



Data Glacier

Your Deep Learning Partner

Project: Hate Speech Detection Using Transformers (Deep Learning)

University: Warsaw University of Technology, Warsaw, Poland

Internship Domain: Data Science

Submitted by: Amir Ali

Batch: LISUM08

Date: 30-June-2022

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

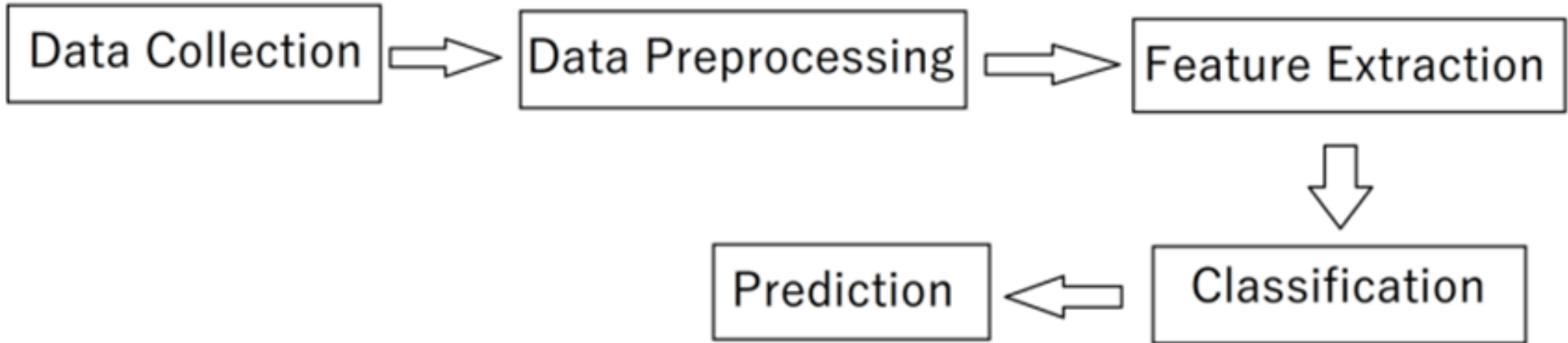


Figure 1: System Architecture

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**

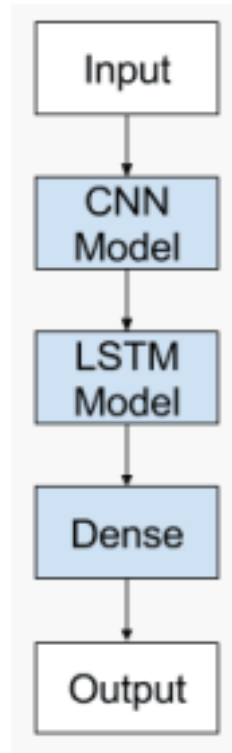
- Tokenization
- Removing Stop Words
- Lemmatization

