**Dataset Statistics**

**Wikipedia Abusive Language Dataset**

*Description*

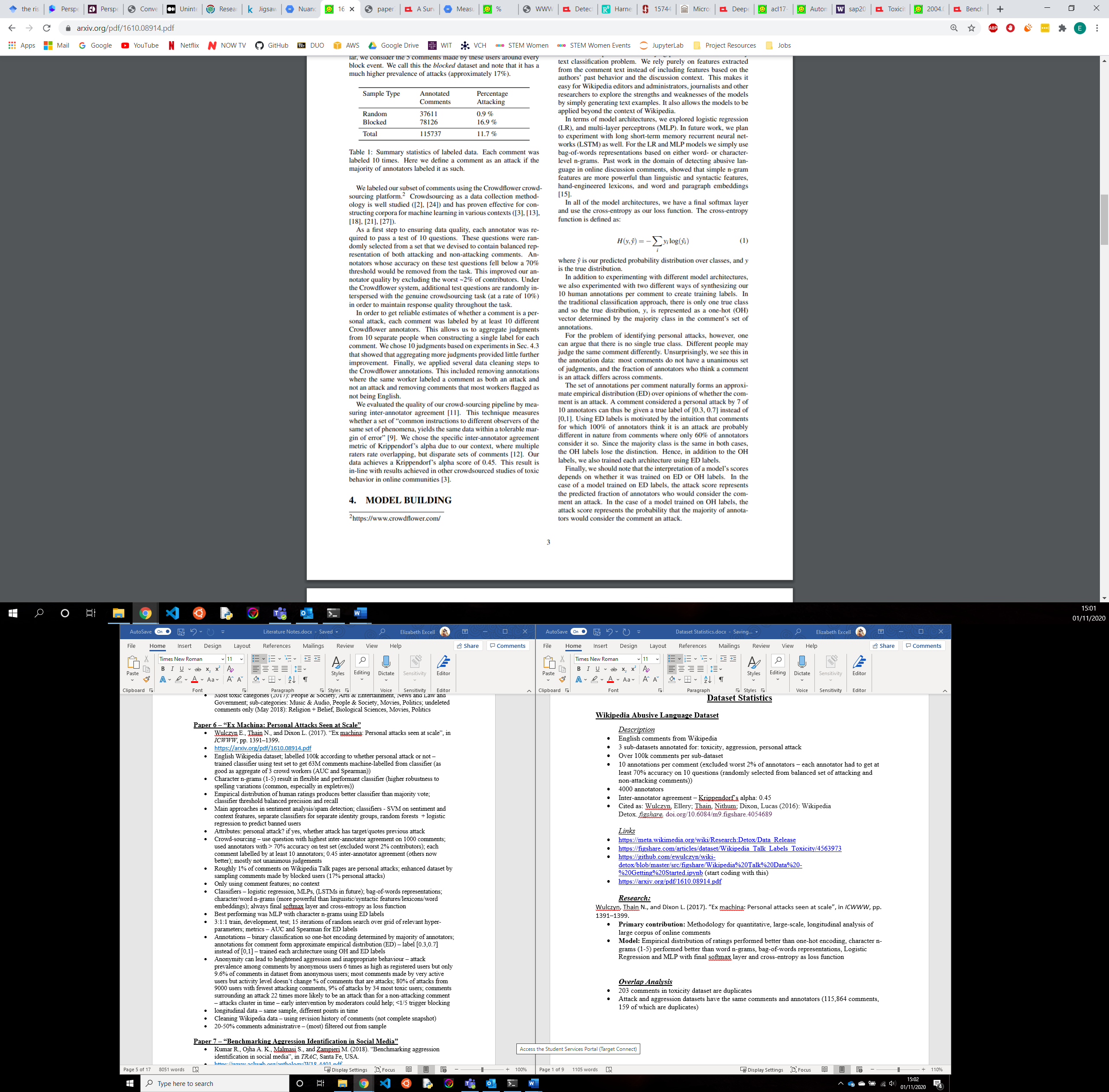
* English comments from Wikipedia
* 3 sub-datasets annotated for: toxicity, aggression, personal attack
* Over 100k comments per sub-dataset
* 10 annotations per comment (excluded worst 2% of annotators – each annotator had to get at least 70% accuracy on 10 questions (randomly selected from balanced set of attacking and non-attacking comments))
* 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://meta.wikimedia.org/wiki/Research:Detox/Fairness>
* <https://figshare.com/projects/Wikipedia_Talk/16731>
* <https://github.com/ewulczyn/wiki-detox> (Code for paper)
* <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>
* <https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data?select=train.csv.zip> (Kaggle competition where types of toxicity in dataset identified)

***Research using dataset:***

Wulczyn, Thain N., and Dixon L. (2017). “Ex machina: Personal attacks seen at scale”, in *ICWWW*, pp. 1391–1399.

* (Personal attack dataset)
* **Primary contribution:** Methodology for quantitative, large-scale, longitudinal analysis of large corpus of online comments
* **Model:** Empirical distribution of ratings performed better than one-hot encoding, character n-grams (1-5) performed better than word n-grams, bag-of-words representations, Logistic Regression and MLP with final softmax layer and cross-entropy as loss function  , set split into 3:1:1 ratio train, development and test, performed 15 iterations of random search over grid of relevant hyperparameters (in <https://github.com/ewulczyn/wiki-detox/blob/master/src/modeling/cv_ngram_architectures.ipynb>)
* **Metrics:** AUC, Spearman

Chu, T., Jue, K. and Wang, M., 2016. Comment abuse classification with deep learning. *Von https://web. stanford. edu/class/cs224n/reports/2762092. pdf abgerufen*.

* (Personal attack dataset)
* **Contribution:** Using deep learning models
* **Model:** RNN with LSTM cell and word embeddings and a CNN with word/character embeddings
* **Metrics:** AUC, F1 Score, Accuracy

Binns, R., Veale, M., Van Kleek, M. and Shadbolt, N., 2017, September. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *International conference on social informatics* (pp. 405-415). Springer, Cham.

* (Toxicity dataset – must be older version as gives different statistics)
* **Contribution:** How do latent norms and biases affect the operation of offence detection systems? Do the norms of offence held by people who contributed to the training data result in classifiers which systematically favour certain norms of offence over others? (built different classifiers for demographically distinct subsets of annotators – focusing on gender)
* **Findings:** Inter-annotator agreement lower for women than men (older version of data – also found female toxicity scores lower which is the opposite of what I found), unfair – more false positives when attempting to replicate women’s collective judgements than men’s
* **Model:** Multiple text classifiers trained on various subsets of data, maximum 10,000 features, TF-IDF vectorizer, Logistic Regression
* **Metrics:** AUC, TPR (sensitivity), TNR (specificity)
* **Notes:** I’m not convinced by the way they split the data to examine men and women so I don’t trust the results (they selected comments with male and female annotators and sampled 10 men/women/both with replacement for each comment and averaged – but more men than women in dataset)

Merayo-Alba, S., Fidalgo, E., González-Castro, V., Alaiz-Rodríguez, R. and Velasco-Mata, J., 2019, September. Use of Natural Language Processing to Identify Inappropriate Content in Text. In *International Conference on Hybrid Artificial Intelligence Systems* (pp. 254-263). Springer, Cham.

* (Personal attack and aggression datasets)
* **Contribution:** detecting violent content on text documents using NLP techniques (no real new contribution)
* **Model:** TF-IDF and bag-of-words, Logistic Regression, SVM, Naïve Bayes and StarSpace (deep learning approach from Facebook)
* **Metrics:** Accuracy

D'sa, A.G., Illina, I. and Fohr, D., 2019. Towards non-toxic landscapes: Automatic toxic comment detection using DNN. *arXiv preprint arXiv:1911.08395*.

* (Toxicity dataset)
* **Contribution:** Investigate different approaches based on different state-of-the-art deep NNs and word representations for automatic toxic comment detection
* **Model:** Mikolov’s word embedding/fastText subword embedding, one-hot encoding, CNN, biLSTM, biGRU, BERT
* **Metrics:** F1 score, RMSE, MAE

Nejadgholi, I. and Kiritchenko, S., 2020. On Cross-Dataset Generalization in Automatic Detection of Online Abuse. *arXiv preprint arXiv:2010.07414*.

* (Toxicity dataset)
* **Contribution:** Exploring topic and task formulation bias in cross-dataset generalisation
* **Model:** BERT-based classifier, linear prediction layer, batch size 16, 2 epochs
* **Metrics:** TPR, TNR, macro-averaged F1 Score, Accuracy

Cecillon, N., Labatut, V., Dufour, R. and Linares, G., 2020. WAC: A Corpus of Wikipedia Conversations for Online Abuse Detection. *arXiv preprint arXiv:2003.06190*.

* (Attack, aggression, and toxicity datasets)
* **Contribution:** Reconstruction of conversations creating a large corpus of 380k annotated messages including context
* **Model:** Hybrid approach – content-based + graph-based
* **Metrics:** Macro Precision, Recall, F1 score

Pavlopoulos, J., Malakasiotis, P. and Androutsopoulos, I., 2017. Deep learning for user comment moderation. *arXiv preprint arXiv:1705.09993*.

* (Attack and toxicity datasets)
* **Contribution:** Creating new dataset of 1.6M moderated user comments, applying deep learning to user comment moderation, considering semi-automatic scenario
* **Model:** CNN on word embeddings, RNN and variants with attention mechanism
* **Metrics:** AUC, Spearman

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., and Asokan, N. (2018). All you need is "love": Evading hate speech detection. In 11th ACM Workshop on Artificial Intelligence and Security, pages 2–12.

* (Attack dataset)
* **Contribution:** Reproducing 7 state-of-the-art hate speech detection models and proving that they only perform well on the same type of data they are trained on and are brittle to certain adversaries
* **Model:** Logistic Regression, MLP on character n-grams, CNN+GRU, LSTM, transfer learning
* **Metrics:** macro-averaged F1 score

Mishra, P., Yannakoudakis, H., and Shutova, E. (2018). Neural character-based composition models for abuse detection. In 2nd Workshop on Abusive Language Online, pages 1–10

* (Attack and toxicity datasets)
* **Contribution:** Design model that can compose embeddings for unseen words
* **Model:** context-aware representations of characters (character n-grams with augmented/context hidden state), RNN – GRU + LR layer (like Pavlopoulos et al., 2017), GloVe vectors
* **Metrics:** macro-averaged F1 score, precision, recall

Dixon L., Li J., Sorensen J., Thain N., and Vasserman L. (2018). “Measuring and Mitigating Unintended Bias in Text Classification”, in *Proceedings of AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society.*

