Handling Bias in Toxic Speech Detection: A Survey

TANMAY GARG, SARAH MASUD, THARUN SURESH, and
TANMOY CHAKRABORTY, IIT Delhi, India

Detecting online toxicity has always been a challenge due to its inherent subjectivity. Factors such as the context, geography, socio-political climate, and background of the producers and consumers of the posts play a crucial role in determining if the content can be flagged as toxic. Adoption of automated toxicity detection models in production can thus lead to a sidelining of the various groups they aim to help in the first place. It has piqued researchers' interest in examining unintended biases and their mitigation. Due to the nascent and multi-faceted nature of the work, complete literature is chaotic in its terminologies, techniques, and findings. In this article, we put together a systematic study of the limitations and challenges of existing methods for mitigating bias in toxicity detection.

We look closely at proposed methods for evaluating and mitigating bias in toxic speech detection. To examine the limitations of existing methods, we also conduct a case study to introduce the concept of *bias shift* due to knowledge-based bias mitigation. The survey concludes with an overview of the critical challenges, research gaps, and future directions. While reducing toxicity on online platforms continues to be an active area of research, a systematic study of various biases and their mitigation strategies will help the research community produce robust and fair models.[1]

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Information systems** → *Social networks*; • **Social and professional topics** → **User characteristics**;

Additional Key Words and Phrases: Toxic speech, hate speech, social networks, unintended bias, bias mitigation, bias shift

---

[1]**Disclaimer:** This article includes examples of toxic speech that contain some profane words. These examples are only included for contextual understanding. We tried our best to censor vulgar, offensive, or hateful words. We assert that we do not support these views in any way.

---

1 INTRODUCTION

Online social networks (OSNs) have enabled a thriving ecosystem where people from diverse backgrounds can share opinions and ideas. Such online forums are safe spaces for many marginalized groups to support each other and gain a sense of community in the face of discrimination. However, anti-social users often use OSNs to harass and intimidate others. Such behavior manifests in content aimed at harming individuals or groups based on personal attributes such as race, gender, and ethnicity. At its extreme, hate speech (a type of toxic speech) can lead to incidents of offline violence [28, 59]. It should be noted that such incidents of violence and hate crime do not occur in isolation. Often negative stereotyping and hostility in the real world get heightened during online interactions. Meanwhile, online hate speech and doxing [102] promote offline brutality and even genocides. The online and offline propagation of toxic behavior is a vicious cycle. Such behavior often hampers the ability of marginalized groups to share their opinions freely and further isolates them [83, 106]. Therefore, the real-world impact of harmful online content and its multifaceted nature have galvanized academic and industrial research toward the early detection and mitigation of such content.

Toxic speech. Throughout this survey, the term toxic speech will be used as an umbrella term to refer to any form of content, including but not limited to hate speech, cyberbully, abusive speech, misogyny, sexism, offense, and obscenity. We follow the definition of toxic speech as given by Dixon et al. [24]: "rude, disrespectful or unreasonable language that is likely to make someone leave a discussion" as an umbrella definition. The literature surveyed here is distributed among various forms of toxic speech, known to adopt different and sometimes misleading terms to refer to equivalent classes [33]. The subjectivity in defining toxicity can also be attributed to the lack of an in-depth understanding of toxicity through the lens of social and psychological science. Machine learning researchers and practitioners often overlook the root cause of toxic behavior on the Internet [95]. Such root cause analysis can help understand the power dynamics of hate [66], the prevalent discrimination, and its evolution on the Internet [30, 58]. Recent studies in implicit hate speech [27, 42] attempt to bridge this gap; however, such practice in computational methods for toxicity detection is not yet well adopted.

Vulnerable groups. In relation to toxic speech, there is always a group (or an individual from the group) at the receiving end of all the negativity and hate. Such groups can be termed as vulnerable/target/protected groups. For bias evaluation, they may be even referred to as subgroups or identity groups. Each of these groups can be identified with negative phrases that are used to refer to the individuals of the particular group. Such phrases are called "identity terms." For example, when analyzing hate speech against Jewish people, "k*k*" is a very commonly used slur.

Unique characteristics of online toxic speech. Toxicity detection is a highly subjective task, and context plays a crucial role in determining whether the content can be flagged as toxic. Different online platforms and legal agencies use varying definitions of toxicity. Meanwhile, existing datasets display a variety of biases via their collection pipeline [24, 113] and are subject to unreliable annotations [81]. This article explores why these characteristics render toxic speech vulnerable to various unintended biases and the methods proposed to handle them.

Scope of the survey. Bias mitigation methods in **natural language processing (NLP)** have been extensively studied [7, 36, 110]. Consequently, the mitigation techniques in NLP have been helpful in tasks such as textual entailment [100] or reading comprehension [35]. It also motivated researchers to apply bias mitigation for toxic speech detection. While the respective methods have progressed, the results are not as effective as expected. The argument goes back to the subjective/objective nature of tasks like toxicity detection vs. textual entailment. It inspired us to conduct a survey that can help better understand the current state-of-the-art bias mitigation in

Table 1. List of Popular Toxic Speech Datasets in the Study of Bias as Discussed in This Article

Dataset	Sampling	Lexical		Annotation	Racial		Gender	
	E	E	M	E	E	M	E	M
Founta [34]	[69, 113]	[113]	[120]		[19, 20, 44]	[85, 116, 120]	[44]	[70]
W&H [109]	[4, 69, 113]	[113]		[64, 108]	[20, 64]			[70, 113]
Wulczyn [115]	[78, 113]	[113]	[5, 24]	[2, 111]				
Davidson [21]			[5]	[64]	[20, 44, 64, 85]	[116]	[44]	
Waseem [108]				[64, 108]	[20, 64]			[70]
GHC [48]			[49]					
Stormfront [22]			[49]					
Evalita2018 [29]								[68]

"E" (Evaluates) indicates when a dataset was used to establish the presence of bias. "M" (Mitigates and Evaluates) indicates if debiasing techniques were performed on the dataset. Here, sampling, lexicon, and annotation are sources of bias, whereas race and gender are targets of bias. Note that this table is a representative sample of the literature surveyed here and is not a comprehensive set.

toxicity detection. The scope of our study is not to discuss the comprehensive research of biases in NLP; instead, we present a thorough analysis of the methods that study *bias as applied to the case of automatic toxicity detection.* Additionally, several surveys have already examined the existing literature in modeling the toxicity detection task [32, 89]. We do not survey all the existing toxicity detection methods; instead, we focus on a subset of them, exploring and mitigating bias in toxic speech detection. Meanwhile, Yin and Zubiaga [117] surveyed the literature addressing the robustness of hate speech detection methods and addressed the subject of bias in hate speech detection. While their discussion remained general commentary, we aim to develop an extensive understanding of these methods.

