Project : Hate speech detection using Transformer (Deep Learning)

Internship Domain : Data Science

Submitted By : Amrapali Mhaisgawali

Batch : LISUM14

12-Oct-2022

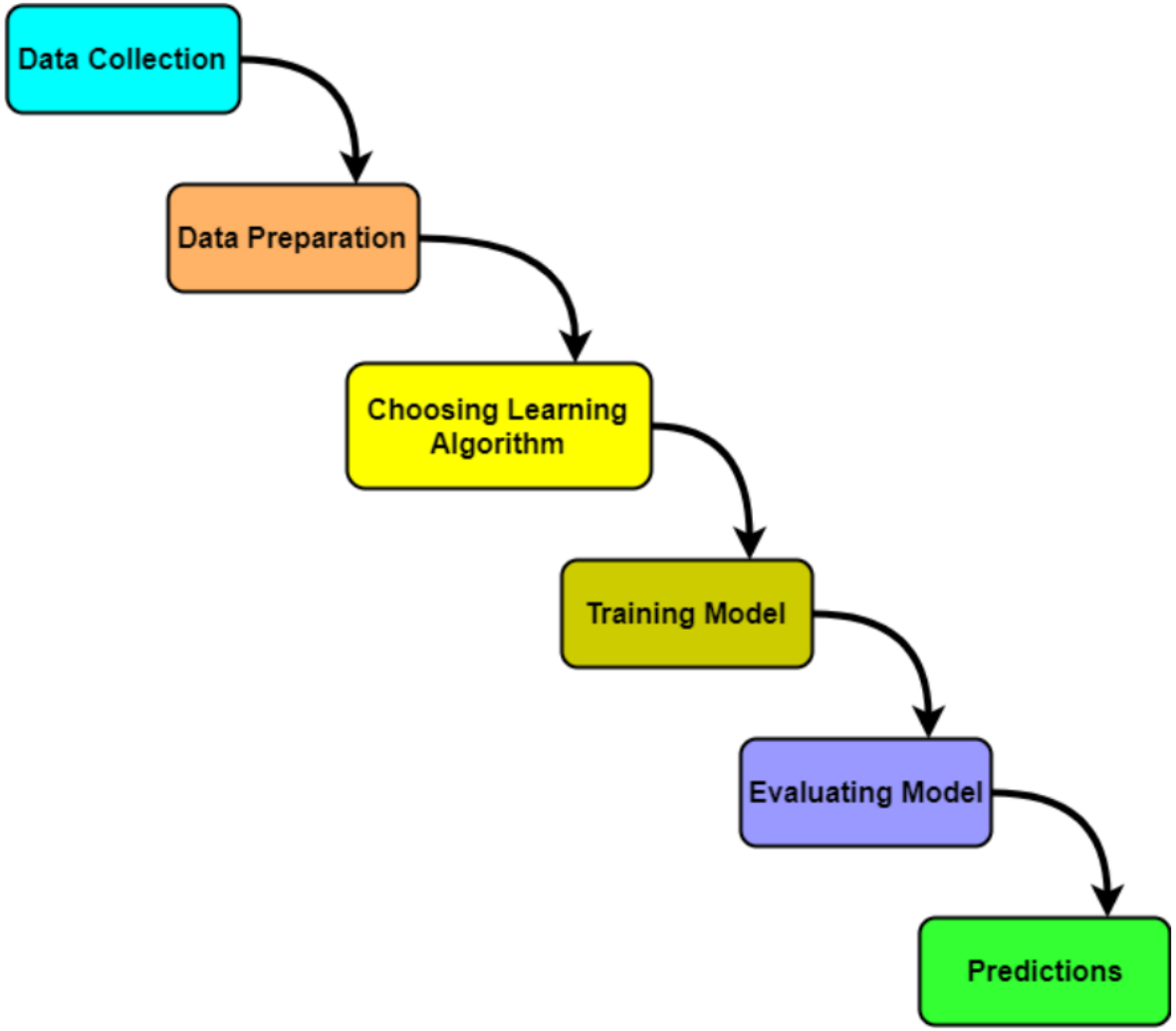
Agenda

- Problem Statement
- Project Architecture
- Feature Engineering
- Feature extraction
- Model Building
- Result Evaluation
- Application Design
- Conclusion
- References