Feature Extraction

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

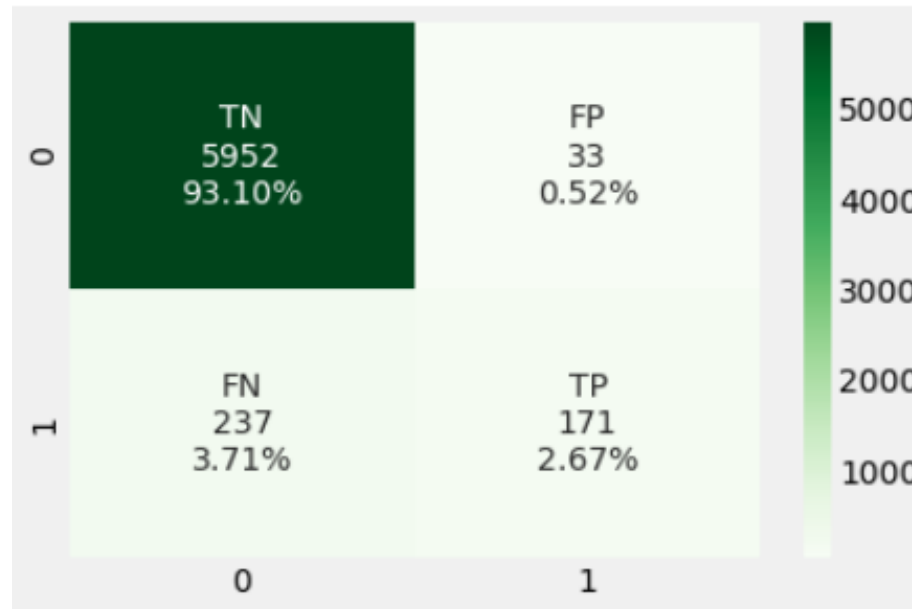
Deep Learning Model

CNN with LSTM Model



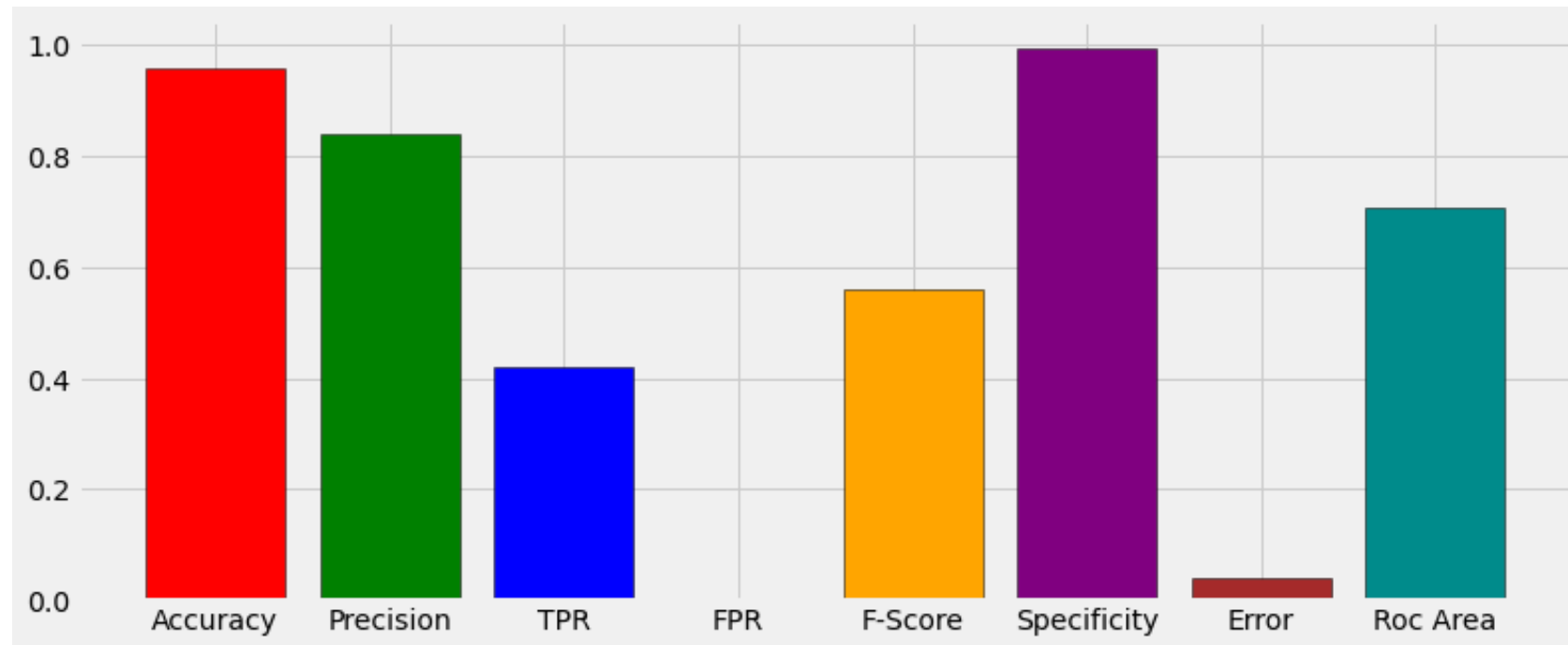
Result Evaluation

Confusion Matrix Visualization



Result Evaluation (Cont'd)

- Result Visualization



Application Design

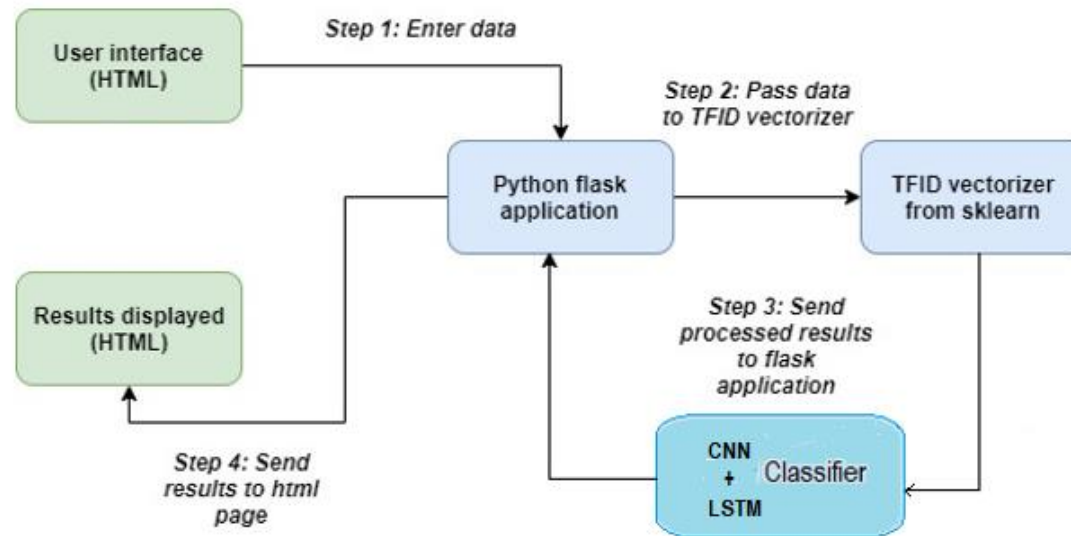


Figure 3: Application Design

Application Design



The mockup features a light beige background with decorative elements of orange and teal circles and dotted patterns in the corners. The main title 'DATA SCIENCE INTERN AT DATA GLACIER' is centered in large, bold, multi-colored letters (blue, green, black, orange). Below it, the project name 'PROJECT: HATE SPEECH DETECTION' is displayed in red and black. To the left of the project name is a red prohibition sign over the words 'HATE SPEECH'. To the right, the author's name 'BY: AMIR ALI' is written in blue, accompanied by a small blue signature. Below the main content area, there is a text input field with the placeholder 'Enter Your Comment Here', a green checkmark icon, and a blue 'Predict' button.

DATA SCIENCE INTERN AT DATA GLACIER

PROJECT: HATE SPEECH DETECTION

HATE SPEECH

BY: AMIR ALI

Enter Your Comment Here

Predict

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 Multimodel Deep Learning (RMDL) Techniques.

Reference

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

[2] <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

[3] Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., & Darrell, T. (2016, May 31). Long-term recurrent convolutional networks for visual recognition and description. arXiv.org. Retrieved May 8, 2022, <https://arxiv.org/abs/1411.4389>

Thank You😊