**Literature Review**

Toxic language detection has attracted significant research interest in recent years as the volume of toxic user-generated content online has grown with the expansion of the Internet and social media networks (Schmidt and Wiegand, 2017). As the features of each corpus and the definitions of toxicity vary significantly between different domains and authors, many research papers have taken disparate approaches to the problem of toxic language detection. Pavlopoulos *et al.* (2020) examined the effect of context on the classification of toxic comments using their own annotated subset of comments from Wikipedia Talk pages. Kolhatkar *et al.* (2020) also created their own corpus, filling theirs with constructive comments taken from news articles to examine the reverse of the classification task, promoting comments that are labelled as constructive rather than deleting comments labelled as toxic. Zhang *et al.* (2018) investigated the linguistic cues that indicate a conversation is turning awry based on dataset of their own creation containing conversations hosted on Wikipedia and Reddit.

While much of the current research in this area is varied in scope, domain, and objective, many authors have had to contend with bias present in their datasets. As the majority of corpora use human annotators to assign toxicity scores to comments, any biases held by the pool of annotators are propagated in the classifier which can lead to non-toxic comments from certain identity groups being identified as toxic, known as false positive bias. (Dixon *et al.*, 2018; Sap *et al.*, 2019). The majority of research into unintended bias has been done to identify the types of bias present in the corpora and measure them (Borkan *et al.*, 2019; Dixon *et al.*, 2018) while other research in the field of natural language processing investigates how classification results are affected by the aggregation of crowd worker annotations, minimizing the diversity of views present in the scoring of a highly subjective task (Arovo and Welty, 2013; Balayn *et al.*, 2018). The area of unintended bias in toxic language detection that was least represented in the literature was an examination of how the demographic makeup of the human annotators can cause bias in the model, which became the motivation for this project. The closest research to this in the field of toxic language detection was conducted by Sap *et al.* (2019), who examined racial bias in Twitter corpora, where the majority white annotators gave higher toxicity scores to tweets with an African American English dialect.

**Terminology and Datasets**

Due to the broad definition of toxic language and it’s highly subjective nature, much of the current research focuses on different subtypes of toxicity, such as hate speech (Sap *et al.*, 2019; Schmidt and Wiegand, 2017), abusive (Nobata *et al.*, 2016; Park and Fung, 2017), and offensive language (Pavlopoulos *et al.*, 2019; Razavi *et al.*, 2010). Davidson *et al.* (2017) noted the importance of distinguishing between different types of toxic language by highlighting the legal and moral implications of hate speech and how much more destructive it can be to the targeted identity groups than commonplace offensive language. This range of terminology has led to the creation of a wide variety of corpora, each annotated using the author’s definition of toxicity and the specific focus of their research question, often meaning that the corpora cannot be reused for further research. Many of these datasets also have a limited size due to the time and expense of annotating large numbers of comments (Wulczyn *et al.*, 2017). This presents a challenge as the majority of datasets are incomparable and not transferable between tasks, meaning no accurate comparison of results between papers can be performed. This is also due to the differences in the domains of the corpora, which contain comments of different lengths from users with different demographics and exhibit different linguistic styles and bias compared to other corpora. The annotations also vary between datasets, in quality, number of annotators and the guidelines given to annotators including the scales to measure toxicity and the definitions of toxicity. This lack of consistency also leads to duplicated research as results are replicated on differing subtypes of toxicity and various corpora (Kumar *et al.*, 2018).

The definition of ‘toxicity’ in this paper is taken to be the same as the definition that the annotators of the chosen corpora based their toxicity scores on, namely ‘*a rude, disrespectful, or unreasonable comment that is likely to make you leave a conversation’* (Borkan *et al.*, 2019).

* [15] S. O. Sood, J. Antin, and E. F. Churchill. Using crowdsourcing to improve profanity detection. In AAAI Spring Symposium: Wisdom of the Crowd, 2012. – first to use crowdsourcing – go into available datasets, crowdsourcing and annotators, why datasets chosen

**Features**

**Classifiers**

* Simple systems first used blacklists and regular expressions (Koratana A. and Hu K. (2019). “Toxic Speech Detection”., Nobata C., Tetreault J., Thomas A., Mehdad Y., and Chang Y. (2016). “Abusive language detection in online user content”, in *ICWWW*, pp. 145–153.)
* Mention Perspective
* Kumar R., Ojha A. K., Malmasi S., and Zampieri M. (2018). “Benchmarking aggression identification in social media”, in *TRAC*, Santa Fe, USA.
* Mention BERT
* Mention versions of LSTM

**(Unintended Bias - Metrics)**

**References**

Aroyo L. and Welty C. (2013). “Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard”, *WebSci2013, ACM*.

Balayn A., Mavridis P., Bozzon A., Timmermans B., and Szlávik Z. (2018). “Characterising and mitigating aggregation-bias in crowdsourced toxicity annotations”, in *Proceedings of the 1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management*, vol. 2276. CEUR.

Borkan D., Dixon L., Sorensen J., Thain N., and Vasserman L. (2019). “Nuanced metrics for measuring unintended bias with real data for text classification”, in *Companion Proceedings of the 2019 World Wide Web Conference*,Association for Computing Machinery, pp. 491–500.

Davidson T., Warmsley D., Macy M., and Weber I. (2017). “Automated hate speech detection and the problem of offensive language”, in *ICWSM*, pp. 512–515.

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.*

Kumar R., Ojha A. K., Malmasi S., and Zampieri M. (2018). “Benchmarking aggression identification in social media”, in *TRAC*, Santa Fe, USA.

Nobata C., Tetreault J., Thomas A., Mehdad Y., and Chang Y. (2016). “Abusive language detection in online user content”, in *ICWWW*, pp. 145–153.

Park J. H. and Fung. P. (2017). “One-step and two-step classification for abusive language detection on twitter”, in *1st Workshop on Abusive Language Online*, pp. 41–45.

Pavlopoulos J., Sorensen J., Dixon L., Thain N., and Androutsopoulos I. (2020). “Toxicity Detection: Does Context Really Matter?”, in *Proc. of 58th Annual Meeting of Association for Computational Linguistics,* pp. 4296-4305.

Pavlopoulos J., Thain N., Dixon L., and Androutsopoulos I. (2019, June). “ConvAI at semeval-2019 task 6: Offensive language identification and categorization with perspective and bert”, in *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 571-576.

Razavi A.H., Inkpen D., Uritsky S., and Matwin S. (2010, May). “Offensive language detection using multi-level classification”, in *Canadian Conference on Artificial Intelligence*, pp. 16-27. Springer, Berlin, Heidelberg.

Sap M., Card D., Gabriel S., Choi Y., and Smith N.A. (2019, July). “The risk of racial bias in hate speech detection”, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1668-1678.

Schmidt A., Wiegand M., (2017). “A survey on hate speech detection using natural language processing”, in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pp. 1–10.

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

Zhang J., Chang J.P., Danescu-Niculescu-Mizil C., Dixon L., Hua Y., Thain N., and Taraborelli D., (2018). “Conversations gone awry: Detecting early signs of conversational failure”, *arXiv preprint arXiv:1805.05345*.