To begin with, we develop a taxonomy of bias based on the sources and targets of harm. Each bias mitigation method can be applied to one or more data transformation stages of a **machine learning (ML)** pipeline [98]. Subsequently, Table 1 provides a summary of popular hate speech datasets and their usage for the study of evaluating and mitigating bias in toxic speech detection. As a part of our literature survey, we also review the reproducibility of existing debiasing methods employed for toxicity detection. The details of the experiments are provided in Appendix C. Meanwhile, we also discover a compelling phenomenon of bias shift in knowledge generalization-based methods and provide a short description of the same in Appendix B.

Survey methodology. Following Yin and Zubiaga [117], we considered Google Scholar as the primary search engine to curate relevant papers. We focused on toxicity debiasing studies done within the past five years and mainly looked at research published in 2016 onwards. We started with relevant keywords such as "bias," "toxic speech," "abusive speech," and "hate speech," shortlisting a seed set of papers through their abstracts. Papers were also collected from recent proceedings of relevant data mining, NLP and web-science conferences (ACL, EMNLP, NAACL, AAAI, WebSci, ICWSM, etc.), journals (TACL, TKDD, PLOS, etc.), and workshops (WOAH etc.). We also visited the citing and cited papers of the seed papers to locate relevant papers further. This shortlisting process was done between September–November 2021.

2 CATEGORIES OF BIAS

All machine learning models, including toxicity detection, assume some bias in the data to perform predictions [24]. However, we do not intend the toxicity prediction to vary based on the speaker's racial background, for example. If a model exhibits such bias, then we call it *unintended bias.* In the rest of the article, we use "bias" to refer to unintended bias in toxic speech detection.
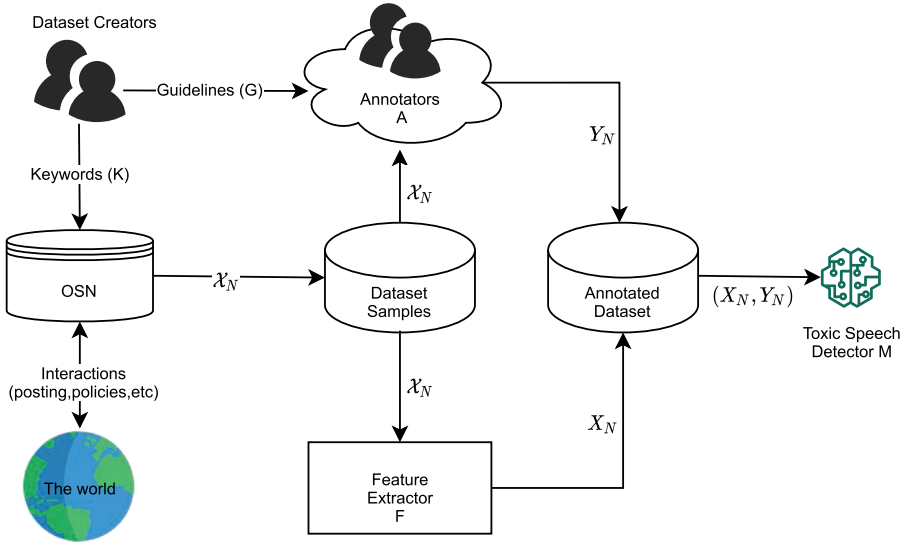
Fig. 1. The pipeline of a toxic speech detection model can be visualized as a sequence of data transformations: (i) the OSN sampler API ($S : K \rightarrow \mathcal{X}_N$) takes as input a set of keywords K (possibly empty) and returns a set of samples $\mathcal{X}_N$; (ii) the annotation function ($A : \mathcal{X}_N, G \rightarrow Y_N$) converts $\mathcal{X}_N$ and the annotation guidelines G to a sequence of annotations Y_N; (iii) the feature extraction function ($F : \mathcal{X}_N \rightarrow X_N$) converts $\mathcal{X}_N$ to a sequence of feature vectors X_N. (iv) Finally, the function ($M : \mathcal{X}_N, Y_N \rightarrow Y'_N$) aims to predict the toxicity label. Biases based on the source of harm are associated with these transformations as (i) sampling bias with S, (ii) lexical bias with (F, A), and (iii) annotation bias with A. Meanwhile, the biases based on the target of harm manifest when the prediction Y'_N of the model M are analyzed.

Background on NLP Bias. In cognitive science, bias is assumed to be a shortcut (valid/invalid)—our brain resorts to when informing our actions and interactions with others. Biases can develop due to a limited worldview and repeated exposure to similar behavior in the surrounding [43]. Stereotypes get further amplified due to the formation of filter bubbles [6] and echo chambers. When such biased interactions are employed to train natural language models, they learn stereotypical statistical associations prevalent on the Internet [15]. The seminal work by Bolukbasi et al. [10] led the initiative of establishing gender (male vs. female) bias in non-contextual embeddings trained on large-scale web corpora. This work prompted similar studies [7, 15] of bias in NLP towards other social/personal attributes.

Blodgett et al. [7] broadly classified the NLP biases as either allocational or representational harm. Sun et al. [96] further granulated representation harm into denigration, stereotyping, recognition, and under-representation when evaluating various NLP tasks and their associated language models for mitigating gender bias. In a tangential approach, Shah et al. [91] and Kumar et al. [52] looked at biases in NLP through the lens of data and modeling pipelines. Shah et al. [91] proposed a pipeline to highlight the origin of various biases and their eventual outcome disparity. Kumar et al. [52] largely classified the biases as related to either data source, label/annotation, and embedding representation.

