Week 11 Deliverables

Orkun Kınay

16 July 2024

Team Member's Details

- **Group Name:** Orkun
- **Name:** Orkun Kınay
- **Email:** orkunkinay@sabanciuniv.edu
- **Country:** Turkey
- **College/Company:** Sabancı University
- **Specialization:** NLP

Problem Description

The goal of this project is to develop a hate speech detection model using Twitter data. The dataset consists of tweets labeled as hate speech or non-hate speech, which will be used to train and evaluate machine learning models. The primary objective is to accurately classify tweets and mitigate the spread of hate speech on social media platforms.

GitHub Repo Link

GitHub Repository URL

EDA Presentation for Business Users

The Exploratory Data Analysis (EDA) provides insights into the dataset and helps in understanding the data distribution, patterns, and relationships. The key findings from the EDA are presented below.

Distribution of Tweet Lengths

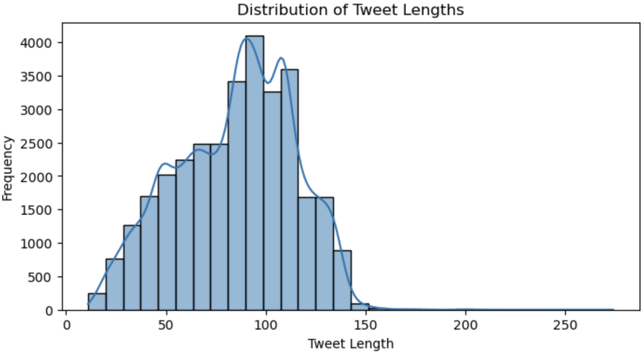

Figure: Distribution of Tweet Lengths

Class Distribution

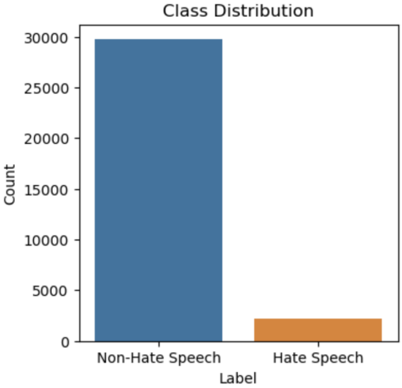

Figure: Class Distribution

Word Clouds

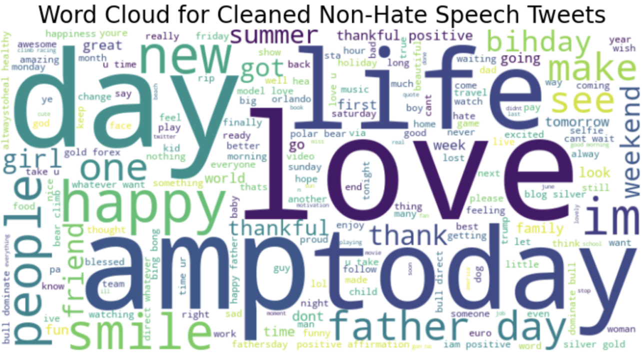

Figure: Word Cloud for Non-Hate Speech Tweets

Word Clouds

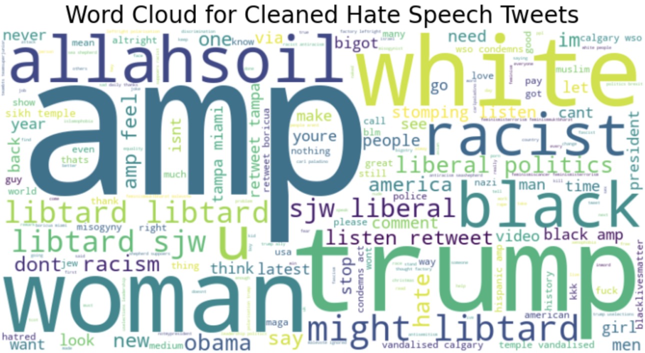

Figure: Word Cloud for Hate Speech Tweets

Statistical Analysis: Tweet Length

```
count 31962.000000
mean 84.739628
std 29.455749
min 11.000000
25% 63.000000
50% 88.000000
75% 108.000000
max 274.000000
```

Statistical Analysis: Common Words

Non-Hate Speech Tweets: Common words include:

- day (2797)
- love (2745)
- happy (1679)
- u (1578)
- amp (1325)
- life (1221)
- time (1205)
- im (1112)
- today (1069)
- get (949)
- like (948)
- positive (932)
- thankful (925)
- father (919)
- new (917)
- bihday (856)
- good (820)
- smile (812)
- make (804)
- people (790)

Statistical Analysis: Common Words

Hate Speech Tweets: Common words include:

- amp (283)
- trump (216)
- white (153)
- libtard (149)
- black (146)
- like (140)
- woman (120)
- racist (109)
- politics (97)
- people (95)
- liberal (92)
- allahsoil (92)
- u (89)
- might (77)
- sjw (74)
- new (71)
- hate (69)
- obama (68)
- retweet (67)
- dont (67)

Recommended Models for Technical Users

Based on the insights from the EDA, the following models are recommended for the hate speech detection task:

- **Logistic Regression:** A simple yet effective model for binary classification tasks. It provides interpretable results and is computationally efficient.
- **Random Forest:** An ensemble learning method that creates multiple decision trees and combines their predictions. It handles imbalanced data well and provides feature importance.
- **Gradient Boosting:** Another ensemble method that builds trees sequentially, each one correcting the errors of the previous one. It is effective for improving prediction accuracy.
- **Support Vector Machines (SVM):** A powerful classification algorithm that works well for text data. It can handle high-dimensional spaces and different kernel functions can be used for non-linear data.

Recommended Models for Technical Users

Based on the insights from the EDA, the following models are recommended for the hate speech detection task:

- **Deep Learning Models:** Using models like BERT (Bidirectional Encoder Representations from Transformers) can capture the contextual meaning of words in the tweets and improve classification performance.

Summary

The EDA has provided valuable insights into the dataset, highlighting the importance of text length and common words in distinguishing hate speech from non-hate speech tweets. The recommended models offer a range of approaches to effectively classify hate speech, with options for both traditional machine learning and deep learning methods.