Problem Statement

- The term hate speech is understood as any type of verbal, written or behavioural communication that attacks or uses derogatory or discriminatory language against a person or group based on what they are, in other words, based on their religion, ethnicity, nationality, race, colour, ancestry, sex or another identity factor. In this problem, we will take you through a hate speech detection model with Machine Learning and Python.
- Hate Speech Detection is generally a task of sentiment classification. So, for training, a model that can classify hate speech from a certain piece of text can be achieved by training it on a data that is generally used to classify sentiments. So, for the task of hate speech detection model, we will use the Twitter tweets to identify tweets containing Hate speech.
- The goal is to classify tweets into two categories, hate speech or non-hate speech. Our project analyzed a dataset CSV file from Kaggle containing 31,962 tweets.

Project Architecture



Data Collection

The Data is about Twitter hate Speech taken from Kaggle [1] which contains the 3 number of features and 31962 number of observations. Dataset using Twitter data, it was used to research hate-speech detection. The text is classified as: hate-speech, offensive language, and neither. Due to the nature of the study, it is important to note that this dataset contains text that can be considered racist, sexist, homophobic, or offensive.

Total number of observations	31962
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	2.95 MB

Table: Data Information

Feature Engineering

- **Text Cleaning**
  - Lowercase
  - Remove Punctuation
  - Remove URLs
  - Remove @tags
  - Remove Special Characters
- **Pre-processing Operations**
  - Tokenization
  - Removing Stop Words
  - Lemmatization

Feature Extraction

- **TF-IDF Model**
  - Creating the histogram
  - frequent words from dictionaries
  - TF Matrix
  - IDF Matrix
  - TF-IDF Calculation

