Final Project – DataGlacier Virtual Internship 2021

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

Name: Manoj Kumar Thangaraj

E-mail: manojthangaraj92@gmail.com

Country: Ireland

College: Dublin Business School

Specialization: Natural Language Processing

Problem Statement:

Hate Speeches are taking over every media, social platform etc., The term hate speech is understood as any type of verbal, written or behavioural 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, colour, ancestry, sex or another identity factor. Designing a model to detect such speeches or posts on these platforms are getting complicated these days with increase in usage different languages.

Exploratory Data Analysis:

https://github.com/manojthangaraj92/NLP-Hate-Speech-Detection-Transformers/blob/master/NLP   Final.ipynb

EDA has been done on the same main notebook.

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.
- Therefore, this has no useful meaning for that word, and it might influence the model in predicting wrong class. Therefore, we can remove that word from the corpus.
- Also, we found many local slang words found in the corpus, which is not correctly spelled in English. Attempt has been 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 dictionary in would be useful in a long term in developing the model.
- Class Imbalance has been found on the data, so for designing the model, the under-sampling 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.