**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, aggression, personal attacks
* total unique comments in each, total unique annotators
* no. comments overlapping datasets?
* %M/F, English, Age groups and education in each dataset
* Average no. annotations for M/F/English/not/……
* Average toxicity score for M/F/…

**Civil Comments Toxicity Kaggle (CCTK)**

*Description*

* ~2M English comments from Civil Comments
* 450,000 comments labelled for identity
* Unintended bias in dataset for subgroups

*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)