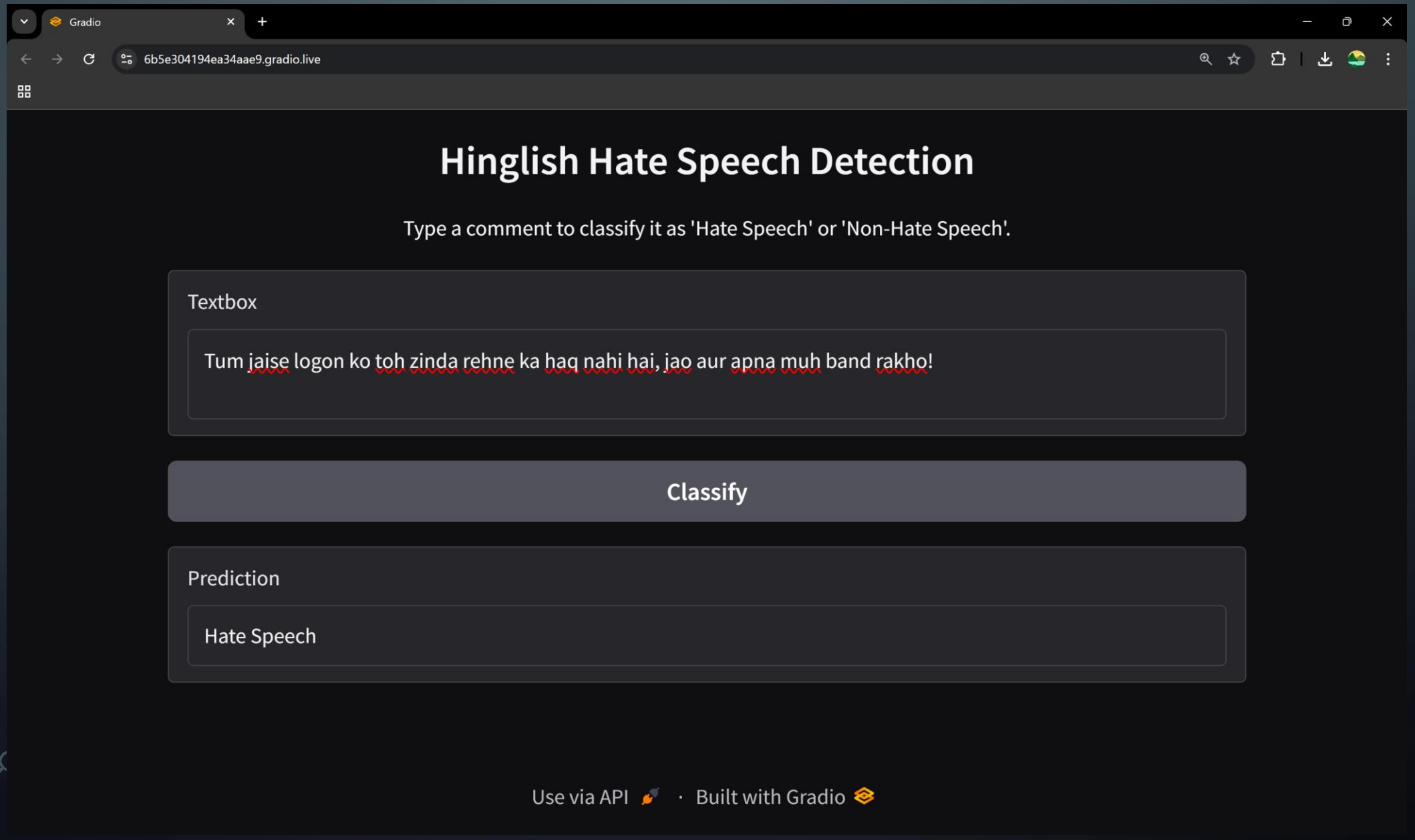
	precision	recall	f1-score	support
Non-Hate	0.75	0.88	0.81	611
Hate	0.63	0.40	0.49	305
accuracy			0.72	916
macro avg	0.69	0.64	0.65	916
weighted avg	0.71	0.72	0.70	916