Relation to NLP Bias. A simple bias mapping pipeline for training a generic NLP model is depicted in Figure 2. Comparing the training pipelines in Figures 1 and 2, we can observe how the first step (data collection and sample) in both workflows is primarily the source of sampling bias. In the case of toxicity detection, biased sampling is performed to gain a higher percentage of toxic samples. Consequently, in broader NLP modeling, it occurs due to a skewed ratio in
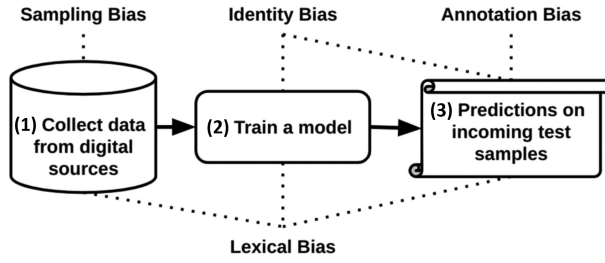
Fig. 2. The pipeline of training and evaluation of NLP models can be visualized as a sequence of (1) data collection, (2) training of models, and (3) employing the models on incoming test samples. Here, collecting data from digital sources can introduce sampling and lexical biases (if specific web pages/digital footprints are ignored). Model training can also introduce linguistic bias when frequently picking co-occurring phrases. Training processes and skewed datasets can introduce identity-based prejudice (race, gender, age, political affiliation, etc.). If predictions on downstream NLP tasks happen on human-labeled gold datasets, then it can be a source of annotation and sampling biases (not depicted here). Both linguistic and identity-based biases manifest during the model evaluation and in-the-wild testing.

terms of quality and quantity of digital footprints of specific topics (such as content in support of LGTBQ vs. against them). Sometimes, the training of NLP models can be unsupervised; in such cases, the annotation bias in the pipeline shifts to the last stage when it is employed on downstream tasks that use gold-label datasets to establish performance. All the shortcomings of human annotation apply to various NLP tasks. Some are more static and objective, leading to less bias and disagreements. Meanwhile, others can be highly subjective and ephemeral. As a result of various sampling and annotation biases, we observe the prevalence of identity-based harm introduced due to spurious lexical correlations. We term the bias/prejudice against a person's demographic or psychographic characteristic or any identifiable physical or mental characteristic as *identity bias*. As such, prejudice against specific identity groups can be analyzed separately based on the target of harm. Inspired by various approaches, we develop the following taxonomy based on the (i) sources and (ii) targets of harm. Biases studied in this survey are not unique to the task of toxicity detection; instead, they are an extension of the biases in NLP applications.

Based on *sources* of harm: We take inspiration from Suresh and Guttag [98] to group the surveyed methods into categories based on the source of downstream harms during the data collection process. The authors defined the process, consisting of: selecting a *population*, selecting and measuring *features*, and *labels* to use. We study categories of bias according to the transformations related to these steps (described in Figure 1) as: sampling, lexical, and annotation bias.

Based on *targets* of harm: The next three categories of bias in toxic speech are each dedicated to a target group of downstream harm (Figure 3): (i) racial bias, (ii) gender bias, and (iii) psychographic bias like political affiliations [45]. However, the study of biases based on psychographic attributes (grouping individuals w.r.t. their beliefs and interests) is yet to gain popularity. Through this survey, we hope to encourage future exploration of bias categories tied with psychographic attributes.

Other categorizations of bias. Some of the papers surveyed here present implicit vs. explicit bias categorization. We refer the readers to Reference [56] for an understanding of this bias stratification. Blodgett et al. [7] use a taxonomy that categorizes between *allocational* and *representational* harms. While this taxonomy is useful to segment existing papers based on their motivation of handling bias, the taxonomy we developed lets us approach the proposed methodologies from an application point of view.
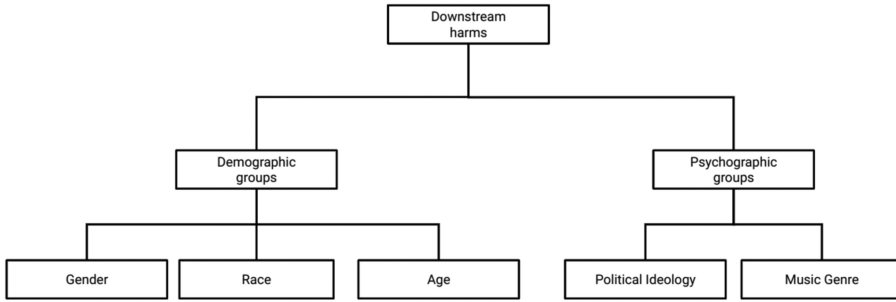
Fig. 3. A taxonomy of bias based on the downstream harm. The harms can be inflicted through the real-world deployment of a toxic speech detection system.

3 BIASES BASED ON SOURCE OF HARM

Overview. This section will cover three main categories of biases that stem from the data processing step. We begin our discussion with sampling and annotation bias and then cover lexical bias as the consequence of sampling and annotation. Sampling bias occurs at the first step of data collection. The workaround employed to obtain a more significant percentage of toxic comments against vulnerable communities eventually manifests as bias against the very community it intends to protect.

Such a dataset may require fine-grained labels obtained at large scale by human-aided annotations. However, performing annotations for subjective tasks such as labeling toxicity is tricky, because there is no standard definition of what can be considered toxic. Here, the cognizance of the annotators about the existing discrimination and current socio-political status quo is essential in obtaining valuable annotations. Especially when looking into toxicity against a particularly vulnerable group such as sexual orientation, it is necessary to consider the annotators' political, religious, and social views before annotations. It should be noted that such social beliefs need not be independently formed. Religion, for example, is not a personal belief system but can be culturally reinforced, sometimes acting as a proxy for race and ethnicity.

