Group Name: Sedats

Name: Sedat CAN (one-member group)

Email: sedatcancan1@gmail.con

Country: Germany

Specialization: Data Science, NLP, Data Analyst

Problem Description

Hate speech is one of the biggest problems that social media users face on daily basis. Especially, Twitter users are harassed and discriminated against because of their religion, ethnicity, nationality, race, color, ancestry, sex, etc.

The hate speech detection model would evaluate and classifies whether tweets include hate speech or not and filters these tweets to offer a better user experience.

Business understanding

Hate speech has a negative effect on society. It can cause stress, social exclusion, and suicide. Moreover, it can trigger social fault lines such as racism, and hate against minorities and sexes. The model that labels the tweet as containing harassment and removes the tweet before it spread on social media has the potential to attract customers.

Project lifecycle along with the deadline

- Data analysis - 26 Nov 2022
- Data Cleansing and Transformation - 2 Dec 2022
- EDA performed on the data - 9 Dec 2022
- EDA Presentation - 16 Dec 2022
- Model Selection and Model Building - 23 Dec 2022
- The Project deadline is 30 Dec 2022

Github Repo link

https://github.com/sedat-can/Final-project-.git

Data Intake Report

Name: Final project
Report date: 19.11.2022
Internship Batch: LISUM14
Version:1
Data intake by: Sedat CAN
Data intake reviewer:<intern who reviewed the report>
Data storage location: https://www.kaggle.com/code/codeserra09/hate-speech-sentiment-analysis-lg-mnb-dt-rf-knn/data

Tabular data details:

Total number of observations	17197
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	4.74 MB

Note: Replicate the same table with the file name if you have more than one file.

Proposed Approach:

- The data may contain data dedup or dedupe and therefore there should be a data cleaning process.
- The Data consists of 3 columns. The last column most probably refers to ‘label’. As the first and last columns contain numbers, the second one is textual data