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.

Data Understanding:

The dataset that we got is of two files. One is 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 wither 0 or 1. The problem with the tweets 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, beautifulsoup libraries to remove any HTML parser and stuffs like that and finally get the finished good quality data to send it for analysis.