Data Glacier

Data Science Intern at Data Glacier

Project: Hate Speech detection using Transformers (Deep Learning)(NLP)

Week 8: Deliverables

Name: Pinakin Prabhakar

University: Lakehead University

Email: pnakinprabhakar8905@gmail.com

Country: Canada

Specialization: Data Science

Batch Code: LISUM17

Submitted to: Data Glacier

Table of Contents:

1. Project Plan

Weeks	Date	plan
Weeks 07	March, 2023	Problem Statement, Data Collection, Data Report
Weeks 08	March, 2023	Data Preprocessing
Weeks 09	March, 2023	Feature Extraction
Weeks 10	March, 2023	Building the Model
Weeks 11	March 16, 2023	Model Result Evaluation
Weeks 12	March 23 , 2023	Flask Development + Heroku
Weeks 13	March 30, 2023	Final Submission (Report + Code + Presentation)

2. Problem Statement

Hate speech is a form of verbal aggression that can have a harmful impact on individuals and groups targeted by such speech. With the rise of social media platforms, hate speech has become more widespread, and detecting and removing it has become a major challenge.

The problem of hate speech detection can be addressed using deep learning models like Transformers, which can learn representations of text that capture semantic relationships and contextual information. The task is to develop a hate speech detection model using Transformers that can accurately classify text as hate speech or not hate speech.

The objective of this project is to build a robust and efficient hate speech detection model that can accurately classify text from various sources, including social media platforms and online forums. The model should be able to handle different languages and dialects and should be scalable to handle large volumes of data. The model should also be interpretable, providing insights into the features that contribute to hate speech classification. The model's performance should be evaluated using standard metrics, such as precision, recall, and F1 score, and compared against existing state-of-the-art models for hate speech detection.

3. Dataset

Dataset link: Twitter hate speech | Kaggle

This dataset, which was collected from Kaggle and includes three characteristics, is all about Twitter hate speech. The hate speech on Twitter is tracked using this dataset. Three sorts of text are distinguished: neutral, hate speech, and offensive language.

- Number of observations: 31963
- Number of n/a's: 0
- Features: 3

4. Data/Text Preprocessing

- Text Cleaning
- Lower Casing
- Removal of Punctuations
- Removal of Stop words
- Removing Special Characters
- Removing URL’s and many more