Hate Speech Detection Using Machine Learning

Presented To:
Md. Mynuddin
Assistant Professor,
Department Of CSE
Presented By: Enforces

Introduction

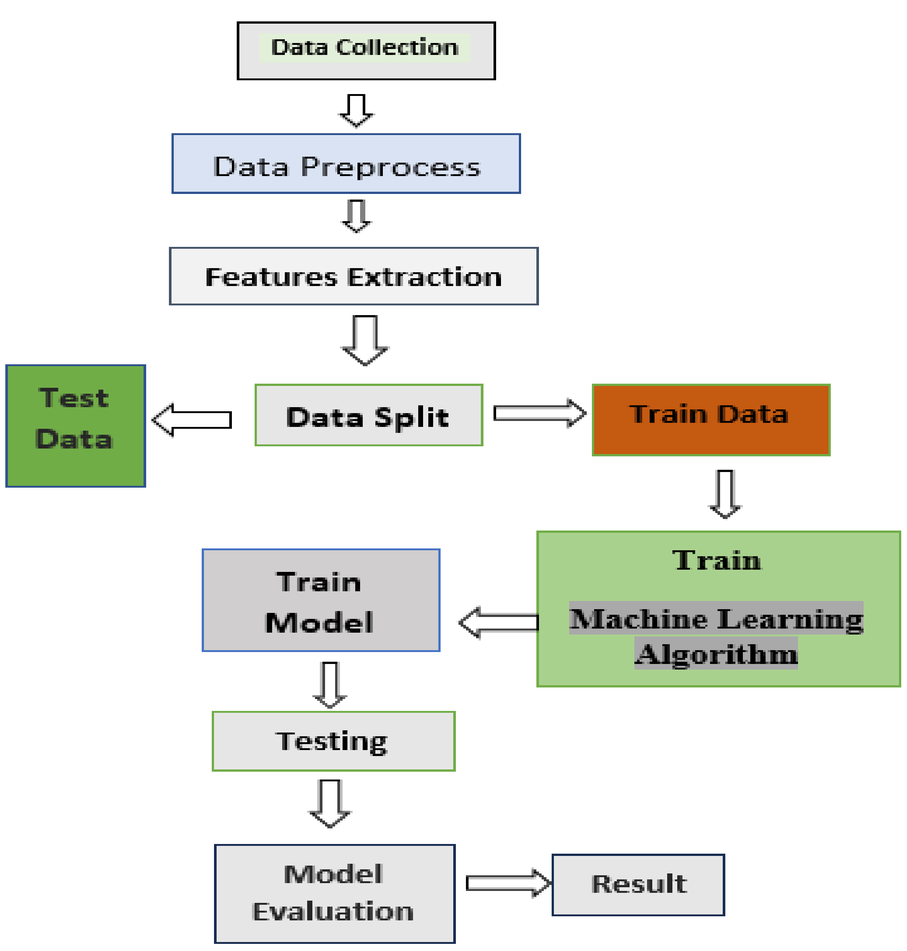
Hate speech is a major concern online, leading to cyberbullying and misinformation. This project uses **Random Forest, Naïve Bayes, and Logistic Regression** to classify text as hate or non-hate speech, aiming to improve detection accuracy and promote safer digital spaces.

Objective:

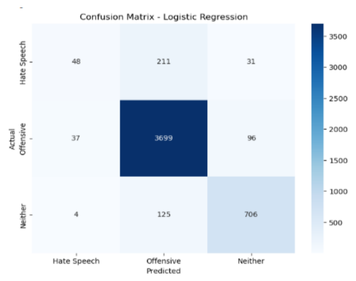
1. Apply machine learning models.
2. Enhance detection accuracy
3. Evaluate model performance

Methodology

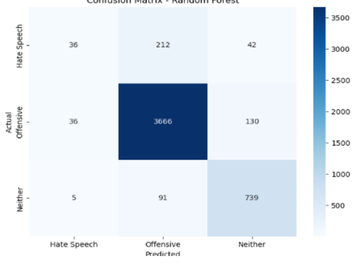
1. Data Collection
2. Preprocessing
3. Feature Extraction
4. Model Training (Random Forest, Naïve Bayes, and Logistic Regression) for classification.
5. Evaluation accuracy, precision, recall, and F1-score.



