Toxic Comment Classification

Team Members:

Jason Parsons	B00642710	jason.parsons@dal.ca
Brandon Poole	B00677266	PooleB@dal.ca
Alexandra Startsev	B00691787	al702917@dal.ca

Problem Statement:

Classifying online comments for their intention to harm others is challenging. Classifying such comments is hard due to a variety of reasons, these include the grammar and syntax of language, the context in which it is used, and the tone the speaker intended. These cannot always be determined in written form, even by some humans. Using natural language processing, data mining, and other machine learning techniques, solutions to this problem have emerged.

We are proposing a project based on the Toxic Comment Classification Challenge open competition being held in conjunction by Google and Jigsaw's Conversation AI team and hosted by kaggle.com. The goal of the competition is to improve on the Conversation AI team's solution of the problem of classifying toxicity in Wikipedia comments. The goal of the competitors is to produce an output file that contains a list of comment ids' that correspond to individual Wikipedia comments, as well as a score from 0-1 for 6 different comment toxicity types (toxic, severe toxic, obscene, threat, insult, and identity hate). These scores determine on which level each comment can be classified in each of the six toxicity types. Our goal would be to produce such an output file using our own implementation solution following this same format, and then compare our results to that of other teams. The Kaggle website has a leaderboard with overall team scores listed that will enable us to make this comparison.

For more information on the competition, you can visit:
https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge.

Current and Possible Approaches:

The competition's website contains documentation and papers from the Google Brain and Jigsaw teams outlining the strategies and shortcomings of their approach. That documentation has made up the bulk of our research for this project proposal. Their code base is not publically available, but a way to interact with it has been made available in the form of the Perspective API.

The current approaches to the problem of comment classification represent a shift from classical algorithms, to the use of machine learning systems. The main difference between these two approaches is the reliance of machine learning algorithms on existing data, unlike its classic counterparts (Viegas, Wattenberg, 2017). Due to this reliance on real world data, models based on machine learning techniques can pick up unintended bias from the data used. This bias increases the difficulty of classifying toxicity in online comments.

The unintended bias is the main shortcoming of the Conversation AI team's current approach. Their current model falsely associates frequently attacked identities with toxicity; a process known as false positive bias (Dixon, Li, Sorensen, Thain, Vasserman, 2017). Thus, a comment that should have a low score in toxicity may be incorrectly assigned a high score because the comment contains the name of a frequently attacked identity/group. This bias is caused by dataset imbalance in which frequently attacked identities are overrepresented in toxic comments made online (Dixon, et. al., 2017). Conversation AI proposes a solution to this problem by adding assumed non-toxic data from Wikipedia articles involving these identities to mitigate the imbalance. The Conversation AI team uses a set of evaluation metrics to measure unintended bias with varying results. The approaches include AUC (Area Under the Receiver Operating Characteristic Curve), Error rate Equality Difference, and Pinned AUC. All three evaluation metrics contain shortcomings, but Pinned AUC has been identified as the best performing metric so far (Dixon, et. al., 2017).

The goal of the toxic classifier competition is to find additional mitigation techniques, as well as create our own classifier based on this mitigation. One possible approach for doing so is to decrease the reliance on bias associated with individual identity words. Instead, we can look at the context in which these words are used in conjunction with the words surrounding them. The Conversation AI team makes a suggestion for future techniques to eliminate the need for unintended bias selection done by the human researchers. They should aim to automate the process of data mining (Dixon, et. al., 2017).

These proposed techniques may be outside the scope of this project, but they can be a part of our possible approach to the problem. Our work may involve exploring Conversation AI's recommendations for the state of the art solutions in the field.

Project Plan:

To prepare for this proposal, we have invested a significant amount of research into the requirements of the toxic classification competition. We also explored the current approach used by Google's Conversation AI team. Our goal for the rest of the project is to create a solution using the competition's rules while keeping Conversation AI's recommendation in mind. Once this is done, we will compare our results to the other teams in the competition. We will relate these findings in our final report and presentation.

References:

Dixon, L., Li, J., Sorensen, J., Thain, N., Vasserman, L. (2017). Measuring and mitigating unintended bias in text classification. Association for the Advancement of Artificial Intelligence.

Dixon, L. (2017). Conversation corpora, emotional robots, and battles with bias. ConversationAI.github.io

Viegas, F., Wattenberg, M. (2017). Fairness in machine learning. https://github.com/conversationai/unintended-ml-bias-analysis/blob/master/presentations/AI-with-the-best%20fairness%20presentation.pdf