**Crowdsourcing Proposal**

**Aims**

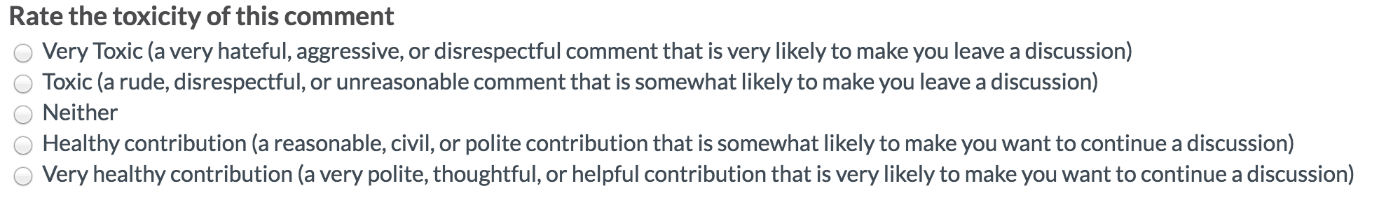
* Augment current dataset on toxic language with new comments annotated with demographic information and toxicity scores
* More annotated toxic comments will lead to a better performing classifier (currently there is a strong class imbalance between toxic and non-toxic comments)
* Can compare annotations to ground truth to assess reliability
* Validate conclusions drawn
* Draw new conclusions from extra demographic information
* Aid other researchers by adding some demographic information to a popular dataset used for understanding bias and validating the annotations in that dataset

**Dataset**

* Subset of Civil Comments Toxicity Kaggle dataset labelled with identities referenced in comments
* **Reasons:** Recent dataset (2019), large (450,000 comments), reasonable number of toxic comments, widely used in recent literature, comments broken down into identities which could provide extra insight, average comment length similar to main dataset (352 vs 401), comments annotated by same number of annotators on average as main dataset (10.4 vs 10.1), reliable as proposed by Google team and used in many implementations, has been used for looking at bias before, although the comments come from a different domain (news sites vs Wikipedia) the Wikipedia datasets either used data from the dataset we want to augment/there were too few toxic comments/the annotations were done by a toxicity classifier rather than humans, this dataset will be used as part of the analysis anyway so provides consistency
* **Selecting data to be annotated:** Choose data where all identity annotators labelled the comment as referencing an identity group (fewer identity annotators so want to ensure comment does contain an identity). Choose comments from identity groups with enough comments (Male, Female, Homosexual, Christian, Jewish, Muslim, Black, White). Use the same number of comments for related identity groups (Male/Female, Christian/Jewish/Muslim, Black/White). From each identity group randomly choose 50% of the chosen number of comments from toxic comments (>x% of annotators labelled toxic, x to be chosen later), 50% from non-toxic comments (0% of annotators labelled as toxic) to mitigate unintended bias.

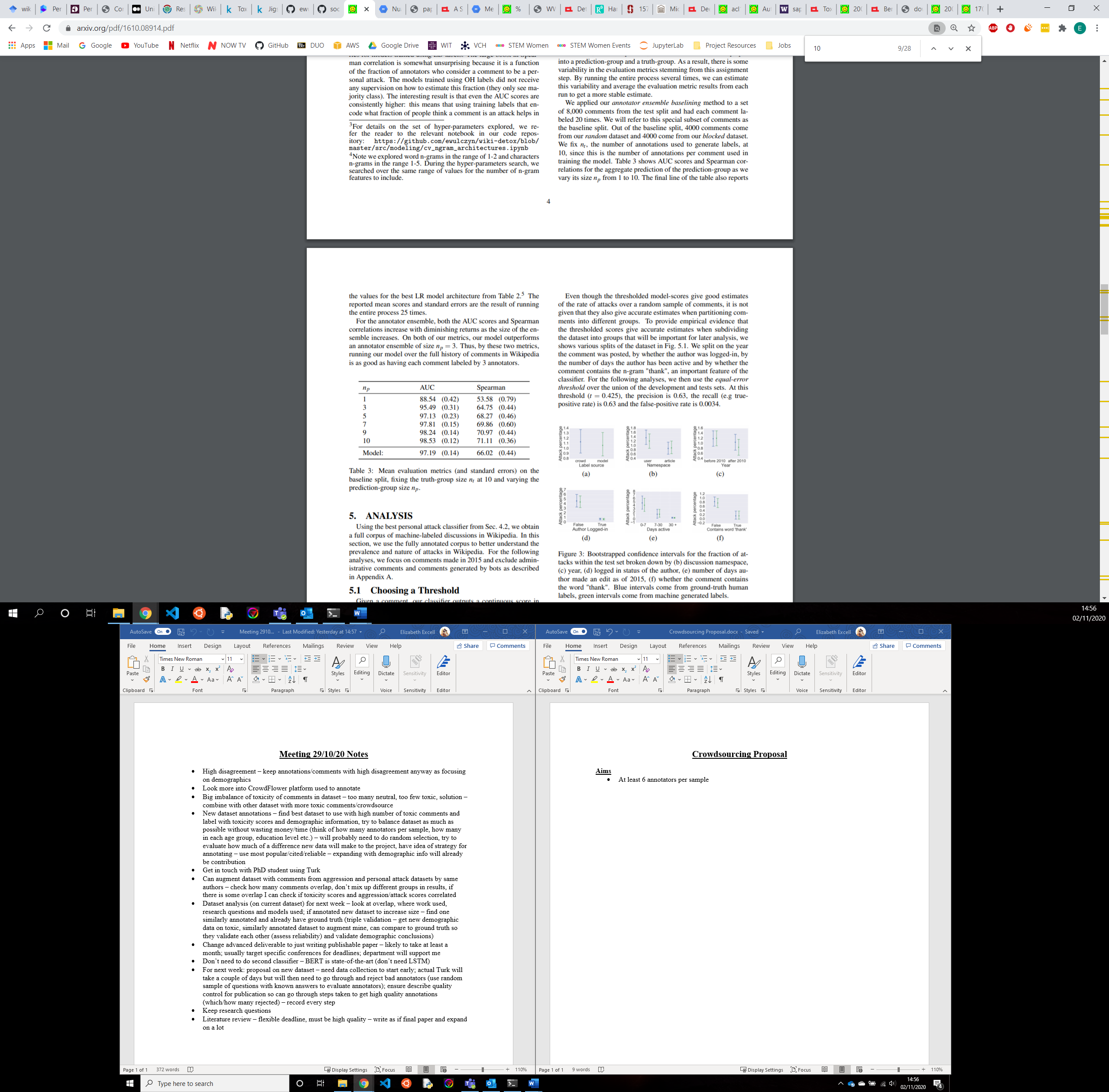
**Strategy**

* Ask annotators their demographic information: male/female/other, whether English is their first language (Yes/No), what age group they fit into (<18, 18-30, 31-45, 46-60, >60), what their level of education is (None, Some, High School, Bachelors, Masters, PhD, Professional), their sexual orientation (Heterosexual, Homosexual, Bisexual/Pansexual, Other), their religious identity (Christian, Jewish, Muslim, Other/None), their race (Black, White, Asian, Latino, Other)
* For each comment, ask the annotators:



(same as question asked for main dataset)

* Want at least 6 annotators per sample (prediction quality drops significantly with fewer annotators)



* Ideally want balanced demographics, especially for gender, religion, first language and race. Sexual orientation is likely to be heavily skewed towards heterosexual, age is likely to be mostly 18-30 and 31-45, and level of education is likely to be mostly Bachelors/high school.
* Assess quality of annotators by using random sample of questions with known answers and remove the worst quality annotators