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. Business understanding

Social media platforms such as Twitter often cause social problems because online platforms are highly contagious. Detecting hate speech and identifying and reviewing or removing potentially harmful content can help alleviate these problems.

From a business perspective, these systems will be able to provide users with a safe social media environment, which will not only increase user satisfaction and engagement, but will also automatically improve the reputation of the platform.

3. 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.

4. Data Intake Report

Name: Hate Speech Detection using Transformers
Report date: Nov. 2023
Internship Batch: LISUM25
Version:<1.0>
Data intake by: Seoyoung Kim
Data intake reviewer:

Data storage location: https://www.kaggle.com/datasets/vkrahul/twitter-hate-speech?select=train_E6oV3lV.csv

Tabular data details:

1. train_E6oV3lV

Total number of observations	31962
Total number of files	1
Total number of features	3
Base format of the file	csv
Size of the data	3.1 MB

2. test_tweets_anuFYb8

Total number of observations	17197
Total number of files	1
Total number of features	2
Base format of the file	csv
Size of the data	1.6 MB

Proposed Approach:

- Mention approach of dedup validation (identification)

1. Explore dataset to identify potential duplicates. Errors during data collection, preprocessing, or merging datasets can result in duplication.

2. When possible, unique identifiers are used such as Tweet ID or User ID, to flag or remove exact duplicates.

3. Implement a text-based deduplication approach to identify and process tweets containing identical or very similar text content. You can use techniques like the cosine similarity.

- Mention your assumptions (if you assume any other thing for data quality analysis)

1. Assume that the provided label (0 or 1) for hate speech is accurate and assigned based on a reliable annotation process.

2. Assume that the text_format properties of the training and test datasets follow a consistent format and that all variations are within acceptable ranges.

3. Assume that common types of noise in Twitter data (e.g. hashtags, mentions, emojis) have been properly handled during preprocessing.

4. Assume that the dataset is representative of real-world Twitter data in terms of language usage, topics, and demographics to ensure the generalizability of the model.

5. Assume that the training and testing datasets are consistent in terms of data distribution, noise, and characteristics, allowing the model to generalize well.

5. Github Repo link

https://github.com/syoungk7/Hate_Speech_Detection