**Ideas**

* (From meeting) – see what happens if mask offensive words from dataset – can we still accurately tell toxic language based on context? (might be innocent if between friends)
* Current findings that performance of classifier doesn’t improve when context-aware due to small number of context-sensitive comments in dataset – use new dataset? Look at how to efficiently annotate larger corpora of comments in context? See if can find additional context beyond title/parent comment? – comments randomly sampled – look into particular topic/tones/frequently target communities?
* Evaluating relative benefit of CCTK dataset compared to sub-string matching of terms that reference an identity (effectiveness of annotator’s recognition of identities)
* Developing full taxonomy of different possible biases + systematic approach for metrics used in diagnosis
* Investigating how relationships between annotators (demographics/ if agreed with comment) /way annotated (multiple questions/asked independently) affects the annotations and classification results
* Whether some terms are predictive of constructive comments / how the context of a conversation affects the constructiveness
* Modelling contributions to discussions other than constructiveness – diverse points of view / healthy levels of disagreement
* Impact of personal attacks/toxic comments on target user’s future contributions?
* Based on conversation, predict likelihood of next comment being an attack

|  |  |
| --- | --- |
| **Idea** | **Feasibility** |
| How does masking offensive words affect the classification decision (with/without context)? Investigate which classifiers can use rest of comment/context to predict toxicity | **Datasets:** Toxicity Detection w/ and w/o context, Synthetic Test Set, CCTK, Wikipedia Machine Annotations of Talk Pages, Davidson et al. (2017), OLID, FDCL18  **Current Research:** Some of above datasets collected comments containing specific hateful words, Paper 1 looked at role of context in classification decisions, some papers look at how users mask offensive language to evade detection (without looking at context) and words semantically related to offensive words: <https://dl.acm.org/doi/abs/10.1145/3243082.3243111>  <https://www.aclweb.org/anthology/W12-2103.pdf>  **Pros:** Have large twitter and Wikipedia datasets filled with offensive comments, some with context; question has not been looked at before with the angle of removing words but adding context  **Cons:** Not hugely novel, not many comments have context in datasets so that angle could lead nowhere  **Score:** **5/10**  **Improve:** Try to make as different to previous work as possible, ensure context can be used effectively |
| How effectively can an annotator recognise whether the target/speaker belongs to a particular identity and how does that change their classification decision? | **Datasets:** CCTK, Aggression-annotated Corpus of Hindi-English Code-mixed Data, OLID, SOLID, Demographic 16  **Current Research:** below examines difference in annotations when annotator is aware of race of author  <https://homes.cs.washington.edu/~skgabrie/sap2019risk.pdf>  **Pros:** Interesting and novel take on annotator bias  **Cons:** Some datasets contain some target identities (would need to confirm if annotations correct/what identities missed) – only 1 looks at likely speaker identity but not annotated; would need to get annotators to annotate data  **Score: 2/10**  **Improve:** Need dataset asking for annotators impression of target/author identities + correct identities– ask some to factor in, others not to |
| Investigate all biases in large datasets and which metrics can be used to best identify them/how to correct them/which classifiers more robust to bias | **Datasets:** CCTK, Wikipedia Toxicity Kaggle, Wikipedia Machine Annotations of Talk Pages (largest datasets), Synthetic Test Set  **Current Research:** <https://storage.googleapis.com/pub-tools-public-publication-data/pdf/66073ca7ac60ee38e93fc1d173a09cab65f2fef3.pdf>  looked at metrics used to investigate unintended bias  <https://arxiv.org/abs/1909.09758>  <https://ieeexplore.ieee.org/abstract/document/9087368>  <https://www.aaai.org/ojs/index.php/ICWSM/article/view/7334>  <https://dl.acm.org/doi/abs/10.1145/3278721.3278729>  <http://pure.tudelft.nl/ws/portalfiles/portal/52000511/paper7.pdf>  <https://homes.cs.washington.edu/~skgabrie/sap2019risk.pdf>  all look at measuring/mitigating unintended bias (and there are more)  **Pros:** Can thoroughly evaluate large datasets for different types of bias and produce analysis of how to avoid it for future publications  **Cons:** Already fairly widely researched, might be hard to do something new, need to have identities of subgroups  **Score: 4/10**  **Improve:** Find new angle on bias |
