Exploratory Data Analysis

Hate Speech detection and analysis

26TH February 2022

Background

Hate speech is a growing problem online and the ability to detect it has become evermore import in the modern era. Here we construct a Natural language processing model to detect hate speech in text.

Problem Statement

The term hate speech is understood as any type of verbal, written or behavioural 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, colour, 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 a 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.

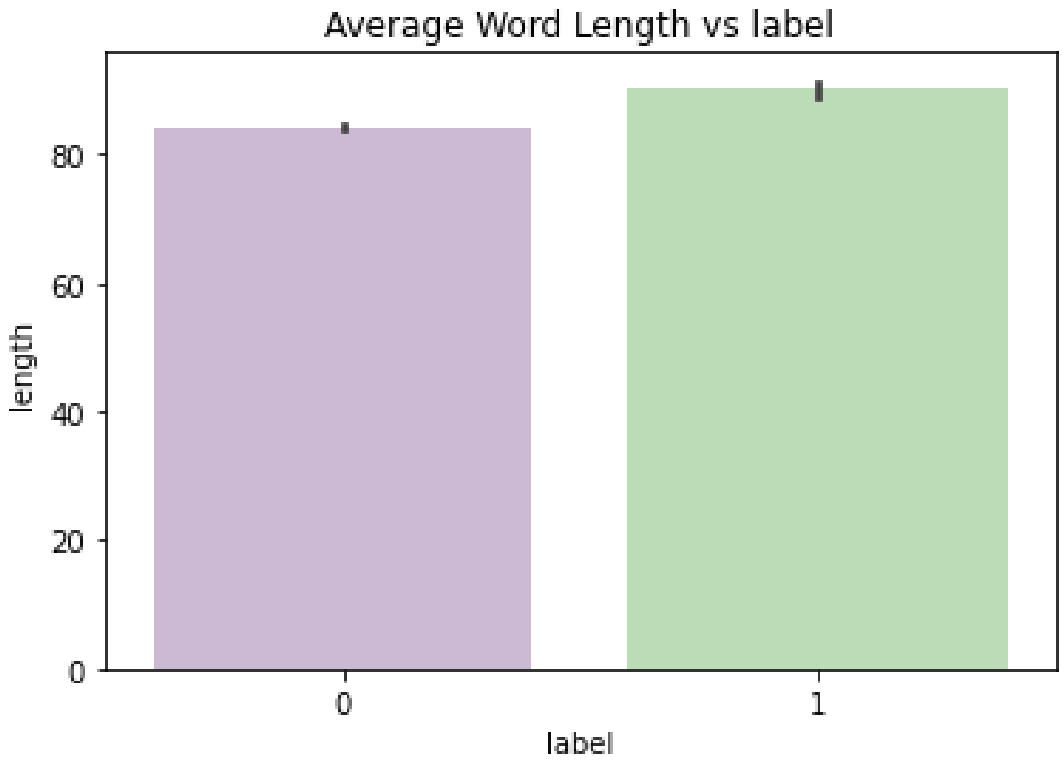
Data Analysis Approach

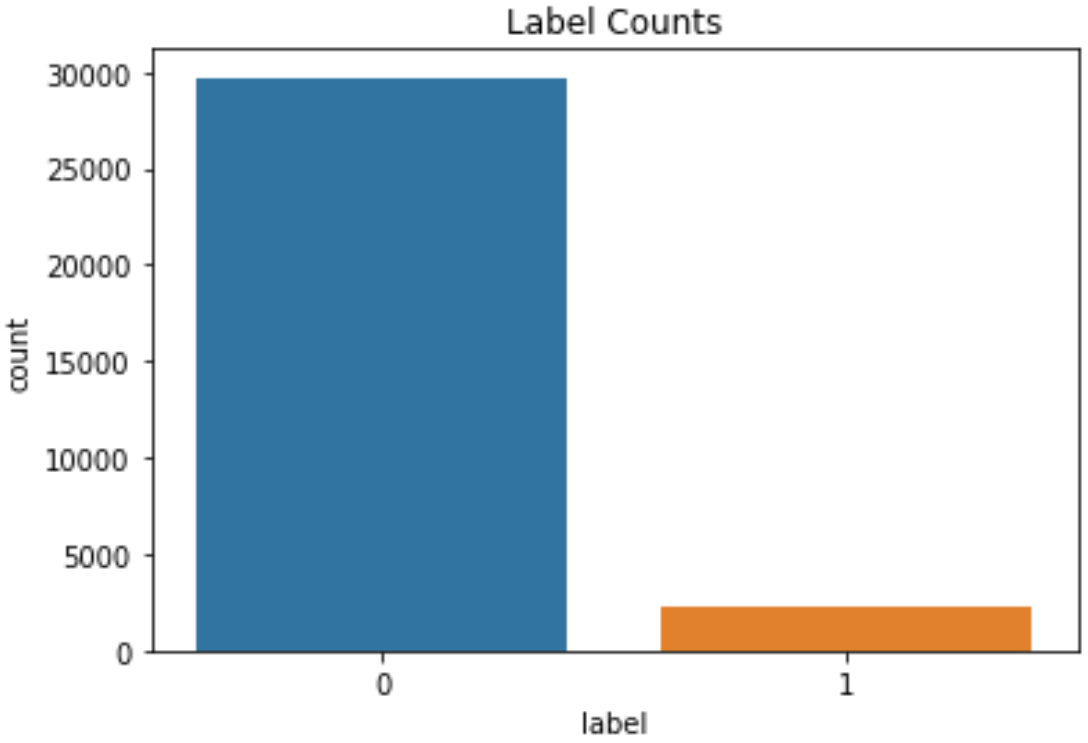
- Explore and Understand the data.
- Prepare and clean the data.
- Analyze the data and find the features/variables that are associated with hate speech
- Give recommendations for the classification model that is to be built to automate the process of hate speech identification.

Data Exploration

- One file used for the dataset
- 3,424 data points
- 75 features/variables (6 derived)

Analysis of Numerical Features





- This is the descriptive analysis of the average word frequency of positive(0) and negative (1) words
- The dataset is imbalanced as there are more positive than negative word counts.

Analysis of Categorical Features

Non-Hate Comments

Hate Comments

- The word maps show the most frequent non hate comment words versus the hate comment words

EDA Summary

- From the Exploratory Data Analysis done, we are able to find how the different features/variables affects drug persistency.

- The dataset is imbalanced as there are more positive than negative word counts.

Recommendations

For the purpose of automating the process of hate speech identification, the following machine learning models can be used:

- **Logistic regression** – It is a type of linear model that is used for binary classification. It predicts output which is a categorical dependent variable. Such predictions are like yes or no, A or B, etc.
- **XGboost**– This is high-performance gradient boosting framework based on decision tree that is used for classification.

Thank You