Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

- Hate Speech Detection using Transformers

Nov. 2023

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

Executive Summary

- The Hate Speech Detection project focuses on leveraging deep learning, specifically the Transformer architecture, to identify hate speech in Twitter data.
- The dataset consists of cleaned tweets after going through a comprehensive preprocessing pipeline including URL removal, user mentions, special characters, tokenization, lowercase, stopword removal, and lemmatization.
- The objective is to develop a model that can effectively classify tweets into hate speech and non-hate speech.

Problem Statement

- As hate speech increases on social media platforms, especially Twitter, efforts are ongoing to protect users and make platforms safe.
- The main purpose of this project is to develop a detection model that can classify tweets containing hate speech.
- These models can help mitigate problems later by identifying and flagging potentially harmful content for review or removal.

Approach

This approach could use transformer-based deep learning models, especially BERT, which are known to be effective in sentiment detection. The model will be fine-tuned according to the training data and evaluated based on test data.

1. **Data preprocessing:** Comprehensive cleaning texts including URL and mention removal, minification expansion, tokenization, stopwords removal, lemmatization, etc.
2. **EDA:** Extensive exploratory data analysis (EDA) to understand dataset characteristics.
3. **Model Selection and building:** As a base model, a traditional model such as SVM and a Transformer model such as BERT for expression learning can be used.
4. **Model evaluation:** Perform model evaluation using metrics such as accuracy, F1 score, ROC curve.

EDA

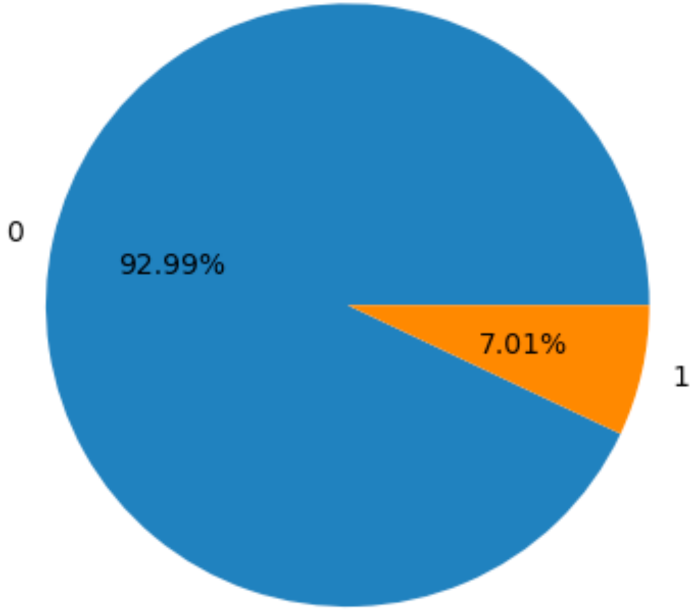
Since the tweet dataset (train_E6oV3lV.cvs) has only one feature except id and label, analysis was mainly done after data cleaning and tokenization for the tweet column rather than analysis of the dataset itself.

1. **Label distribution** : Examine the distribution of hate speech labels and non-hate speech labels to understand class balance.
2. **Text Length**: Analyze the distribution of tweet lengths and compare the lengths of hate speech tweets and non-hate speech tweets.
3. **Word Frequency**: Identify frequently used words in both types of tweets using word clouds.

EDA Summary

- **Data Exploration**:
  - ✓ Dataset: train_E6oV3lV.cvs (converted to dataframe using pandas)
  - ✓ Type of the dataset: <class 'pandas.core.frame.DataFrame'>
  - ✓ Keys: Index(['id', 'label', 'tweet'], dtype='object')
  - ✓ Data shape - (31962, 3)
  - ✓ Data types – id: int64, label: int64, tweet: object, dtype: object

