**Dataset Statistics**

**Wikipedia Abusive Language Dataset**

*Description*

* English comments from Wikipedia
* 3 sub-datasets annotated for: toxicity, aggression, personal attack
* Over 100k comments per sub-dataset
* 10 annotations per comment (excluded worst 2% of annotators)
* 4000 annotators
* Inter-annotator agreement – Krippendorf’s alpha: 0.45
* Cited as: Wulczyn, Ellery; Thain, Nithum; Dixon, Lucas (2016): Wikipedia Detox. *figshare.* [doi.org/10.6084/m9.figshare.4054689](https://doi.org/10.6084/m9.figshare.4054689)

*Links*

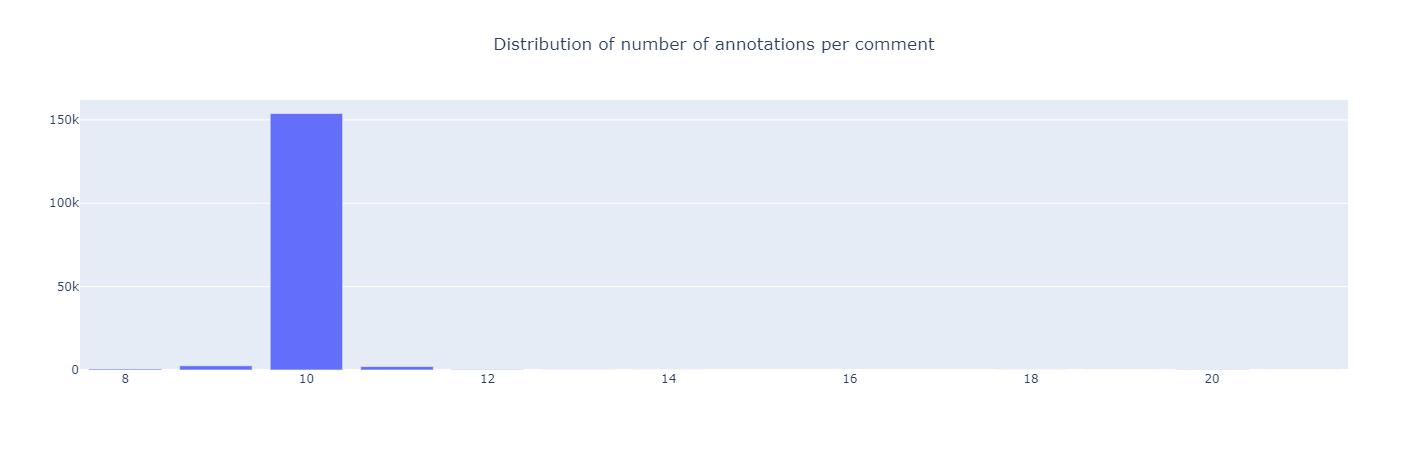
* <https://meta.wikimedia.org/wiki/Research:Detox/Data_Release>
* <https://figshare.com/articles/dataset/Wikipedia_Talk_Labels_Toxicity/4563973>
* <https://github.com/ewulczyn/wiki-detox/blob/master/src/figshare/Wikipedia%20Talk%20Data%20-%20Getting%20Started.ipynb> (start coding with this)
* <https://arxiv.org/pdf/1610.08914.pdf>

