Toxicity Classification - Project Proposal

Colan Biemer — Avinash Guledagudda — Manisha Sharma

A. Introduction

Fifty years ago we did not have the omnispresent problem of cyberbullying. It can take place over SMS, apps, social media, forums, gaming, and more [1]. It can be harmful and dangerous. As a result, it is in our best interest to investigate it and find potential solutions. In this project, we propose to use a toxicity data set, which we explore in section -C, to train a toxicity scoring model. We intend to use a pre-trained model that we will fine tune with a single output layer to score toxicity of a sentence or set of sentences, see section -D for details. In addition, we will be training several baseline models to compare to the pre-trained model. Finally, we will take the best model and build a web server that allows users to input a twitter name. The server will take the most recent tweets and return an estimate of how toxic the user is, see section -E.

B. Related Work

Jigsaw opened a Kaggle Competition [2] which focused on toxicity classification and it had 2,624 teams participate. In addition to the competition they hosted, they have an open API called Perspective [3] which can be called to rate the toxicity of a given sentence. Details on their model and its abilities are not available.

Outside the context of toxicity, we have a task that is not classification [1] which would have a discrete number of possible classes (e.g. not toxic, mildly toxic, and toxic). Instead, we aim to rate the input on a continuous scale between 0 and 1, where 1 would be the most toxic. This is to our advantage in terms of usability since Belz and Kow found that a continuous scale is preferred by human raters due to the additional information provided [2]. In terms of potential accuracy, the task is likely to be more difficult in continuous space.

C. Data Set

The data set provided by Kaggle is a csv file containing 45 columns. The first column is the id with the second being the target score for the toxicity. Following that is the input text itself. Afterwards, what follows is a long list of scores relating to groups and individuals that tend to be unfairly biased against in models. For example, take the word "gay." This word on the internet tends to be used as an insult but a sentence like, "I am a gay man" should not receive a toxic score. However, due to unintended bias in models this sentence will typically receive a negative score. A goal in using these extra fields is to reduce unintended bias.

D. Models

As an experiment we ran a naive bayes classifier with a bag of words [3] implementation as a baseline approach. In addition we simplified the expected output to three classes of not toxic, mildly toxic, and toxic. We found that our baseline classifier had under 52% accuracy [4]. Therefore, we believe that to get decent results, a state of the art model must be trained.

To remedy our lack of resources, we will use BERT [4] a pre-trained model that we can fine-tune with a single output layer. We use a single output layer instead of multiple since recent work shows that adding additional layers after BERT will result in minimal gains, and a heavy increase training time [5].

To evaluate our fine-tuned BERT model, we will not only compare it to Kaggle participants but we will also train several other baseline models. At the moment we intend to look into logistic regression and support-vector machines [6]. Depending on time, we will look to add more baseline models such as a RNN [7].

E. Server

We intend to make our model available publicly through a website, hosted with Heroku [5], where users can input the username of a Twitter user. The server

[1] https://www.stopbullying.gov/cyberbullying/what-is-it/index.html

[2] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

[3] https://www.perspectiveapi.com/#/home

[4] https://github.com/bi3mer/CS6120_NLP_Project/blob/master/naive_bayes.ipynb

[5] https://www.heroku.com/

will collect recent tweets through Twitter's REST API [6] with a Python library called Tweepy [7]. The server will collect the most recent tweets that the user has made and rate the toxicity and return a mean across the set of retrieved tweets.

We intend for this to be a fun way to see how toxic public figure and friends can be. We do, however, hope that users will also view their own toxicity scores. If their toxicity scores are high, then hopefully they will reflect and try to lower their scores.

REFERENCES

[1] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, pp. 3–24, 2007.

[2] A. Belz and E. Kow, "Discrete vs. continuous rating scales for language evaluation in nlp," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, 2011, pp. 230–235.

[3] J. H. Martin and D. Jurafsky, *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall Upper Saddle River, 2009.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[5] S. Gupta and E. Hulburd, "Exploring neural net augmentation to bert for question answering on squad 2.0," *arXiv preprint arXiv:1908.01767*, 2019.

[6] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

[7] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh annual conference of the international speech communication association*, 2010.

[6] https://developer.twitter.com/en/docs

[7] https://www.tweepy.org/