**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, filled with constructive comments taken from news articles to examine the reverse of the standard 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 containing conversations hosted on Wikipedia and Reddit. After a thorough examination of all of the fields involved in toxic language detection, the focus of this project was decided to be an examination of the role the demographics of annotators play in toxicity classification.

**Unintended Bias**

While much of the current research in this area is varied in scope, domain, and objective, many authors have had to contend with the 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 mislabelled as toxic, an effect known as false positive bias. (Dixon *et al.*, 2018; Sap *et al.*, 2019).

The most cited research in the area of unintended bias in toxic language detection has focused on identifying the types of bias present in the corpora and measuring 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 most recent research into unintended bias uses the metrics presented in Borkan *et al.* (2019) to build classifiers that can detect the identity groups mentioned in the comments as well as measuring and reducing the bias present in the toxicity scores produced by the classifiers (Hamida *et al.*, 2019). Reichert *et al.* (2020) sought to reduce this bias by balancing the toxic examples referencing common identity terms with example comments created through natural text generation, while Zhang *et al.* (2020) formalized the problem as a type of selection bias and proposed a debiasing framework based on instance weights for a set of pre-defined identity terms. Vaidya *et al.* (2020) compared the robustness of some of the top performing toxic language detection classifiers to unintended bias towards commonly attacked identity groups, applying an attention-based multi-task learning approach.

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. Similar research to this in the field of toxic language detection has been 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. In addition to this, Binns *et al.* (2017) explored methods for detecting potential bias by building classifiers trained on comments whose annotators came from different demographic groups, namely gender.

**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 language (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. The definition of ‘toxicity’ in this paper was chosen to be the same as the definition that the annotators of the selected corpus 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).

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 varying lengths from users with different demographics that exhibit different linguistic styles and forms of bias compared to other corpora. In addition, the annotations vary between datasets in quality, number of annotators and the guidelines given to annotators, including the system used to measure toxicity and the chosen definition of toxicity. This lack of consistency can lead to duplicated research as results are replicated on differing subtypes of toxicity and various corpora (Kumar *et al.*, 2018).

In recognition of the issue of oversaturation of toxic language corpora, some large datasets have recently appeared that generalise well to multiple tasks and have been widely adopted by the literature. The first of these datasets is taken from the Civil Comments Toxicity Kaggle challenge, a set of 2M comments taken from news sites and annotated for toxicity and all of its subtypes, with over a fifth of comments being annotated for mentions of commonly targeted identities, making the corpus highly useful for evaluating unintended bias. The corpus is seen as reliable due to having a large number of annotators per comment and because it was put forward by the Conversation AI team at Google, which accounts for its widespread use in the literature. (give examples and compare to how those papers describe datasets)

The second such dataset was created by Wulczyn *et al.* (2017), who created 2 datasets, the first containing over 160k comments with toxicity annotations, and the second containing over 100k comments labelled with personal attack and aggression annotations, roughly 78k of which were also in the toxicity dataset. The main benefit of this corpus, other than its size, is the inclusion of the demographic identities of the crowd workers, as it is the only publicly available corpus to do so, making it valuable for investigating how the demographics of annotators affect toxicity classifications, as this paper aims to do. As such, the larger and more broadly defined toxicity dataset will be used in this paper. (give examples – go into more reasoning? – look at own examination)

(quickly evaluate all datasets found in literature/any notable ones)

list some papers datasets used in, go into available datasets (talk about research on chosen datasets), why datasets chosen

**Need to include section on gender debiasing?**

**Crowdsourcing**

Crowdsourcing was first introduced as a technique for tailoring profanity detection to different corpora and domains in 2012 by Sood *et al.* Since then, it has become a popular technique used to gauge community opinions on the toxicity of comments, in addition to other variables such as identifying toxicity subtypes or references to identity groups. This also helps modern classifiers to overcome the challenges posed by list-based systems such as deliberate spelling mistakes used to enable toxic comments to evade detection as the semantics of the misspelled words are understood by the crowd workers and are included in the training data.