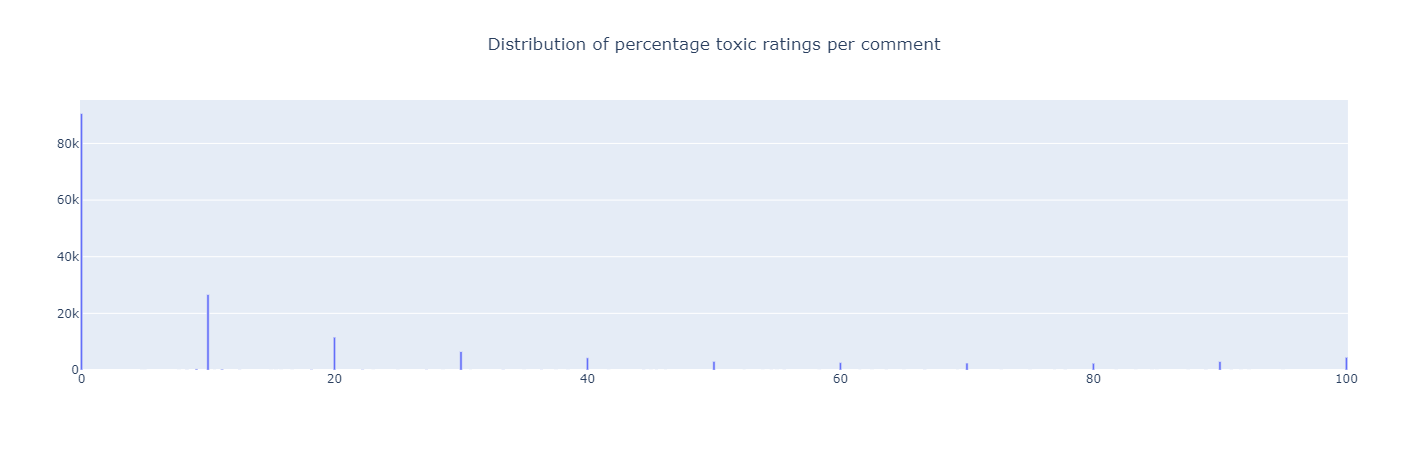
*Schema*

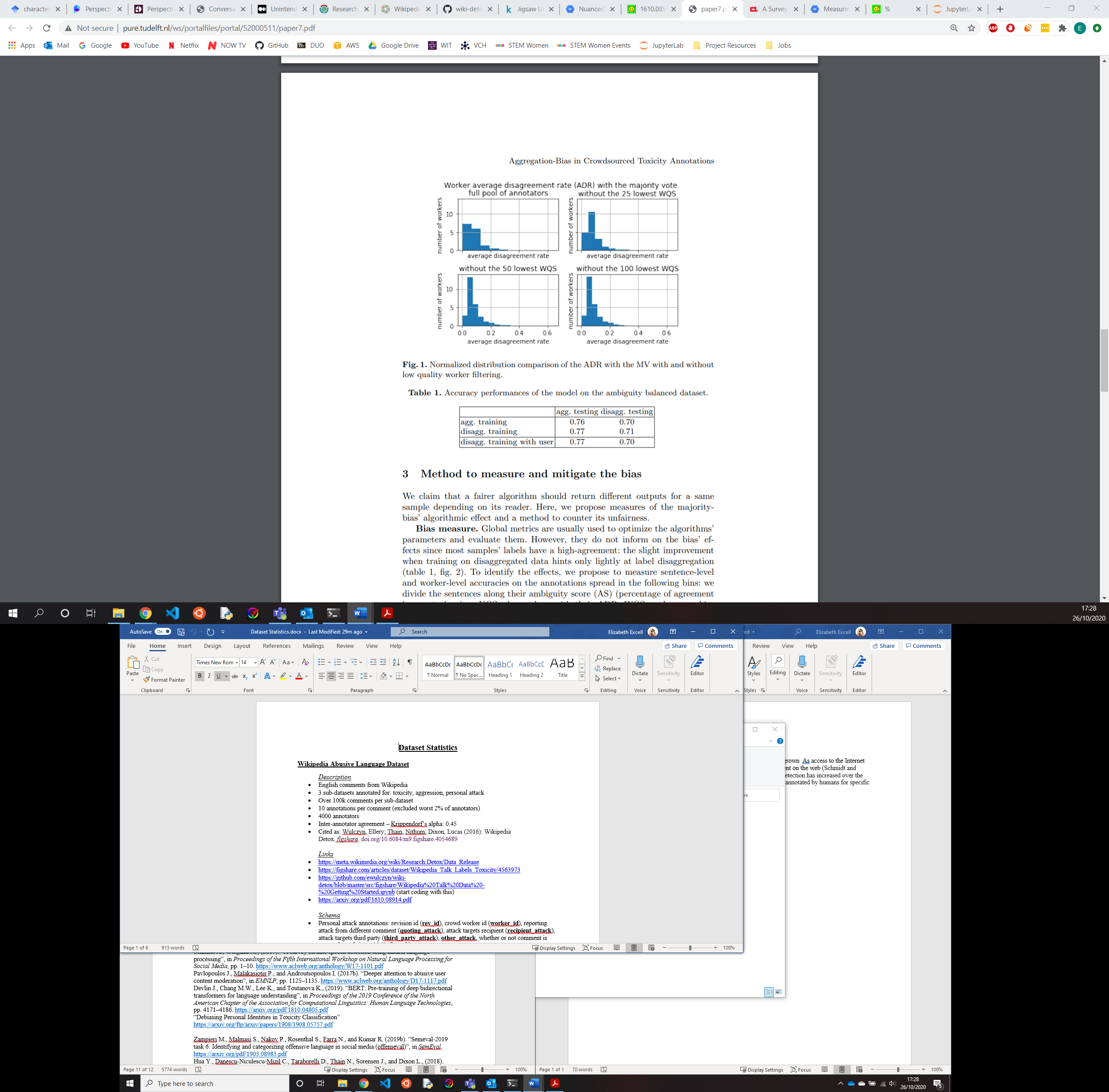
* Personal attack annotations: revision id (**rev\_id**), crowd worker id (**worker\_id**), reporting attack from different comment (**quoting\_attack**), attack targets recipient (**recipient\_attack**), attack targets third party (**third\_party\_attack**), **other\_attack**, whether or not comment is personal attack (**attack**) (1 if attack)
* Aggression annotations: **rev\_id**, **worker\_id**, **aggression\_score** (-2 very aggressive to 2 very friendly), **aggression** (1 if comment has aggressive tone)
* Toxicity annotations: **rev\_id, worker\_id, toxicity\_score** (-2 very toxic to 2 very healthy), **toxicity** (1 if score<0)
* Annotated comments: **rev\_id, comment** - concatenation of content in edit of talk page (markup and HTML stripped out, NEWLINE\_TOKEN, TAB\_TOKEN and ‘ remain), **year, logged\_in** (if author logged in), **ns** - namespace of discussion page (user or article), **sample** (from random sampling of all comments/from random sampling of 5 comments around a block event), **split** (train, dev or test)
* Demographics: (some fields unanswered – need to strip out) **worker\_id, gender, English\_first\_language, age\_group** (‘Under18’, ‘18-30’, ‘30-45’, ‘45-60’, ‘Over60’), **education** (none, some, hs, bachelors, masters, doctorate, professional)

