**Project Specification**

**project specification**, including a**Gantt chart** (or similar) which outlines your time plan throughout the year (both these documents are formatively assessed). The project specification is expected to be **max. 2 pages long** and should include:

* Preliminary preparation
  + what do you need to prepare / understand better?
* Deliverables
  + Basic / Intermediate / Advanced
* References

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

**Name:** Elizabeth Excell

**Supervisor:** Noura Al-Moubayed

**Possible Research Questions**

* How do the demographics of the annotators in toxicity datasets affect the results of toxic language classification?
* How do the demographics of annotators and the identities in comments create bias in toxicity datasets?
* What can be done to reduce the bias in toxic language classifiers by examining the demographics of the annotators and their impact on toxicity scores for comments that reference commonly targeted identity groups?
* How do the demographics of the annotators of toxicity datasets and their relationships to identity groups commonly targeted in toxic comments affect the bias present in toxic language classifiers?

**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 conversation doesn’t turn awry (Hosseini et al., 2020).

However, the majority of the research has had to deal with 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…

**Deliverables**

*Basic*

* Build toxicity classifier…

*Intermediate*

* Add information from other dataset…

*Advanced*

* New annotated dataset…

**Project Plan**

blah blah

GANTT CHART

**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)