Comment classification for offensive/toxic comments
Machine Learning for Natural Language Processing 2020

Ralph Saidi
ENSAE
ralph.saidi@ensae.fr

Abstract

I take on the challenge of classifying Wikipedia comments along 6 possible hateful speech dimensions, we train a bag of words model with a random forest classifier as well as a continuous bag of words model with the fastText embedding matrix.

1 Problem Framing

The internet is arguably the most revolutionary invention of our lifetimes. However when you put this magical tool in the hands of peole who have a lot of free time and a hint of anonymity, a lot of them will inevitably resort to expressing hateful speech. As a consequence, it is useful to have an automatic process to sort out the hate, without having someone go through every reported comment. This is the challenge I decided to take on. I have procured a dataset containing more than 150K comments labeled along 6 dimensions: toxic, severe toxic, obscene, threat, insult, identity hate.

2 Experiments Protocol

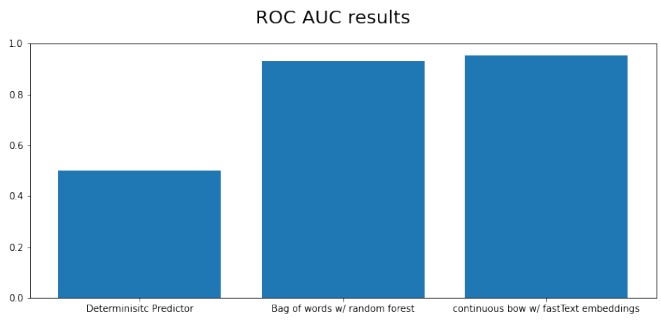
The first stap was to reduce the size of the training data by dividing the original dataset by 20 and using the remaining data as the test data. The training data constitutes about 8000 comments and the test data about 150 000. This was done in order to reduce the computing time, after all I'm mainly interested in comparing algorithms with the same amount of training.

The first baseline model was a deterministic predictor that is completely desensitized, i.e consider all comments as safe. This shouldn't yield horrible accuracy since most comments are safe. The second was a bag of words model with a random forest model. Lastly, we have a continous bag of words model with a fastText embedding matrix.

The algorithms are trained and compared along their ROC AUC scores.

3 Results

The first deterministic predictor has a ROC AUC of 0.5, which basically means it cannot deduce anything on that data, nothing surprising here. The bag of words with a random forest has ROC AUC of 0.93 and finally, with the fastText embedding matrix we get a score 0.95 Having had the intention of implementing BERT for this problem, I was surprised that a simple bag of words on a relatively small training sample could perform so accurately.



4 Discussion/Conclusion

I've always been skeptical when it comes to AI and language, especially with this kind of problem. On one hand, the complexity of language allows for a wide range of variability along many dimensions, including sarcasm and humor, things that us humans still don't have a firm grasp on. On the other hand, relying on such algorithms without having balanced data in terms of subjects, could perpetuate the same social biases and lead to more discrimination.

.

References