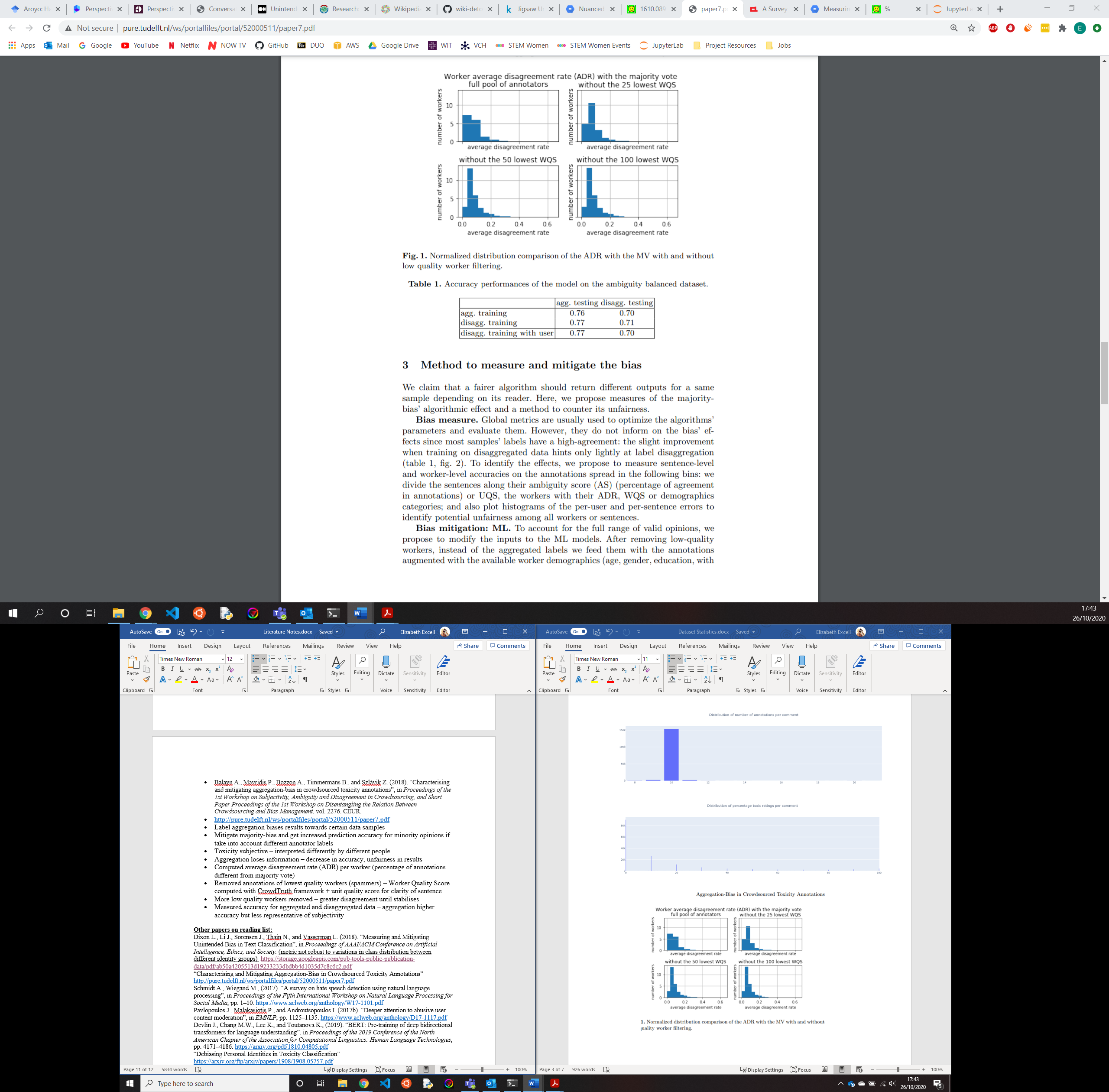
*Demographics*

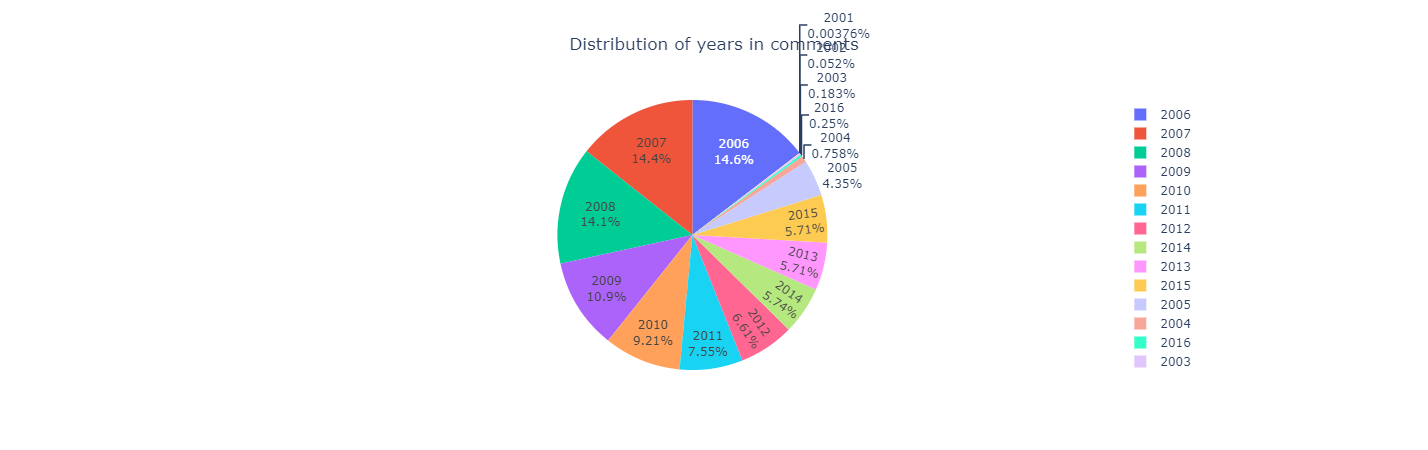
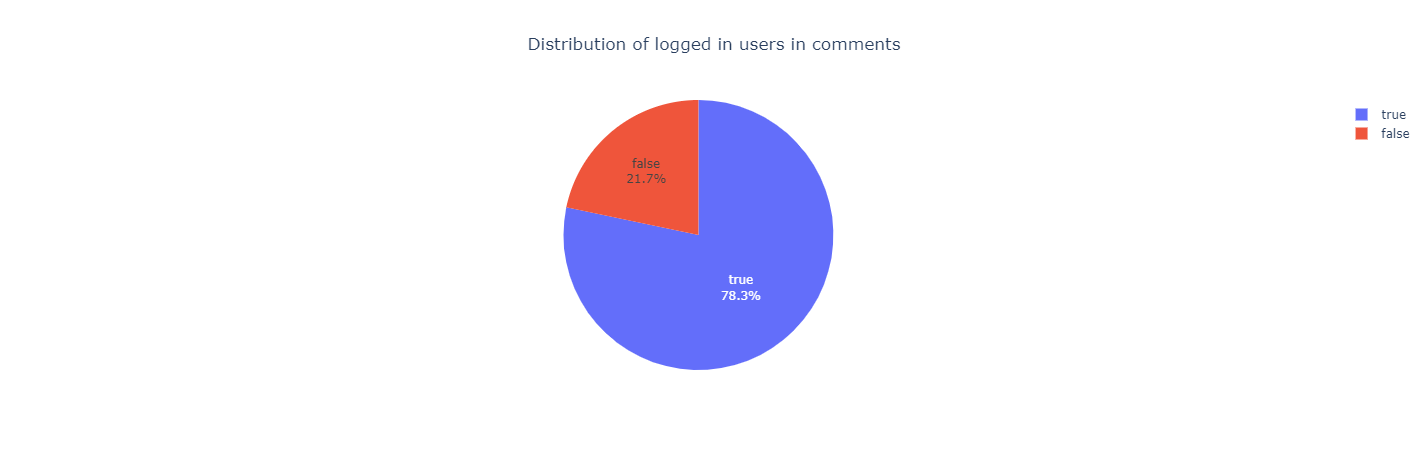
* Toxicity – 1,598,289 annotations, 159,686 comments, 4301 annotators, 371.61 average annotations per annotator, 10.01 average annotations per comment (majority had 10 annotators)
* Not much agreement on toxicity of comments (see second graph)
* Inter-annotator agreement – Krippendorf’s alpha: 0.45
* Removing lowest quality workers causes greater disagreement with majority vote (see third graph)
* Performs better with aggregated testing (looking at majority vote)
* Years comments made reasonably distributed
* 78.3% of users who made comments logged in
* 59.5% of comments from user pages, 40.5% from article pages
* Half of comments are randomly sampled, other half randomly sampled from the 5 comments surrounding block events (ensuring a certain amount of toxicity in the dataset)
* 14.5% of annotations labelled comments as toxic
* 49% of annotations labelled comments ‘Neutral’, 35% “Healthy”, 12% “Toxic”, 3% “Very Toxic”, 1% “Very Healthy”
* Roughly 1/3 of annotations labelled by women
* Female annotators annotated fewer comments on average than male annotators (362.11 vs. 382.14)
* Female annotators more likely to label a comment as toxic than male annotators (15.7% vs. 13.9%) and give lower scores [-2 = very toxic, 2 = very healthy] (female average score = 0.18, male average score = 0.23)
* Despite being a dataset filled with English comments, only 17.4% of annotations were made by someone whose first language was English
* People whose first language was English annotated fewer comments on average than those whose first language wasn’t English (351.04 vs. 380.57)
* People whose first language was English judged slightly more comments to be toxic than those whose first language wasn’t English (15.2% vs. 14.4%) but also gave comments a higher rating on average (indicating healthier comments) (English score 0.23, other score 0.21) (not significant result – could suggest non-native English speakers more likely to pick neutral)
* 54% of annotations were made by people aged 18-30, 36% by 30-45, 8% by 45-60, 2% by <18s and 1% by >60s
* Mean no. annotations follows (roughly) same pattern: 18-30 (380.33), 30-45 (375.67), <18s (356.46), 45-60 (348.4), >60s (323.9) (may be that trends indicate willingness to do job)
* Descending order of toxicity judgements of age groups: 45-60 (16.6% toxic), >60 (15.7% toxic), 30-45 (14.9% toxic), 18-30 (14.1% toxic), <18 (12% toxic)
* Ascending order of toxicity scores of age groups (decreasing perceived toxicity): >60 (0.17), 18-30 (0.2), 45-60 (0.22), 30-45 (0.23), <18 (0.27) (difference to previous order could be down to 45-60s being less likely to vote neutral than 18-30s)
* 40% of annotations made by people who completed their Bachelors, 28% by people who completed high school, 15% by people who completed their Masters, 13% by people with professional qualifications, 3% by people with some schooling, 1% by people with a PhD and <1% by people with no schooling
* Descending order of toxicity judgements by education: bachelors (14.9% toxic), masters (14.7% toxic), professional (14.3% toxic), high school (14.25% toxic), some (13.3% toxic), none (13.2% toxic), PhD (13.1% toxic)
* Ascending order of toxicity scores by education (decreasing perceived toxicity): masters (0.19), bachelors (0.205), professional (0.21), high school (0.227), some (0.23), PhD (0.25), none (0.62)
* Mean no. annotations (descending): some (385.4), professional (381.99), masters (376.24), bachelors (374.89), high school (371.84), PhD (355.68), no education (308.33)
* group as % of total and mean number of annotations for group correlated as both indicate level of interest from that demographic
* **Note:** There are also personal attack/aggression datasets that use the same format and are likely to be composed of similar demographics – could join sets/examine statistics of those as well

