Project - Data collection – Milestone 4

Madhumathy Kumaran
kumaran.m@emailaddress.edu

Percentage of Effort Contributed by Student :_____97_____
Signature of Student :__Madhumathy Kumaran_____
Submission Date:___ _____Mar 5 2023_____

Twitter Sentiment Analysis – Hate and Abusive speech Analysis

Madhumathy Kumaran

Master of science in Computer Science,

Northeastern University, Seattle

***Abstract*— Any communication that disparages a target group of people based on a trait like race, color, ethnicity, gender, sexual orientation, nationality, religion, or another feature is usually referred to as hate speech. The volume of hate speech is continuously rising as a result of social media's enormous growth in user-generated content. Along with the phenomenon's effects on society, interest in online hate speech identification and the automation of this activity, has risen steadily over the past few years. This study describes a dataset of hate speech that includes thousands of words that have been manually classified as containing or not hate speech.**

I. INTRODUCTION

In recent years, various studies looked at a variety of vulgar and unfriendly expressions for people based on various race, ethnicity issues. utilizing Twitter data between October 22 and October 28 and the Tweet Binder analytics application. Prior to Musk's acquisition, the seven-day average of Tweets employing the researched hate phrases was never more than 84 times per hour. However, during the hours of midnight to noon on October 28, 2022 (shortly after Musk's acquisition), 4,778 tweets containing the hate speech in question were sent out. "The notion of loosening social media moderation has always fueled the propagation of prejudice and conspiracies. This is especially risky for young individuals using platforms, according to Bond Benton, a professor at Montclair who worked on this study and studies online extremism. The outlying problem targets the platforms with less to no lax or moderation towards hate speech and abuse and the effect it has on the users. The dataset used in this analysis is [1].

Problem Statement

To automatically detect hate speech in twitter and classify them as abusive and non-abusive and make respective models with the same.

Data Used

Dataset overview: The following data set is used https://github.com/t-davidson/hate-speech-and-offensivelanguage/tree/master/data.

Data fields: The data is stored as CSV and each data has five columns:
Count: number of users who tweeted(min is 3, sometimes more users tweeted
hate_speech = number of users who judged the tweet to be hate speech

offensive_language = number of users who judged tweet as offensive
neither = neither offensive nor hate speech
class = 0- hate speech 1 – offensive language 2 – neither tweet = comment made by the user.

Data Preprocessing and transformation

To obtain useful data, the data frame is preprocessed to remove punctuation, stop words, URLs, PoS tagging, hashtags and emojis. Further the words are tokenized, lemmatized and stemmed to obtain the most frequent words in the data.
The notebook containing preprocessing feature selection and extraction is given here[3].

2. HYPOTHESIS

This paper has a direct focus on classifying speech into abuse and non-abuse speech. This project has nearly 30000 tweets from the public repository. The data preprocessing involves two steps, Bag of words and Term Frequency Inverse Document Frequency (TFIDF). The **bag-of-words** approach is a simplified representation used in natural language processing and information retrieval. In this approach, a text such as a sentence or a document is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity [2]. **TFIDF** is a numerical statistic that is intended to reflect how important a word is to a document in a collection. It is used as a weighting factor in searches of information retrieval, text mining, and user modeling [2]. This approach can help in decreasing the hate and abuse comments or at least remove them from the public view giving a healthy environment to voice out the ideas and thoughts of the users.

.

REFERENCES

[1] https://huggingface.co/datasets/hate_speech18

[2] https://towardsdatascience.com/detecting-hate-tweets-twitter-sentiment-analysis-780d8a82d4f6

[3] http://localhost:8888/notebooks/NLP-Project/Project-Milestone3.ipynb#Feature-Extraction