**Project Specification**

**Title:** Toxic Language Detection using Machine Learning

**Name:** Elizabeth Excell

**Supervisor:** Noura Al-Moubayed

**Project Description**

As access to the Internet has grown over time, so has the volume of toxic user-generated content on the web (Schmidt and Wiegand, 2017). As a result, the research interest in toxic language detection has increased over the past few years, leading to the construction of many different corpora annotated by humans for specific research purposes. The most recent papers on toxic language detection have focused on a wide range of issues, including the role of context in toxicity classification (Pavlopoulos et al., 2020) and the role of constructive comments in ensuring a polite conversation remains on track (Hosseini et al., 2020).

However, the majority of the research has not addressed the issue of bias caused by the annotators who decide the toxicity of the comments fed into the classifier. This can cause bias to be propagated in the classifier and lead to innocent comments from certain demographic groups being misclassified as toxic due to the inherent bias in the training data (Borkan et al., 2019). While there has been some research dedicated to mitigating the bias caused by crowdsourced toxicity annotations (Balayn et al., 2018), no research has been conducted on how the demographics of the annotators affect the toxicity scores and classification results, or how obfuscating the comments by masking terms linked to identity groups may reduce the unintended bias in the toxicity corpora.

The aim of this project is to investigate how the classification accuracy of toxic language classifiers can be improved by examining the demographics of the annotators and how possessing certain characteristics may make an annotator more or less likely to assign a high toxicity score to a given comment, especially comments that reference commonly targeted identity groups. This will be done by thoroughly examining the data in the Wikipedia Talk Corpus (Wulczyn et al., 2016), the largest toxicity dataset to contain the demographic information of the annotators, and using the insights gained to improve the performance of a classifier.

**Preliminary Preparation**

* Understanding of main NLP techniques and ML classifiers
* Understanding of composition of chosen datasets
* Knowledge of Jupyter Notebooks, python, and pandas

**Deliverables**

*Basic*

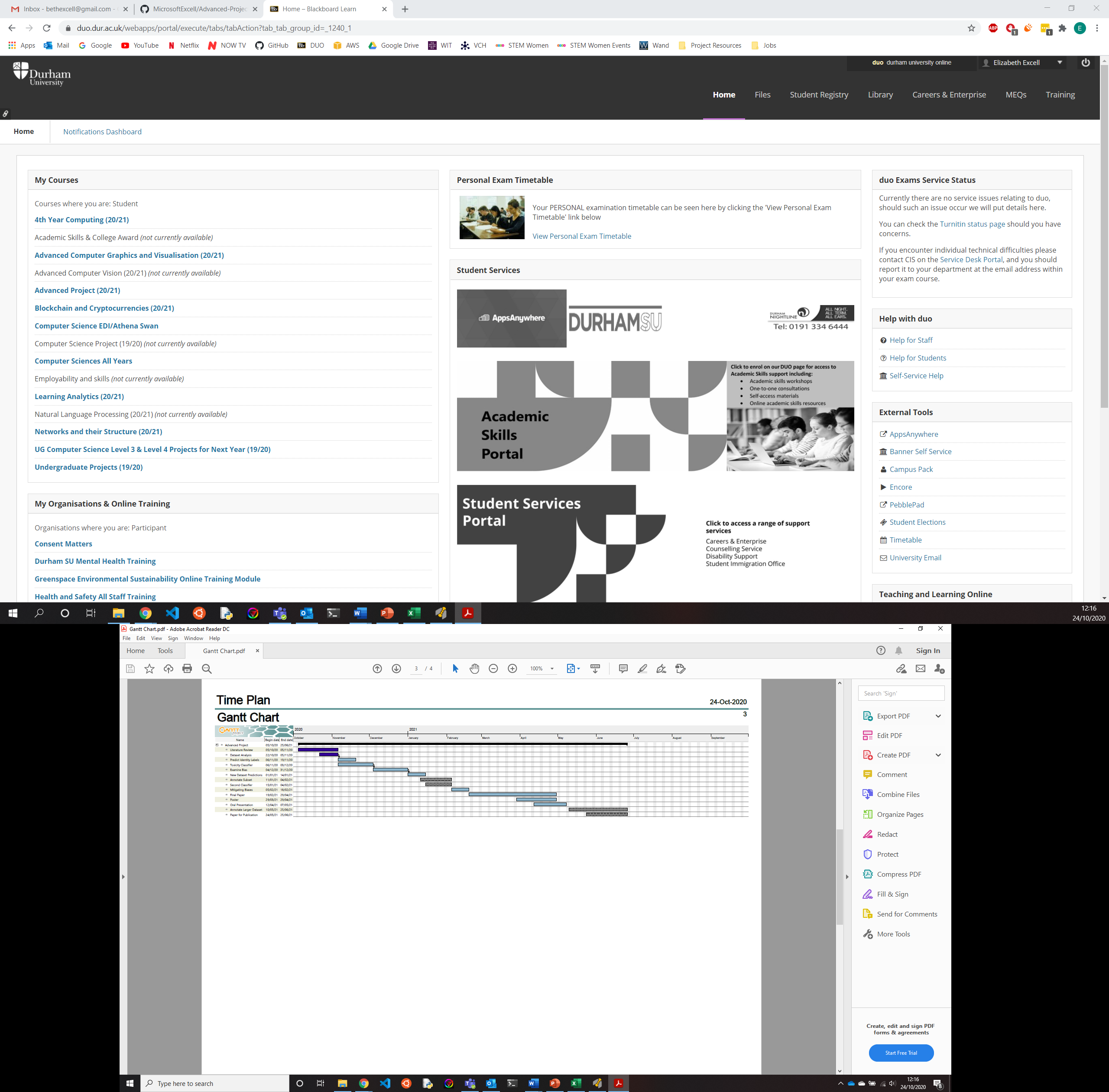
* Analyse demographic dataset, looking at how toxicity score changes with the demographics of the annotators and whether there’s a variety of demographics or if most annotators come from same identity groups.
* Predict whether commonly targeted identity groups are referenced in comments (using identity labelled dataset to find terms related to identity groups).
* Use predictions to see how toxicity scores change for demographics that are included in the comments.
* Build toxicity classifier (one of state-of-the art models such as BERT or a version of LSTM).
* Examine bias by 1) altering demographics of annotators in training set (e.g. try just using female annotators for comments referencing women/weight their opinions higher), 2) including the demographic information in the classifier, 3) masking/changing identities in test set to see how affects toxicity scores (change female to male pronouns, flip tokens like gay to man and names like Mohammed to David)

