Hate Speech Detection

Exploratory Data Analysis and Model Proposal

09/05/23

Agenda

Problem Statement

Assumptions

EDA and Featurization

Proposed Models

Business Understanding

- Hate Speech Detection is a task of sentiment classification.
- Censor hate speech posts.
  - These aren’t in line with our policy.
  - Defined as discriminatory messages based on identity.
- Earn user’s trust as safe and accessible platform.
- Raise advertiser confidence in brand image and platform.
  - Increase ad revenue.

Dataset and Assumptions

- The data is derived from real tweets.
- The training data is labelled correctly.
- The training and test data are from the same domain.
- The amount of hate speech compared to non-hate speech reflects the proportion on the platform (see below)..
- Tweets are below the limit of 200 characters and are formatted in the same way.

test_tweets.csv

Total number of observations	16130
Total number of files	1
Total number of features	2
Base format of the file	.csv
Size of the data	1.56 MB

train.tweets.csv

Total number of observations	29530
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	2.96 MB

Dataset split into training and testing data.

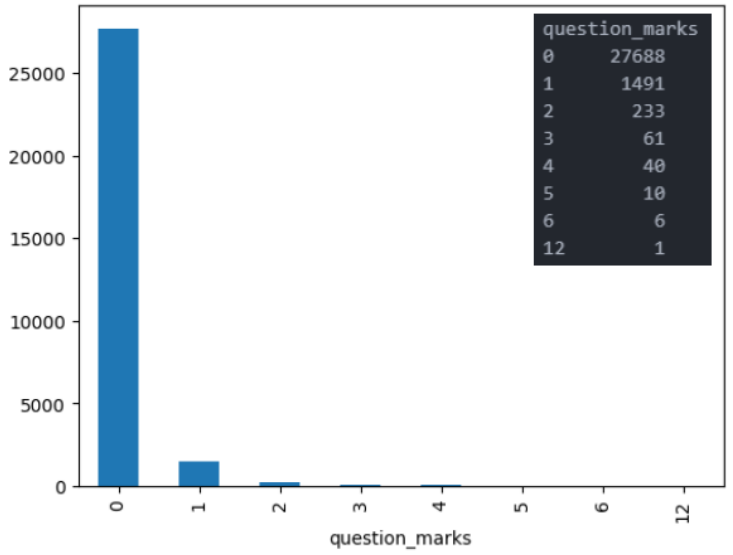
Extra Training Features

What factors are indicators of hate speech?

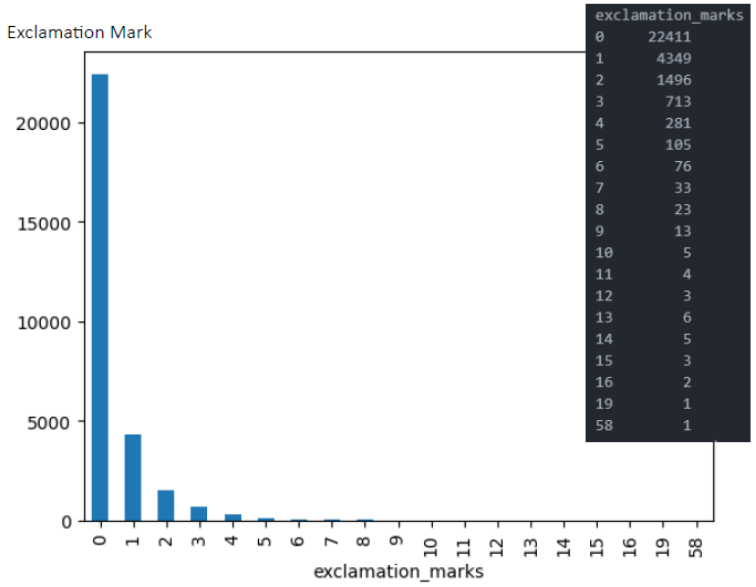
- Word count and avg. length:
  - some speech patterns are indicative of anger.
- Hashtags:
  - hashtags might be associated with hate speech.
- Exclamation marks:
  - can be an indicator of rage.
- Question marks:
  - people often use rhetorical questions to show anger.
- Uppercase usage:
  - can be an indicator of anger.
- Sentiment:
  - there might be a link between use of negative words and hate speech.

