CS464: Introduction to Machine Learning
Project Proposal - Group 13

Furkan Demir 21802818, Asya Doğa Özer 21803479, Ufuk Palpas 21702958,
Lara Özyeğen 21902086, Doruk Karakaş 21901467

Project Title: Written Hate Speech Detector
Project Name: Big Brother

Description of the Question

Detecting hate speech became an issue with the increase in online content. It is mostly seen in social media posts. With a hate speech detection system, social media platforms like Twitter, YouTube, etc. can automatically interfere with hate speech posts and prevent cyberbullying. The aim of the project is to find an efficient way to automatically detect hate speech containing comments on social media platforms. We will use natural language processing techniques to analyze sentences and train our algorithm accordingly.

Description of the Data

We have selected our datasets which contain comments imputed by the users from various social media sites in English. These datasets are composed of users' remarks and their label that indicates whether it is considered to be hate speech or not.
Twitter hate speech dataset contains 29530 tweets as a train set and half of those tweets are binary labeled as containing hate speech while the other half does not. As a test set 16130 tweets are included [1].
Dynamically generated hate speech dataset has 19826 different texts as a train set and 4943 different texts as a test set which are binary labeled as offensive or not [2].
Ethos hates speech dataset has over 1000 sentences that are labeled in binary as containing hate speech or not. Ethos also has a multi-labeled dataset that classifies sentences as racist, sexist, nationalist, etc. But we will focus on binary samples as our aim is to detect hate speech, not categorize [3].

Plan and Milestones

We are planning to create and train a model with a portion of the datasets mentioned as mid-term progress and test the accuracy. Then for the final demo, our aim is to increase the accuracy of our model. Also, we are planning to merge the datasets by converting all the labels into the same format (binary) and train our model with a larger dataset for increased precision. The deliverable will be detecting whether the given text includes offensive language/ hate speech.

References

[1] "Twitter hate speech". https://www.kaggle.com/vkrahul/twitter-hate-speech?select=train_E6oV3lV.csv.
[2] "Dynamically generated hate speech". https://www.kaggle.com/usharengaraju/dynamically-generated-hate-speech-dataset.
[3] "Ethos Hate Speech Dataset". https://github.com/intelligence-csd-auth-gr/Ethos-Hate-Speech-Dataset/tree/master/ethos/ethos_data.