Building further on the context of sexual orientation, if the annotators are provided with metadata and background of the speakers participating in the comments, then it can be a double-edged sword. On the one hand, it can provide the context to correctly label such comments as non-toxic despite containing seemingly offensive terms. On the other hand, it can trigger latent biases in the annotators. They can willfully provide erroneous labels and label the interaction as hateful and harmful, leading to suppression of the voices of the minority.

Once the data is collected and annotated, it is ready for pre-processing and modeling. Once modeled, we can retrospectively analyze the models for downstream harms and biases it incorporates against different identity groups such as race, gender, or both. One such bias is lexical. Lexical bias can be evaluated at both dataset (source of harm) and modeling level (downstream harm). For this survey, we include it with the source of harm, as lexical biases may or may not be identity-specific and can be generally spurred from the presence/absence of words in both toxic and non-toxic labels. It should be noted that all biases we discuss in the coming sections, i.e., sampling, annotation, and lexical biases, directly/indirectly reflect the varying linguistic and social attributes and subjective nature of toxic content. An abbreviation of various terms used in our discussion of biases is enlisted in Table 2. A complete description of the various bias metrics and their usage is provided in Appendix A.

Table 2. Abbreviations and Expansions Used throughout the Article

Abbreviation	Expansion
SBET	Synthetic Bias Evaluation set
BSW	Bias Sensitive Words
SOC	Sampling and Occlusion
L_{CE}	Cross-entropy Loss
AAE	African-American dialectal English
NHB (NHW)	Non-Hispanic Black (White)
FPR (FNR)	False Positive (Negative) Rate
FPED (FNED)	False Positive (Negative) Equality Difference
pAUCED	pinned AUC Equality Difference
subAUC	Subgroup AUC
BPSN (BNSP)	Background Positive (Negative) Subgroup Negative (Positive) AUC
GMBAUC	Generalized Mean of the Bias AUCs
pB	pinned Bias
FPR (TPR)	False (True) Positive Rates
NPV (PPV)	Negative (Positive) Predictive Value
WEAT	Word Embedding Association Test
SEAT	Sentence Encoding Association Test

3.1 Sampling Bias

The first step towards toxicity detection is to curate data samples from across the web. Among social media content, harmful content is not a majority class. On Twitter, they cover no more than 3% of the tweets [34]. A random sampling of data thus requires going through a considerable amount of non-toxic content to get a substantial amount of toxic samples. Hence, datasets are curated by adopting heuristics that achieve a higher density of toxicity in the samples. These heuristics can introduce spurious correlations of linguistic features with toxicity labels if not tracked, therefore termed as *sampling bias*. In this section, we will give an overview of the two broad methods of sampling data from web and how to measure and mitigate the sampling bias.

Sampling techniques and effect of text source for toxicity datasets. In *boosted random sampling*, heuristics are applied to increase the density of harmful content in an initial random sample. This sample is usually labeled for toxicity, and the keywords or user names in these samples are used to extract more samples. Meanwhile, *topic-biased sampling* [84] starts by acquiring samples based on a predetermined set of keywords (hate lexicons) and socio-political topics that are commonly known to be toxic. Dataset creators have popularly used topic-biased sampling to quickly acquire samples with a higher toxicity concentration. Wiegand et al. [113] studied six datasets (three boosted and three topic-wise sampled) and used an abuse lexicon to mark examples as explicitly or implicitly abusive. They found the boosted randomly sampled datasets to contain a lesser concentration of overall toxicity; out of them, a higher concentration was explicit abuse. Moreover, a higher explicit abuse ratio was seen to aid in downstream performance, as *verbal cues for explicitly toxic labels are easier to model.* Their work was further investigated by Razo and Kübler [78] by conducting sampling experiments on the same underlying dataset. While broadly consistent with previous studies, they discovered that, all things being equal, the text source has a more significant impact on the quality of the toxic dataset curated than the underlying sampling strategy.

Metrics to measure sampling bias. To capture the topical and lexical biases in the curated datasets, Ousidhoum et al. [69] developed two bias metrics. These measures make use of predefined hate lexicons $w' \in W'$. While B_1 captures the average similarity between all topical words and

the hate lexicons, B_2 captures how likely each topic contains a keyword. The robustness in B_1 scores on increasing topic count indicates the new topic words to be similar to the keywords, pointing towards a topic bias. For instance, the Waseem & Hovy dataset (referred to as the W&H dataset henceforth) [109] showed robustness to topics, which can be attributed to its high topic and user biases [4, 113]. An Arabic dataset [3], which was curated based on religious keywords, gave comparable B_1 values for all three target attributes, suggesting its coverage of origin and gender-based topics as well. The authors further showed that these biases and variations in sampling techniques impact cross dataset generalization. In short, one can say that models trained on a dataset sampled for a set of topics do not generalize.

Mitigation via proxy topic-sampling. Van Rosendaal et al. [105] created a dataset by automating the pre-selection of keywords to reduce sampling bias. They used Reddit to find keywords from posts with more or less equal amounts of upvotes and downvotes. Such controversial topics can be used as a proxy to collect toxic speech from OSNs [40]. A small sample from the tweets filtered based on these keywords was verified as more toxic than an equal-sized sample of unfiltered tweets. However, this difference was minimal, suggesting the inability of the proposed method to create a dataset with the desired density of toxic comments. Moreover, Van Rosendaal et al. [105] did not validate their initial claim of a lower bias in the filtered dataset.