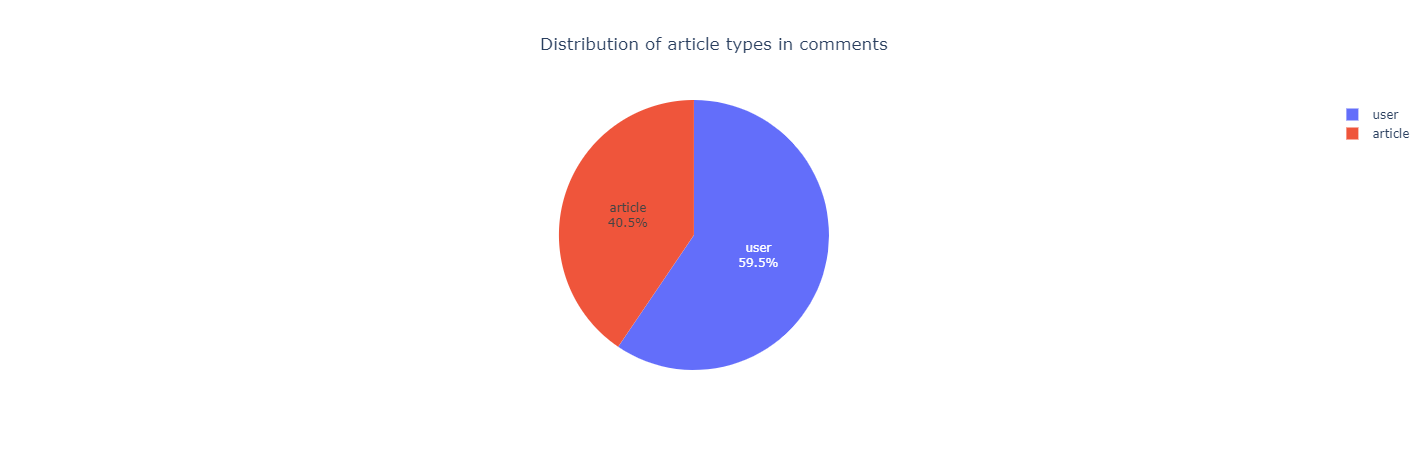
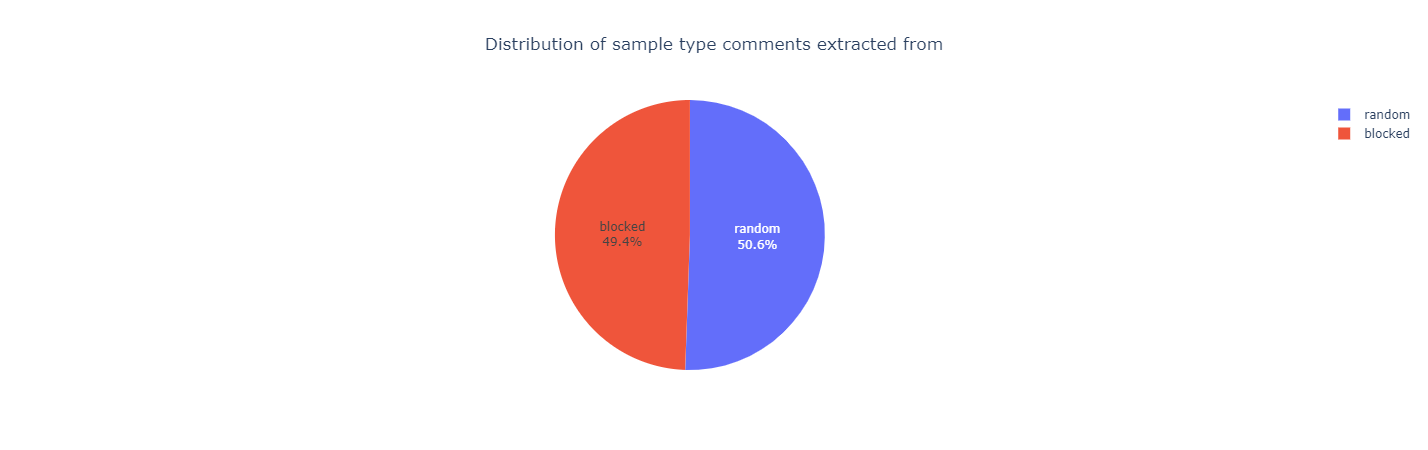


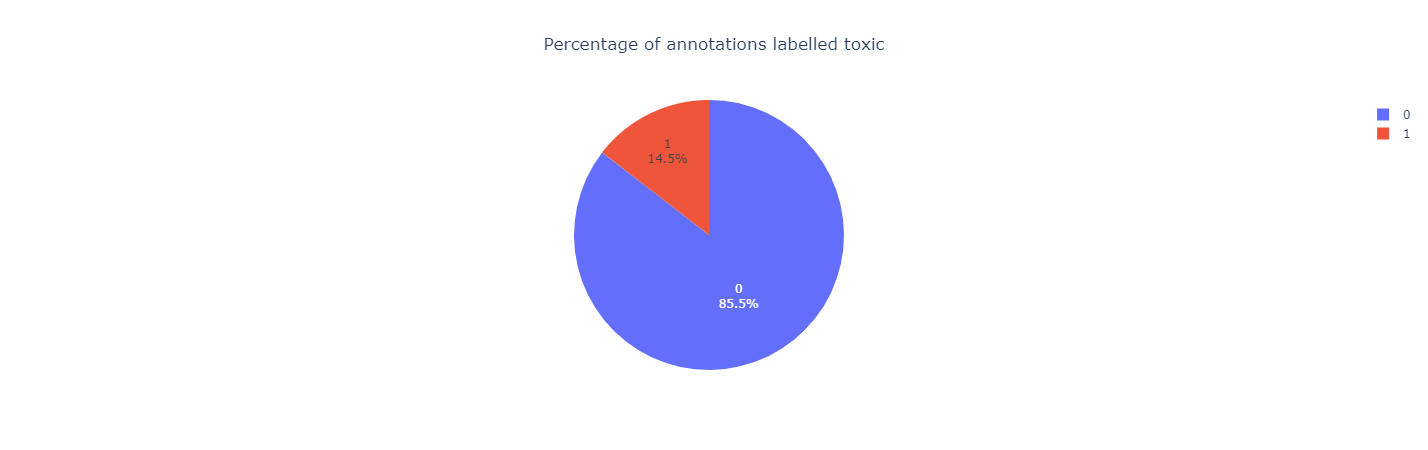
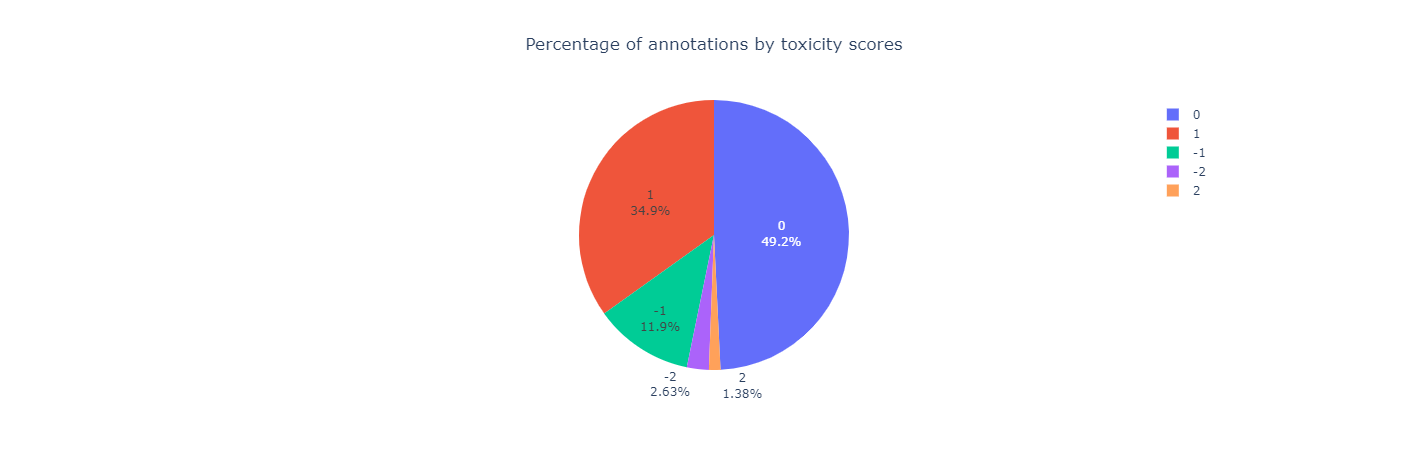