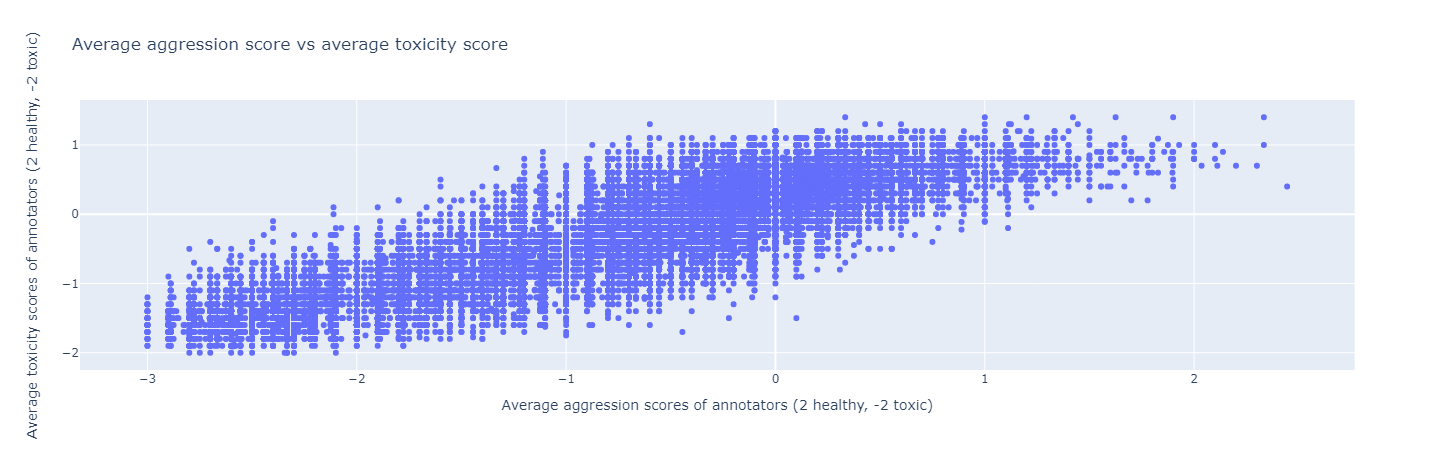
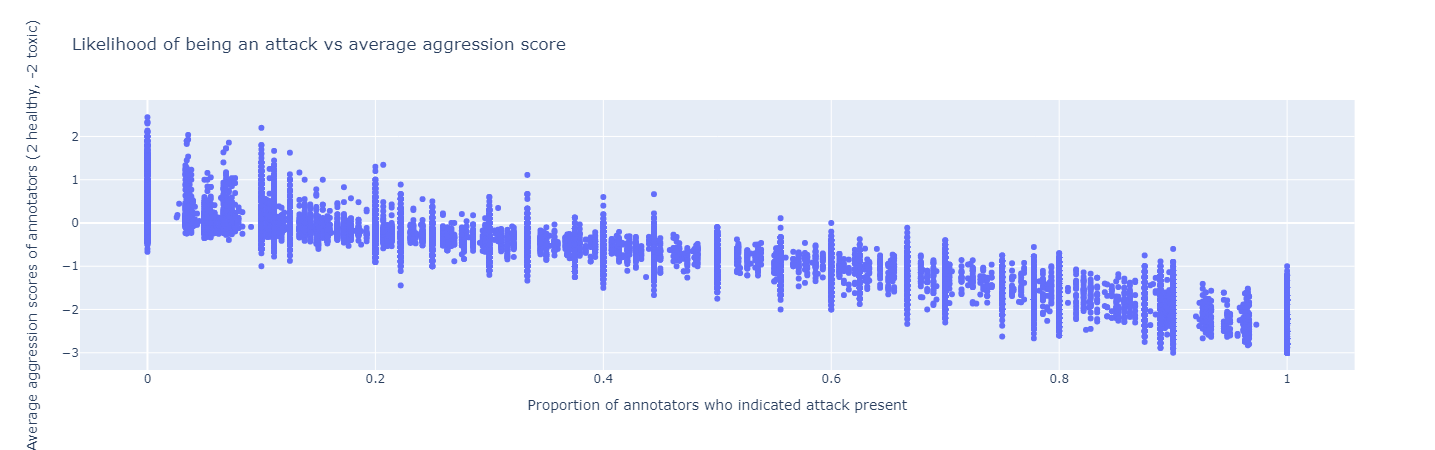
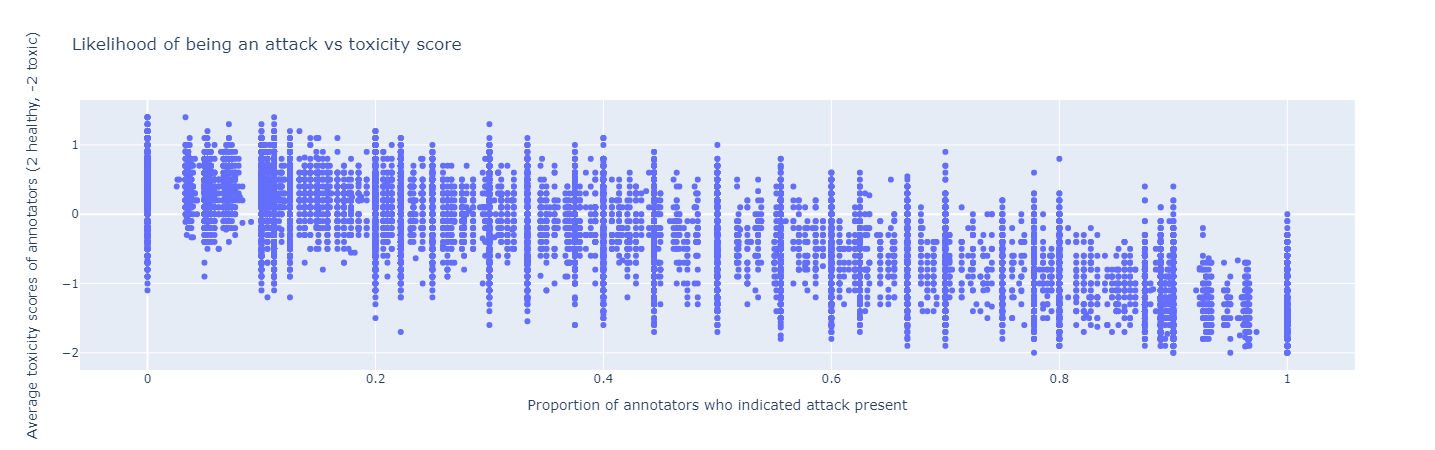
* (Attack and toxicity datasets)
* **Contribution:** Measuring and mitigating unintended bias in machine learning models by producing new metric
* **Model:** CNNs, Keras, Tensorflow
* **Metrics:** Pinned AUC

Magu, R., Hossain, N. and Kautz, H., 2018. Analyzing uncivil speech provocation and implicit topics in online political news. *arXiv preprint arXiv:1807.10882*.

* (Attack, aggression, and toxicity datasets)
* **Contribution:** Develop methods to study the emergence of incivility within the reader communities in news sites
* **Model:** Logistic Regression
* **Metrics:** AUC, precision, recall, F1 score, accuracy

***Overlap Analysis***

* 203 comments in toxicity dataset are duplicates
* Attack and aggression datasets have the same comments and annotators (115,864 comments, 159 of which are duplicates)
* 77,972 comments are in both toxicity and attack/aggression datasets (49% of toxicity dataset)
* (of non-overlapping comments) 8,051 comments with average aggressiveness score < 0
* 2,866 comments with average aggressiveness scores <-0.25, 1,343 with score<-0.5