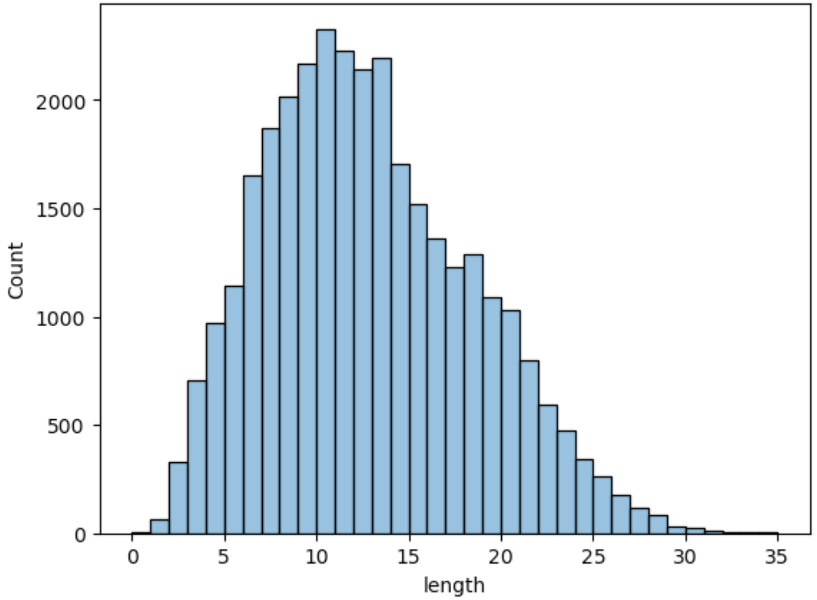
- **Label distribution**:
  - ✓ about 7% tweets are classified with hate sentiment.

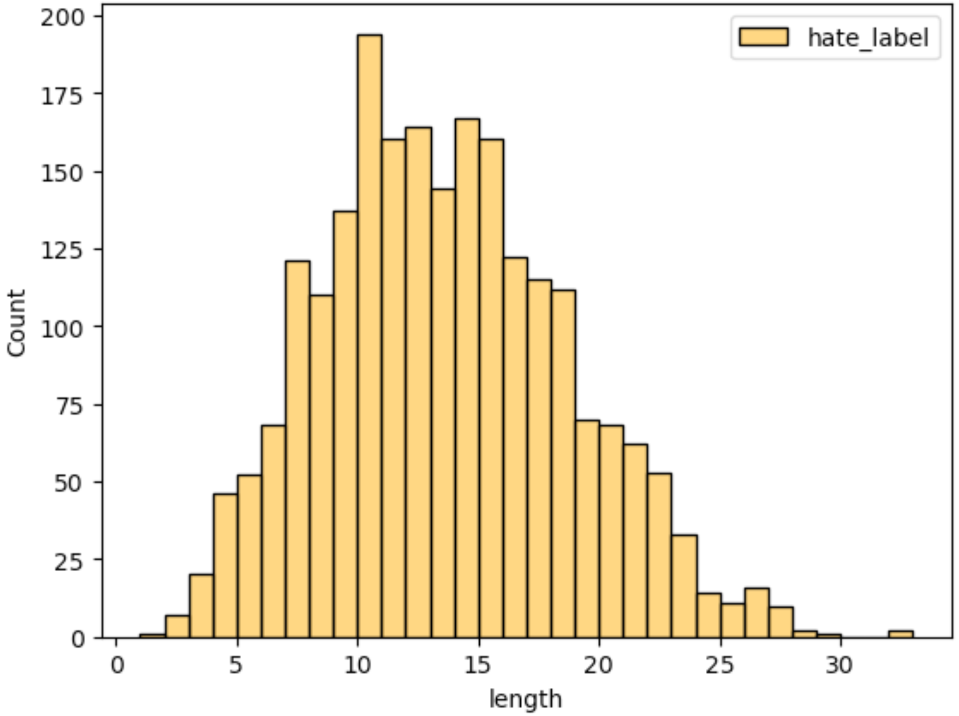


EDA Summary (cont.)

- **Text Length**: Count the number of tokens

  While the text length of hate speech tweets shows a normal distribution, in the case of all tweets, it can be seen that there are more tweets with shorter lengths. This shows that people may use more words to express hate sentiment.