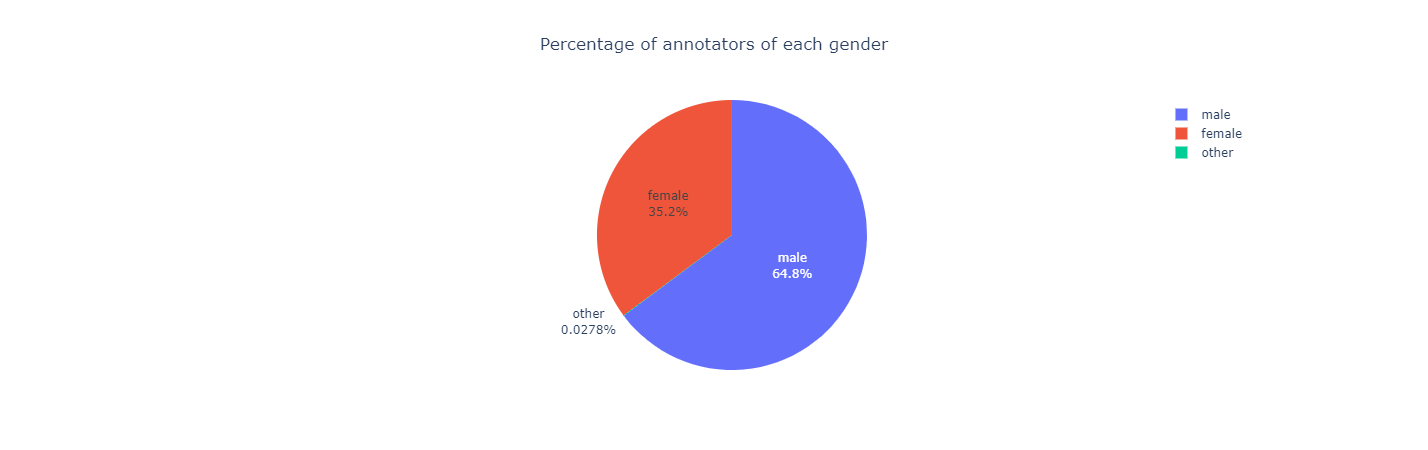
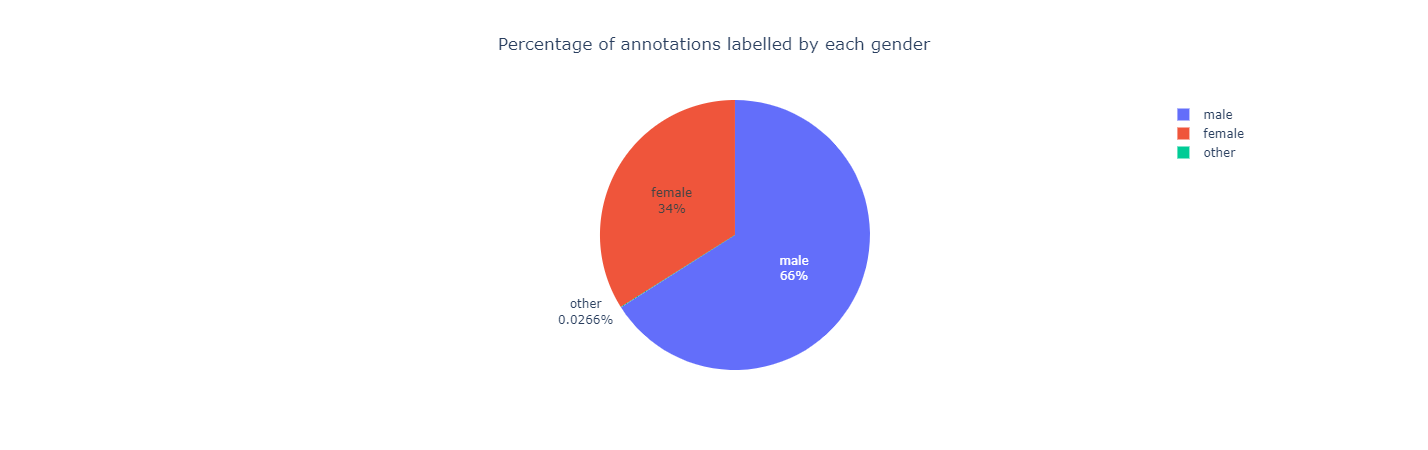