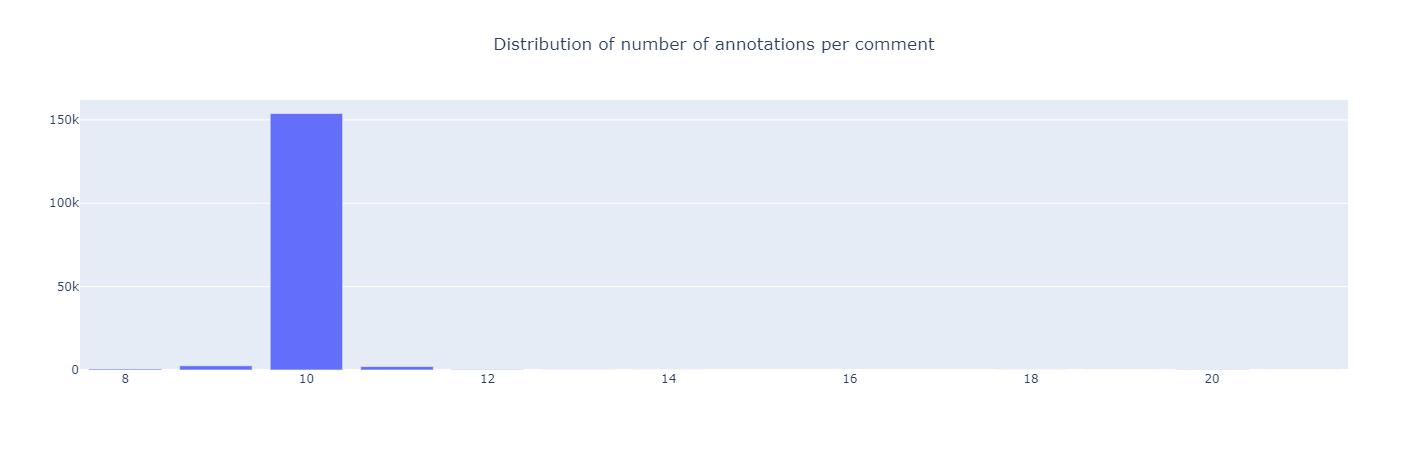


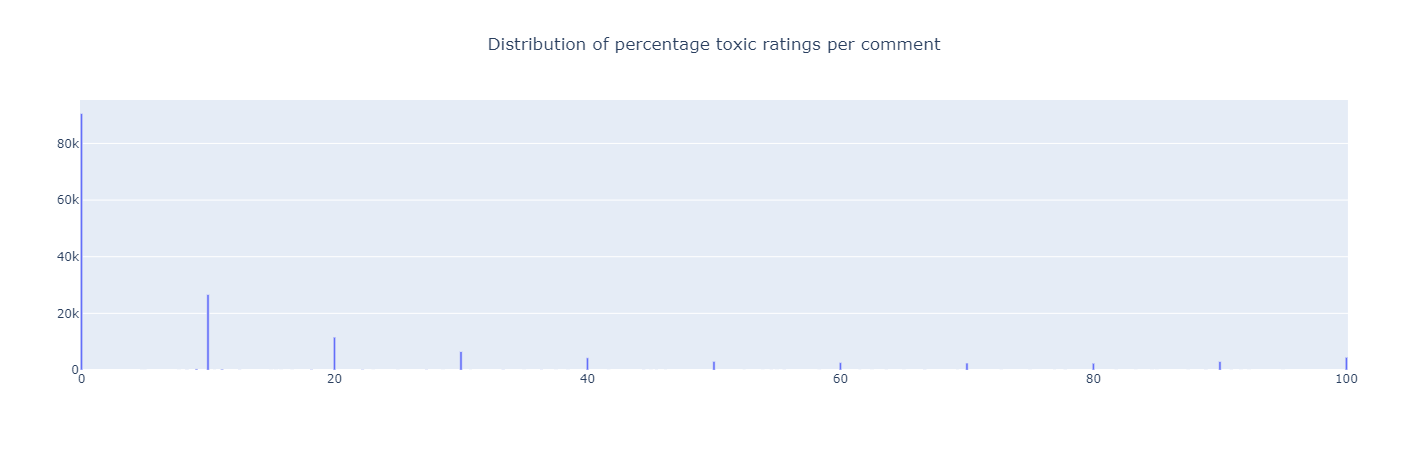
*Schema*

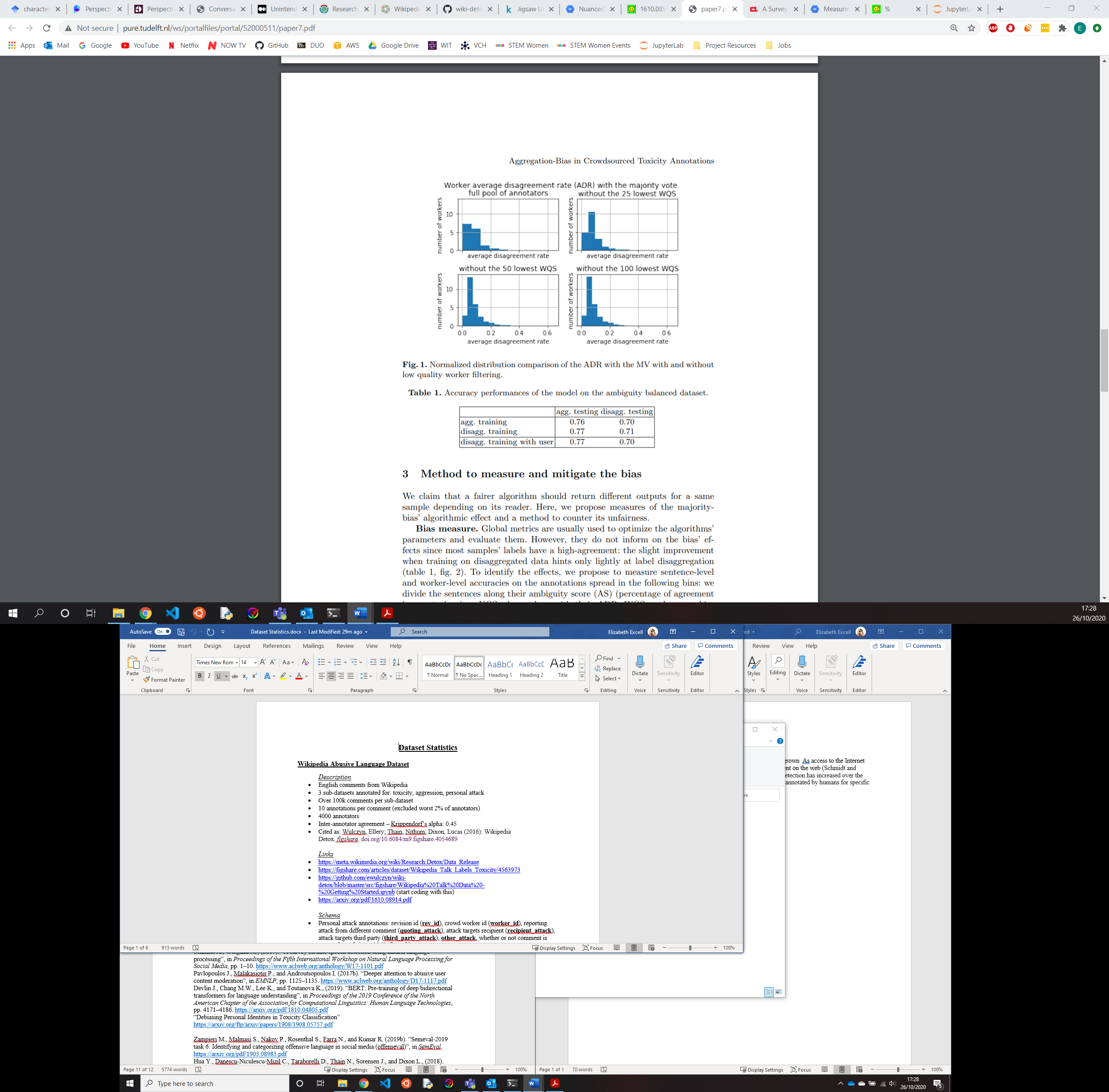
* Personal attack annotations: revision id (**rev\_id**), crowd worker id (**worker\_id**), reporting attack from different comment (**quoting\_attack**), attack targets recipient (**recipient\_attack**), attack targets third party (**third\_party\_attack**), **other\_attack**, whether or not comment 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)

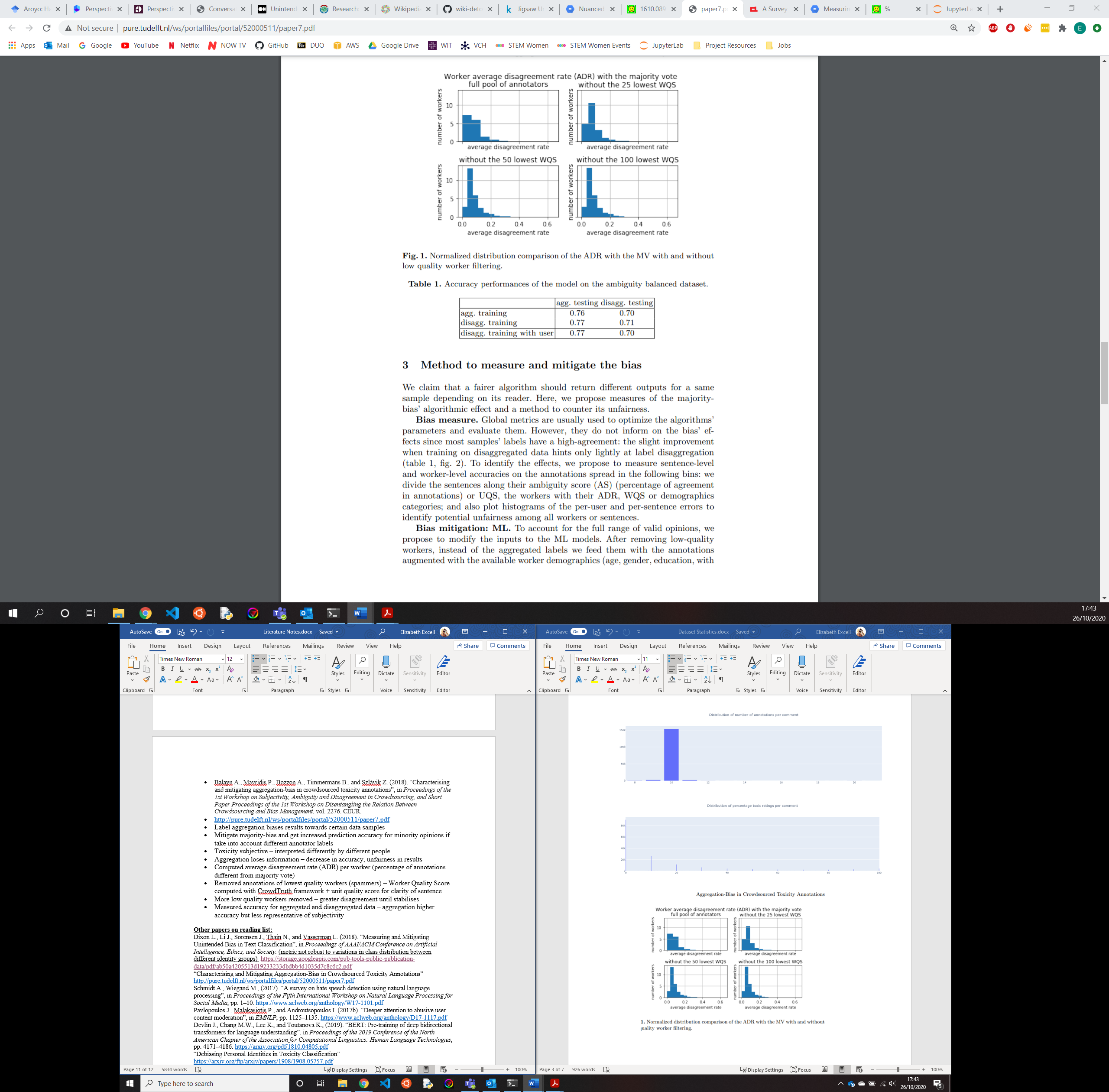
*Demographics*

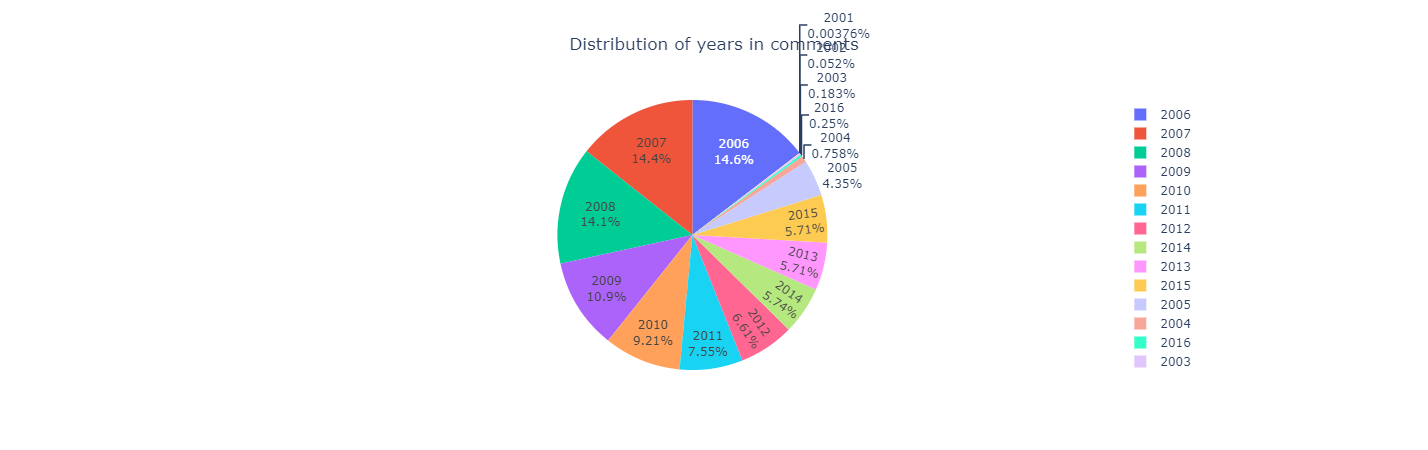
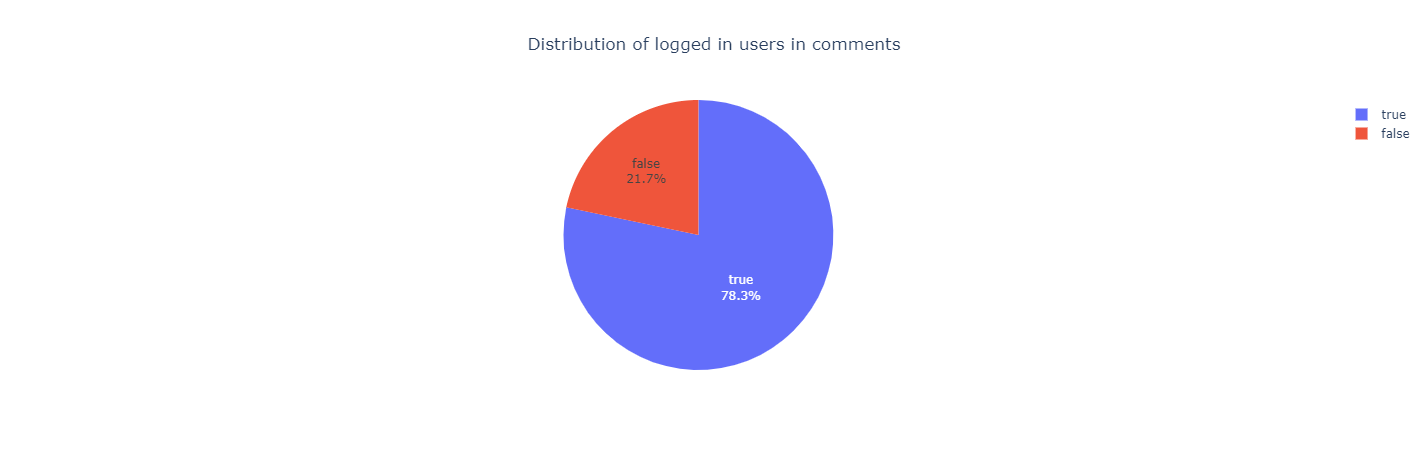
* Toxicity – 1,598,289 annotations, 159,686 comments, 4301 annotators, 371.61 average annotations per annotator, 10.01 average annotations per comment (majority had 10 annotators), average comment length 400.77 (after removing new line and tab tokens)
* Not much agreement on toxicity of comments (see 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

