**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.

Gender bias specifically has gained more attention than other forms of bias against identity groups in a variety of Natural Language Processing (NLP) tasks. Zhao *et al.* (2018) investigated gender bias in coreference resolution, applying gender-swapping and name anonymisation to their dataset in order to balance the use of gender-specific words in the dataset with the goal of eliminating under-representation bias. This technique was highlighted by Sun *et al.* (2019) as an effective way of debiasing models and measuring gender bias in predictions, using the False Positive Equality Distance (FPED) and False Negative Equality Distance (FNED) metrics (Dixon *et al.*, 2018) to measure the difference in performance for gender-swapped sentences, although there is a possibility that this method will introduce noise or nonsensical sentences to the dataset. In line with these findings, the dataset to be annotated in this project will be augmented with gender-swapped versions of the original comments to examine how different demographic groups of annotators regard the toxicity of identical comments referencing different genders.

Another common source of bias is word embeddings, which can form associations between identity groups and stereotypical terms based on their prevalence in the literature used to train the word embeddings. Bolukbasi *et al.* (2016) demonstrated the presence of gender bias in occupations in word embeddings and proposed a system to debias word embeddings by isolating the gender subspace before utilising hard or soft debiasing to remove the gender bias from terms identified as being gender neutral. This work was extended by Manzini *et al.* (2019) to encompass racial bias, transforming the binary classification task of identifying gender-specific and gender neutral terms into a multiclass debiasing problem.

**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. As such, the definition of ‘toxicity’ in this paper was taken 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 are 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 was formulated for the Civil Comments Toxicity Kaggle challenge by Borkan *et al.* (2019), and is composed 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 as it is large, widely used in literature concerning bias in toxicity classification (Reichert *et al.*, 2020; Zhang *et al.*, 2020; Pavlopoulos *et al.*, 2020), and was published by the Conversation AI team at Alphabet. Due to the size of the corpus and it’s unique property of annotating the identity groups referenced in comments, a subset of this data will be annotated in this project to understand the relationship between the demographic identities of the annotators and the demographic groups referenced in the comments, as well as validating the toxicity scores in the corpus.

The second such dataset was created by Wulczyn *et al.* (2017), who created 2 datasets, the first of which contains over 160k comments annotated with toxicity scores and has been used in a wide variety of recent literature for investigating bias (Binns *et al.*, 2017; Balayn *et al.*, 2018; Dixon *et al.*, 2018), creating new task-specific corpora (Cecillion *et al.*, 2020) and developing deep learning approaches to toxicity detection (Pavlopoulos *et al.*, 2017; Mishra *et al.*, 2018; D’sa *et al.*, 2019). The second dataset created by Wulczyn *et al.* (2017) contains over 100k comments labelled with personal attack and aggression annotations, almost 78k of which were also in the toxicity dataset, commonly used to examine the subtypes of toxicity (Gröndahl *et al.*, 2018; Magu *et al.*, 2018). 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.

**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**

Classical machine learning models such as Logistic Regression, Bayesian models, Support Vector Machines (SVMs) and Random Forests depend heavily on the features of the corpus in order to make accurate predictions (Koratana and Hu, 2019). Kolhatkar *et al.* (2020) highlights the importance of choosing features carefully by demonstrating that naïve models may overestimate the importance of features such as length, enabling longer toxic comments to go undetected and creating a system of limited practical value. Deep learning models such as CNNs are shown to more robust to such imbalanced features in a dataset, resisting overfitting to any length-based features and proving their adaptability to different datasets and tasks. This is because they perform automatic feature extraction and can often find patterns in the data that are invisible to humans (Koratana and Hu, 2019).

Nobata *et al.* (2016) evaluated the effectiveness of the most widely used features, finding that using all available features improved classifier performance the most, but n-grams were the most predictive features, with character n-grams performing better than token n-grams due to their additional ability to recognise when the characters of offensive words were being replaced with characters that looked similar in order to avoid detection by toxicity classifiers. They also surveyed linguistic features, which identify useful characteristics in analysing the toxicity of a comment, such as the average length of words in a comment, the number of punctuation marks, capitalised letters, one letter tokens, and URLs in a comment, as well as more informative features such as checking the words in a comment against word lists to count the number of polite words, insulting words and words that can’t be identified in an English dictionary. Kolhatkar *et al.* (2020) added to these linguistic features by counting the number of named entities in the comments and including TF-IDF weighted phrases, text quality features such as the number of spelling mistakes, and content quality features such as coherence scores and spam probabilities.

