Project - Data collection – Milestone 5

Madhumathy Kumaran
kumaran.m@emailaddress.edu

Percentage of Effort Contributed by Student :_____97_____
Signature of Student :__Madhumathy Kumaran_____
Submission Date:___ _____Mar 26 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.

Proposed Approach

To begin with we are trying various models trained after N-grams and TFIDF and the results are compared using the various classifier models. To begin with we split the data set into train and test dataset and train them using various Baseline Models and proposed methods :

Baseline Models: To start with we start the modeling with Bog of words models as base,

1. Naïve Bayes model,
2. Logistic Regression
3. Random Forest classifier.

Neural Models : Some of the neural models this dataset can be compared are :

1. CNN
2. LSTM and Bi LSTM
3. FastText
4. Transformers

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

In the data preprocessing stage, we combine the three datasets used for this work. The tasks involve removal of unnecessary columns from the datasets and enumerating the classes. For the third dataset, we retrieve the tweets corresponding to the tweet-ID present in the dataset. We convert the tweets to lowercase and remove the following unnecessary contents from the tweets:

• Space Pattern
• URLs
• Twitter Mentions
• Retweet Symbols
• Stopwords

We use the Porter Stemmer algorithm to reduce the inflectional forms of the words. After combining the dataset in proper format, we randomly shuffle and split the dataset into two parts: train dataset containing 70% of the samples and test dataset containing 30% of the samples.

C. Feature Extraction

We extract the n-gram features from the tweets and weight them according to their TFIDF values. The goal of using TFIDF is to reduce the effect of less informative tokens that appear very frequently in the data corpus. Experiments are performed on values of n ranging from one to three. Thus, we consider unigram, bigram and trigram features. The formula that is used to compute the TFIDF of term t present in document d is:

$$tf\ idf(d, t) = tf(t) * idf(d, t)$$

Also, both L1 and L2 (Euclidean) normalization of TFIDF is considered while performing experiments. L1 normalization is defined as:

vnorm = v |v1| + |v2| + ... + |vn|
where n in the total number of documents.
Similarly, L2 normalization is defined as:
vnorm = v p v 2 1 + v 2 2 + ... + v 2 n
We feed these features to machine learning models.

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.

Modeling

This dataset is considered using major machine learning models like Naïve Bayes, Logistic Regression and Random Forest classifier and the results are compared. The accuracy of every model is compared using the confusion matrix and the accuracy score is deduced using ruc score.

.

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