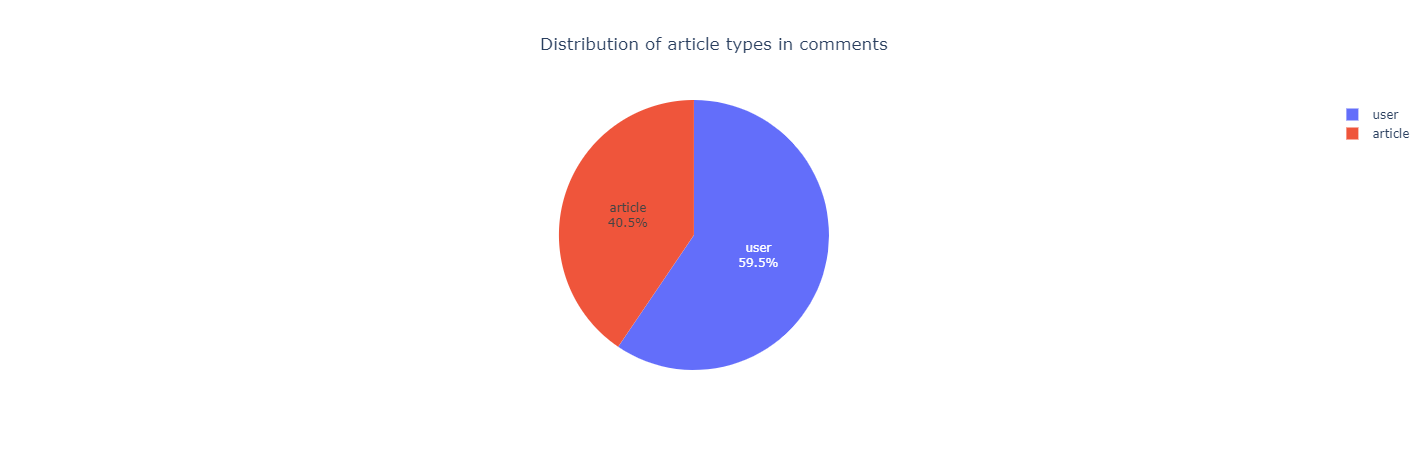
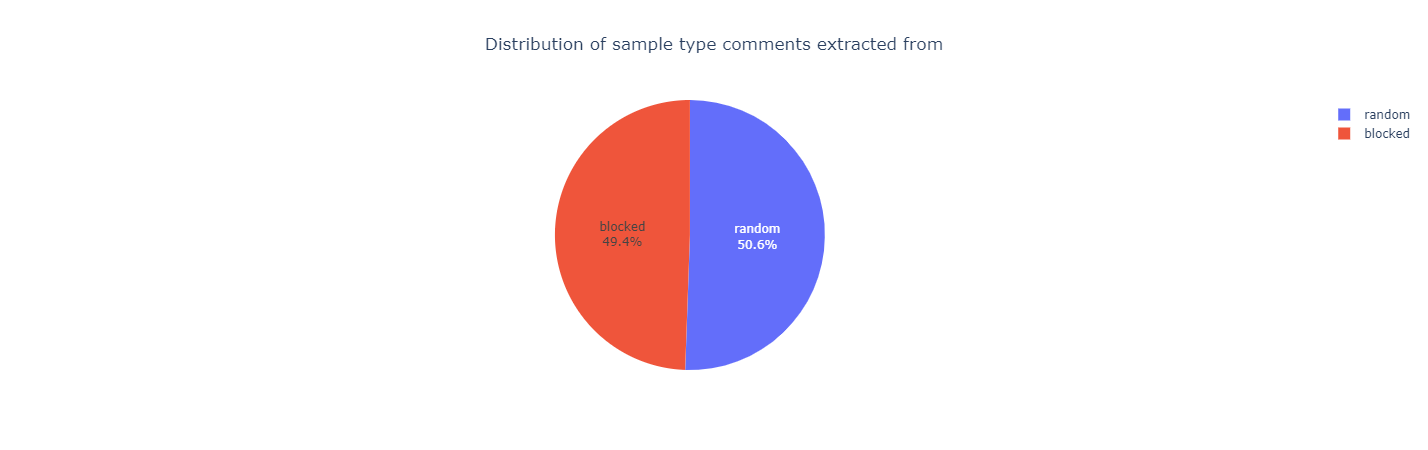


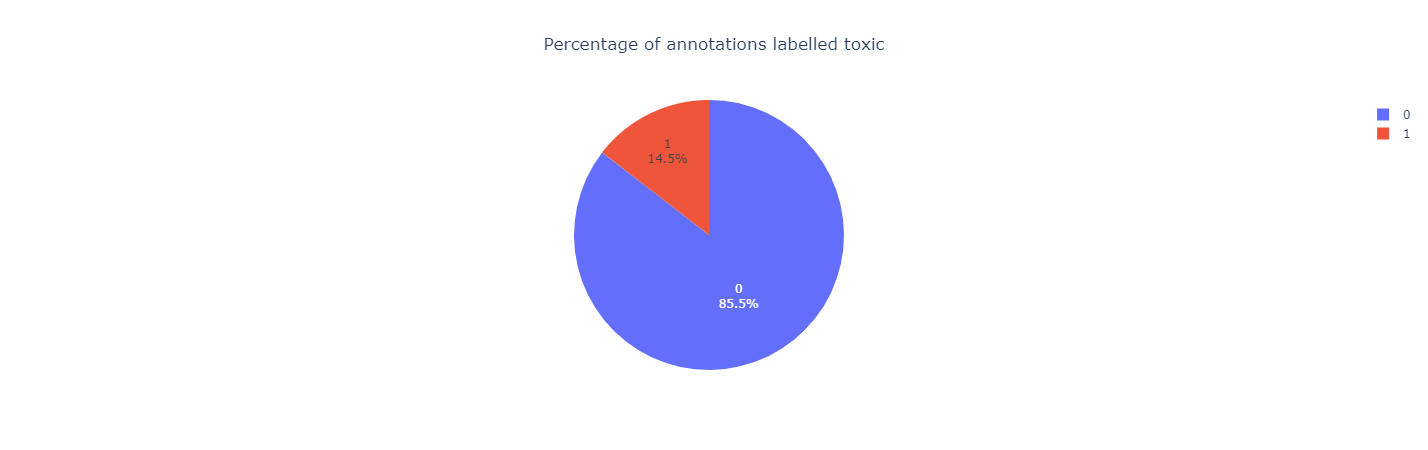
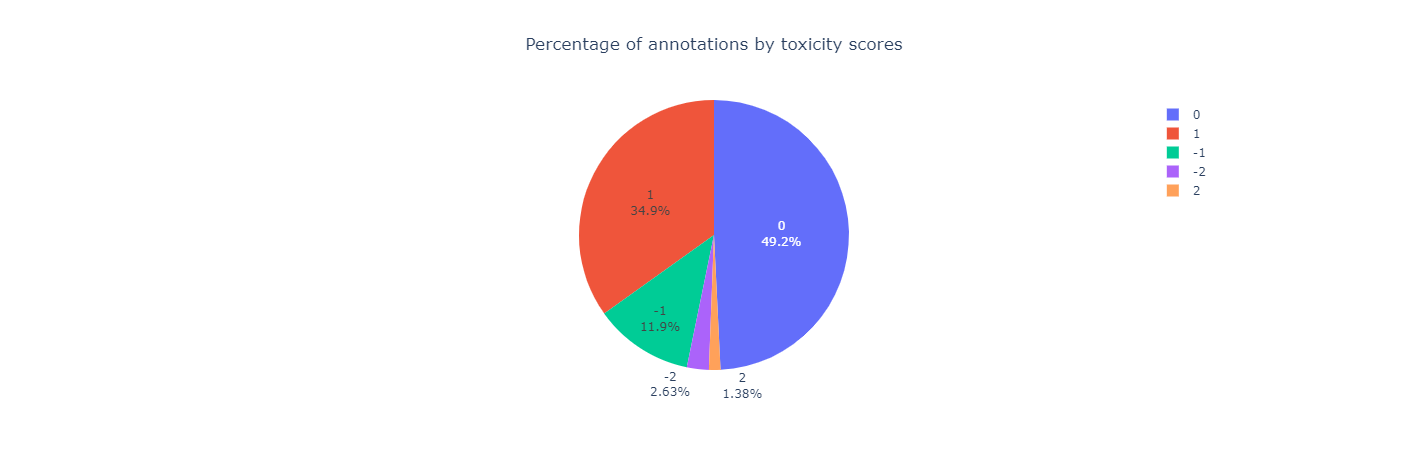


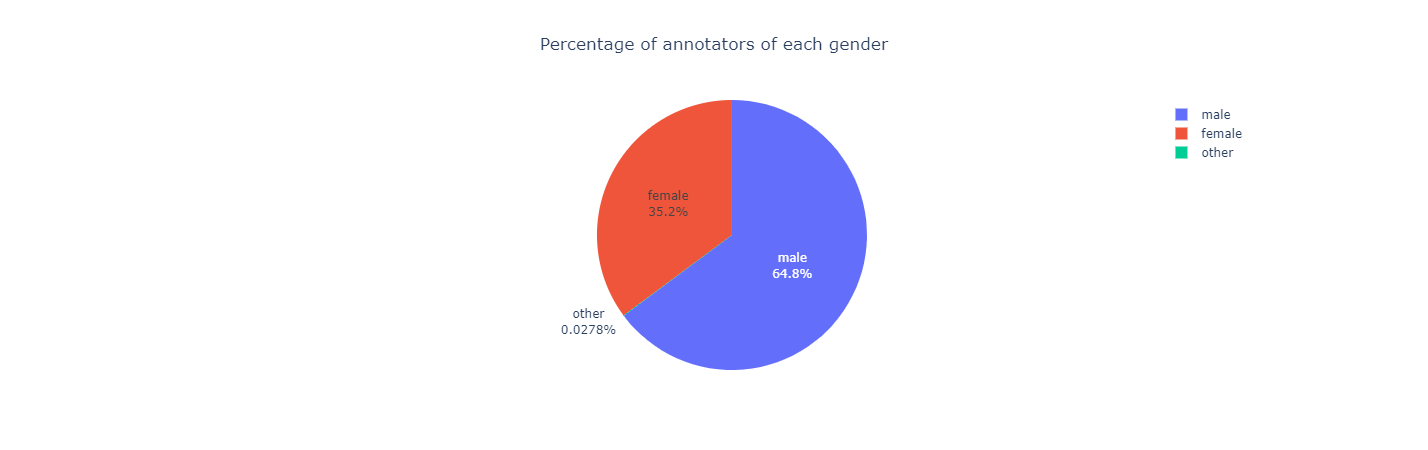
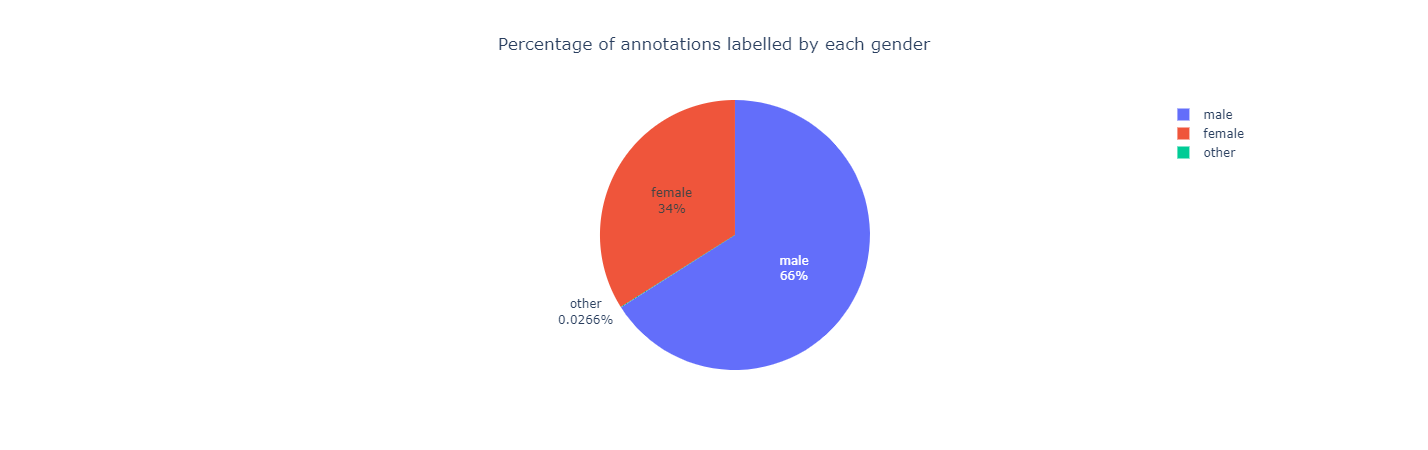


