

Comment classification NLP Project

Discussing things you care about can be difficult. The threat of abuse and harassment online means that many people stop expressing themselves and give up on seeking different opinions. Platforms struggle to effectively facilitate conversations, leading many communities to limit or completely shut down user comments.

One area of focus is the study of negative online behaviors, like toxic comments (i.e. comments that are rude, disrespectful or otherwise likely to make someone leave a discussion).

In this project, you're challenged to build a multi-headed model that's capable of detecting different types of toxicity like threats, obscenity, insults, and identity-based hate. You'll be using a dataset of comments from Wikipedia's talk page edits. The model you will create will hopefully help online discussion become more productive and respectful.

Disclaimer: the dataset for this competition contains text that may be considered profane, vulgar, or offensive.

DATASET

You are provided with a large number of Wikipedia comments which have been labeled by human raters for toxic behavior. The types of toxicity are:

toxic

severe_toxic

obscene

threat

insult

identity_hate

You must create a model which predicts a probability of each type of toxicity for each comment, using Deep Learning and NLP techniques.

File descriptions

train.csv - the training set, contains comments with their binary labels

test.csv - the test set, you must predict the toxicity probabilities for these comments. To deter hand labeling, the test set contains some comments which are not included in scoring.

test_labels.csv - labels for the test data; value of -1 indicates it was not used for scoring;