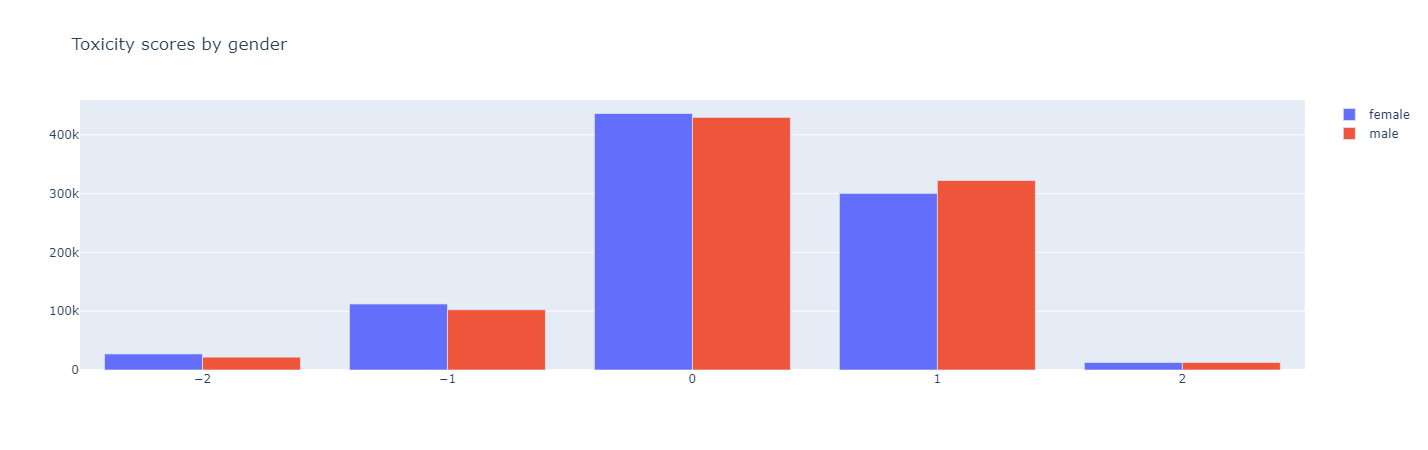


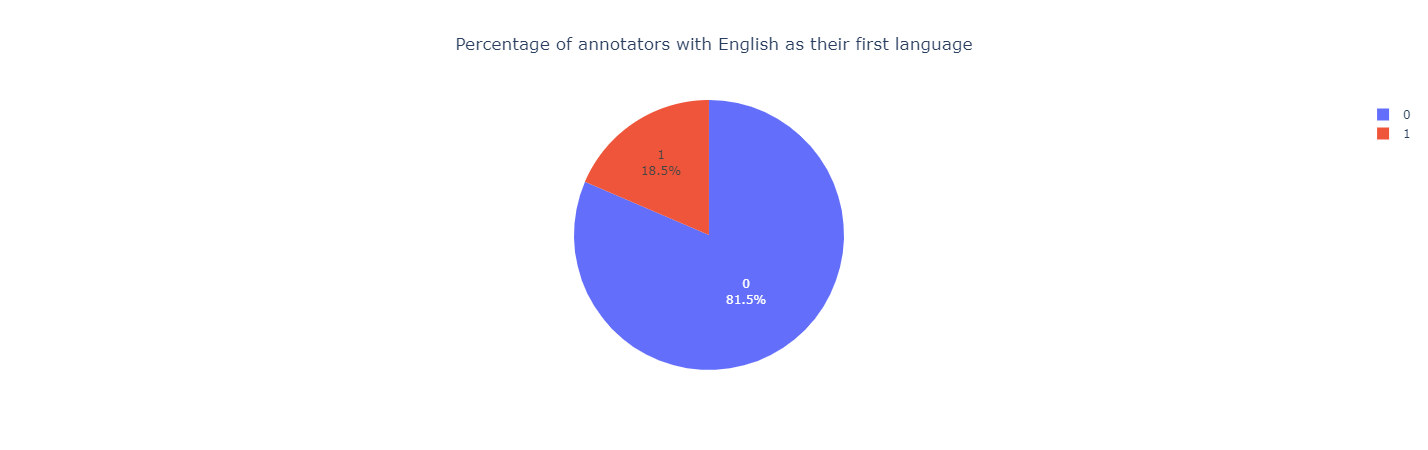
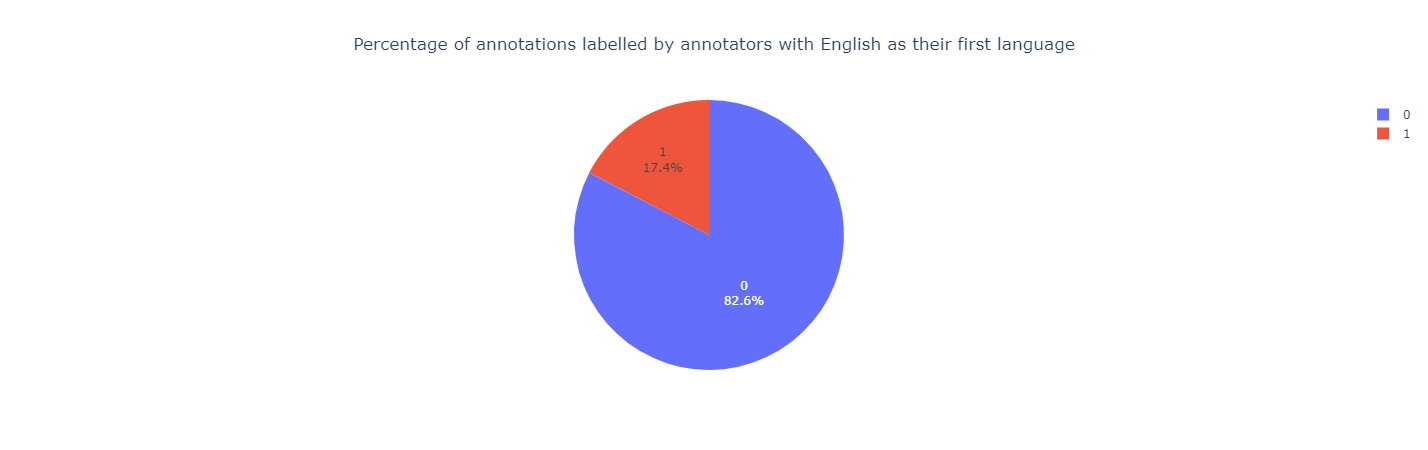
 