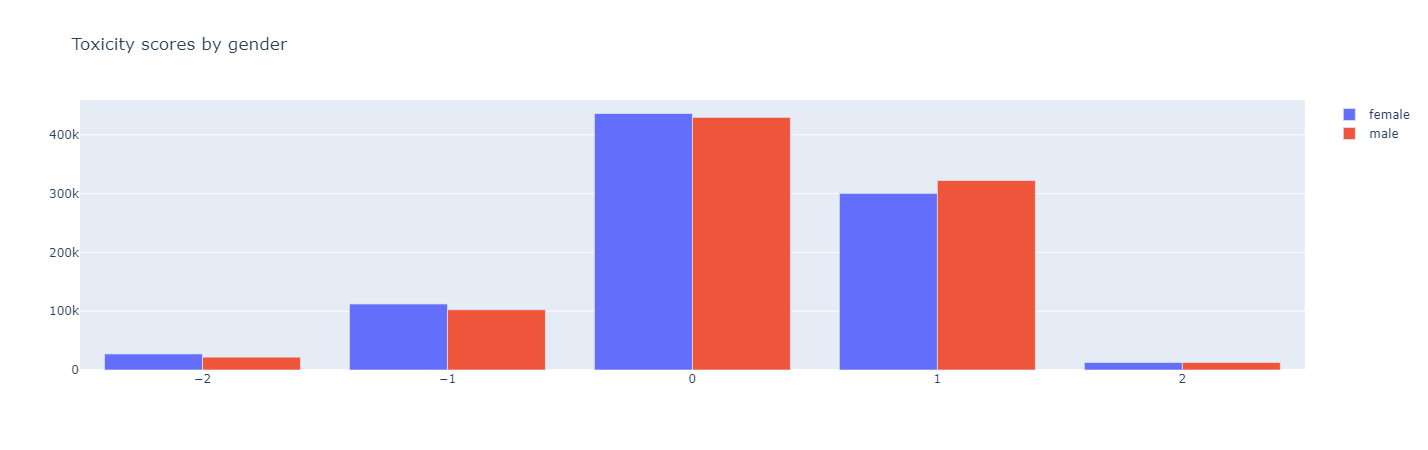


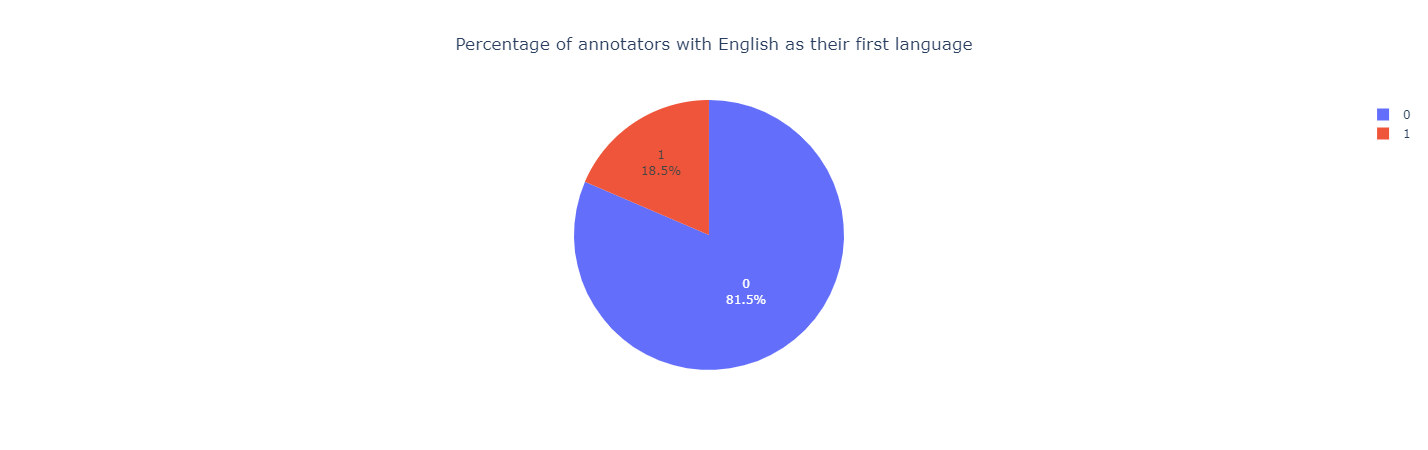
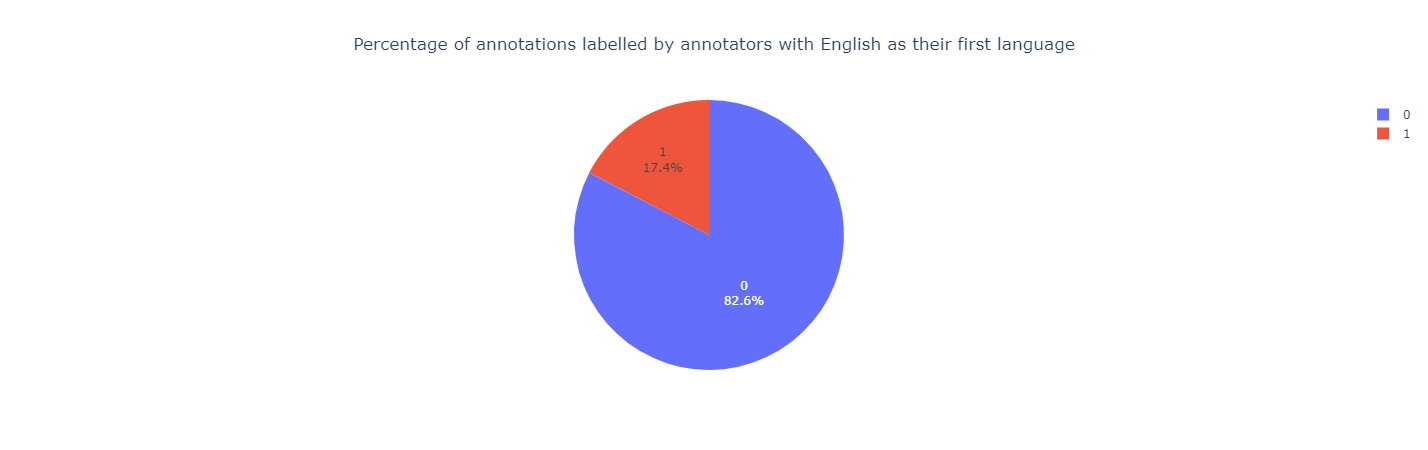
 

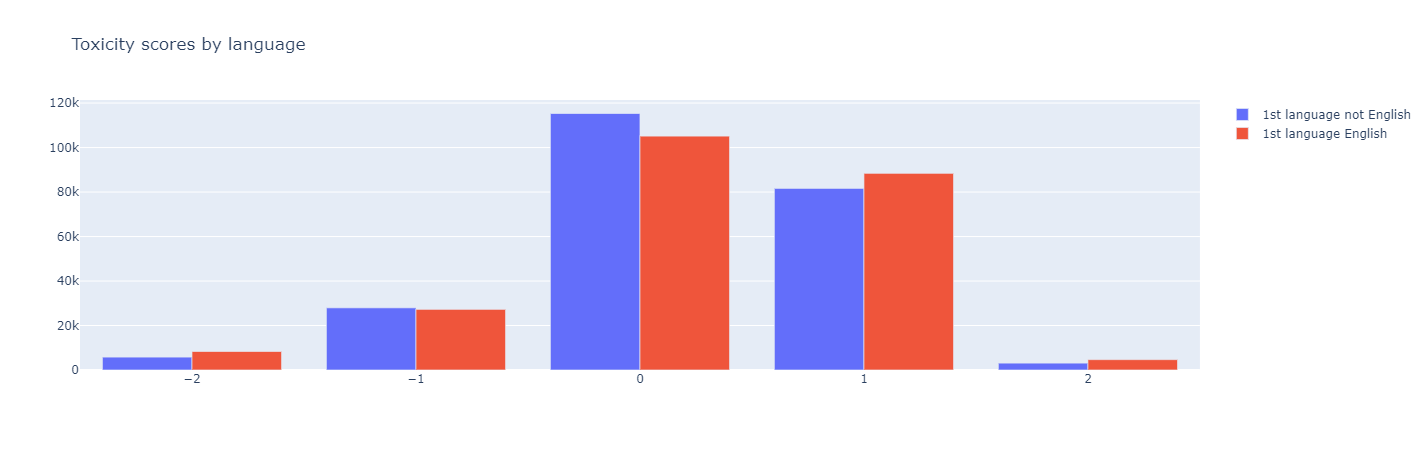
 

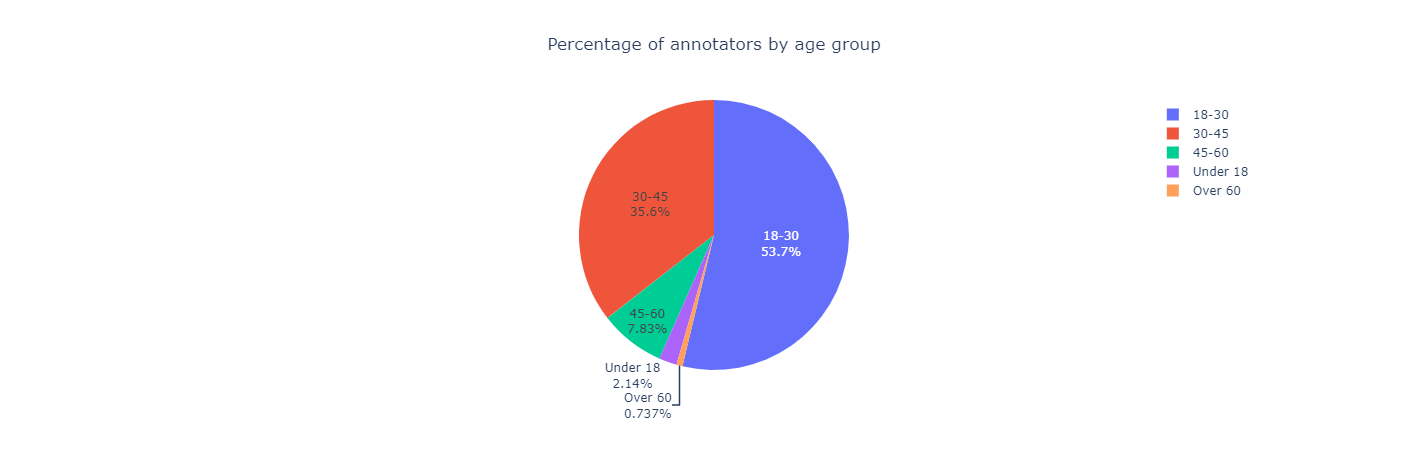
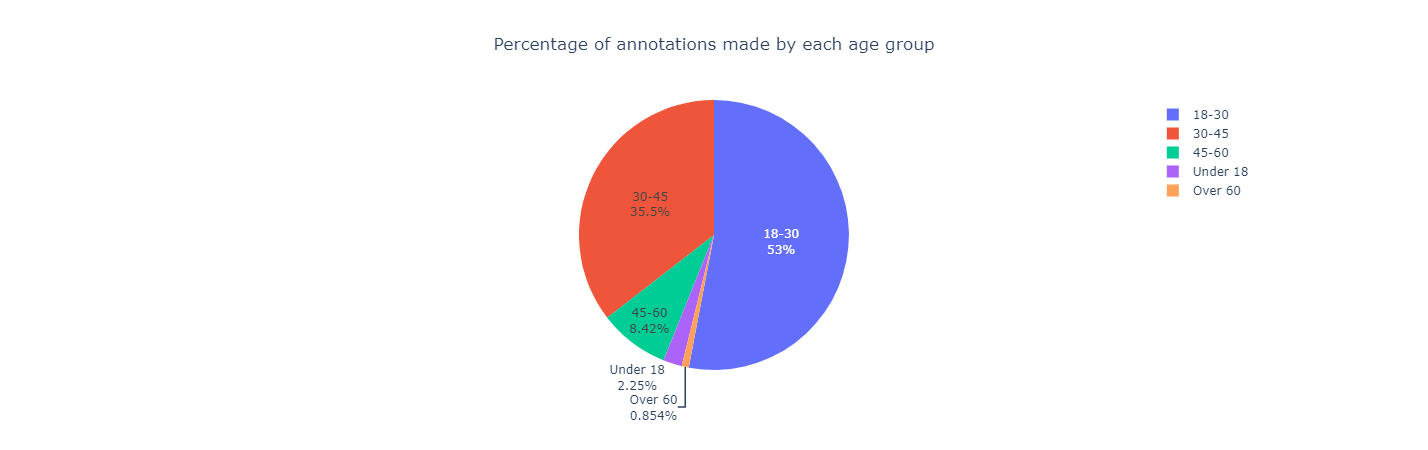
 

