Exploratory Data Analysis

Hate Speech Detection using Transformers

Nov. 2023

Agenda

- Problem Description
- Project Plan
- Data Understanding
- Data Cleansing and Transformation
- Model Selection
- Model Building & Training
- Model Performance Evaluation
- Model Test
- LSTM Model Building & Training
- Model Deployment
- Model Inference
- Conclusion and Limitations
- Future Work

Problem Description

- Rise of hate speech on social media
- Importance of addressing hate speech for a safe online environment
- Derived from a sentiment classification task on labeled Twitter data

Project Plan

- Understanding the problem with a focus on sentiment classification
- Extensive data cleaning and normalization
- Leveraging Transformer-based architecture for expression learning
- Development of a deep learning model for hate speech detection
- Thorough performance evaluation and reporting
- Deployment of models for real-time inference

Data Understanding

- Hate speech dataset from Twitter with attributes: label and text_format
- Challenges addressed: noise in text, data cleaning strategies
- No missing values, outliers, or class imbalance

Data Cleansing and Transformation

- Implementation of text cleaning techniques
- Tokenization, lowercasing, and handling contractions
- Removal of special characters and punctuation
- Word frequency analysis and word clouds for cleaned data

Model Selection

- Consideration of various classifiers:
    1. Multinomial Naive Bayes
    2. Random Forest
    3. Decision Tree
    4. Logistic Regression
    5. SVM
    6. KNN
    7. SGD
    8. XGBoost
- Feature extraction technique: TF-IDF

Model Building & Training

- Data preprocessing steps, including oversampling for class balance
- Training and evaluation of multiple machine learning models
- In-depth evaluation metrics: accuracy, precision, recall, F1 score, ROC-AUC

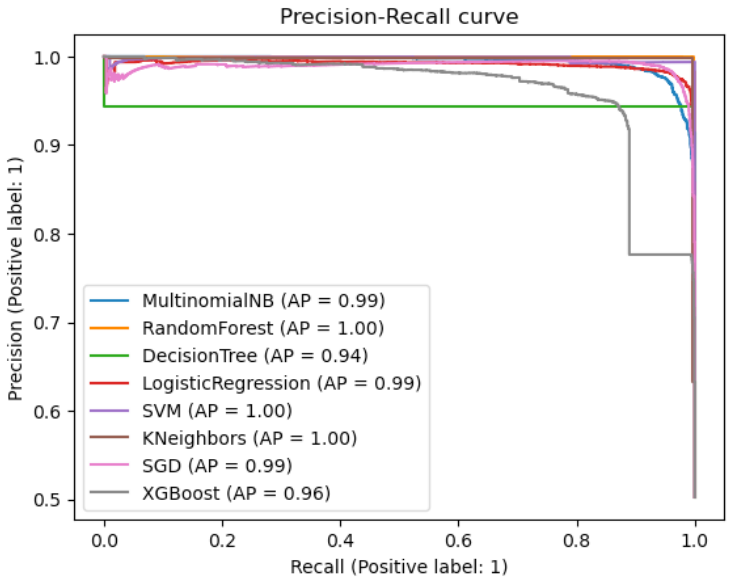
Model Performance Evaluation

- F1-scores for 8 models
- Precision-recall curves for 8 models
- ROC scores and curves for 6 models