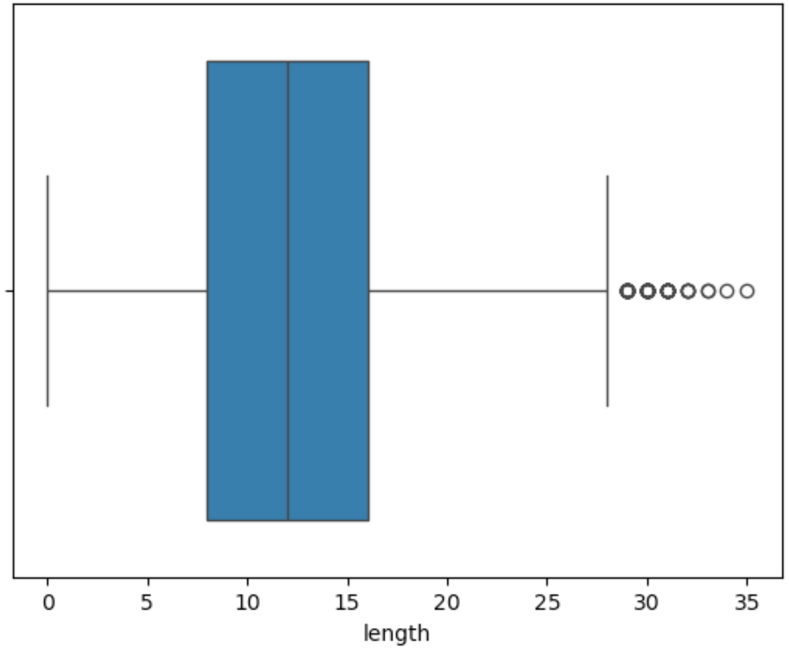


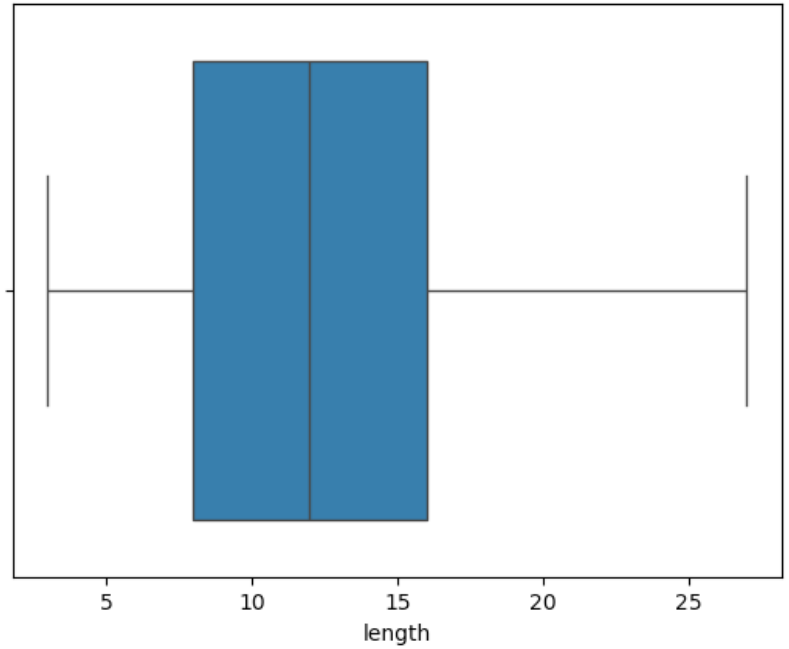
EDA Summary (cont.)

- **Text Length (cont.)**

  Outlier detection by length: Use Boxplot

  ✓ Remove outliers without length between 3 and (Q3 + 1.5 * IQR) -> 551 rows were removed.





EDA Summary (cont.)