*Intermediate*

* Utilise new dataset with comments labelled with identities.
* Predict the demographic identities of the annotators/predict toxicity score annotators with known demographics from old dataset would give comments in the new dataset (and run classifier with new comments).
* Annotate subset of new dataset using Amazon Mechanical Turk (ask for toxicity scores and demographics for set of comments with known identity groups). Uses: as validation set, confirm previous results examining relationship between annotator demographics and identity groups referenced in comment, changing some identities (e.g. Mohammed to David, gay to man) and seeing if affects toxicity scores.

*Advanced*

* Build second state-of-the-art toxicity classifier to see if more robust to bias and to provide greater comparison for results.
* Investigate reducing biases in model without changing demographics or masking identities.
* Annotate much larger dataset with identities, annotator demographics and toxicity scores for future research/extending paper into publication.

**Time Plan**

**References**

* Balayn A., Mavridis P., Bozzon A., Timmermans B., and Szlávik Z. (2018). “Characterising and mitigating aggregation-bias in crowdsourced toxicity annotations”, in *Proceedings of the 1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management*, vol. 2276. CEUR.
* Borkan D., Dixon L., Sorensen J., Thain N., and Vasserman L. (2019). “Nuanced metrics for measuring unintended bias with real data for text classification”, in *Companion Proceedings of the 2019 World Wide Web Conference*,Association for Computing Machinery, pp. 491–500.
* Pavlopoulos J., Sorensen J., Dixon L., Thain N., and Androutsopoulos I. (2020). “Toxicity Detection: Does Context Really Matter?”, in *Proc. of 58th Annual Meeting of Association for Computational Linguistics,* pp. 4296-4305.
* Schmidt A. and Wiegand M. (2017). “A survey on hate speech detection using natural language processing”, in *Proceedings of the Fifth International workshop on natural language processing for social media*, pp. 1-10.
* Wulczyn E., Thain N. and, Dixon L. (2016): Wikipedia Detox. *figshare.* [doi.org/10.6084/m9.figshare.4054689](https://doi.org/10.6084/m9.figshare.4054689)