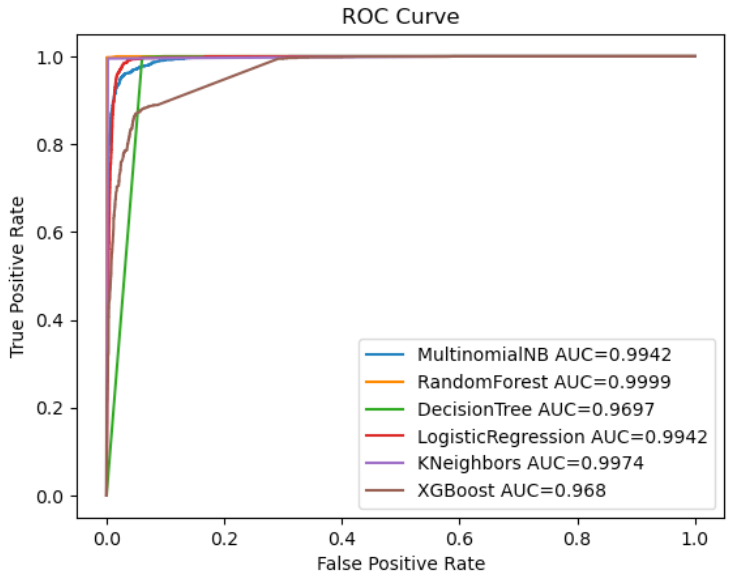
f1-scores for 8 models

0	1	accuracy	macro avg	weighted avg	model
0.950320	0.954137	0.952305	0.952229	0.952371	multinomialnb
0.984192	0.984838	0.984522	0.984515	0.984527	randomforestclassifier
0.948978	0.954175	0.951716	0.951577	0.951841	decisiontreeclassifier
0.974783	0.975899	0.975353	0.975341	0.975363	logisticregression
0.994473	0.994590	0.994532	0.994532	0.994533	svc
0.406469	0.730741	0.629542	0.568605	0.689607	kneighborsclassifier
0.969085	0.970163	0.969633	0.969624	0.969640	sgdclassifier
0.912087	0.905473	0.908900	0.908780	0.909037	xgbclassifier

precision-recall curves for 8 models

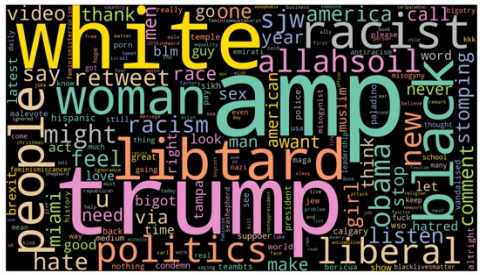


ROC scores and curves for 6 models



Model Test

- Results on the training dataset, including word frequency analysis
- Highlights of significant terms associated with hate speech
- Results on the test dataset and the consistency of identified terms

	Not hate speech in tweets	Hate speech in tweets
Training Dataset Result (total 31962 tweets)	{'love': 2766, 'day': 2735, 'happy': 1679, 'amp': 1319, 'life': 1221, 'time': 1205, 'today': 1066, 'get': 949, 'like': 945, 'positive': 932}	{'amp': 283, 'trump': 216, 'white': 153, 'libtard': 149, 'black': 146, 'like': 140, 'woman': 120, 'racist': 109, 'people': 106, 'politics': 97}
		
Test Dataset Result (total 17197 tweets)	{'love': 1523, 'day': 1417, 'happy': 937, 'amp': 708, 'time': 666, 'life': 627, 'today': 580, 'new': 533, 'get': 509, 'positive': 489}	{'trump': 196, 'amp': 117, 'white': 111, 'black': 92, 'woman': 91, 'racist': 72, 'people': 64, 'like': 59, 'libtard': 57, 'politics': 51}
		

LSTM Model Building & Training

- Addressing imbalanced data through oversampling
- Tokenization, padding, and loading GloVe embeddings for LSTM model
- Defined sequential neural network architecture
- Training process with validation accuracy progression

