

Deliverables

Team Member's Details

Group Name : Data Junkies

- Name : Farheen Fatima
- Email: farheen.fatima07@gmail.com
- Country: United States
- College/Company : California State University, Sacramento
- Specialization : Data Science

- Name: Mohamed Derbeli
- Email: derbelimohamed1@gmail.com
- Country: Spain
- College/Company : Data Glacier
- Specialization: Data Science

- Name: Mohamed Akrem Ben Jemia
- Email: akrem008.benjamie@gmail.com
- Country: Spain
- College/Company: École National d'ingénieur de Carthage
- Specialization: NLP

Problem Description & Business Understanding

The term hate speech is understood as any type of verbal, written or behavioral 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, color, 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 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.

Project lifecycle along with deadline:

- week 1: week of 19 September 2022 - Problem Understanding and Data Cleaning
- week 2: week of 26 September 2022 - Data preprocessing and data cleaning with different methods
- week 3: Week of 2 October 2022 - Making recommendations based on data analysis and selecting the model that will be applied after.
- Week 4: week of 16 October 2022 - Create a presentation for the past week
- Week 5: Week of 16 October 2022 - EDA presentation and proposed technique for modeling.
- Week 6: Week of 23 October 2022 - Select a base model and then try a different model and compare results.
- Week 7: week of 30 October 2022 - Write a final report for the project and make a presentation

Data Intake Report:

Data Intake Report

Name: Hate Speech Detection using Transformers

Report date: 09/18/2022

Internship Batch: LISUM12

Version: 1.0

Data intake by : Farheen Fatima, Mohammed Derbeli, Mohammed Akrem Ben Jemia

Data intake reviewer: Mohamed Derbeli, Mohamed Akrem Ben Jemia

Data storage location: https://www.kaggle.com/datasets/vkrahul/twitter-hate-speech?select=train_E6oV3IV.csv

Tabular data details:

Total number of observations	31962 rows
Total number of files	2 – train and test
Total number of features	3
Base format of the file	.csv
Size of the data	4.52 MB

Github Repo link: https://github.com/Mohamed-derbeli/Hate_Speech_Detection_Using_Transformers

