ToxicityProcessingTraining

A repo for learning how to compute toxicity scores. The model used for computing toxicity is https://github.com/unitaryai/detoxify.

How to run

1. pip install requirements.txt: `pip install -r requirements.txt`
2. run `python toxicity.py` with arguments found in Command Line Arguments section below.

Example: `python toxicity.py --input=C:/Users/user/Documents/GitHub/ToxicityProcessingTraining/sample.txt --output=C:/Users/user/Documents/GitHub/ToxicityProcessingTraining/output.csv --model=original`

Command line arguments

`--input` : The input file. Must be a full path to a txt file. For example, `--input=C:/Users/user/Documents/GitHub/ToxicityProcessingTraining/sample.txt` .

`--output` : The output file. Must be a full path to a csv file. For example, `--output=C:/Users/user/Documents/GitHub/ToxicityProcessingTraining/output.csv` .

`--model` : The type of model to use. Must be one of `original` , `unbiased` , or `multilingual` .

For more information on model choices, see Model Choices below.

What is toxicity?

Toxicity can be defined as the quality of being very harmful or unpleasant in a pervasive or insidious way. For social media posts, this is often the case when the post is attacking or brining down a certain person or group.

Computational toxicity is a measure of how toxic a comment is. The computational toxicity score is a probability between 0 and 1; 0 meaning the comment is not toxic, and 1 meaning the comment is toxic.

Detoxify Model

Source: https://github.com/unitaryai/detoxify

The Detoxify toxicity model was designed to predict toxic comments on 3 Jigsaw challenges: Toxic comment classification, Unintended Bias in Toxic comments, Multilingual toxic comment classification.

Model Choices

original

predicted labels: 'toxicity', 'severe_toxicity', 'obscene', 'threat', 'insult', 'identity_hate'

unbiased

predicted labels: 'toxicity', 'severe_toxicity', 'obscene', 'identity_attack', 'insult', 'threat', 'sexual_explicit'

multilingual

predicted labels: 'toxicity'

The multilingual model has been trained on 7 different languages so it should only be tested on: english, french, spanish, italian, portuguese, turkish or russian

Limitations and Ethical Considerations

If words that are associated with swearing, insults or profanity are present in a comment, it is likely that it will be classified as toxic, regardless of the tone or the intent of the author e.g. humorous/self-deprecating. This could present some biases towards already vulnerable minority groups.

The intended use of this library is for research purposes, fine-tuning on carefully constructed datasets that reflect real world demographics and/or to aid content moderators in flagging out harmful content quicker.

Some useful resources about the risk of different biases in toxicity or hate speech detection are:

The Risk of Racial Bias in Hate Speech Detection Automated Hate Speech Detection and the Problem of Offensive Language Racial Bias in Hate Speech and Abusive Language Detection Datasets

*Source: https://github.com/unitaryai/detoxify