Predicting Toxic Wikipedia Comments

Data Scientist Role Test
Project X LTD
Andrew Nagel
January 9, 2023

Overview

The objective of this work was to utilize a dataset of Wikipedia comments that have been labelled for toxic behaviour to create a model that predicts the probability of each type of toxicity for a given comment. Each comment was human-labelled with 0 to 6 of the toxicity classifiers (i.e. toxic, severe toxic, obscene, threat, insult, identity hate). The dataset was first cleaned by removing new lines, non-characters, capitalization, and stop words in the Wikipedia comments. Next, the comments were transformed into TF-IDF vectors for subsequent training of a logistic regression model. The dataset was then split, such that 75% of it was utilized for training and 25% for testing. Since the dataset contains multi-targeted classifiers, a logistic regression model was simply trained independently for each of the 6 toxicity classes. The model was then used to predict class labels on the testing set, and confusion matrices were generated for each toxicity class to convey model performance.

Results

The confusion matrices plotted in Fig. 1 illustrate that the logistic regression model classifies the testing set to at least 95% accuracy for each of the toxicity comment types. However, there is a large class imbalance in the data set, with a much larger distribution of non-toxic comments than toxic ones. Thus, the model actually classifies the non-toxic comments to a high degree of accuracy but fails to classify the toxic comments to an acceptable level of error. For instance, when testing for severe toxic comments, the model correctly predicted 93 severe toxic comments but missed 306 severe toxic comments. This corresponds to a detection of only 23% of the severe toxic comments. The regression model performed the best on obscene comments, correctly predicting around 69% of this class in the testing set. To reduce these errors in the regression model, resampling of the training set could be conducted by either under-sampled the non-toxic comments or over-sampled the non-toxic comments to balance out the training set.

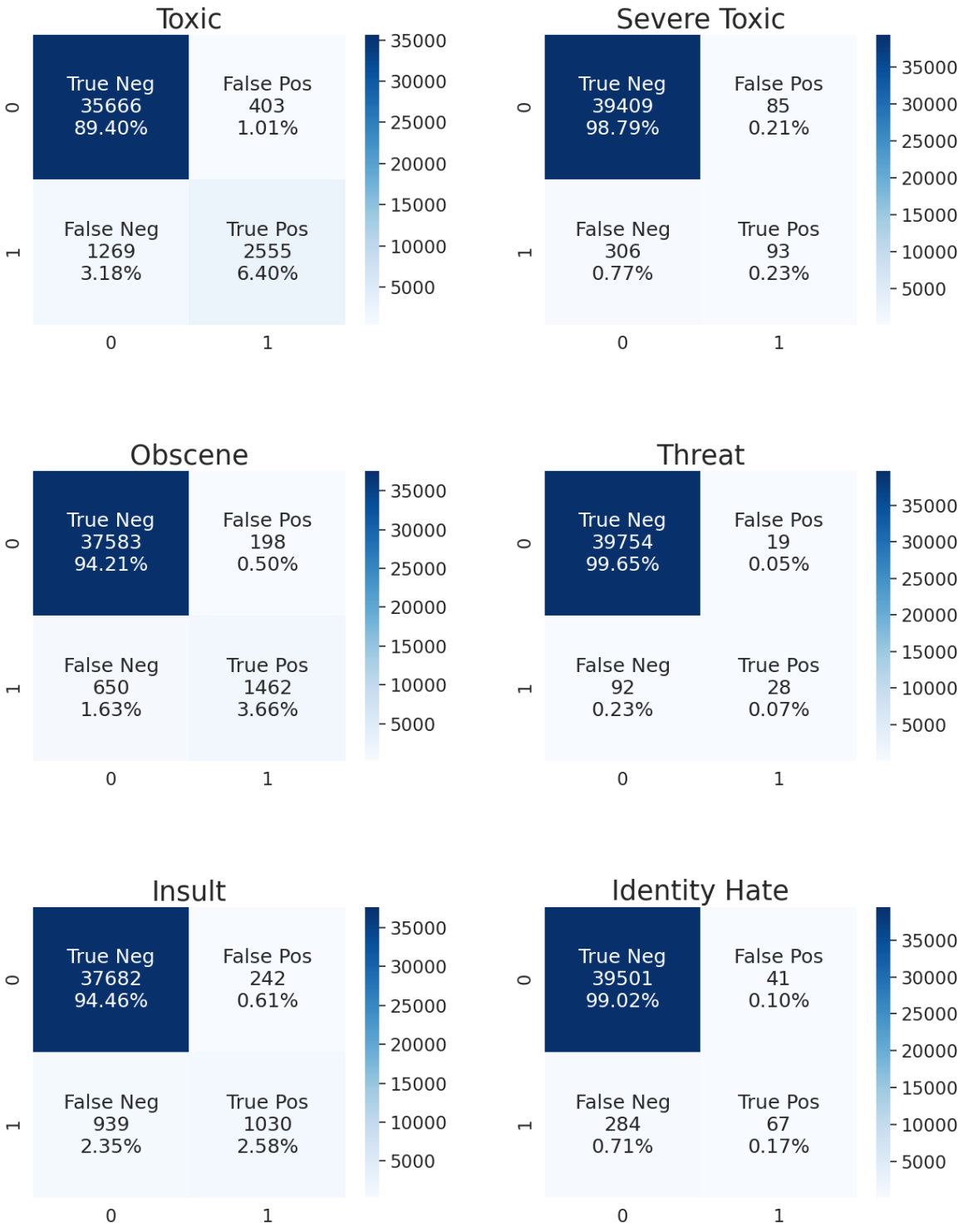

Figure 1: Confusion matrices illustrating the regression model's performance when classifying types of toxicity in Wikipedia comments.