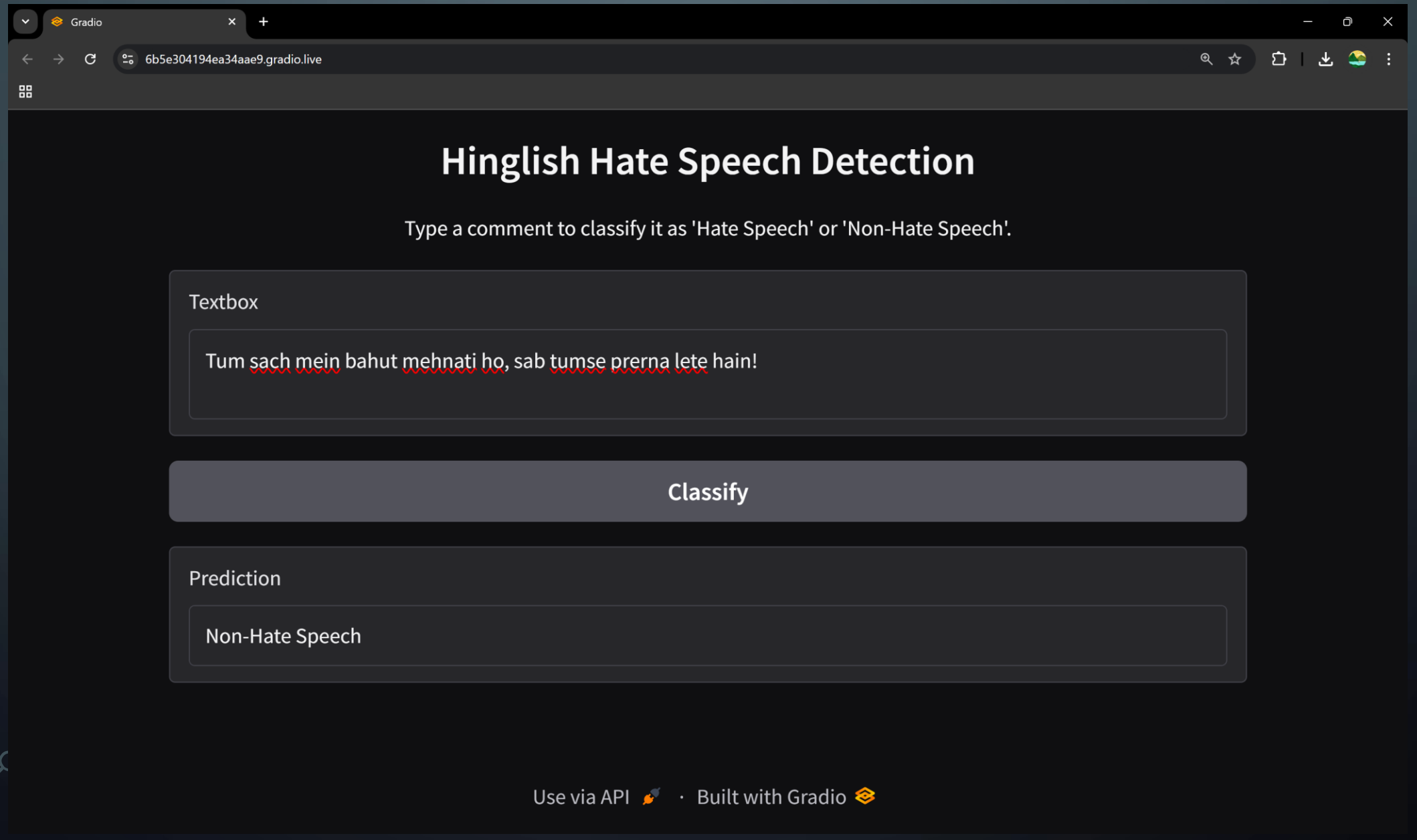
Model Interpretability with LIME