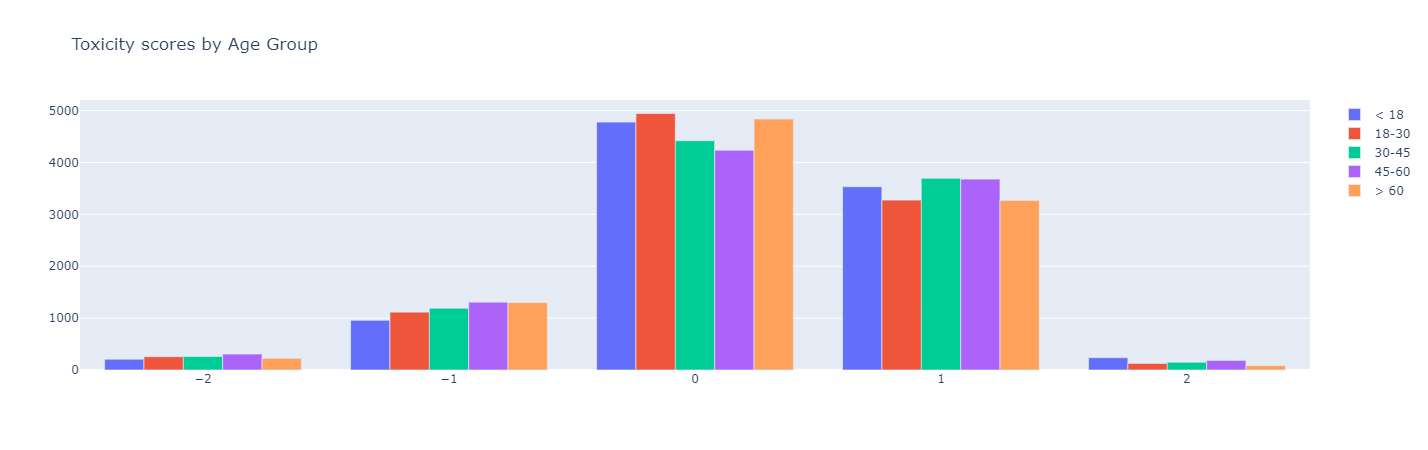
 

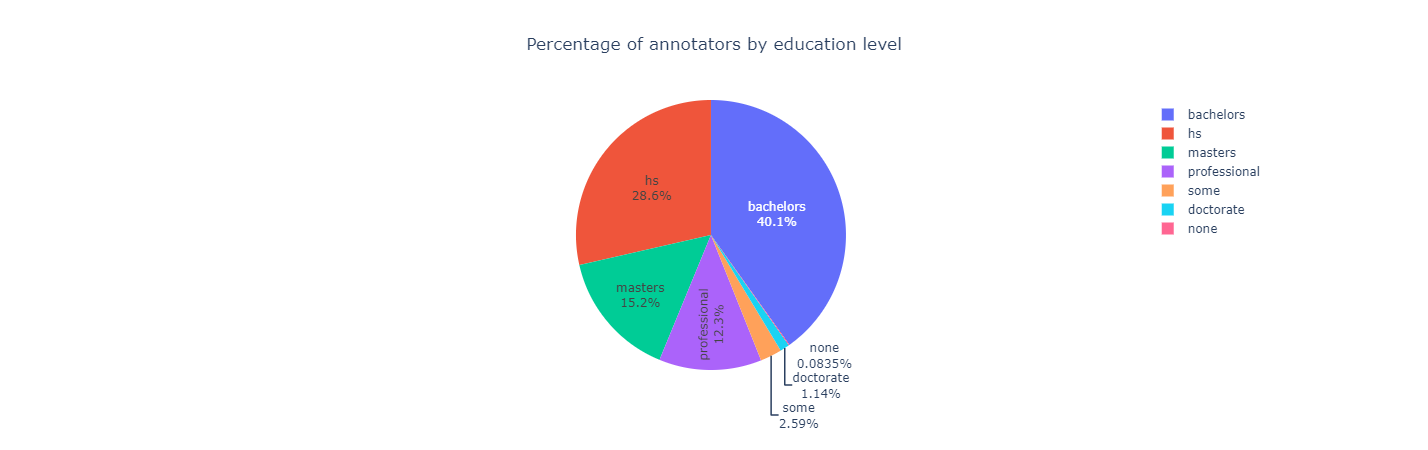
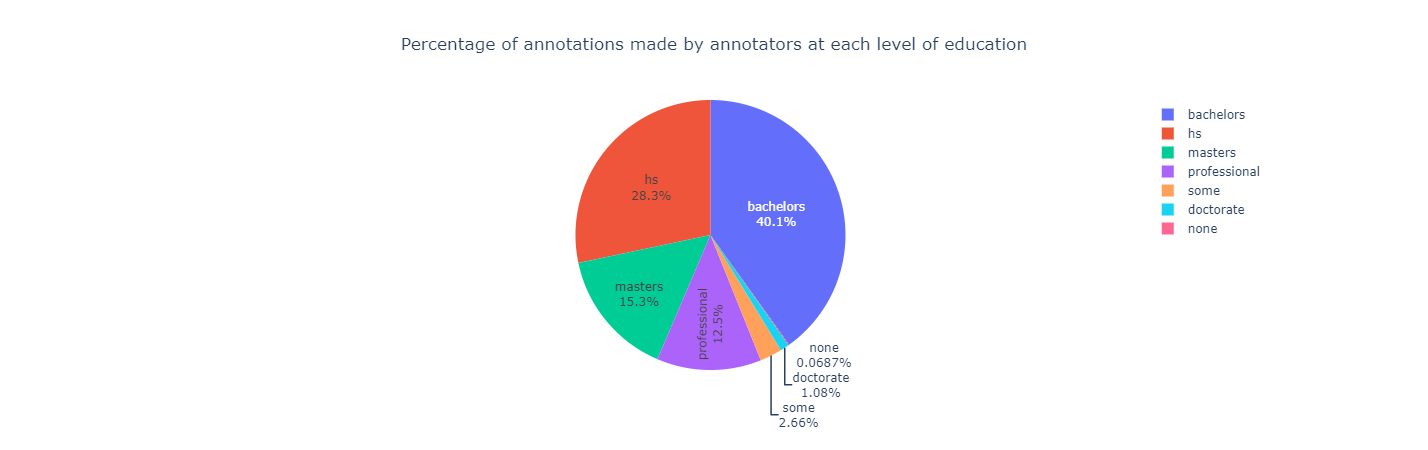


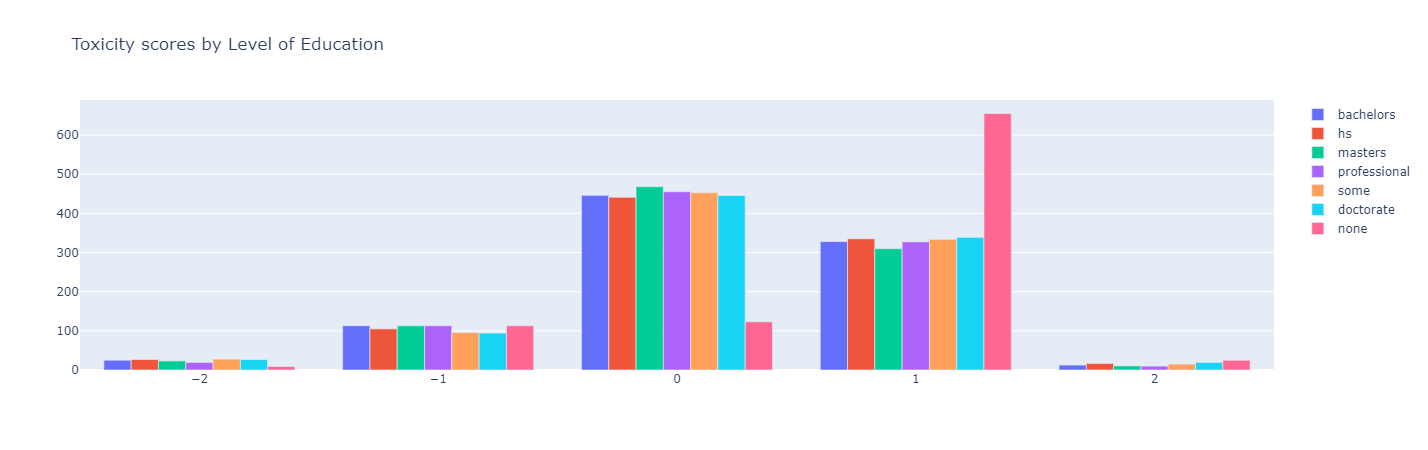
 







**Civil Comments Toxicity Kaggle (CCTK)**

*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*

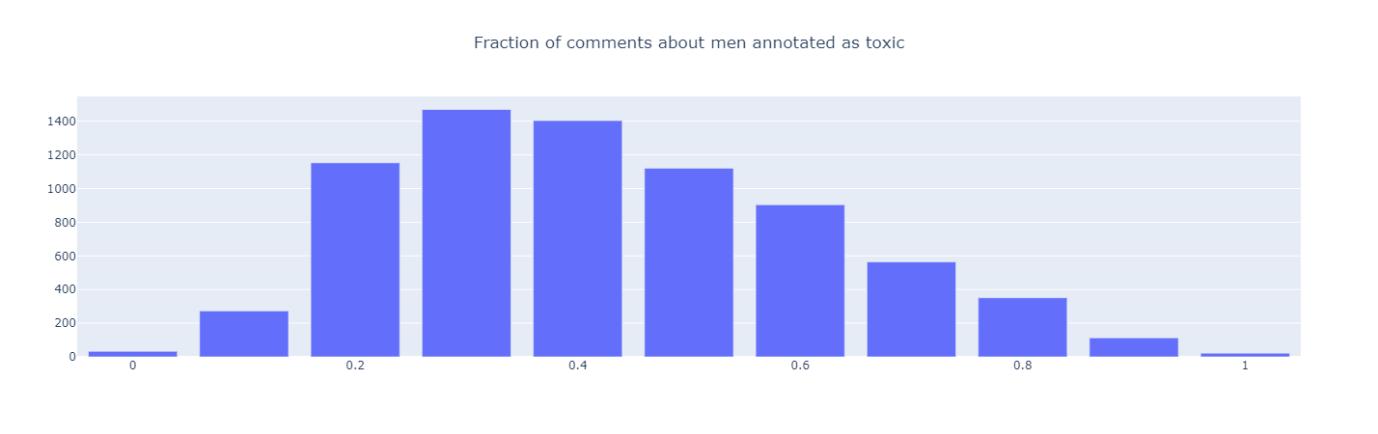
* **id, comment\_text, target** (fraction of annotators who believed comment toxic), **severe\_toxicity, obscene, threat, insult, identity attack, sexually explicit,** fraction of annotators who said identities mentioned in comment: **male, female, transgender, other\_gender, heterosexual, homosexual\_gay\_or\_lesbian, bisexual, other\_sexual\_orientation, christian, 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*