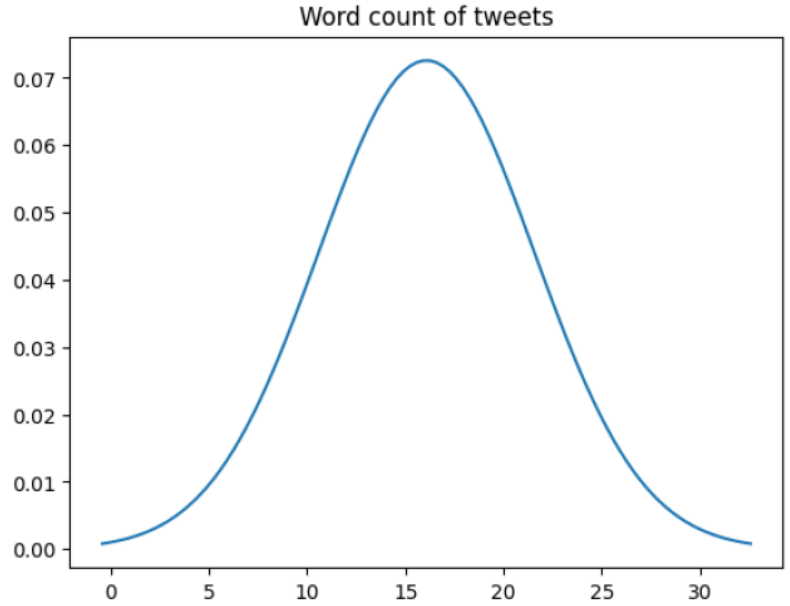
Extra Training Features

Question Marks



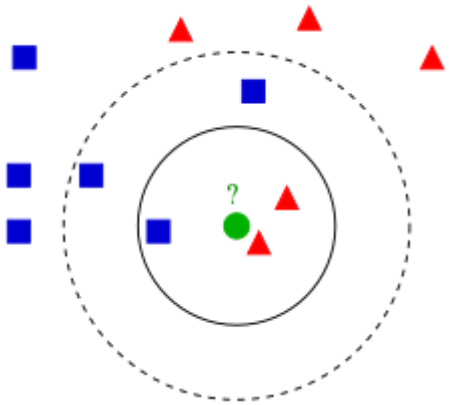
Exclamation Mark



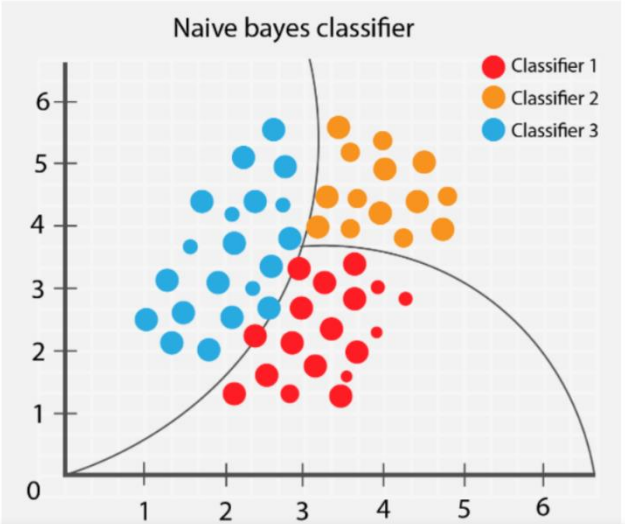


Proposed Models

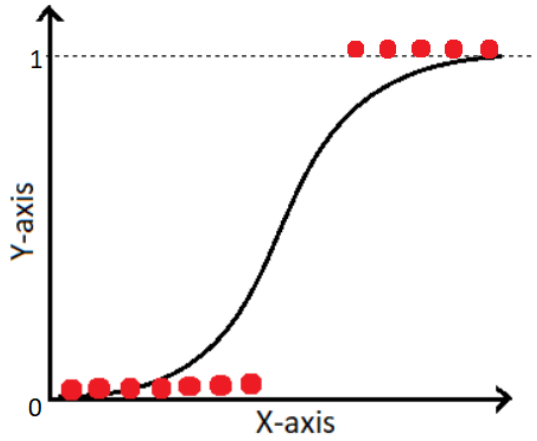
1. K-Nearest Neighbour



2. Naïve Bayes



3. Logistic Regression



Thank You