



Data Glacier

Data Science Intern at Data Glacier

Project: Hate Speech Detection using Transformers (Deep Learning)

Week 13: Final Report

Name: Anitha Venkatachalam

Email: anitha.Venkat83@gmail.com

Country: Canada

Specialization: Data Science

Batch Code: LISUM14

Date: 28Dec 2022

Submitted to: Data Glacier

Outline

- Problem Statement
- System Architecture
- Result Evaluation
- Application Design
- Conclusion
- Reference

Problem Statement

The term hate speech is understood as any type of verbal, written or behavioral 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, color, 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 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 data that is used to classify sentiments. So, for the task of the hate speech detection model, we will use Twitter tweets to identify tweets containing Hate speech.

System Architecture

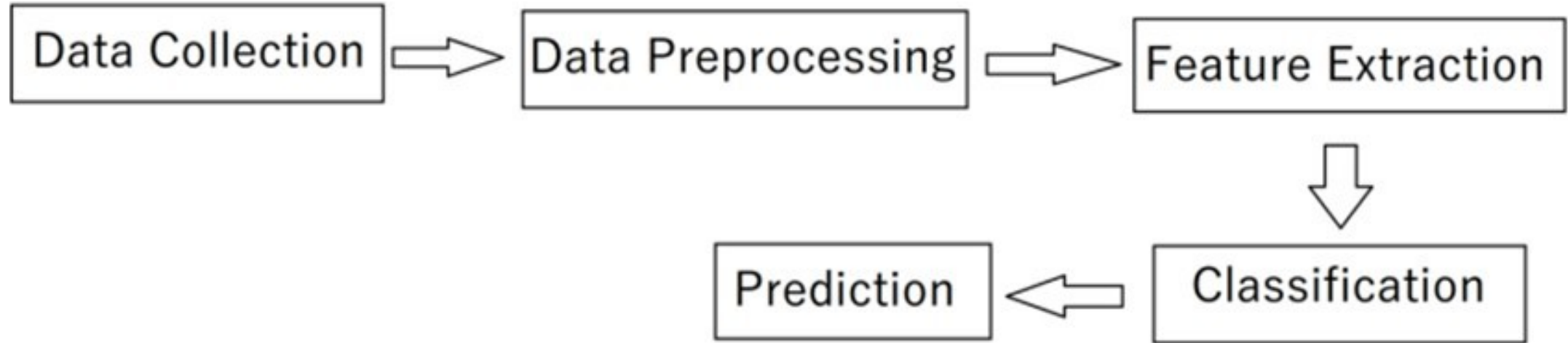


Figure 1: System Architecture

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.

Table 1: Data Information

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

Data Preprocessing

- **Text Cleaning**

- Lowercase
- Remove Punctuation
- Remove URLs
- Remove @tags
- Remove Special Characters

- **Preprocessing 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

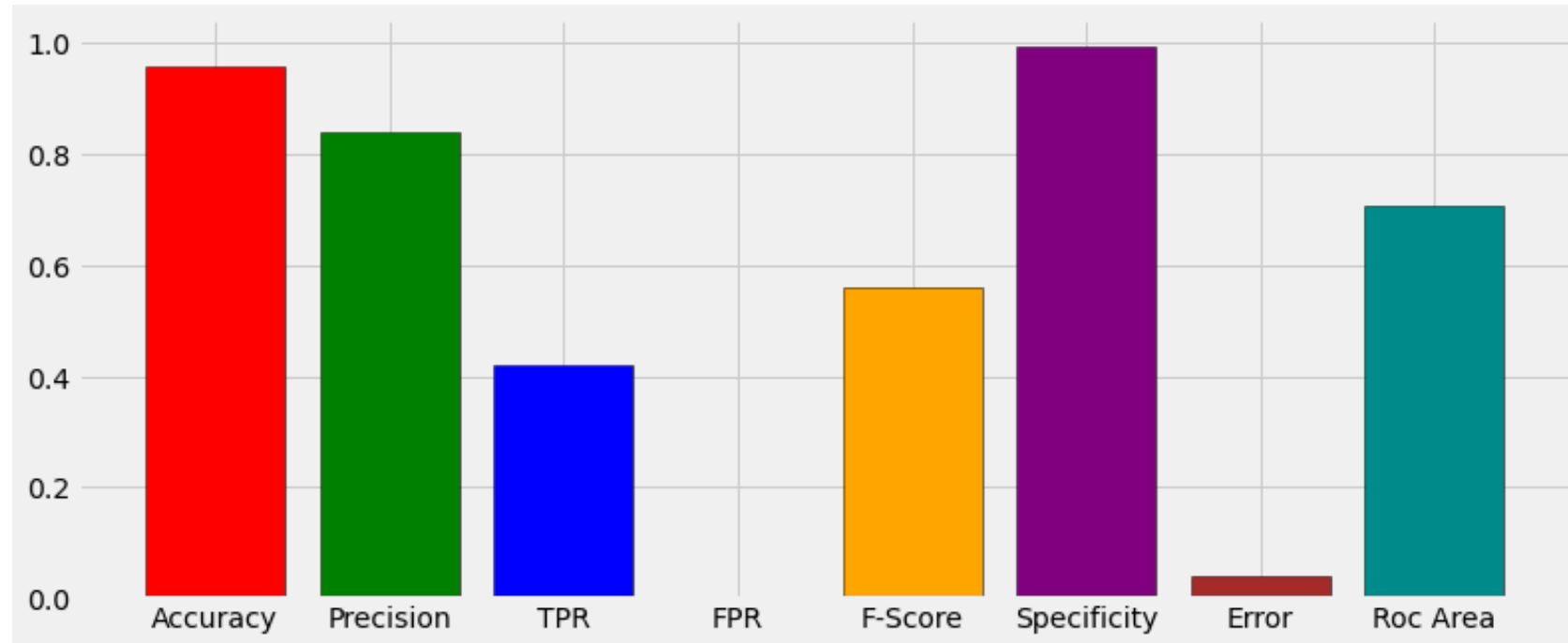
Result Evaluation

Confusion Matrix Visualization

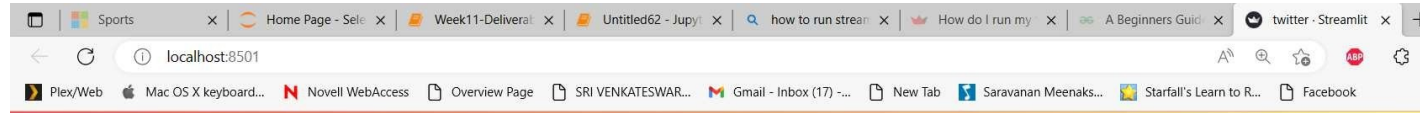


Result Evaluation (Cont'd)

- Result Visualization



Application Design



Hate Speech Detection

Enter any Tweet:



Conclusion

The goal of this project was to find capable methods and settings that could be used to help detection of Hate and Free Speech on twitter. The error rate of the model is not zero, so still, some incorrect can be classified as true by the model. In future, we will enhance this work by implementing Temporal Convolutional Network (TCN) and Random Multimodal Deep Learning (RMDL) Techniques.

Thank You😊