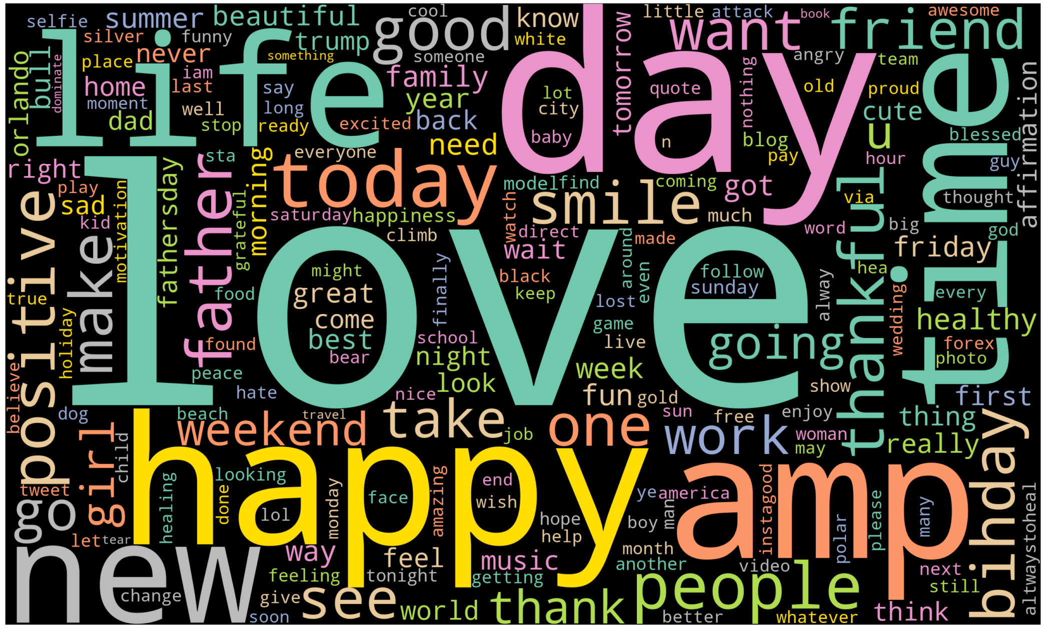
- **Word Frequency**

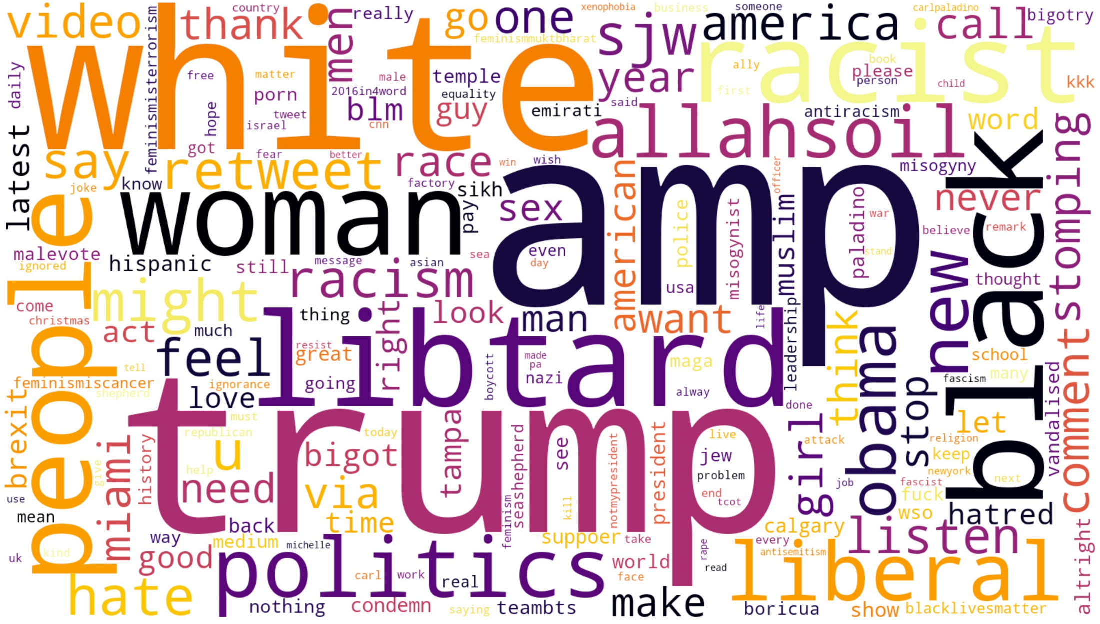
Total words' top 10

{'love': 2795, 'day': 2749, 'happy': 1691, 'amp': 1602, 'time': 1235, 'life': 1235, 'like': 1085, 'today': 1083, 'get': 1000, 'new': 988}

Hate tweets' top 10

{'amp': 283, 'trump': 216, 'white': 153, 'libtard': 149, 'black': 146, 'like': 140, 'woman': 120, 'racist': 109, 'people': 106, 'politics': 97}





Final Recommendation

1. Since hate speech tweets account for only 7 percent of all Twitter data, analysis using N-grams as well as words should be added.
2. On Twitter, hatred toward a specific event or person may become a pattern (e.g., Trump-Obama in hate speech tweets).
3. Given the complex and contextual nature of such hate speech, models such as Transformers' BERT could be used with traditional models like SVM.
4. Then, measure the performance of the hate speech detection model using appropriate evaluation metrics such as accuracy, precision, recall, and F1 score.
5. Finally, it is important to consider ethical implications and potential bias in data and model predictions. Careful validation and interpretation of model predictions are essential.