| How do the demographics of the annotators affect the annotations and classification results? (could try to find toxic comments containing “she” or other words related to women and see if toxicity scores given by women higher, same for xenophobia and English not first language/education and ageism and age group? – investigate bias in annotator demographics (mostly middle aged white men?) and see how classification affected when toxic comments annotated by victimised identities) – could even use a recommender system to predict annotator’s score for new comments given old scores by combining with identity-labelled dataset? | **Datasets:** Wikipedia Abusive Language Data Set (gender, English first language, age group, education)  **Current Research:** Some papers briefly mention the demographics of their annotators (e.g. 75% white) but not much research into annotators themselves  <http://pure.tudelft.nl/ws/portalfiles/portal/52000511/paper7.pdf>  looks into aggregation bias from crowdsourcing  **Pros:** Interesting, novel, could try to find correlations between demographics and toxicity scores – e.g. women may find sexist remarks more toxic than men, but if the annotators are mostly men they could overrule the female vote  **Cons:** Only have 1 dataset, contains demographics of annotators but only toxicity score of comments – no identities so would have to manually find sexist examples for women, ageist examples for older groups  **Score: 6/10**  **Improve:** need to ensure there is enough data on the demographics of the annotators, would be good if could get identities within annotated comments |
| Which terms are predictive of a comment being constructive and how does context help in telling the constructiveness of a comment? | **Datasets:** Constructive comments corpus, SOCC, YNACC, SENSEI  **Current Research:**  some research on how conversations turn awry and how constructive comments aid a conversation  <https://arxiv.org/pdf/2004.05476.pdf>  <https://www.aclweb.org/anthology/W17-3002.pdf>  <https://www.aclweb.org/anthology/W17-4218.pdf>  <https://www.aclweb.org/anthology/P19-1250.pdf>  **Pros:** Not much research into constructive comments  **Cons:** Not much context, may not add much to above papers  **Score: 4/10**  **Improve:** get more context for wider variety of conversations, ensure adding to current research, not replicating it |
| Investigating factors that contribute to a healthy conversation (diverse points of view, healthy disagreement, politeness/respect) | **Datasets:** Constructive comments corpus, SOCC, YNACC, SENSEI  **Current Research:**  some research on how conversations turn awry and how constructive comments aid a conversation  <https://arxiv.org/pdf/2004.05476.pdf>  **Pros:** Novel, hasn’t been looked at before  **Cons:** Unsure how to measure these factors in a conversation and unsure if data is there to investigate these factors  **Score: 2/10**  **Improve:** could combine with above idea just looking at constructiveness in general – not very novel |
| How being exposed to toxic comments in a conversation affects the rest of the users in the conversation? (likely to return attack/withdraw from conversation/try to be polite) | **Datasets:** Constructive comments corpus, SOCC, YNACC  **Current Research:**  <https://dl.acm.org/doi/abs/10.1145/3366423.3380074>  lots of research looking at triggers – not much focus on the effects of toxic comments  **Pros:** Novel  **Cons:** Might be difficult to get/connect user information – would need to figure out who was in the conversation, who it was directed at, who continues to contribute and see if patterns different for users across conversations (unsure if have data)  **Score: 2/10**  **Improve:** get more information on users in conversation |
| Predicting likelihood of next comment being toxic given conversation | **Datasets:** Constructive comments corpus, SOCC, YNACC  **Current Research:** Research focuses on toxicity of comments given rather than likely toxicity of next comment in conversation  some research on how conversations turn awry  <https://arxiv.org/pdf/2004.05476.pdf>  **Pros:** Novel, have conversation data from many different threads  **Cons:** Unsure how good predictions will be or if idea too basic?  **Score: 6/10**  **Improve:** More data on authors of comments, add to idea to make more complex and expand contribution |

**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?