Group name: Arjohi

Name: Hiten Chadha

Email: hitenchadha1995@gmail.com

Country: Denmark

College/Company: Technical University of Denmark

Specialization: NLP

GitHub repository link:
https://github.com/hitenchadha1910/DG-week7

Problem description & Business understanding:

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.

We will analyze a dataset CSV file from Kaggle containing 31,935 tweets. The dataset was heavily skewed with 93% of tweets or 29,695 tweets containing non-hate labeled Twitter data and 7% or 2,240 tweets containing hate-labeled Twitter data. We will try different classification algorithms after the preprocessing and data cleaning steps.

Project Life Cycle:

19 July - 26 July:

Describe the problem. Understand the data, provide an overview of the data and investigate it, clarify the data type we have got, explain the approaches applied on the data set to overcome problems like NA value, outlier etc.

27 July - 2 August:

Focus on data cleansing and transformation done on the data. Try different data cleansing approaches.

3 August - 9 August:

Perform exploratory data analysis on the dataset, use the work from Week 2 of the internship as inspiration. Make recommendations regarding the project.

10 August - 16 August:

Create a presentation which plainly describes and visualises the work from the prvious week on EDA for non-technical business users and also present the final recommendations. On the final slide, include the recommended model for this dataset, which will be useful for technical users.

17 August - 23 August:

Select your base model and then explore 1 model of each family if its classification problem then 1 model for Linear models, 1- Model for Ensemble, 1-Model for boosting and other models if you have time.

24 August - 30 August:

Write a report for the project and also include a PowerPoint presentation.

Data Intake Report

Name: Advance NLP : Hate Speech detection using Transformers - Group Project

Twitter Hate Speech Dataset

Report date: 18.07.2022

Internship Batch: LISUM10

Version:1.0

Data intake by: Hiten Chadha

Data intake reviewer: Hiten Chadha

Data storage location: Local hard drive

Tabular data details:

Total number of observations	32k from train and 17k from test dataset
Total number of files	2 (train and test datasets)
Total number of features	3
Base format of the file	.csv
Size of the data	2 MB

Proposed Approach:

- The duplicated() command will be used in Python to identify duplicate entries.
- Assumptions: No assumptions have been made so far.