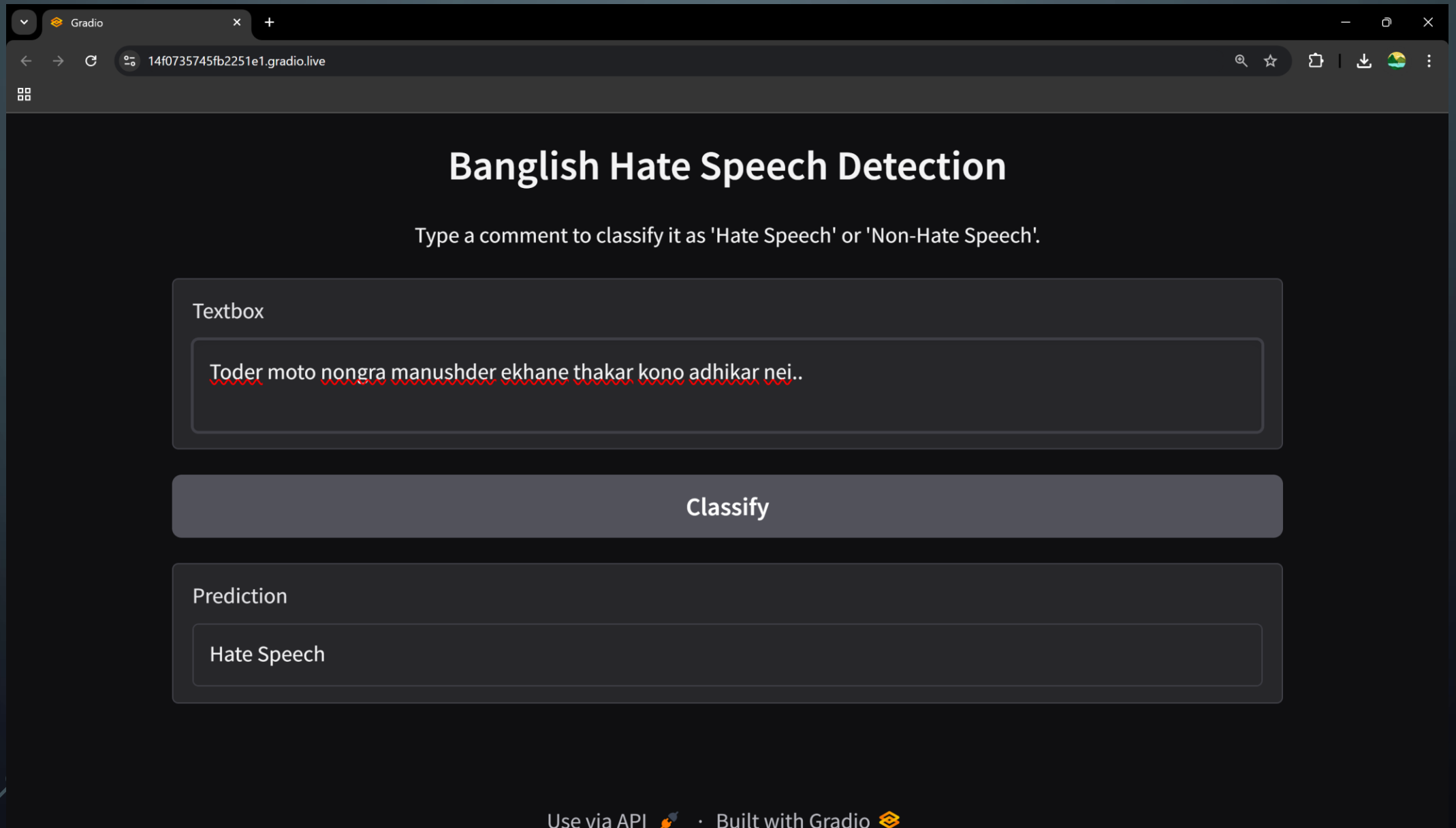
Live Gradio Demo (Hinglish Hate Speech)



