CSCI 8360 final project proposal: Jigsaw Unintended Bias in Toxicity Classification

Abolfazl Farahani, Jonathan Myers, Saed Rezayi

April 2019

1 Synopsis

The continuing alteration of modern society through its rapid embrace of social media fuels rising concerns regarding uncertainty and potential regulation. Bias remarks serve as a prime example where commonality centered algorithmic approaches for categorizing people compounds underlying bias. The Kaggle problem for this project looks to automate recognition of such biases in remarks. We will take advantage of the open status of this Kaggle competition to help evaluate our progress.

2 Introduction

The Jigsaw bias classification problem started on March 29, 2019 and will continue to accept submissions till June 19, 2019. The work for this Kaggle completion came from the Conversation AI team at Jigsaw and Google using data gathered, evaluated, and publicly released by the Online Hate Index Research Project in the D-Lab at UC Berkeley (no publication citations on Kaggle). For this challenge, we will need to take sentences and derive a score between 0 and 1, where 0 implies nothing offensive and 1 implies very offensive – a measure they label as "toxicity".

The authors of the competition quickly point out previous work failed due to independent work evaluation. Basic identity terms, such as "man" or "woman", don't intrinsically carry offensive meanings, but their use could easily shift implication from benign to offensive. Simply adding the word "not" could up end the entire sentences' meaning and/or interpretation.

Classification problems like this have been around for countless years in computer science, and, as such, this problems give us an opportunity to look into ways classification algorithms have been used before, what data science methodologies would be applicable for analyzing the classification results, and use the test set supplied by Kaggle to evaluate possible approaches for deriving "toxicity". The following sections discuss how we will determine which approaches for solving this problem are worthy of evaluation, how we will enact software to solve this problem, and how we will present the final results as our course project.

3 Research Plan

The Latent Dirichlet Allocation (LDA)[Blei et al., 2003] of 2003 now has 26,196 citations and serves as a highly influential classification algorithm. The three-level hierarchical Bayesian model provides probabilities for the inclusion of text provided into different categories, or "topics". One good research

approach would be to look into the many ways people have used LDA for word, segment, and sentence classification over the years – there likely exist many freeware options, and that will be an initial research goal.

The Kaggle problem also provides percentages for inclusion of topics (such as "male", "female", "christian", etc.) in the statements – only for the training set, submission only asks for toxicity. We could include the percentages supplied for building the LDA and explore incorporating their predictions into the LDA's prediction of toxicity.

Deriving the sentiment of sentences has been a computer science challenge for decades, providing many opportunities to evaluate approaches built by others. T Wilson's derivation on contextual polarity from phrase analysis in 2005 received 2,929 citation and appears to build on a statistical approach for predicting polarity, which we could use to derive toxicity[Wilson et al., 2005]. Y Kim's application on CNN to classify sentences in 2014 received 3,909 citation and would serve as a helpful guide for determining an appropriate manner for applying CNN to sentences[Kim, 2014].

We also have access to the book Sentiment Analysis and Opinion Mining, by Bing Liu in 2015[Liu, 2015]. The books supplies a wide range of techniques that target sentence level sentiment analysis and it should arrive at UGA by April 9, 2019. Although it might not supply the best approach, like many textbooks, it certainly provides a wide list of ideas and starting points for further research.

4 Software Plan

As alluded to in the research section, the number of possible options will continue to grow as we explore the works of others, but actual implementation and evaluation of software is the fundamental goal of this project. As such a pragmatic approach would be to look for the techniques cited in publications that have practical implementations. For example, algoritmia.com has implementation of LDA and LDA Mapper (with example usage https://blog.algorithmia.com/lda-algorithm-classify-text-documents/).

Others have implemented Python software specifically for sentiment analysis. C. J. Hutto and Eric Gilbert of Georgia Tech published the Valence Aware Dictionary and sEntiment Reasoner (VADER) approach along with implementation in Python, Java, C#, and other languages [Hutto and Gilbert, 2014] (https://github.com/ cjhutto/vaderSentiment). In 2016, Tensorflow released SyntaxNet, a Natural Language Understanding (NLU) system, which provides accurately parsed sentence structure – another block of data we could use to derive intent (https://github.com/tensorflow/models/tree/master /research/syntaxnet). Although the Concept Checker (ConCH - https://github.com/clips/conch) focuses mainly on clinical text, the authors build a model that maps multiple word sequences and orderings to concepts, which may allow us to constitute for how basic word "chunks" carry different intent[Tulkens et al., 2019]. There also exist tutorials and multiple Python libraries that implement Latent Semantic Analysis.

In the end, we will generate documents and software as a GitHub project, just like the past three projects in CSCI 8360. We intend to provide Python code that takes the Keggle data and generates predictions. Science intrinsically finds value in facts, so we will provide access to techniques that work, those that don't, and options/combinations for further experiments. We will also provide the exact details regarding how we generated results along with instructions to ensure other users will generate the same results.

References

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[Hutto and Gilbert, 2014] Hutto, C. J. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*.

[Kim, 2014] Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

[Liu, 2015] Liu, B. (2015). *Sentiment analysis: mining options, sentiments, and emotions*. Cambridge University Press.

[Tulkens et al., 2019] Tulkens, S., Suster, S., and Daelemans, W. (2019). Unsupervised concept extraction from clinical text through semantic composition. *Journal of Biomedical Informatics*.

[Wilson et al., 2005] Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.