Model Building

- **Deep Learning Model – RNN(LSTM)**

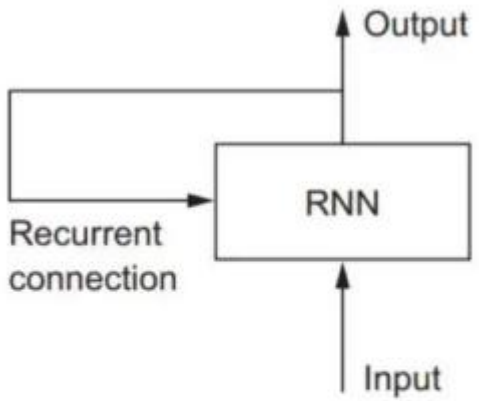


Fig 2 RNN Model

Result Evaluation

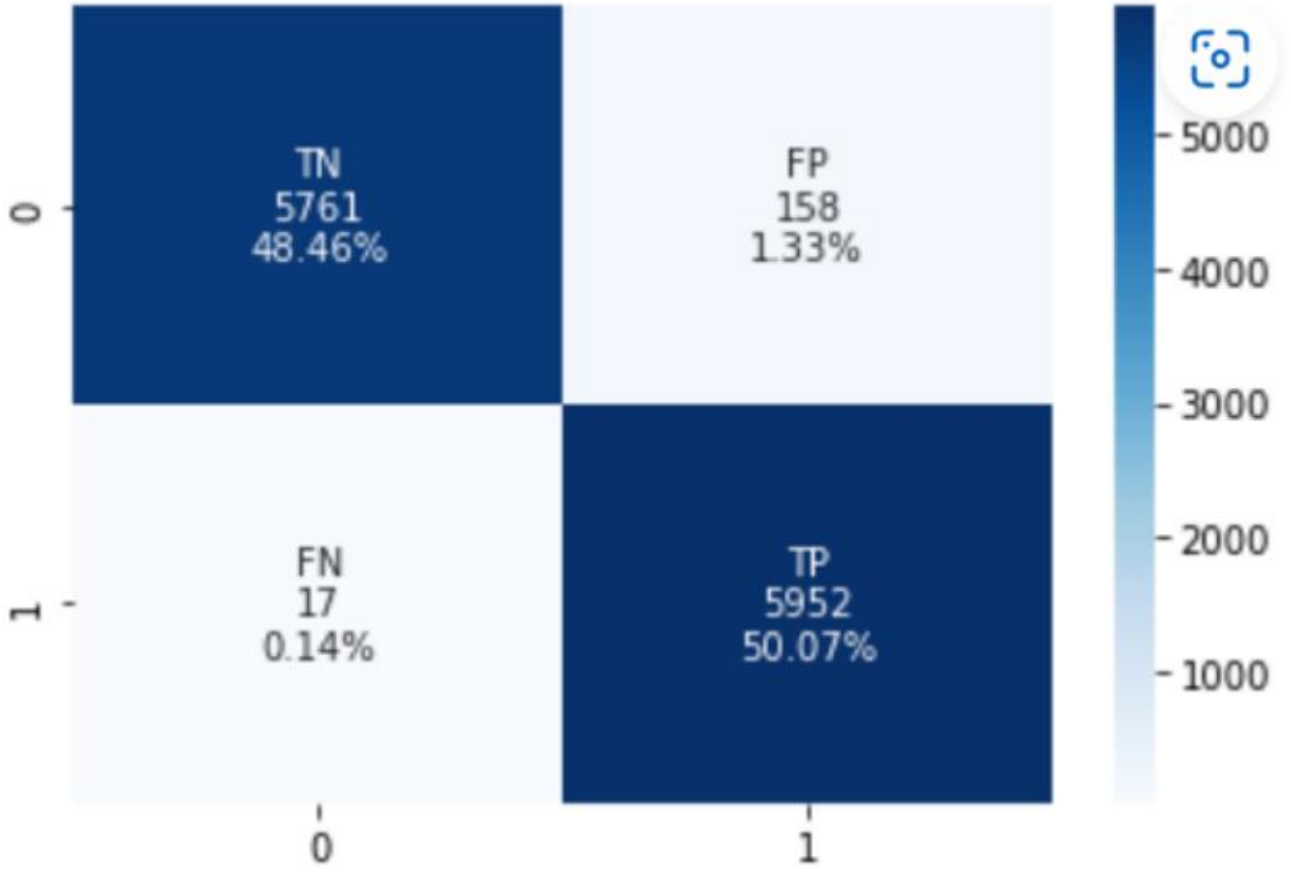


Fig 3 Confusion Matrix

Result Evaluation

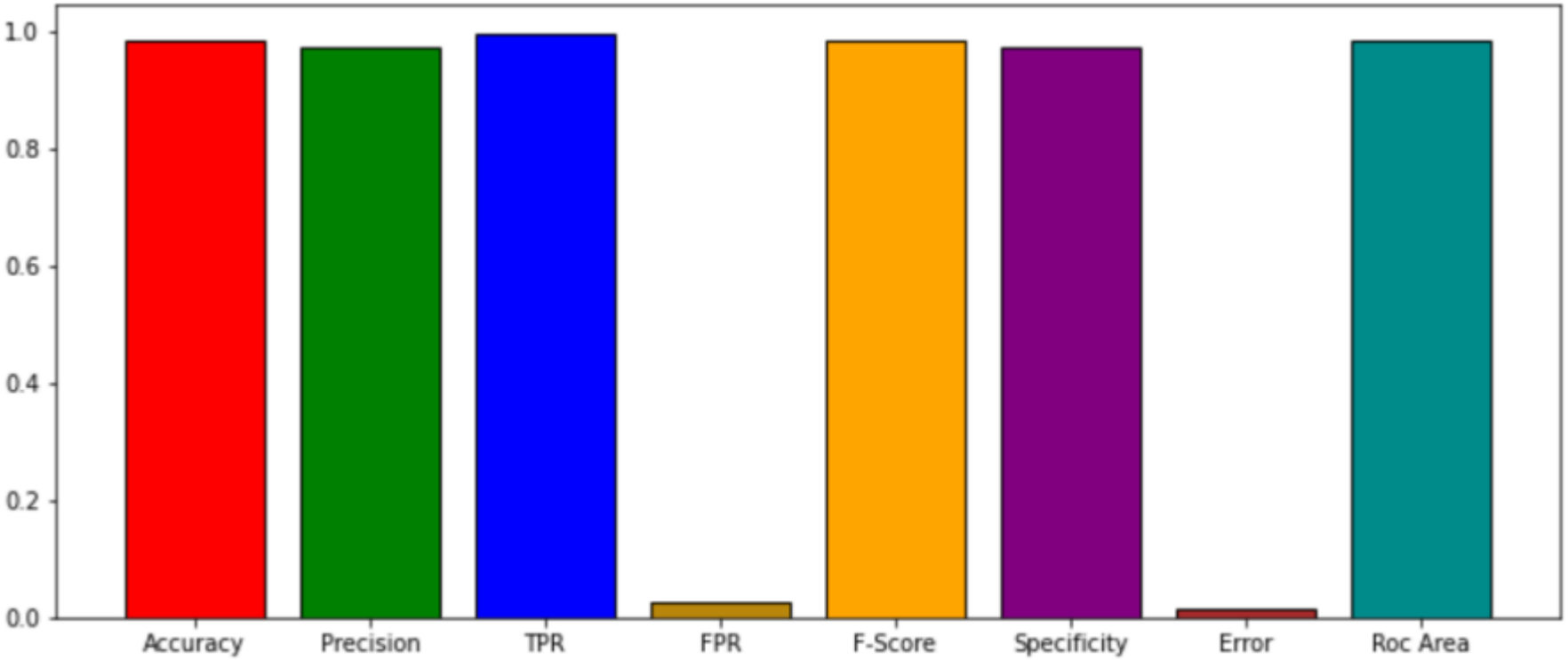