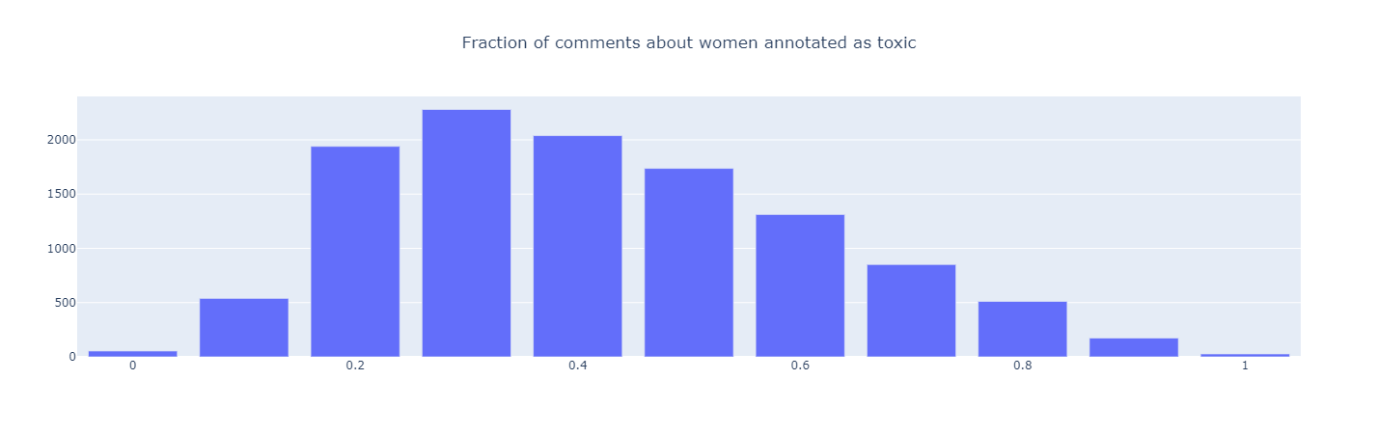
* 6.4 identity annotators on average, 10.4 toxicity annotators on average, mean comment length 351.74
* **18**% of identity comments labelled **male**, **18% female, 4% gay/lesbian, 14% Christian, 4% jewish, 5% muslim, 5% other religion, 5% black, 9% white, 4% Asian, 4% other race/ethnicity, 4% psychiatric/mental illness**
* **20,899 comments referencing male identity (591 with >=80% annotators labelling toxic, 1155 with >=70%, 2059 with >=60%)**
* **33,146 female (850 with >=80%, 1701 with >=70%, 3013 with>=60%)**
* **5487 homosexual (637 with >=70%, 1114 with >=60%, 1742 with >=50%)**
* **18,981 Christian (663 with >=70%, 1237 with >=60%, 2106 with >=50%)**
* **4405 Jewish (259 with >=70%, 475 with >=60%, 793 with >=50%)**
* **13,967 Muslim (644 with >=80%, 1271 with>=70%, 2182 with >=60%)**
* **8468 black (590 with >=80%, 1178 with >= 70%, 2077 with >=60%)**
* **14,357 white (715 with >=80%, 1586 with >=70%, 2901 with>=60%)**

Taking >= 10 annotators, >=80% annotators agree on identity

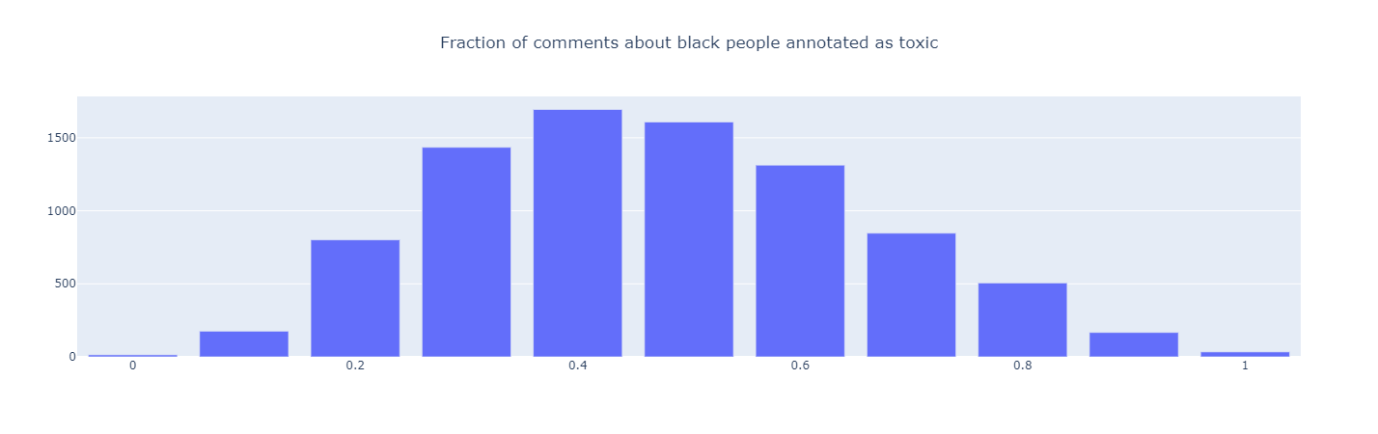
* **12,479** comments referencing **male** identity – toxicity a lot more ambiguous (low agreement)



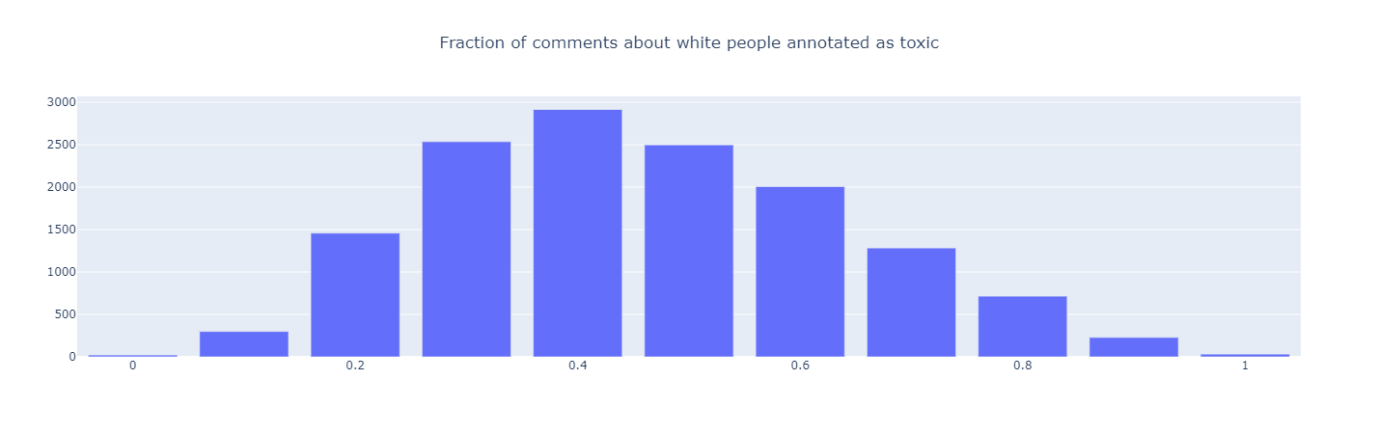
* **16,713** comments referencing **female** identity – toxicity ambiguous (low agreement)



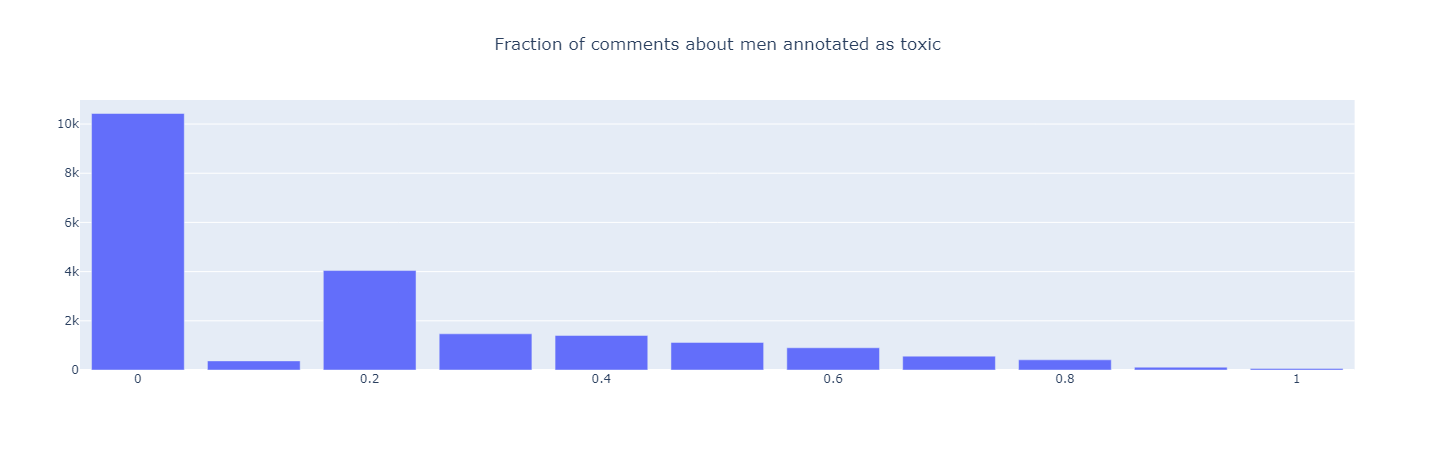
* **8595** comments referencing **black** people – toxicity ambiguous (low agreement)

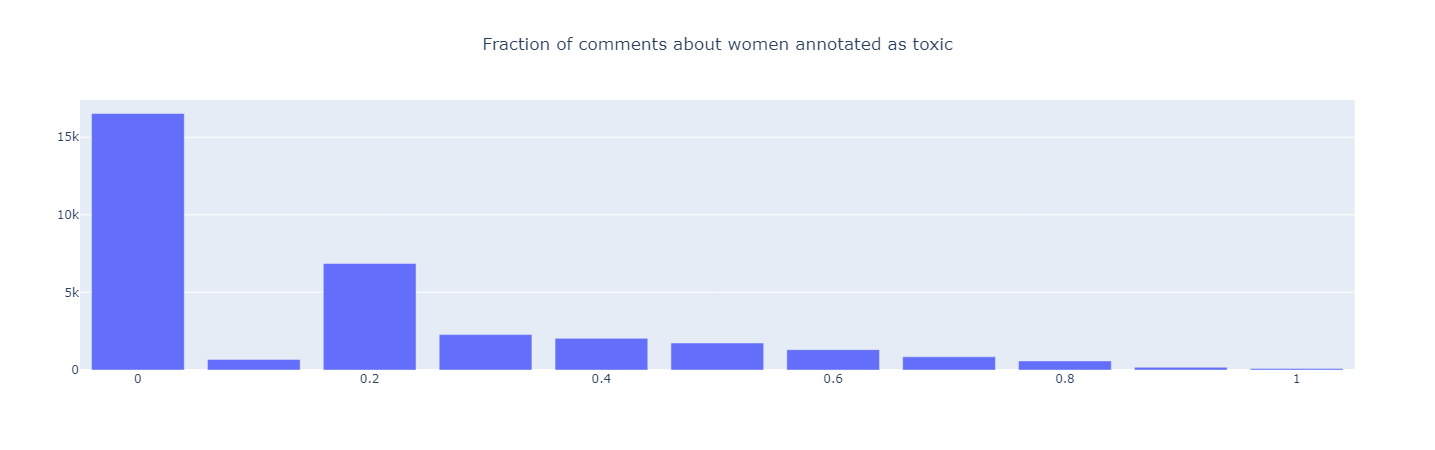


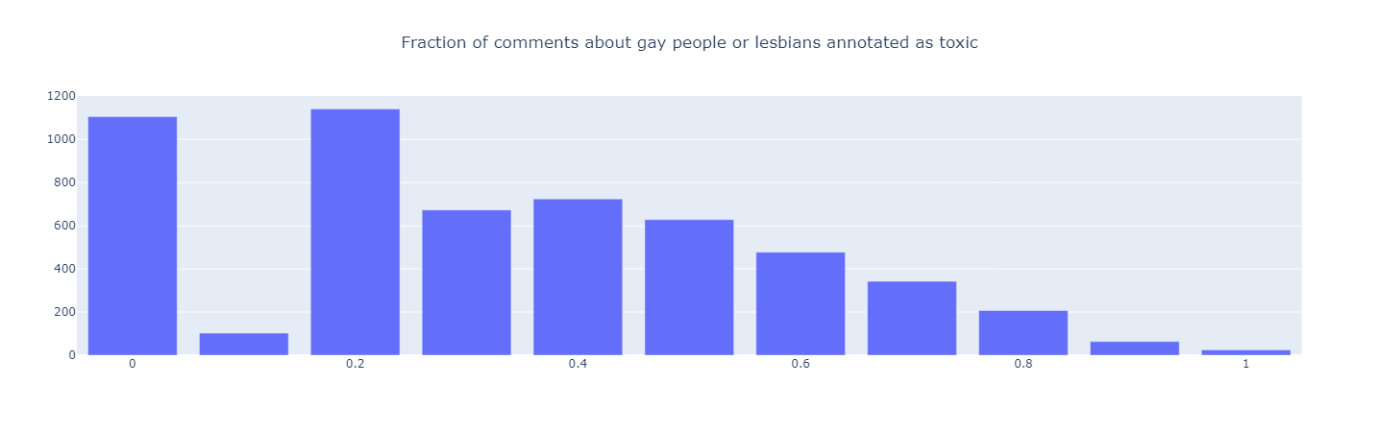
* **13,994** comments referencing **white** people – toxicity ambiguous (low agreement)