In addition to the above linguistic features, Nobata *et al.* (2016) described useful syntactic features by identifying part-of-speech (POS) tags and linking comments to their parent, grandparent, and child comments, aiming to capture long range dependencies that might otherwise evade detection. Similar techniques were examined by Schmidt and Wiegand (2017), who also examined typed dependency relationships to create offensiveness level scores, in addition to detecting imperatives and pre-defined polite words in order to assess the tone of each comment. They also investigated a similar branch of work, sentiment analysis, to assess the practicality of estimating the polarity of a comment by counting the number of positive, neutral, and negative words as an additional step in a toxicity classifier.

Other studies have used the meta-information available to gain additional clues about the toxicity of a comment, including user-based information such as the number and average length of comments by the author, as well as their prior history of hateful or polite comments (Kolhatkar *et al.,* 2020). Schmidt and Wiegand (2017) also noted that knowing the gender of the user may help as men are more likely to post hate speech than women. In addition to this, they suggested the use of multimodal information such as images and videos, focusing on the tags and pixel-level features associated with them, as predictive features to supplement the classification of any associated text.

Some studies such as Koratana and Hu (2019) and Schmidt and Wiegand (2017) have also discussed the use of knowledge-based features, which rely on up-to-date knowledge of stereotypes as well as the context in which the comment was written to draw conclusions about the toxicity of the comment. However, knowledge-based features are rarely used due to the difficulty of manually coding and continuously updating the knowledge base.

As machine learning methods have gained popularity over feature-based methods in recent years, word embeddings have become commonplace in many models, the most popular of which are the word2vec, GloVe and FastText embeddings, employed by many of the models in Kumar *et al.* (2018). Word embeddings assign a vector representation to each word, giving similar vectors to semantically similar words, and providing more contextual information than the standard one-hot encoding vectors that note the words present in a comment but not their relationships to other known terms (Schmidt and Wiegand, 2017). In order to utilise the word embeddings for the entire comment, the word vectors for a comment are often averaged, although Nobata *et al.* (2016) notes this limits the effectiveness of the word embeddings due to the loss of word order sensitivity and semantics. An alternative approach put forward by Le and Mikolov (2014) proposed the creation of a paragraph vector rather than individual word vectors to ensure that the sematic information in each comment is retained and this is shown to outperform the basic bag-of-words approach.

**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.

Bolukbasi T., Chang K.W., Zou J.Y., Saligrama V., and Kalai A.T. (2016). “Man is to computer programmer as woman is to homemaker? debiasing word embeddings”, in *Advances in neural information processing systems*, pp. 4349-4357.

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.

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*.

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*.

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.*

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*, pp. 2–12.

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*.

Koratana A. and Hu K. (2019). “Toxic Speech Detection”.

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

Le Q. and Mikolov T. (2014, January). “Distributed representations of sentences and documents”, in *International conference on machine learning*, pp. 1188-1196.

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

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

Manzini T., Lim Y.C., Tsvetkov Y., and Black A.W. (2019). “Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings”, *arXiv preprint arXiv:1904.04047*.

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

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., Malakasiotis P., and Androutsopoulos I. (2017). “Deep learning for user comment moderation”, *arXiv preprint arXiv:1705.09993*.

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. and 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.

Sun T., Gaut A., Tang S., Huang Y., ElSherief M., Zhao J., Mirza D., Belding E., Chang K.W., and Wang W.Y. (2019). “Mitigating gender bias in natural language processing: Literature review”, *arXiv preprint arXiv:1906.08976*.

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.

Zhao J., Wang T., Yatskar M., Ordonez V., and Chang K.W. (2018). “Gender bias in coreference resolution: Evaluation and debiasing methods”, *arXiv preprint arXiv:1804.06876*.

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*.