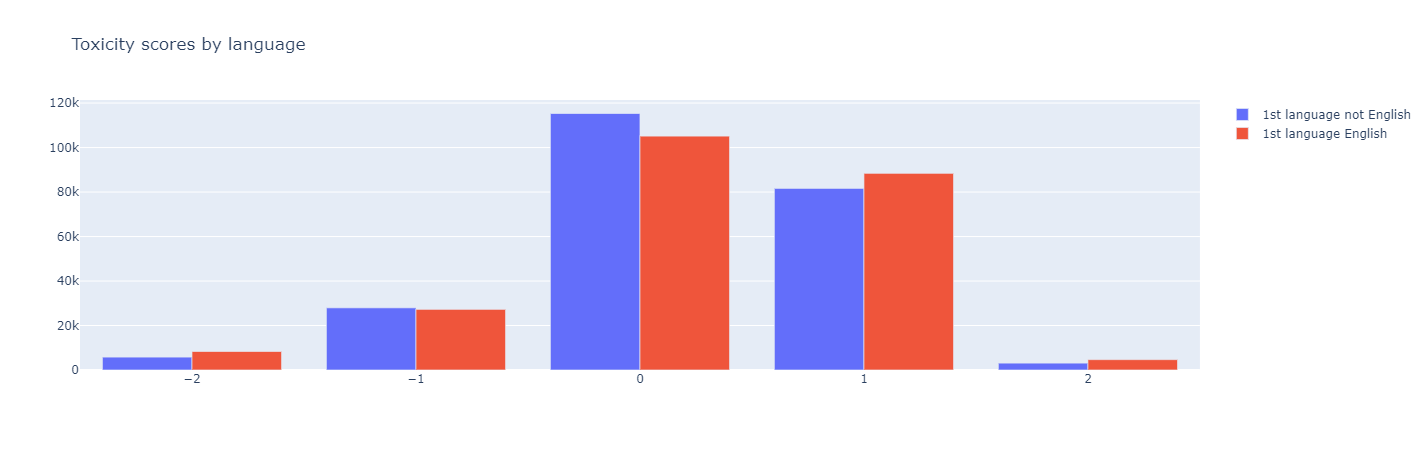
 

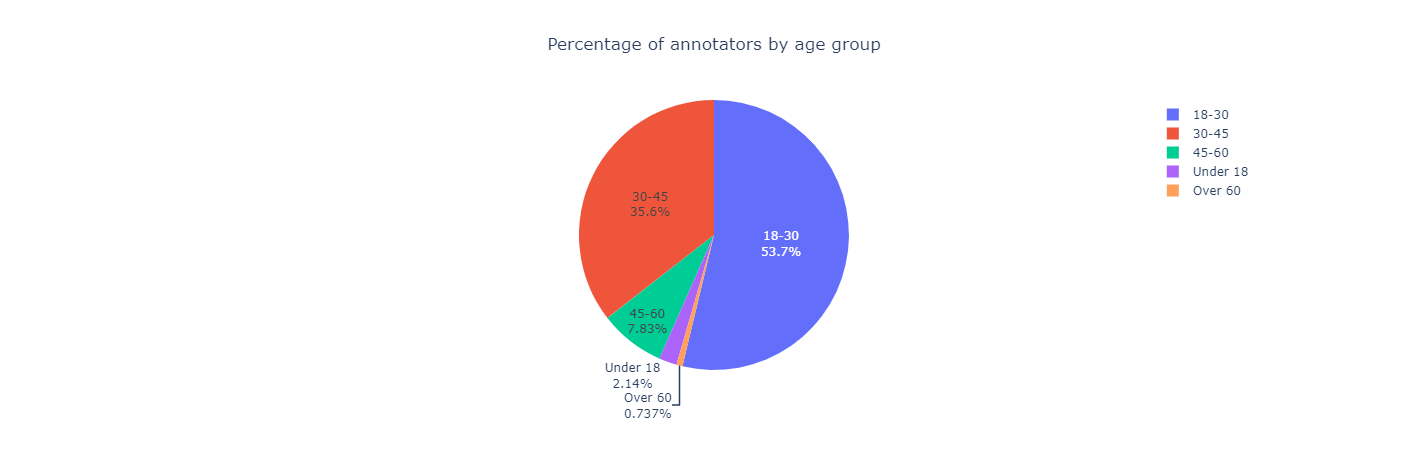
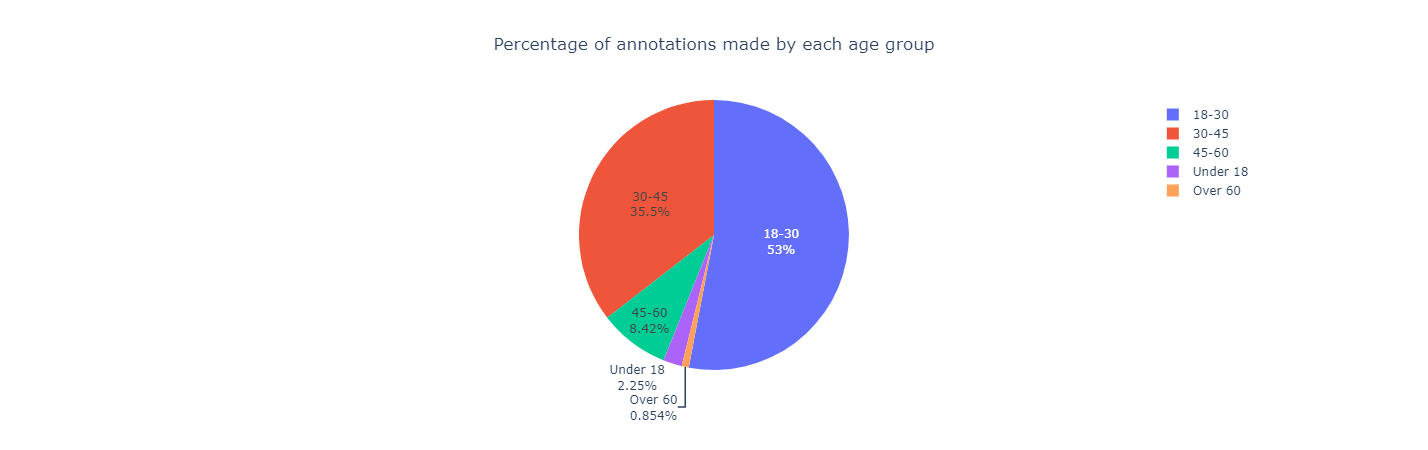
 

