Offensive Comment Classification

Ryan Scovill, Sina Zareian and Marcelo Paco Zepeda

Simon Fraser University, Department of Computing Science

{rscovill, szareian, mpacozep} {@sfu.ca}

Introduction

- ➢ Profanity and disrespectful comments online can become a serious issue if not tackled, especially with the ability to be anonymous.
- ➢ With the increasing surge of social media users, the moderation task cannot be done manually.
- ➢ The classification task is to correctly classify obscene comments
- ➢ Malicious users may find clever ways around moderation by using variations of spellings or character replacements, so we will address this by simulating noise

Main Objectives

- ➢ Add “noise” to the dataset to mimic intentional misspellings
- ➢ Explore models to improve accuracy of the baseline model
- ➢ Implement a new model to improve the accuracy of the new noisy baseline

Data

- ➢ The training dataset provided by Jigsaw [1] contains 160k of Wikipedia comments classified as toxic, severe_toxic, obscene, threat, insult, or identity_hate via crowdsourcing [2]
- ➢ The test set contains 153k comments

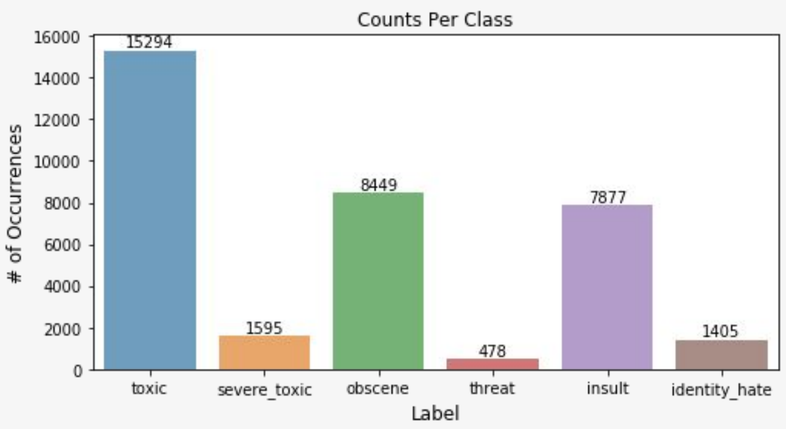

Figure 1: Distribution of class labels

Adding Noise

- ➢ Add optical character recognition (OCR) error on training set
  - Specifically OCRAug from nlpaug package
- ➢ OCRAug replaces target characters with predefined mapping table
- ➢ Example of “clean” vs noisy sentence:
  - Clean:
    “The quick brown fox jumped over the lazy dog”
  - Noisy:
    “The qoick 6rown f0x jomped ovek the la2y d0g”

Model

Denoising model

- ➢ Concatenate the 300-dim one-hot vector with the 50-dim embedding matrix
- ➢ Applied to a bidirectional LSTM Model with 2 epochs
- ➢ Used a stochastic gradient descent optimizer (SGD)

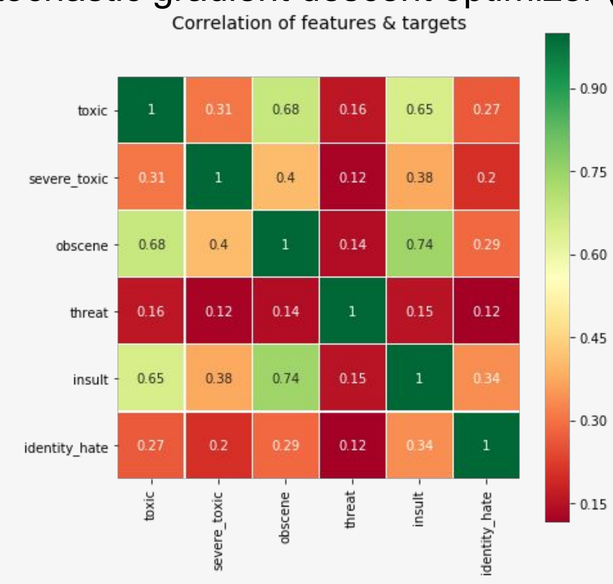

Figure 2: Demonstrates relationships between labels

Results

- ➢ Using our model, we increased the noisy baseline accuracy by 3.0%.
- ➢ All the accuracy measurements are done using Kaggle’s mean column-wise ROC AUC metric.

	Baseline	Improved
Baseline w/Noise	0.685	**0.715**

Conclusions

- ➢ By introducing noise into the dataset we created a new baseline in which we slightly improved accuracy
- ➢ Demonstrates our model can be used for better classifying offensive comments with noise

References

[1] Toxic Comment Classification Dataset
https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data
[2] E. Wulczyn, N. Thain, and L. Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In Proceedings of the 26th International Conference on World Wide Web (WWW '17), 1391-1399. DOI: https://doi.org/10.1145/3038912.3052591

Acknowledgements

We would like to acknowledge Anoop Sarkar for his expertise and guidance in this project.