Mitigation via data mixing. Topic/author bias degrades downstream performance—the W&H dataset [109] is a *case in point.* Wiegand et al. [113] noticed that removing top PMI terms from the W&H dataset for the toxic labels led to a considerable drop in classification performance, suggesting a high topic bias. The dataset also contained high author bias, with more than 70% sexist tweets from only two authors and about 99% racist tweets from a single author. This high author bias can be attributed to the improvements in performance when encoding the user meta-information as features [63, 75]. It begs whether the classification model trained on the W&H dataset learns to discriminate the hate styles or the hateful authors' styles. Arango et al. [4] tried to correct the author bias in this dataset by limiting the number of toxic comments to 250 per user while augmenting more toxic comments from the Davidson dataset [21]. They found a substantial increase in generalization on the unseen Hateval dataset [46]. *However, we note that they did not account for the topical homogeneity between the training and test datasets that could have led to an increase in performance.*

3.2 Annotation Bias

As toxicity is subjective, it is challenging to train machine learning models to learn what can be potentially toxic. To obtain these labels, we require large-scale human-aided annotations. Fortuna et al. [33] showed that toxic speech datasets use ambiguous and misleading labels to refer to similar underlying categories introducing subjectivity into the annotation process. The subjectivity and discrepancy in the labeling process can lead to *annotation bias*. In this section, we will broadly discuss how the quality of data labels is impacted by the annotation guidelines and the annotator's prejudice. It should be noted that there is no direct way of mitigating annotating bias via modeling alone. Obtaining unbiased annotations requires sensitization of the annotators and the society at large. We hope this section highlights the ambiguous and challenging nature of generating standard labels for toxicity.

The effect of guidelines on the quality of annotations. Annotation pipelines need to consider the subjective nature of hate speech [82]. Ross et al. [81] explored the effect on the reliability of annotations before and after providing the guidelines. Both annotations (guided vs. non-guided) showed a high Pearson's correlation coefficient, indicating that they captured the same underlying construct. Similarly, the Jigsaw[2] team adopted a labeling schema that is subjective to individual

---

[2]https://jigsaw.google.com/the-current/toxicity/.

interpretation [24]. The team found that more annotators could agree on the toxicity of comments based on a more generic guideline. On the contrary, Fortuna et al. [31, 33] suggested using hierarchical multi-class annotation schemes.

The effect of annotators' implicit biases on quality of annotations. Waseem [108] studied the influence of an annotator's expertise on toxicity labeling. They formed two groups of annotators—experts and amateurs—and found a low agreement within these groups across all labels on the W&H dataset. They also found that the amateurs were more likely to assign a toxic label to the sample, leading to a loss in downstream performance. The ratings by the amateurs were also closer to the W&H dataset's crowdsourced annotations. *We need to investigate the impact of the expertise and sensitivity of annotators on the final labels obtained.*

Al Kuwatly et al. [2] investigated the effect of the annotator's demographics on the quality of resulting classifiers. They used the annotators' demographic information (*gender*, *age*, *education*, and *native English speaker*) present in the Wulczyn personal attacks dataset [115]. They found no differences between the two genders. Whereas, models trained on native English speakers' annotations outperformed the ones for non-native speakers in terms of $F1$ score and sensitivity. Similar differences were observed across both groups for age and education. *While this investigation relied on the per-annotator demographics collected during the annotation process, this information can be tricky to collect and limited to the demographic strata considered during the data creation phase.*

Meanwhile, Wich et al. [111] adopted an unsupervised method to group the behavior of annotators. Annotators and the inter-annotator agreements were converted into a graph with the annotators as nodes and their agreement scores as edge weights. Next, a community detection algorithm [9] was used to group similar annotators. They trained toxicity detection systems on annotations by groups with low inter-rater agreements, which resulted in low-performance on the test datasets labeled by other groups. Similar results were seen for a group with higher inter-rater agreement but low agreement with the other groups, indicating a bias in this group's combined annotation. However, the study did not mention obtaining multiple observations for a group, opening the question of variance in performances due to random model initialization. *An exciting extension to this study would be analyzing the type of biases shown by such anomalous groups and if their ratings should be penalized or contrarily given higher weights to improve the inclusion of the demographic represented by them.*

The need to incorporate disagreement in toxicity modeling. In an intriguing study to account for annotation disagreements, Gordon et al. [39] developed a disagreement deconvolution setup to estimate the lower bounds on model performance accounting for variation in human annotations. They compared the drop in performance against the oracle system for respective models trained on datasets covering subjective tasks like toxicity vs. objective tasks like image detection. Researchers observed that when disagreements are not accounted for (i.e., taking a single label per sample), the performance drop is more for subjective tasks than objective ones. Consider a crowdsourced toxicity sample that receives 5 annotations (3 non-hate, 2 hate). When we go by majority voting, the model will consider predicting the label vector as [1,0]. Meanwhile, the probability distribution it should learn is [0.6,0.4]. Framing subjective tasks via majority labeling fails to capture the latent distribution of variable human understanding of toxicity. *The study crucially points out the importance of accounting for disagreements to project the realistic performance of models when deployed in real-world toxicity detection.*

3.3 Lexical Bias

The keyword-based heuristics in data sampling, coupled with a lack of context and human predilection, can cause toxic speech classifiers to learn spurious lexical correlations. The models produce high toxicity ratings for non-toxic texts due to a disproportionately high presence of certain terms

(or phrases) in the content labeled as toxic. This phenomenon is called *lexical bias*. We refer to terms contributing to lexical bias as ***bias sensitive words*** **(BSW)** [5]. This section highlights mechanism to mitigate lexical bias that can be mainly achieved via (a) correcting data labels and (b) debiasing at the modeling level.

Deciding on the target BSW. While Dixon et al. [24] and Kennedy et al. [49] relied on the manual creation of a list of identity terms, Vaidya et al. [103] used the identity attributes annotated in the Civil comments dataset[3] Borkan et al. [12]. Mozafari et al. [64] used **local mutual information (LMI)** to find the terms highly correlated with the toxic labels. Meanwhile, Badjatiya et al. [5] developed unsupervised methods to find the set of BSW.

Mitigation through data correction and filtering. Dixon et al. [24] and Badjatiya et al. [5] proposed new methods for mitigation of lexical bias. However, Zhou et al. [120] investigated methods that have been effective on other NLU debiasing tasks [54, 99].

- **Length-sensitive upsampling:** Dixon et al. [24] found that toxic samples for an identity term are non-uniformly distributed across different lengths. Subsequently, the non-toxic label was upsampled using statements from Wikipedia articles across the length buckets. The authors also observed an improved **pAUCED (pinned AUC)** value on a **synthetic bias evaluation set (SBET)** containing samples equally distributed across the toxicity labels for each identity term. Moreover, the **FPR (False Positive Rate)** values across the identity terms went down without an increase in the variance of the **FNR (False Negative Rate)** values, indicating a reduction in false-positive bias while not affecting the false-negative bias. *Note that these observations were made on SBET, which employs a static target-attribute pair list. Since this list had a different distribution than the initially debiased data, calculations of bias based on SBET may not capture the changing dynamics/semantics when the samples are debiased.* We note the same in our case study of bias-shift (Appendix B).

