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

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. Manual monitoring and moderation are impractical given the vast scale of user-generated content on platforms like Twitter. 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.

Exploratory Data Analysis:

Recommendations:

- The EDA has been done on the textual data to see which words used the most on each class and found that the word ‘User’ has been used much on both classes. This word is nothing, but the common word used by Twitter. For positive words, ‘love’ comes second, whereas for negative words, ‘trump’ comes next.’

- Also, we found many local slang words in the corpus that are not correctly spelled in English. Attempts will be made to replace those words by building a dictionary with key as the slang words and their correct word as value and replacing them across the corpus. Building such a dictionary in would be useful in the long term in developing the model.

- Class Imbalance has been found on the data, so for designing the model, the undersampling technique adopted as oversampling may overfit the model. There would be some loss in the model, but it can be compensated in a long run by acquiring more data with the other class in future.