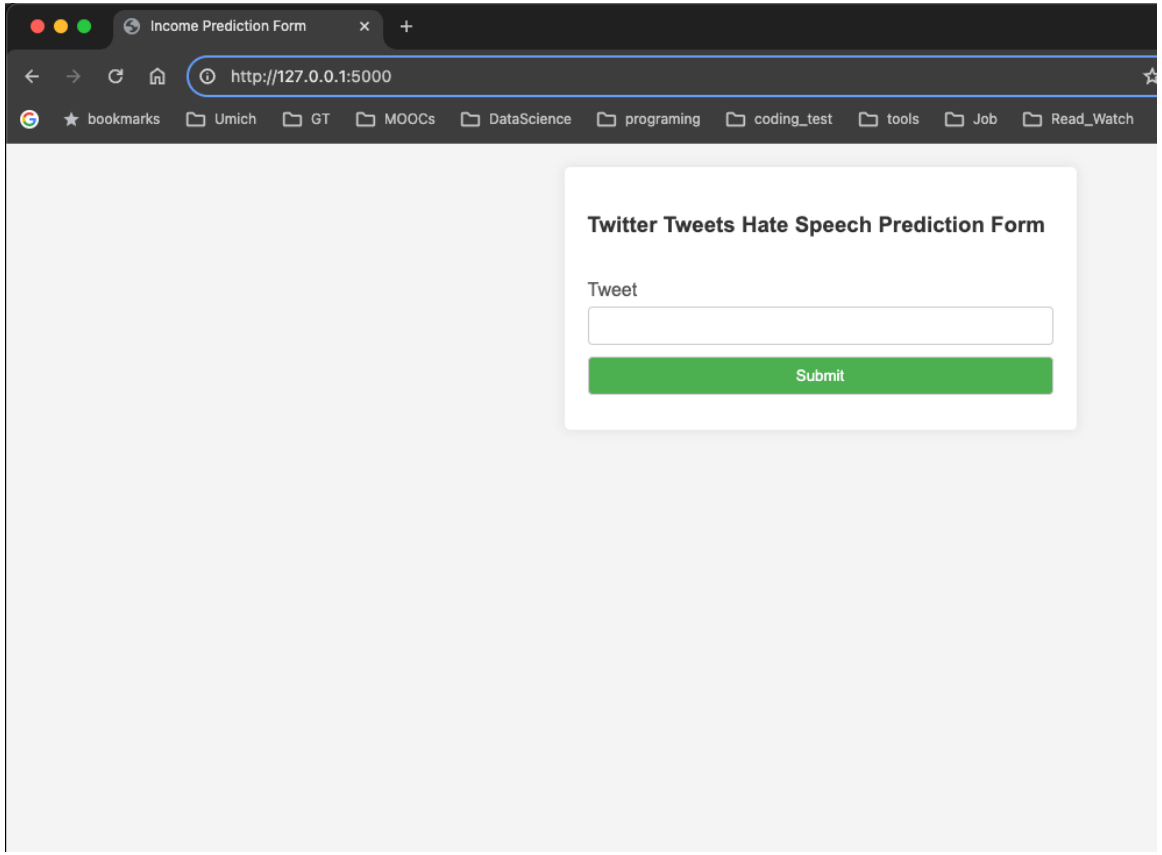
```
Model: "sequential_3"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_3 (Embedding)     (None, 30, 100)           1000000

 lstm_6 (LSTM)               (None, 30, 64)            42240

 lstm_7 (LSTM)               (None, 64)                33024

 dense_4 (Dense)             (None, 64)                4160

 dropout_3 (Dropout)         (None, 64)                0

 dense_5 (Dense)             (None, 2)                 130

=================================================================
Total params: 1079554 (4.12 MB)
Trainable params: 79554 (310.76 KB)
Non-trainable params: 1000000 (3.81 MB)
_________________________________________________________________
```

```python
# Train and Evaluate model

model.fit(train_padded, train_df['label'], epochs=10, validation_data=(val_padded, val_df['label']))

val_loss, val_accuracy = model.evaluate(val_padded, val_df['label'])

print(f'Validation Accuracy: {val_accuracy * 100:.2f}%')
```

```
Epoch 1/10
1486/1486 [==============================] - 57s 36ms/step - loss: 0.3129 - accuracy: 0.8761 - val_loss: 0.1990 - val_accuracy: 0.9224
Epoch 2/10
1486/1486 [==============================] - 52s 35ms/step - loss: 0.1740 - accuracy: 0.9397 - val_loss: 0.1507 - val_accuracy: 0.9418
Epoch 3/10
1486/1486 [==============================] - 50s 34ms/step - loss: 0.1100 - accuracy: 0.9657 - val_loss: 0.0895 - val_accuracy: 0.9688
Epoch 4/10
1486/1486 [==============================] - 52s 35ms/step - loss: 0.0806 - accuracy: 0.9771 - val_loss: 0.0809 - val_accuracy: 0.9759
Epoch 5/10
…
Epoch 10/10
1486/1486 [==============================] - 51s 35ms/step - loss: 0.0265 - accuracy: 0.9928 - val_loss: 0.0569 - val_accuracy: 0.9862
372/372 [==============================] - 4s 12ms/step - loss: 0.0569 - accuracy: 0.9862
Validation Accuracy: 98.62%
```

