Data Science Intern at Data Glacier

Project: Hate Speech Detection Using Transformers (Deep Learning)

Week 8: Project Deliverable 3

Team Name: Fibo

Team Members: Mahyar Arani

Email: arani.mahyar@gmail.com

Batch code: LISUM19

Submission date: May 2, 2023

Submitted to: Data Glacier

Problem Description:

The problem at hand is detecting hate speech in Twitter tweets using a machine learning model. Hate speech is a type of communication that attacks or uses derogatory or discriminatory language against a person or group based on their religion, ethnicity, nationality, race, color, ancestry, sex, or other identity factors. The goal is to develop a model that can accurately classify whether a tweet contains hate speech or not.

Data Cleansing and Transformation

In order to prepare the data for classification modeling, we employed various data cleansing techniques. Our first step was sentence tokenization followed by word tokenization to break down the text data into smaller units for analysis. We then removed symbols, special characters, and URLs to create a cleaner dataset.

To extract additional features from the text data, we used VADER lexicon for sentiment analysis. This approach allowed us to obtain four float scores which we included in our modeling process. We labeled the sentiment of each tweet using a cutoff of 0.33 based on the compound score.

To further enhance the quality of our data, we applied various text cleansing methods using regex. By comparing the results of different approaches, we were able to identify the best method for modeling the text data.

Since this project was performed individually, peer review was not possible. However, we welcome any feedback from other teams that could help us improve our techniques and results.

Project Lifecycle

Weeks	Date	plan
Weeks 07	Apr 19, 2022	Problem Statement, Data Collection, Data Report
Weeks 08	Apr 26, 2022	Data Preprocessing
Weeks 09	May 2, 2022	Feature Extraction
Weeks 10	May 9, 2022	Building the Model
Weeks 11	May 16, 2022	Model Evaluation
Weeks 12	May 23, 2022	Flask Development + Heroku
Weeks 13	May 30, 2022	Final Submission (Report + Code + Presentation)