Project: Hate Speech Detection using Transformers

Name: Seoyoung Kim
Email: idellio007@gmail.com
Country: USA
College: University of Michigan
Specialization: NLP
https://github.com/syoungk7/Hate_Speech_Detection

1. Problem description

With the advent of social media, people can express their opinions at any time, regardless of time and space. However, this freedom of expression has made it possible to use derogatory or discriminatory language against individuals or groups. Racial and social discrimination continues online, too. Detecting and mitigating such negative and hate-related speech will help create a safe and inclusive online environment.

2. Github Repo link

https://github.com/syoungk7/Hate_Speech_Detection/data_preprocessing/

3. EDA performed on the data

- **Data Exploration**:

Dataset: train_E6oV3lV.cvs (converted to dataframe using pandas)

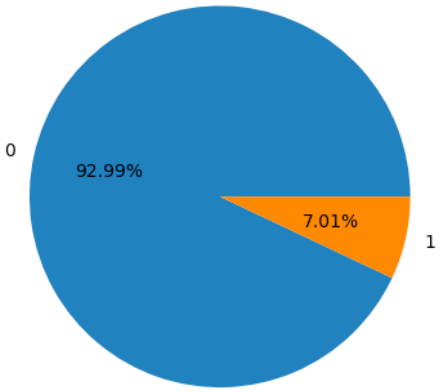
Type of the dataset: <class 'pandas.core.frame.DataFrame'>

Keys: Index(['id', 'label', 'tweet'], dtype='object')

Data shape - (31962, 3)

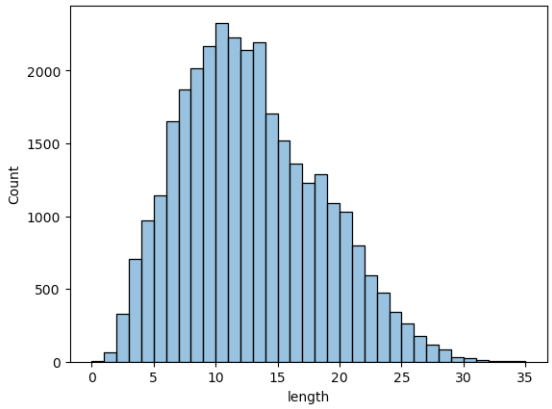
Data types – id: int64, label: int64, tweet: object, dtype: object

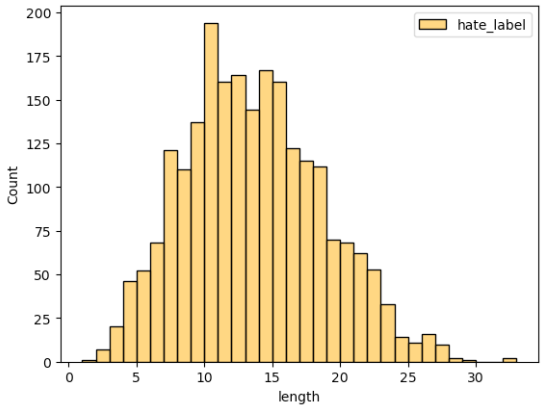
- **Label distribution**: about 7% tweets are classified with hate sentiment.



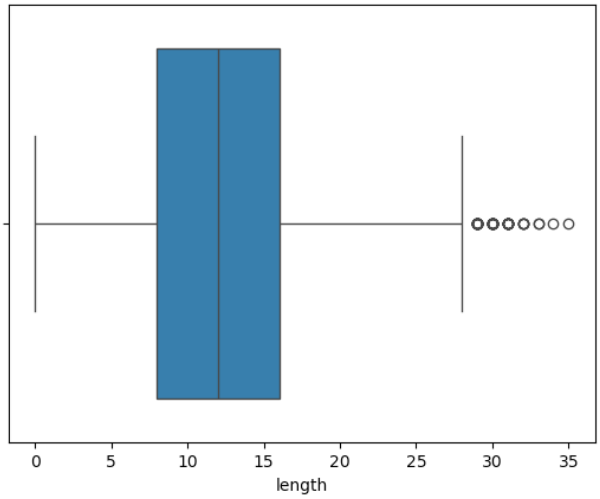
- **Text Length**: Count the number of tokens

  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.