The most popular crowdsourcing platforms used in the problem of toxic language detection are Amazon Mechanical Turk and Figure Eight (formerly CrowdFlower). Buhrmester e*t al.* (2011) found that Amazon Mechanical Turk participants were more demographically diverse than other sample groups, provided a reliable source of data , and could be recruited rapidly and inexpensively. Tetreault *et al*. (2010) produced similar findings, showing that Amazon Mechanical Turk was as effective as using trained annotators for the task of grammatical error annotation, at a fraction of the time and cost.

However, in practise, the effectiveness of crowdsourcing appears to be mixed for much of the literature, with Kolhatkar *et al.* (2020) noting that expert annotators only agreed with the majority opinion of the crowdsourced annotations 87% of the time in the context of evaluating the constructiveness of comments, a verdict also reached by Nobata *et al.* (2016), who concluded that workers on the Amazon Mechanical Turk platform exhibited a much worse inter-annotator agreement than the in-house annotators in the task of abuse classification. Wulczyn *et al.* (2017) found that human annotators were too costly and inefficient, and so annotated comments using a classifier trained on crowdsourced annotations to annotate the rest of their dataset, concluding that the classifier had the same performance as the majority vote of 3 crowd workers. Balayn *et al.* (2018) discussed the bias present as a result of the aggregation of crowd worker annotations, highlighting that many models are skewed towards the opinions of workers who agree with the majority vote, disregarding the opinions of other annotators even in the case of low inter-annotator agreement. The solution to this was shown to be using disaggregated data and transforming the problem from the binary classification of toxicity to the prediction of the proportion of annotators who would classify a comment as toxic, a strategy also proposed by Aroyo and Welty (2013) and adopted by Wulczyn *et al.* (2017). In addition, Balayn *et al.* (2018) evaluated the role of spammers among crowd workers and analysed the quality of the workers and their annotations to remove the lowest quality workers from the sample.

Regardless of the issues posed by crowdsourcing tasks, crowdsourcing still remains the cheapest and most effective way to gauge public opinion on a large dataset. Crowdsourcing and its challenges are especially relevant to the task at hand as this paper wishes to examine the effect of the demographics of crowd workers on the classification results. As such, Amazon Mechanical Turk will be used to gather crowdsourced data to supplement and validate the results taken from public corpora.

**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
* 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. – found bert to be best model for toxicity dataset

**Conclusion**

**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.

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.

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.

Buhrmester M., Kwang T., and Gosling S.D. (2011). “Amazon’s mechanical turk a new source of inexpensive, yet high-quality, data?”, in *Perspectives on psychological science*, vol. 6, issue 1, pp. 3–5.

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.*

Hamida C.B., Ge V., and Miranda N. (2019). “Toxic Comment Classification and Unintended Bias”.

Kolhatkar V., Thain N., Sorensen J., Dixon L., and Taboada M., (2020). “Classifying Constructive Comments”. *arXiv preprint arXiv:2004.05476*.

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.

Reichert E., Qiu H., and Bayrooti J. (2020). “Reading Between the Demographic Lines: Resolving Sources of Bias in Toxicity Classifiers”, *arXiv preprint arXiv:2006.16402*.

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.

Tetreault J., Filatova E., and Chodorow M. (2010, June). “Rethinking grammatical error annotation and evaluation with the Amazon Mechanical Turk”, in *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 45-48.

Vaidya A., Mai F., and Ning Y. (2020, May). “Empirical Analysis of Multi-Task Learning for Reducing Identity Bias in Toxic Comment Detection”, in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, pp. 683-693.

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

Zhang G., Bai B., Zhang J., Bai K., Zhu C., and Zhao T. (2020). “Demographics Should Not Be the Reason of Toxicity: Mitigating Discrimination in Text Classifications with Instance Weighting”, *arXiv preprint arXiv:2004.14088*.

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*.