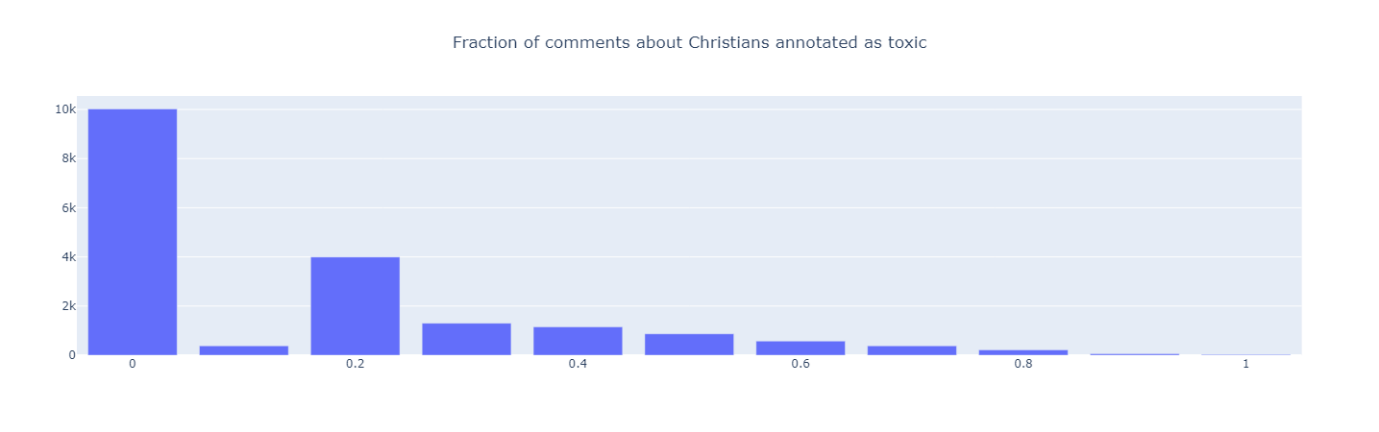


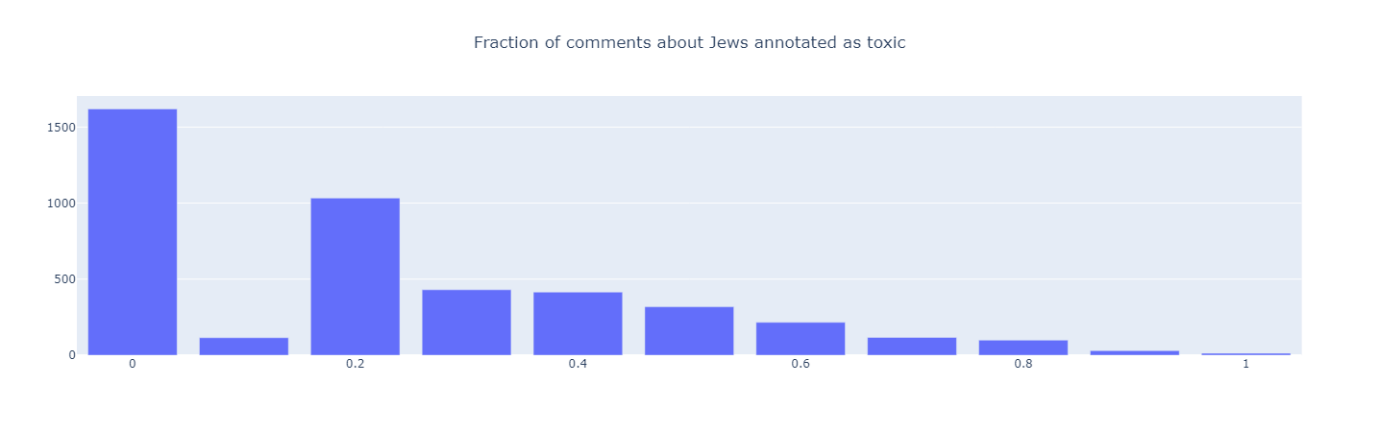
* (39471 comments overall if choose these thresholds)
* **5227** comments reference **men** and **women**
* **1028** comments reference **men** and **black** people
* **861** comments reference **women** and **black** people
* **2342** comments reference **men** and **white** people
* **1153** comments reference **women** and **white** people
* (3937 comments reference both gender and race)
* **3146** comments reference **black** and **white** people
* **Analyse:** Men with/without debiasing (without race or female overlap), women with/without debiasing (without race or male overlap), men and women overlap with/without debiasing (without race overlap), black people with/without debiasing (without white people or gender overlap), white people with/without debiasing (without black people or gender overlap), black and white people overlap with/without debiasing (without gender overlap), gender and race overlap with/with gender/with race/without debiasing
* **Annotate all comments twice (with + without debiasing) and annotate 3937 race and gender comments 4 times (with/with race/with gender/without debiasing) – would need to annotate 85,234 comments at these thresholds**

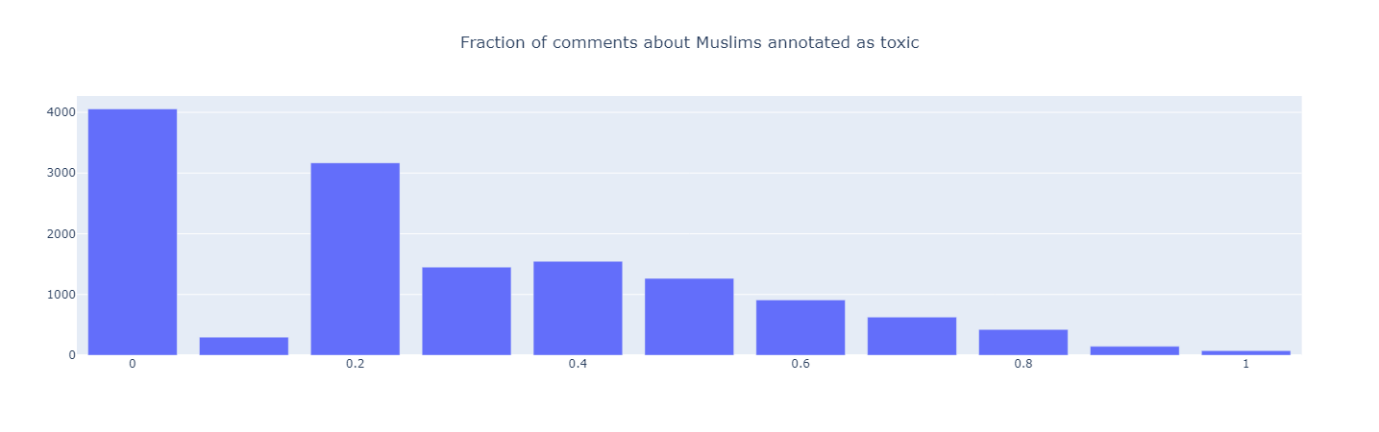


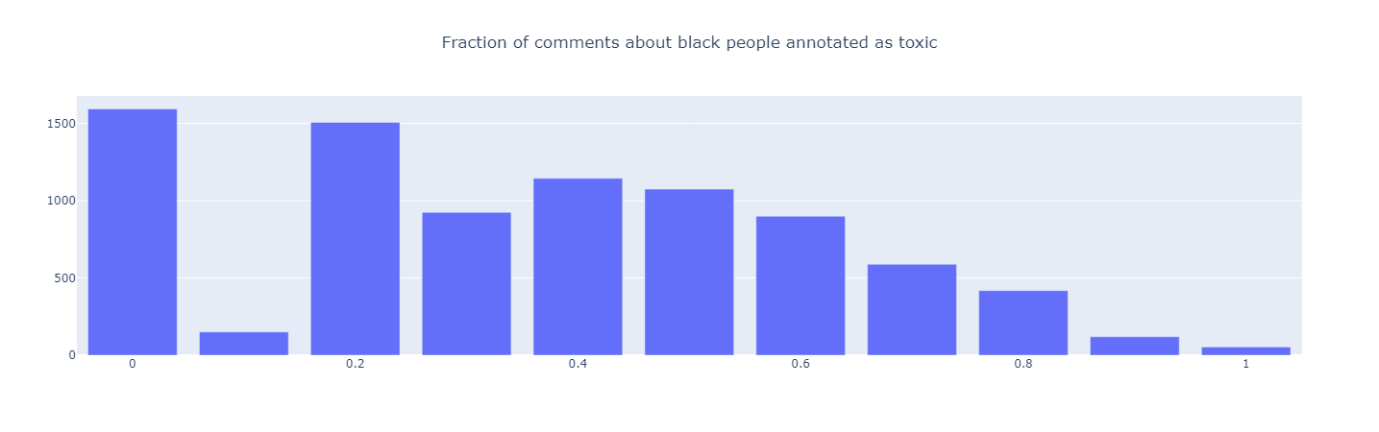


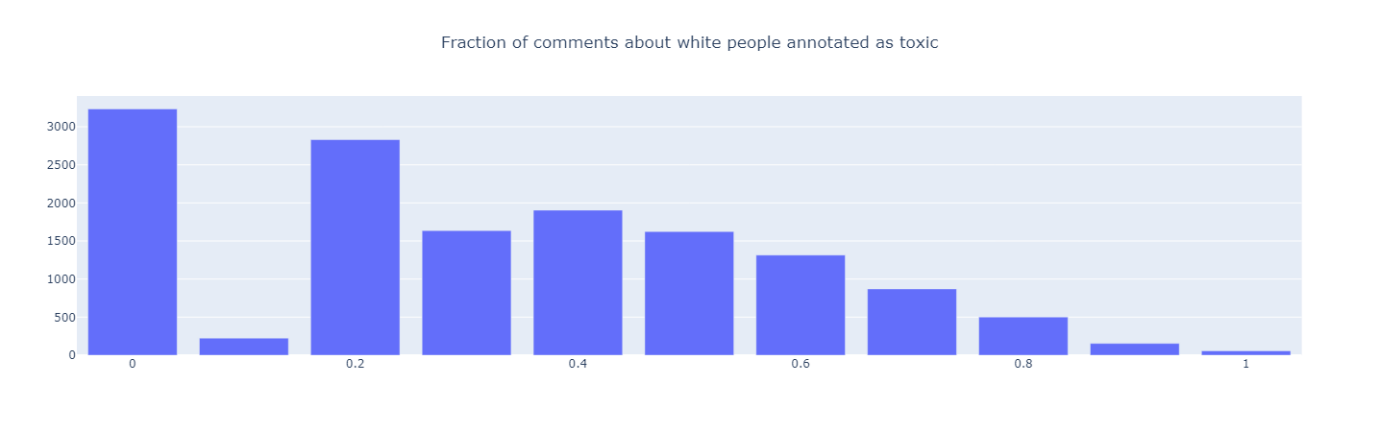












* Word cloud of most common terms associated with each identity