- **Knowledge-based generalizations:** Badjatiya et al. [5] developed mitigation techniques based on knowledge-based generalizations. They propose lexical database generalization; replacing the BSW occurrence with an ancestor in the wordnet [62] hypernym-tree (e.g., *black* with *color*). Similar to Dixon et al. [24], Badjatiya et al. [5] did not account for the newly acquired biases in the dataset post-mitigation. We present a case study related to this observation in Appendix B.

- **Data filtering:** Zhou et al. [120] explored two automated data filtering approaches (AFLite [54] and DataMaps [99]) to obtain the set of training samples that will lead to better generalizability and reduce bias as a by-product. The correlation of the toxicity label in the Founta dataset [34] with BSWs reduced with both AFLite and DataMaps, corroborating a reduction in the lexical bias in the resulting filtered datasets.

Mitigation through the debiased training of downstream models. Kennedy et al. [49], Vaidya et al. [103], and Zhou et al. [120] studied the mitigation of lexical bias through model-level debiasing.

- **Regularizing importance scores:** Kennedy et al. [49] reduced the weightage a model assigns to identity terms by regularizing cross-entropy loss (L_{CE}):

$$L = L_{CE} + \alpha \sum_{w \in x \cap S} [\phi(w)]^2, \tag{1}$$

---

[3]During dataset creation, apart from the toxicity labels, samples were also annotated with the presence of lexical markers of 24 identity attributes.

where x is the set of input sequence words, S is the set of identity terms, $\phi(w)$ is the attention score for term w, and α is the strength of regularization. The BERT models trained on the GHC [48]and Stormfront [22] datasets with regularized loss showed a reduction in FPR on the test datasets. Using Sampling and Occlusion [47] as an explanation algorithm, the authors also validated a reduction of the importance score given by the regularized model to the identity terms, suggesting a decrease in lexical bias.

- **Multi-task learning:** Vaidya et al. [103] used **multi-task learning (MTL)** to learn both the toxicity and identity attribute labels for the Civil comments dataset. This loss L accounts for the cross-entropy loss L_{CE_k} for each identity $k \in [0 - 9]$ in addition to the loss for the toxicity task L_{CE}:

$$L = \sum_{n=1}^{N} \beta_n [\alpha L_{CE}(\hat{y}_n, y_n) + (1 - \alpha) \sum_{k=1}^{9} L_{CE_k}(\hat{y}_n^k, y_n^k)], \qquad (2)$$

where β_n is a sample weight, given a higher value for a non-toxic example n with at least one identity attribute present. Evaluation of the MTL model on the test split showed an improvement in terms of: (i) AUC, suggesting a better overall model performance, (ii) **Generalized Mean of Bias AUC (GMB AUC)**, suggesting better bias reduction, and (iii) **Background Positive, Subgroup Negative (BPSN AUC)** and **Subgroup AUC (subAUC)** for each identity term, suggesting a reduction in the false-positive bias while not increasing the false-negative bias. *However, this mitigation method relied on the annotation of identity attributes, which is not easy to extract and scale.*

- **Ensemble-based debiasing:** Zhou et al. [120] explored the ensemble-based debiased training method on the Founta dataset. They trained an ensemble of two models: an SVM classifier and a RoBERTa-based classifier. The idea was to let the naive model learn to predict the toxic label based on the biased features, encouraging the robust model to rely on the other unrelated features. Finally, once the ensemble was trained, the naive model was discarded. The final model showed lower *FPR* values on the BSWs containing samples from the Founta dataset, indicating better debiasing. However, it came at the cost of loss in accuracy and *F*1 score.

4 BIASES BASED ON TARGET OF HARM

Overview. The biases introduced during data collection and annotation directly/indirectly impact the ability of toxicity models to predict toxicity against identity groups such as race, gender, age, and so on. Such unintended biasing of models against identity attributes manifests due to specific lexical cues present in the data that are erroneously correlated with class labels. As depicted in Figure 4, these can be linked to either the presence of *non-toxic terms* in high proportion in *toxicity labels* or to the presence of *abusive terms* in high proportion in *non-toxic labels*. For example, the co-occurrence of terms such as "Muslims" and "Islam" in comments marked as Islamophobic leads the model to correlate Muslims with toxicity spuriously. Meanwhile, the presence of phrases like "f*** the world, be you," even though intended to be non-toxic, ends up being annotated or predicted as toxic due to the previously learned relation of slur terms with toxicity.

Meanwhile, the focus on abusive terms or reclaimed slurs falsely causes the models to classify non-toxic content as toxic. The primary goal of mitigating biases against target groups thus aims to reduce the attention assigned to the presence/absence of specific terms and shift the focus to the overall context in the content under consideration. One way of achieving this is by providing external context as meta-data or incorporation of present world knowledge databases. In current literature, toxicity detection and debiasing employ large language models that are well known to
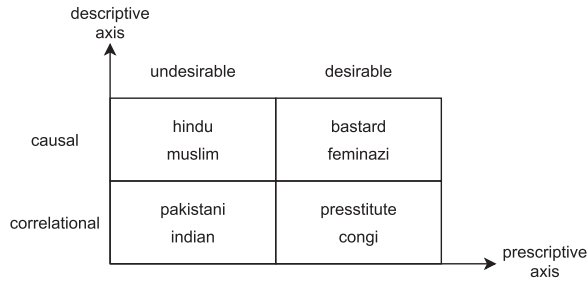
Fig. 4. Ghosh et al. [37] proposed two axes, (i) descriptive and (ii) prescriptive, to segment the associations of the *over-represented* terms for the model under bias evaluation. Here, we consider the cross-section of transferring models trained on English toxicity datasets to toxic phrases in code-mixed Hinglish. The terms in the undesirable quadrants (left column) are those that, despite their frequent presence in toxic classes, are non-toxic/neutral and should not contribute towards toxic label prediction. Meanwhile, the terms in the desirable quadrants (right column) are offensive terms well known in English or colloquial demographic (Indian). For example, while we want the model to correctly predict abusive terms such as bastard or feminazi, or commie to capture toxic intentions, we also want the model to extend to novel lexical terms known to be offensive in the Indian context. For example, "congi" is an offensive connotation to Indian National Congress, a left-leaning political group in India.