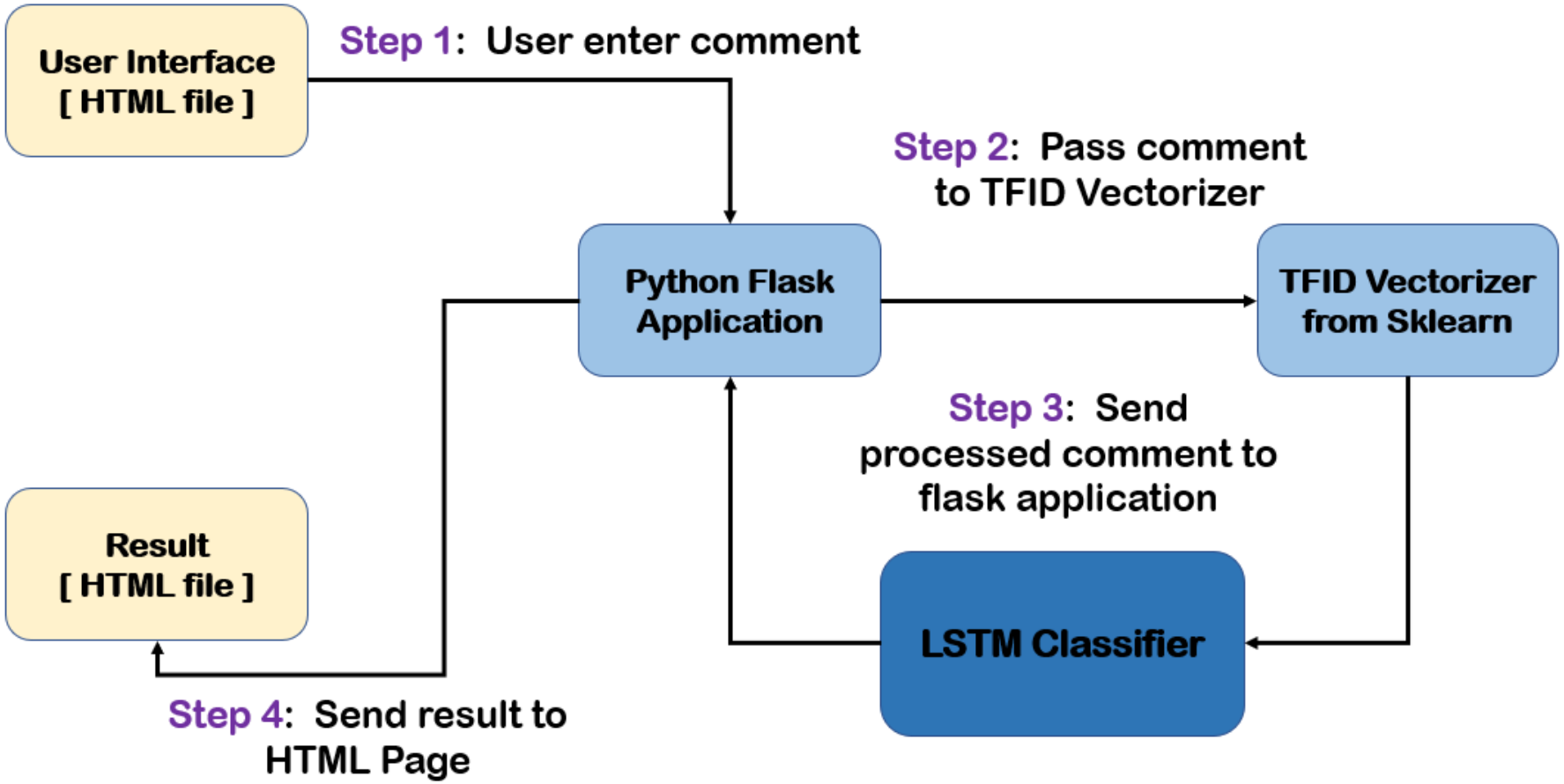
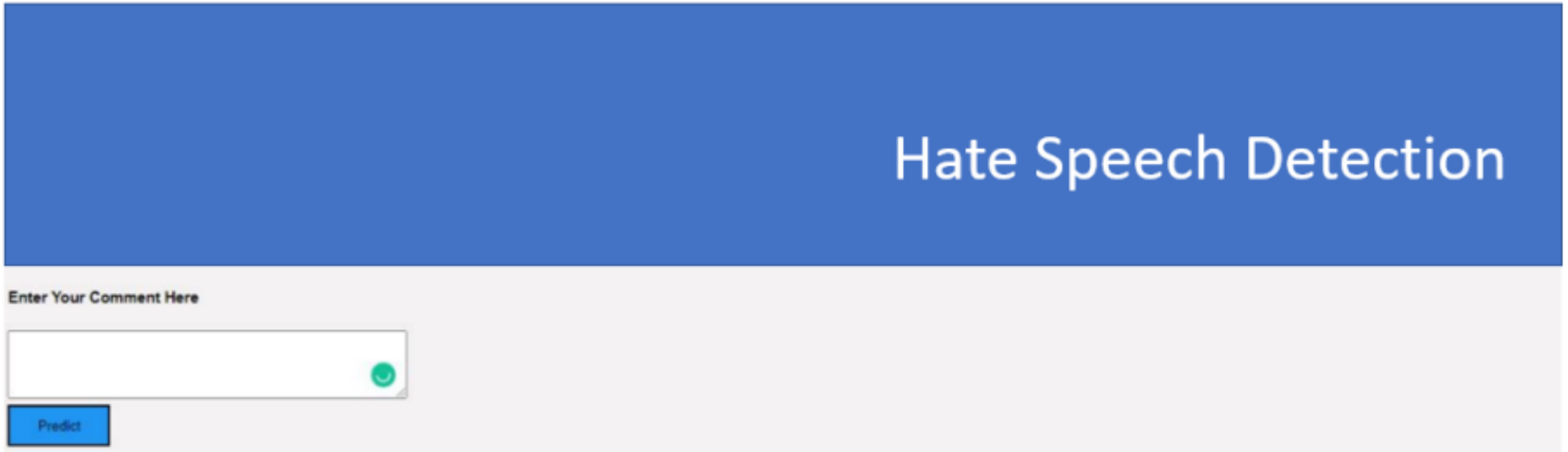


Fig 4 Visualization of Results

Application Design



Application Design



Conclusion

The goal of this project was to find capable methods and settings that could be used to help the detection of Hate and Free Speech of twitter. The error rate of the model is not zero, so still, some incorrect can be classified as true by the model

References

[1] https://www.kaggle.com/datasets/vkrahul/twitter-hate-speech?select=train_E6oV3lV.csv

Thank You