**NLP Report**

**Abstract**

The problem of filtering toxic online comments requires a strategy for dealing with unintended bias. This report presents two classification methods, Random Forest and Recurrent Neural Networks, as potential solutions for classifying toxicity. The Random Forest classification scheme showed a high degree of accuracy in classifying the split data set. The Recurrent Neural Network has yet to provide final results and is still a work in progress. The goal of trying multiple approaches is to compare and contrast the various machine learning techniques that can be applied to natural language processing tasks, and to increase our knowledge in these areas. This report details the methodology used to create both of our classifiers, and a detailed breakdown of the results. We also touch on the work currently being done in this field by other groups such as the Google Brain team.

**1 Introduction**

The problem of unintended bias in online comments is one of the main issues in filtering toxicity in online comments. Bias can occur when words associated with gender, race, or sexual orientation are labelled toxic regardless of the context in which they’re used. This bias results in skewed accuracy results when classifiers are run on data sets made up of comments made online. The… <Fill in more here>

**2 Related Work**

The subject of this project is based upon the Toxic Comment Classification competition held on the Kaggle competition website. The competition was hosted by the Google Brain team, with an aim towards improving upon their own method for classifying online comments. <Fill in more here>

**3 Problem Definition and Methodology**

Classifying levels of toxicity in online comments is made difficult by the challenge of handling unintended bias in current classification schemes. <Fill in more here>

**3.1 RandomForest Classifier**

<Brandon, insert your methodology here>

**3.2 Recurrent Neural Network Classifier**

The Recurrent Neural Network classification method is set up to… <Still a work in progress>

**4 Experimental Design**

The design of… <Fill in more here>

**4.1 Results of RandomForest Classifier**

<Brandon, insert your results here.>

**4.2 Results of Recurrent Neural Network Classifier**

At the time of this writing, the Recurrent Neural Network has not produced any results. It remains a work in progress and further research will be needed to bring it to fruition. Decisions on how many neurons, hidden layers, and hyperparameters to use and tune are non-trivial. In order to better understand how to <Still a work in progress>

**5 Conclusion**

The accuracy of the RandomForest classification scheme suggests that this method may provide a way forward for classifying toxic comments. The Recurrent Neural Network also shows promise, but has a higher degree of complexity. This is due to the inherent nature of neural networks which often contain multiple hidden layers and require fine tuning of the hyperparameters used. These classification methods are only two among many different machine learning techniques that can be applied to natural language processing tasks.

The difficulty in classifying toxic comments online is due to unintended bias present in current classification methods. Bias presents itself when words that are gendered, racial, or sexual orientated in nature are present. If a positive comment using one of these words is classified as toxic simply due to its presence in the sentence, then this is a failure of our classification system. This report presents two possible methods of mitigating this unintended bias and producing a classification system that handles context in a responsible way.

Toxicity continues to be a major issue in online communities. In order to tackle this issue, we need more robust tools for detecting toxic comments. Due to the costs in time and money involved in having human beings manually policing comment sections, an automated system for handling toxicity is needed. The RandomForest and RNN solutions are a first step towards tackling this complex issue.