be biased in their understanding of identity groups. *Thus, any debiasing of toxicity models is lower bounded by the debiasing of the underlying language frameworks and embeddings employed in the detection models.*

4.1 Racial Bias

Racial bias in machine learning. In their seminal work, Buolamwini and Gebru [14] pointed out that facial recognition systems are biased against dark-skinned faces. Similar issues were observed against Black dialects in automatic speech recognition [101] and the negative association of Black names in popular word embeddings [16]. The issue of biases in word embedding is a much-researched topic in NLP, and its impact extends to downstream tasks of toxicity detection as well. Researchers observed that content posted by Black users or content employing non-white dialects is more likely to be mislabelled as toxic [20]. Thus, already marginalized groups are at a greater risk of being flagged and discriminated against by toxicity detection models. While some of the broader literature around dialect detection is important to applications such as chatbots and auto-mated speech systems, recognizing the dialect based on written content is tricky and contended. The initial work in dialect detection in text revolved around pointing out shortcomings of parsing systems that fail the grammar of diverse dialects; they have been recently adopted to study the impact of dialect in the classification of sentiment and toxicity.

Unfortunately, most literature addressing racial bias primarily focuses on **African-American dialectal English (AAE)**. In our discussion of racial bias in toxicity detection, we will discuss research engaged in evaluating and mitigating racial discrimination against posts attributed to Black users. Several resources have been developed to estimate a poster's race, as these attributes are only sometimes available.

Disclaimer on dialect detection. While some of the broader literature around dialect detection is important to applications such as chatbots and automated speech systems, recognizing the dialect based on written content is tricky and contended. The initial work [25, 26] in dialect detection in text revolved around understanding the changing patterns linguistics; they have been recently adopted to study the impact of dialect in the classification of sentiment and toxicity [85, 120].

However, one has to note that detecting dialect can be a proxy for race and opens up the ethical implication of dialect detection and its application. The implication of the work by References [8] and [44] around automatic dialect detection can thus be problematic unless the context of the user's background is considered. For example, one can easily set up spam accounts on social media to imitate and defame a specific demographic region.

- **Blodgett LM and dataset [8]:** This language model was trained on tweets located in the USA and matched them with the USA census' demographic data for the four largest racial/ethnic groups (namely, those of **non-Hispanic Whites (NHW)**, **non-Hispanic Blacks (NHB)**, Hispanics, and Asians). Out of these, the authors used the language model to filter $1.1M$ (and $14.5M$) Black-oriented (and White-oriented) tweets by likely NHB (and NHW) users. *Using Twitter as a source to train for dialects, Blodgett **LM (Language Model)** could very well employ the same negative word association and stereotyping when predicting dialects, as is the case with toxicity detection models. Utilizing such LMs as black-box for future dialect prediction can exponentially impact the biases against non-white speakers [20]. Methods that use human-annotated dialect labels or do not require ground truth for measuring bias [1] should be favored over dialect-proxy LM-based models.*
- **Pietro dataset [74]:** It is a corpus of $5.4M$ tweets from $4,132$ survey participants who reported their race/ethnicity ($3,184$ NHW and 374 NHB).
- **Huang dataset [44]:** It is a multilingual dataset of tweets combining existing datasets from five languages annotated for toxicity. Four demographic attributes (including race) of the users who posted the comments were also annotated using their Twitter profiles.

In the remainder of the article, we refer to tweets marked as NHB-related (or NHW-related) as Black-oriented (or White-oriented) tweets. We let go of common disputed terminologies such as **Standard-American English (SAE)**, Mainstream US English, and so on, to not promote minority exclusion [80, 120]. Huang et al. [44] noted that racial information encoded in a tweet could introduce downstream harm. They trained models on the English subset of the Huang dataset to predict the author's race solely from the tweet's text. Contrary to the findings of Davidson et al. [20], they found that words like *n*gga* and *b*tch* were more significantly related to the NHW class, suggesting a derogatory use of these terms. *It is an exciting finding, as the Huang dataset is one of the few works not relying on the Blodgett LM for proxy of the user's race.*

Racial information can lead to annotation bias. Black-oriented tweets often get mislabeled as toxic. Sap et al. [85] showed a positive correlation between $p(NHB|tweet)$ and the toxic labels for the Davidson and Founta100k. Similarly, Davidson and Bhattacharya [19] applied **structural topic modeling (STM)** [79] on the Founta100k dataset and identified a latent topic containing terms prevalent in AAE (e.g., *ss, n*gga). to be more likely to be flagged as toxic.

Addressing the biased annotations. Sap et al. [85] and Zhou et al. [120] explored the mitigation of racial bias by addressing annotation bias.

- **Annotator priming:** Sap et al. [85] discovered that annotators were more likely to label a sample as "non-toxic" when primed with the dialect or race of the author. However, it must be noted that annotator priming is a tricky task. Sometimes, it can nurse the annotators' implicit human biases instead of suppressing them.
- **Dialect aware label correction[4]:** Zhou et al. [120] utilized the few-shot capabilities of the GPT-3 [13] by supplying a few seed examples [92] for Black-oriented to White-oriented dialect conversion. They used the Founta100k dataset and the Blodgett LM to find the *two *-oriented tweet divisions*. A Black-oriented sample is marked as non-toxic if its White

---

[4]Zhou et al. [120] cautioned against a real-world application of this method due to limitations of GPT-3.

counterpart is labeled non-toxic. The language model trained on the relabeled dataset led to a lower false positive rate on the Black-oriented test samples.

