

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. Project Plan

- ~~1. **Understand the problem:** We understand the nature of hate speech, its impact, and the importance of addressing it on online platforms. We define the problem as a sentiment classification task based on labeled Twitter data.~~
- 2. Data cleaning and normalization:** Preprocess the dataset by cleaning and normalizing the text data. This includes handling noise, removing irrelevant information, and ensuring consistency in the representation of tweets.
- 3. Expression learning:** We leverage Transformer-based architecture for representation learning. Transformers have achieved significant success in natural language processing tasks and are well-suited to capturing contextual information in text data.

4. **Model building and training:** We develop a deep learning model for hate speech detection using preprocessed data. The model is trained on the training data set to optimize accuracy and minimize false positives and false negatives.
5. **Performance evaluation and reporting:** Evaluate model performance on the test dataset using relevant metrics such as accuracy, precision, recall, and F1 score. Provides a comprehensive report on the strengths and limitations of the model.
6. **Model Deployment:** Deploy the trained model into a production environment, ready to integrate with online platforms to actively detect and filter hate speech.
7. **Model inference:** By implementing the model's inference mechanism, we can classify new tweets in real time to identify and flag potential instances of hate speech.

3. Data Understanding

- What type of data you have got for analysis
 - a. The provided dataset for hate speech detection from Twitter consists of two files: test_tweets_anuFYb8.csv and train_E6oV3lV.csv.
 - b. The data attributes include a label (0 or 1) and text_format, which contains the original tweets with noise.
- What are the problems in the data (number of NA values, outliers , skewed etc.)
 - a. Missing Values (NA): no
 - b. Outliers in Text Length: no
 - c. Class Imbalance: no
 - d. Noise in Text: *need to handle*
- What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc. and why?
 - a. **Text Cleaning and Normalization:** Implement text cleaning techniques to remove noise, special characters, and irrelevant information. Normalize the text by converting to lowercase, handling contractions, and addressing variations in language.

- b. **Tokenization and Padding:** Utilize tokenization to convert text into smaller units suitable for input to Transformer models. Apply padding to ensure uniform input lengths, which is crucial for efficient model training.
- c. **Exploratory Data Analysis (EDA):** Conduct a comprehensive exploratory data analysis to understand the distribution of key variables, detect anomalies, and make informed decisions on data preprocessing strategies.
- d. **Feature Engineering:** Consider additional feature engineering if necessary, such as extracting information from tweet timestamps, user mentions, or hashtags, to enhance the model's ability to capture relevant patterns.
- e. **Validation and Cross-Validation:** Use validation sets and cross-validation techniques to assess the model's performance robustly, ensuring it generalizes well to unseen data.

4. Github Repo link

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