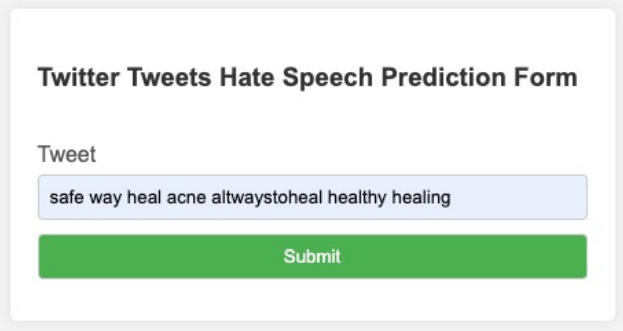
Model Deployment

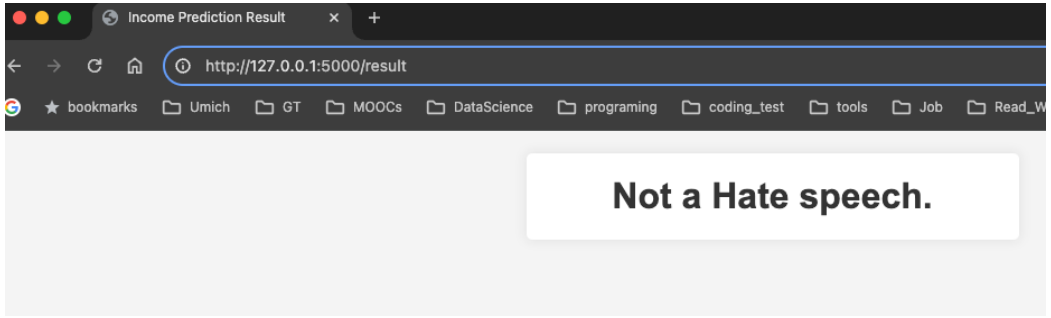
- Deployment using Flask for real-time hate speech detection
- Challenges faced during deployment, including TensorFlow import issues
- Exploration of alternative deployment platforms for a potential solution

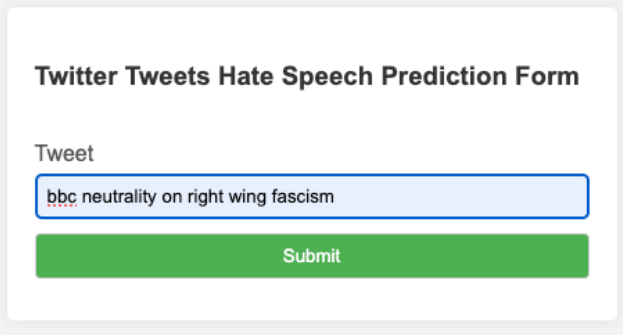


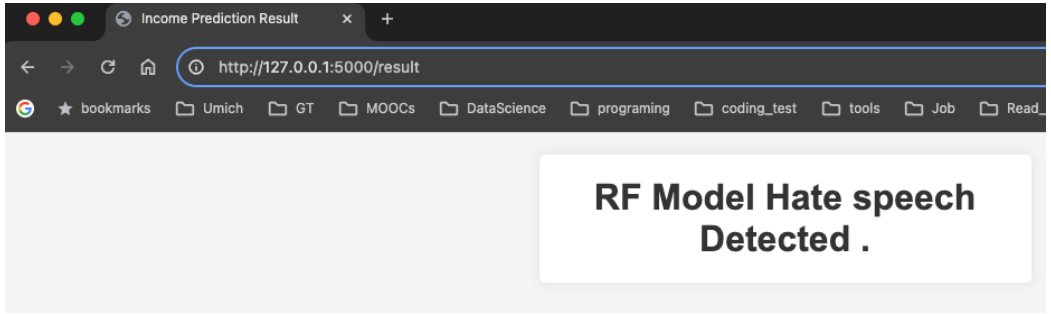
Model Inference

- Real-time results and their significance for hate speech detection
- Demonstrates the practical application of the models on live data









Conclusion and Limitations

- Successes in hate speech detection, including the development of an accurate LSTM model
- Recognition of challenges during deployment, emphasizing TensorFlow import issues and console opening failures on deployment platforms
- Future recommendations for deploying in controlled environments and exploring alternative platforms

Future Work

- Recommendations for alternative deployment platforms
- Consideration of controlled environments during development
- Suggestions for refining and enhancing the project based on encountered challenges

GitHub Repo Link:
https://github.com/syoungk7/Hate_Speech_Detection/

Thank You