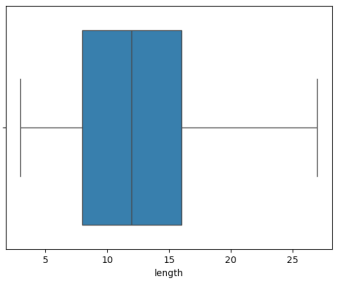




Outliers detection by length: Use Boxplot



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

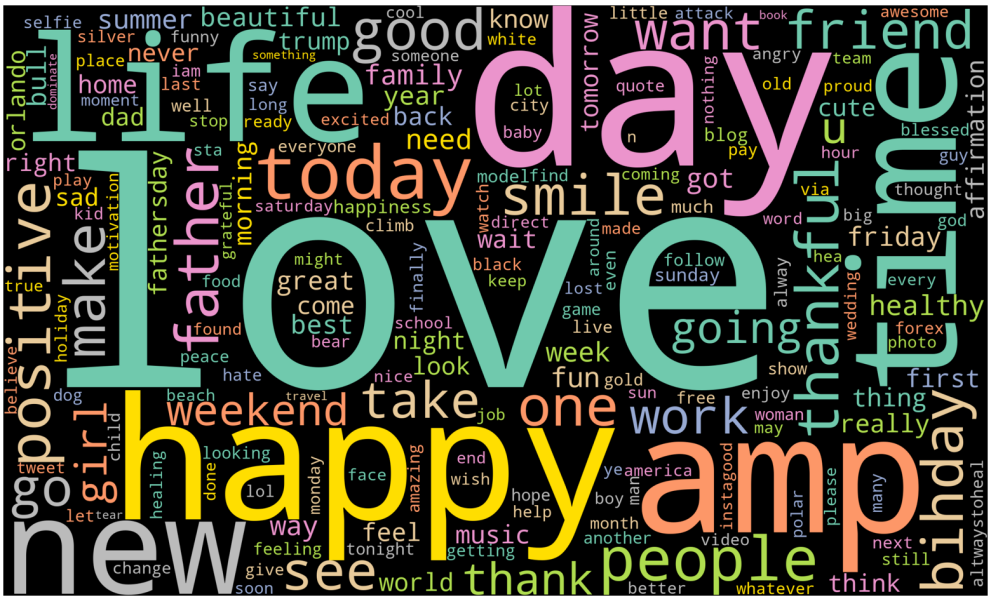


- **Word Frequency**

  Total words' top 10

  {'love': 2795, 'day': 2749, 'happy': 1691, 'amp': 1602, 'time': 1235,

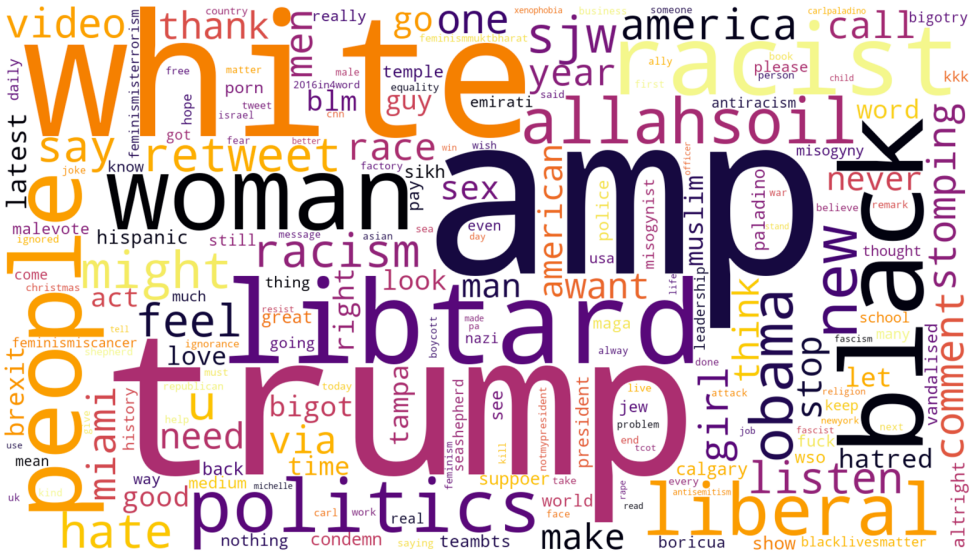
  'life': 1235, 'like': 1085, 'today': 1083, 'get': 1000, 'new': 988}



  Hate tweets' top 10

  {'amp': 283, 'trump': 216, 'white': 153, 'libtard': 149, 'black': 146,

  'like': 140, 'woman': 120, 'racist': 109, 'people': 106, 'politics': 97}



4. Final Recommendation

- Since hate speech tweets account for only 7 percent of all Twitter data, analysis using N-grams as well as words should be added.
- On Twitter, hatred toward a specific event or person may become a pattern (e.g., Trump-Obama in hate speech tweets).
- Given the complex and contextual nature of such hate speech, models such as Transformers' BERT could be used with traditional models like SVM.
- Then, measure the performance of the hate speech detection model using appropriate evaluation metrics such as accuracy, precision, recall, and F1 score.
- 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.