Data Science Internship

at Data Glacier

Project: Hate Speech Detection using Transformers

(Deep Learning)

Week 8: Deliverables

Name: Amrapali Mhaisgawali

University: -

Email: amrapali10@gmail.com

Country: United Kingdom

Specialization: Data Science

Batch code: LISUM14

Date: 25/11/2022

Submitted to: Data Glacier

Index

1. Problem Statement:

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. In this problem, we will take you through a hate speech detection model with Machine Learning and Python.

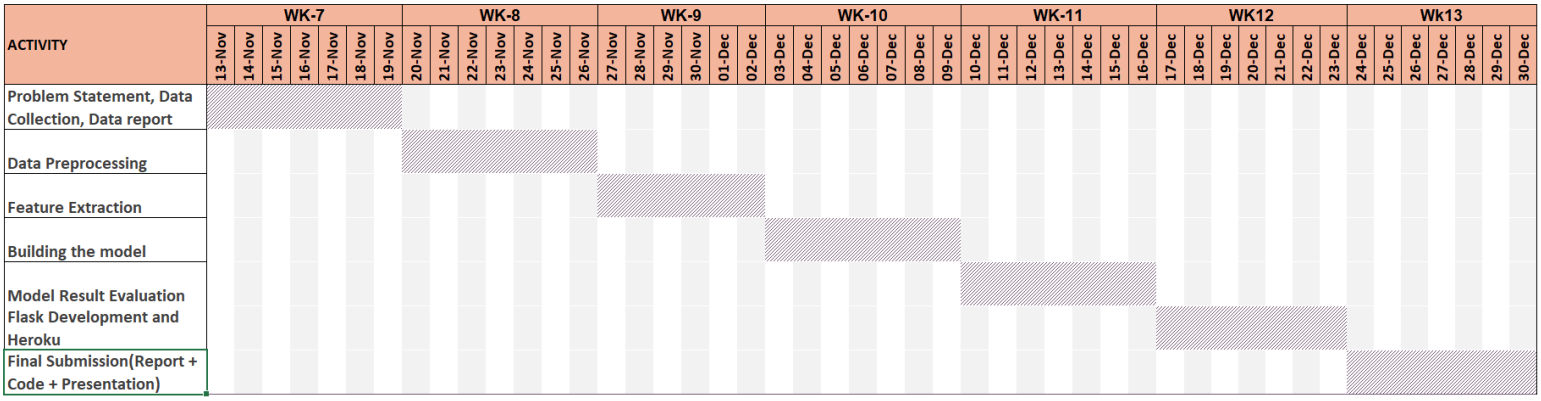
Hate Speech Detection is generally a task of sentiment classification. So, for training, a model that can classify hate speech from a certain piece of text can be achieved by training it on a data that is generally used to classify sentiments. So, for the task of hate speech detection model, we will use the Twitter tweets to identify tweets containing Hate speech.

Our goal is to classify tweets into two categories, hate speech or non-hate speech. Our project analyzed a dataset CSV file from Kaggle containing 31,962 tweets.

2. Business Understanding:

Social media has experienced incredible growth over the last decade, both in its scale and importance as a form of communication. The nature of social media means that anyone can post anything they desire, putting forward any position, whether it is enlightening, repugnant or anywhere between. Depending on the forum, such posts can be visible to many millions of people. Different forums have different definitions of inappropriate content and different processes for identifying it, but the scale of the medium means that automated methods are an important part of this task. Hate-speech is an important aspect of this inappropriate content.

3. Project Lifecycle



4. Data Collection

The Data 'Twitter hate Speech' is used for detection of hate speech taken from Kaggle [1] which contains the 3 features and 31962 number of observations. Data from Twitter website was used to research hate-speech detection. The text is classified as: hate-speech, offensive language, and neither. Due to the nature of the study, it is important to note that this dataset contains text that can be considered racist, sexist, homophobic, or offensive.

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	2.95 MB

Table: Data Information

5. Data Pre-processing

Following are the steps in pre-processing

5.1 Text Cleaning

Text cleaning is the first step in data pre-processing

5.1.1 Lowercase

Data is converted to lower case. Words like User and user mean the same but when not converted to the lower case those two are represented as two different words in the vector space model (resulting in more dimensions). Therefore, data was converted into lower case letter.

5.1.2 Remove Punctuation

It is important to remove the Punctuation as they are not important. Punctuation are removed using regular expression.

5.1.3 Remove URLs

URLs are not important to detect the hate and free speech so they are removed from text.

5.1.4 Remove @tags

@tags are removed which basically used to mentioned someone. So, it's doesn't concern to detect hate speech therefore, they are removed by using regular expressions.

5.1.4 Remove Special Characters

Special Characters which basically don't have meaning. Therefore, they are removed. In order to remove we use python isalnum method.

References

[1] https://www.kaggle.com/datasets/vkrahul/twitter-hate-peech?select=train_E6oV3lV.csv