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.

Business Understanding:

We are using twitter tweets for developing the model. Detecting a hate speech on high level looks like a difficult task. But it is one of the simplest tasks in NLP. It basically a sentiment classification task. Developing a deep learning model that can predict with the set of words that's being tweeted is containing hate speech or not.

As per the project title, transformers have been chosen to build the model with deep learning concepts.

Project Life Cycle with deadlines:

The project is composed to different stages in which each contributes to the project, the same level. The estimated deadline for each stage has been given in the following.

1. Data Cleansing & Normalization – 26th July – 30th July
2. Exploratory data analysis –26th July – 30th July
3. Representation learning – 2nd August – 7th August
4. Model Building and Training - 2nd August – 7th August
5. Model Evaluation – 9th August – 14th August
6. Model Deployment in Amazon SageMaker – 9th August – 14th August
7. Model Inference – 15th August
8. Project Report – 15th August

Data Intake Report

Name: Manoj Kumar Thangaraj

Report date: 20th July 2021

Internship Batch: LISUM01

Version:<1.0>

Data intake by: Manoj Kumar Thanagaj

Data intake reviewer:<intern who reviewed the report>

Data storage location: Github

Tabular data details:

Training file: train_E6oV3lV

Total number of observations	31962
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	3mb

Testing File: test_tweets_anuFYb8.csv

Total number of observations	17198
Total number of files	1
Total number of features	2
Base format of the file	.csv
Size of the data	1.55mb

Proposed Approach:

The training dataset is used for training the model, testing, and validation. The test dataset is sent through the model and get the results of those tweets.

Assumptions:

The tweets are not the raw text. It comes with noise.