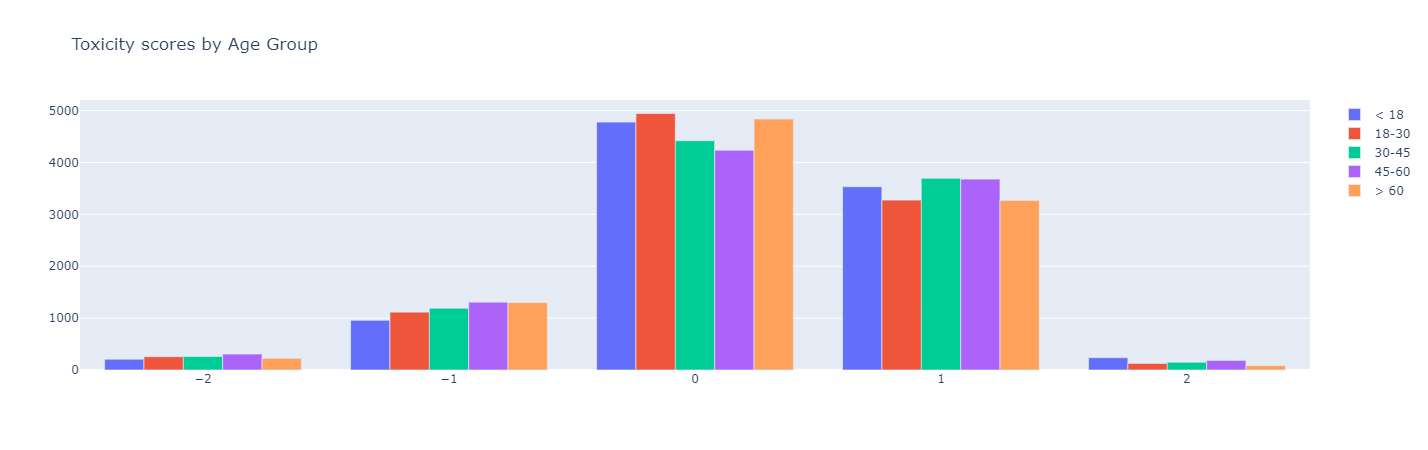
 

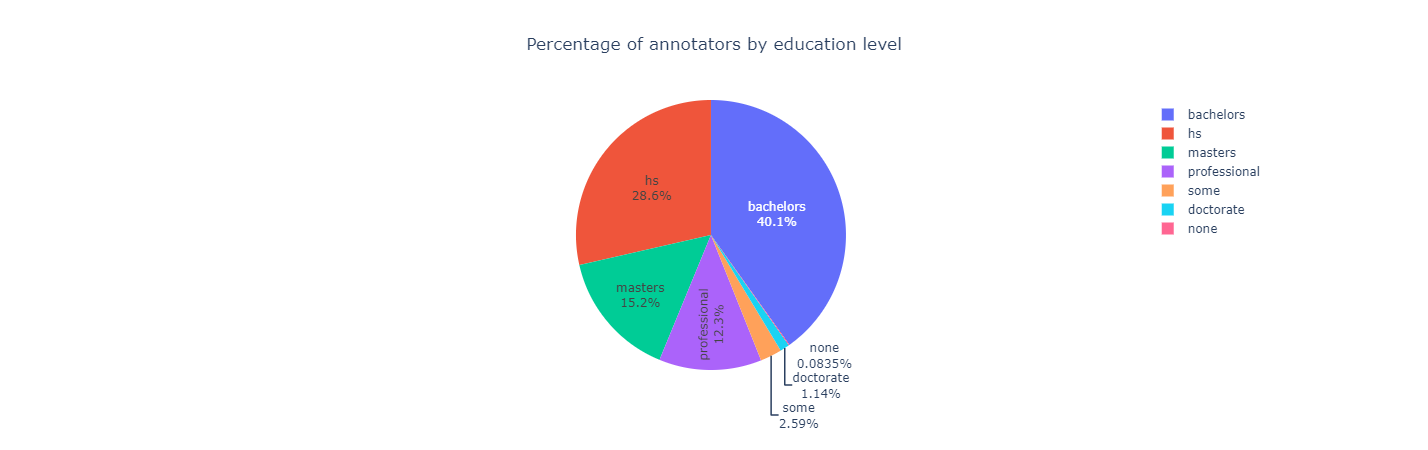
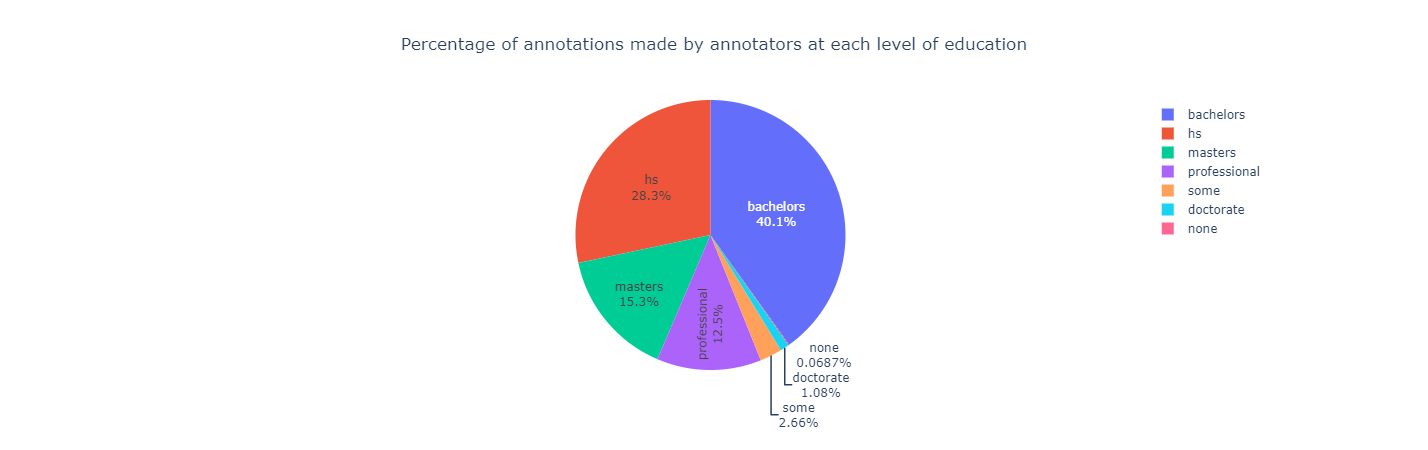


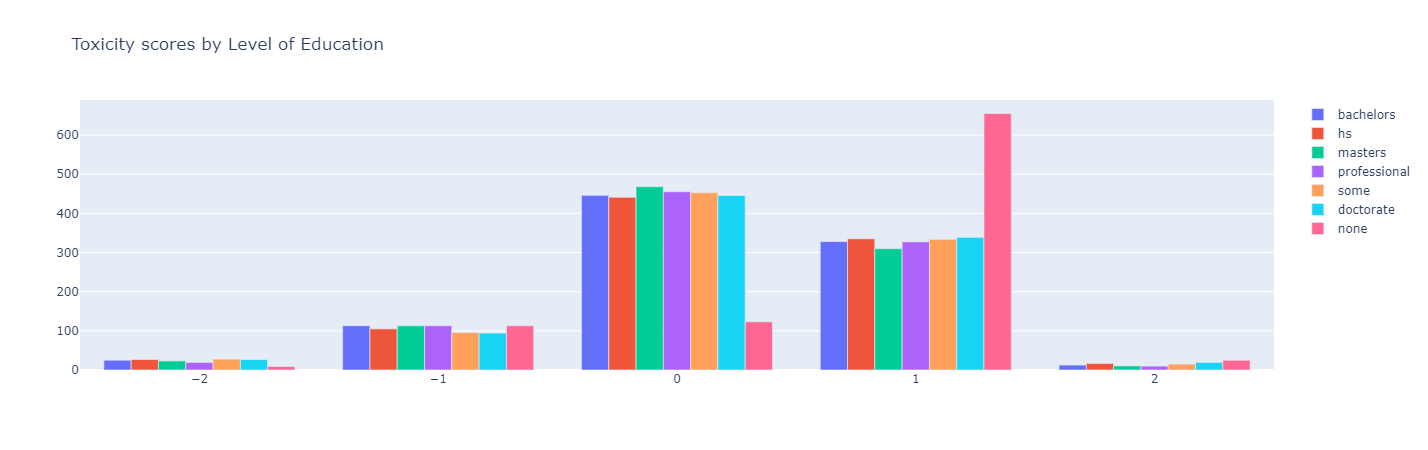
 







**Civil Comments Toxicity Kaggle (CCTK)**

*Description*

* ~2M English comments from Civil Comments
* 450,000 comments labelled for identity
* Unintended bias in dataset for subgroups
* Some comments contain same text but annotated as containing different identities

*Links*

* <https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification>
* <https://conversationai.github.io/research.html>

*Schema*

* **id, comment\_text, target** (fraction of annotators who believed comment toxic), **severe\_toxicity, obscene, threat, insult, identity attack, sexually explicit,** fraction of annotators who said identities mentioned in comment: **male, female, transgender, other\_gender, heterosexual, homosexual\_gay\_or\_lesbian, bisexual, other\_sexual\_orientation, chrisitan, jewish, muslim, hindu, buddhist, atheist, other\_religion, black, white, Asian, latino, other\_race\_or\_ethnicity, physical\_disability, intellectual\_or\_learning\_disability, psychiatric\_or\_mental\_illness, other\_disability, toxicity\_annotator\_count, identity\_annotator\_count, created\_date, split, publication\_id, parent\_id, article\_id,** (civility) **rating, funny, wow, sad, likes, disagree**

*Demographics*

* **18**% of identity comments labelled **male**, **18% female, 4% gay/lesbian, 14% Christian, 4% jewish, 5% muslim, 5% other religion, 5% black, 9% white, 4% Asian, 4% other race/ethnicity, 4% psychiatric/mental illness**
* average no. annotators (toxicity and identity) for identity comments
* average toxicity scores for identities
* % of identity comments labelled with each of subgroups (above)
* Word cloud of most common terms associated with each identity