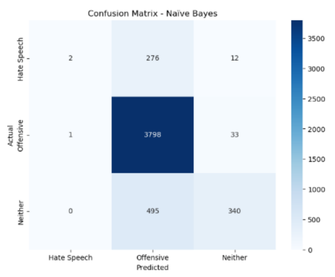
Result



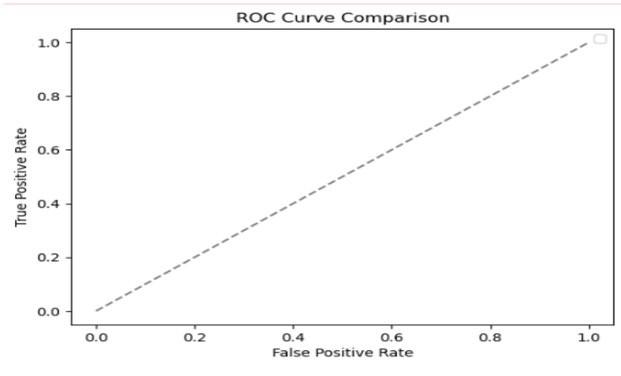

```
Logistic Regression Accuracy: 0.8983
Classification Report:
              precision    recall  f1-score   support

           0       0.54      0.17      0.25       290
           1       0.92      0.97      0.94      3832
           2       0.85      0.85      0.85       835

    accuracy                           0.90      4957
   macro avg       0.77      0.66      0.68      4957
weighted avg       0.88      0.90      0.88      4957
```



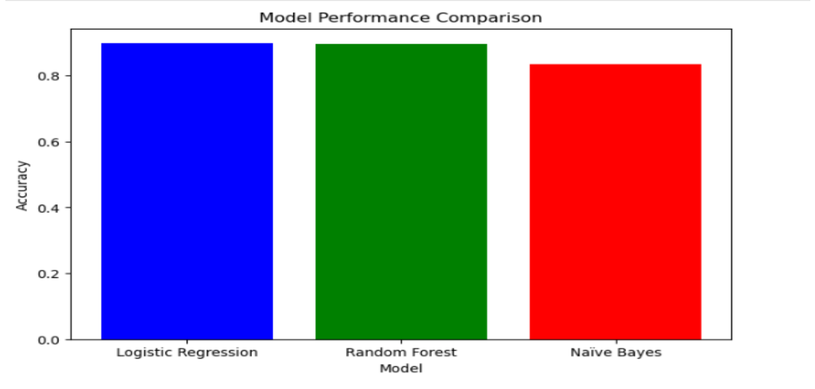

```
Random Forest Accuracy: 0.8959
Classification Report:
              precision    recall  f1-score   support

           0       0.47      0.12      0.20       290
           1       0.92      0.96      0.94      3832
           2       0.81      0.89      0.85       835

    accuracy                           0.90      4957
   macro avg       0.73      0.66      0.66      4957
weighted avg       0.88      0.90      0.88      4957
```




```
Naïve Bayes Accuracy: 0.8352
Classification Report:
              precision    recall  f1-score   support

           0       0.67      0.01      0.01       290
           1       0.83      0.99      0.90      3832
           2       0.88      0.41      0.56       835

    accuracy                           0.84      4957
   macro avg       0.79      0.47      0.49      4957
weighted avg       0.83      0.84      0.79      4957
```



Evaluation



Random Forest and Logistic Regression achieved the highest accuracy

Conclusion

The models help automate content moderation, contributing to safer online spaces.

Future Work

Exploring **deep learning models** and real-time detection for improved accuracy and scalability.