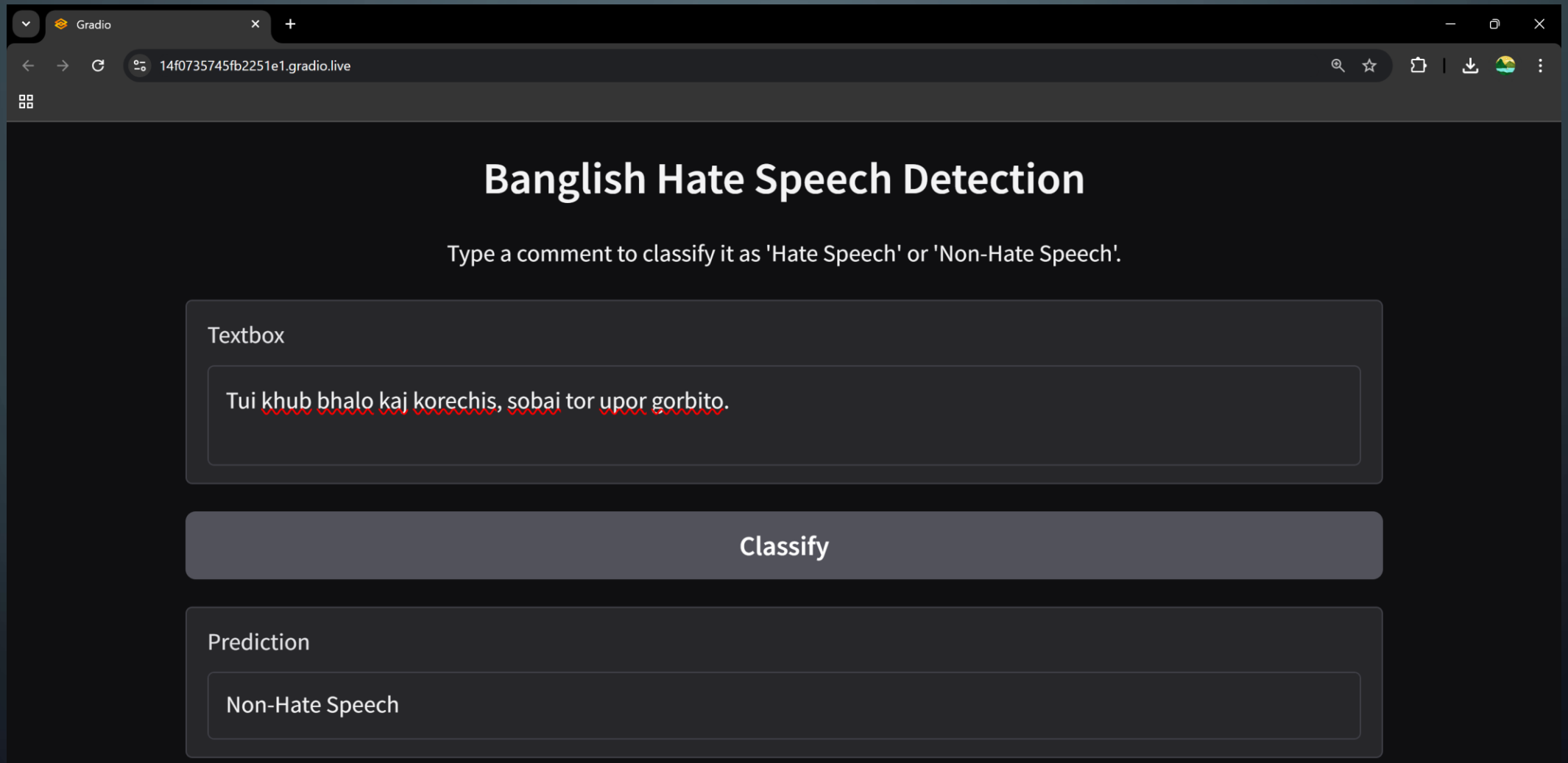
Live Gradio Demo (Hinglish Non-Hate Speech)



Live Gradio Demo (Banglish Hate Speech)



Live Gradio Demo (Banglish Non-Hate Speech)



Preliminary Insights

1. The dense layer classifier (FFNN) outperformed traditional machine learning models, achieving an accuracy of over 70% compared to around 60% for the latter.

2.The Banglish model has demonstrated superior performance compared to the Hinglish model. This can be attributed to the higher quality and consistency of the Banglish dataset, which is less noisy and exhibits better linguistic structure than the Hinglish dataset.

3. Moreover, handling code-mixed data requires effective preprocessing and tokenization techniques to capture the intricate linguistic patterns present in such text.

Future Directions

- Hyperparameter tuning will be performed on the existing models to enhance performance.

- Plan is to experiment with variations of BERT models, such as **LSTM + BERT embedding**, **MConv-LSTM + BERT embedding**, etc, to explore alternative architectures and improve results.

- A comparative analysis of different models will be conducted to refine the approach and optimize detection capabilities.

Conclusion

The project developed hate speech detection models for Hinglish and Banglish using BERT-based architectures, achieving **72%** and **75%** accuracy, respectively. While Hinglish has seen some research, Banglish remains largely unexplored despite being widely used by millions. Additionally, such cross-language comparisons are rare, making this work a significant step in addressing underrepresented languages in multilingual NLP. These efforts contribute to making social media platforms safer for speakers of all languages.

Thank You