Final Project – DataGlacier Virtual Internship 2024

Project Title: NLP - Twitter Hate Speech Detection with Transformer (Deep Learning)

Name: Ankita Manna

E-mail: ankitamanna1@gmail.com

Country: United Kingdom

Specialization: Natural Language Processing

Problem Statement: In recent years, the proliferation of social media platforms, particularly Twitter, has facilitated global communication and the rapid exchange of ideas. However, this accessibility has also enabled the spread of hate speech, defined as any form of communication that demeans, discriminates, or incites hostility toward individuals or groups based on characteristics such as race, religion, ethnicity, gender, sexual orientation, or other identity markers.

Hate speech not only fosters a toxic online environment but also poses significant societal risks, including psychological harm, societal polarization, and real-world violence. The challenge of combating hate speech is compounded by the increasing diversity of languages, regional dialects, and slang used online, as well as the subtlety of implicit hate and the evolution of coded language.

Given the vast scale of user-generated content on platforms like Twitter, manual monitoring and moderation are impractical. This necessitates the development of automated systems using Natural Language Processing (NLP) techniques to detect and flag hate speech. However, existing approaches often struggle with handling multilingual content, understanding context, and differentiating between hate speech and non-hateful but contentious discussions.

This project aims to address these challenges by designing an advanced hate speech detection model leveraging NLP techniques to effectively identify and mitigate harmful content on Twitter, thereby promoting safer and more inclusive digital spaces.

Data Understanding: There are two files in the main dataset. One is the train dataset and test dataset. The train dataset contains three columns i.e., 'id', 'tweet', and 'label'. The test data contains two columns i.e., 'id' and 'tweet'. In both datasets, the tweets are in the form of text data and labels are either 0 or 1.

The problem with the tweet data is that it comes with larger noise. We need some approach to remove these noises. The approach am taking in this problem is use regex, beautiful soup, `word_tokenizer, stopwords` libraries to remove any HTML parser and stuffs like that and finally get the finished good quality data to send it for analysis.