Diagnosing bias in downstream models. Sap et al. [85] showed that a downstream model trained on either Davidson or Founta100k dataset produced: (i) higher FPR values for the black-oriented comments and (ii) higher FNR values for the White-oriented comments of the test-split. Higher positive rates were also observed for Black-oriented tweets from the Blodgett and Pietro datasets. Similarly, Huang et al. [44] observed higher FPED and FNED values by downstream models trained and tested across all languages of the Huang dataset. Meanwhile, Davidson et al. [20] calculated the proportion of Black-oriented (p_{black}) and White-oriented (p_{white}) tweets from the Blodgett LM predicted as toxic by the models trained on respective training set. A t-test on five popular Twitter toxic speech datasets revealed a significant disparity against the Black-oriented tweets. This disparity reduced (yet persisted) when the t-test was repeated conditioning on the presence of identity terms. The authors noted that while this prevailing disparity could have been due to other terms not conditioned upon, it was also possibly the result of the model correlated other subtler details of AAE with the toxicity.

Mitigation through the debiased training of downstream models. Mozafari et al. [64], Xia et al. [116], Zhou et al. [120] explored mitigation of racial bias through model debiasing.

- **Comment re-weighting in loss:** Following Reference [90], Mozafari et al. [64] defined a bias score s_j^c for each label class c and bigram term t_j as:

$$s_j^c = \frac{\sum_{i=1}^n I_{t_j \in x_i} I_{y_i = c}(1 + \alpha_i)}{\sum_{i=1}^n (1 + \alpha_i)}, \tag{3}$$

  where α_i is the weight of the ith sample, and I is the indicator function. These sample weights α were learned by solving the following optimization:

$$min\left(\sum_{j=1}^{|V|} max_c\left(s_j^c\right) + \lambda||\alpha||_2\right), \tag{4}$$

  where V is the set of bigrams in the training data, and λ is a hyperparameter. Though the debiasing was applied to the complete vocabulary, mitigation only in racial bias was evaluated. Reduction in bias against the Black-oriented tweets across all toxicity labels was observed for models trained on the W&H and Davidson datasets.
- **Adversarial training:** Xia et al. [116] used an adversarial training [53] on the Founta100k dataset. One classification head (C) predicted the toxicity label, while the adversary head (D) predicted the protected attribute $p(NHB|tweet)$. This method was shown to be effective when lexical cues between the two attributes are closely related (e.g., terms like n^*gga and b^*tch correlate with both the AAE dialect and toxicity labels [20]). The model showed a reduction in FPR and an increase in macro-$F1$ scores.
- **Ensemble-based debiasing:** Similar to lexical bias (see Section 3.3), Zhou et al. [120] utilized LearnedMixIn for mitigating racial bias on the Founta dataset. They used the four $p(*|tweet)$ predictions from the Blodgett LM as features for the naive model. A drop of FPR on the Black-oriented comments was observed, along with a drop in macro-F1. *However, an insignificant drop in disparity against the Black-oriented tweets was observed when tested on the Pietro dataset, suggesting the inability of mitigation methods on out-of-domain data.*

4.2 Gender Bias

A note on gender in NLP. Bolukbasi et al. [10] pointed out the gendered bias in word embeddings considers a pair of masculine or feminine attributes and job occupations. Since then, multiple studies have evaluated gender bias [96, 104] pinned on the binary gender identity. Only a few studies have explored this phenomenon from a non-binary perspective [118]. Talking specifically of toxicity detection systems, Jigsaw[5] is probably the only dataset that considers more inclusive and fine-grained labels for gender and sexual orientation as the target of toxicity. Unfortunately, the study of non-binary gender bias is missing from the toxicity literature.

Evaluating gender bias. Waseem [108] showed that a model trained on the W&H dataset led to a performance gain by including the user's gender as a feature [86]. It stems from the fact that abusive gendered terms such as "b*t**," "wh***" are more likely to be observed in samples labelled as toxic. Park et al. [70] compared the FPED and FNED values using an SBET [97] model trained on the W&H and Founta datasets. They found that using pre-trained embeddings improved toxic speech detection performance at the cost of more significant equality difference scores. The authors also observed higher bias values for the W&H dataset than the Founta dataset, attributable to Founta's bias-aware design (Section 3.1).

Mitigation techniques. Based on their observations, Nozza et al. [68] and Park et al. [70] further proposed mitigation techniques for gender bias:

- **Debiased embeddings:** Park et al. [70] applied three strategies for gender bias mitigation on the W&H dataset: (i) using debiased-word2vec embeddings [10], (ii) gender-swap data augmentation (using gender pairs identified by Reference [119]), and (iii) transfer learning by first fine-tuning on a less biased large-scale dataset (Founta [34]). The best debiasing performance was achieved by combining debiased embeddings and gender-swapping, reducing FPED and FNED. Meanwhile, the transfer learning approach led to the highest AUC loss. *We attribute this to incompatible labels between the source and the target datasets (abusive vs. sexism).*
- **Length-sensitive upsampling:** Nozza et al. [68] identified 12 common female identity terms in the Evalita2018 [29] dataset. Next, they upsampled tweets from the W&H dataset to balance the count of occurrences of each identity term across the toxic and non-toxic classes of the training dataset. Upsampling while considering the tweet length ranges (similar to Dixon et al. [24]) led to the best gender debiasing. This debiasing, however, came at the cost of a small AUC drop on the test split. *The W&H dataset is known to contain a host of biases that can affect the resulting upsampled dataset.* [113]

4.3 Intersectional Bias

Kim et al. [51] study the intersection of racial and gender bias for AAE. They used the Founta100k dataset and labeled race using the Blodgett LM (cf., Section 4.1). They also marked the dataset with gender and party (political inclination). The party information is used as a control variable. While tweets belonging to the Black community are more likely to be classified as abusive, Black males are significantly more likely to be classified as hateful. In toxic speech detection, it is essential to evaluate such intersectional biases more carefully.

4.4 Cross-geographic Bias

Ghosh et al. [37] observed that most literature in toxicity detection focuses on the English language; this concentrated attention towards a few geographies creates a *knowledge gap* and can

---

[5]https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/.

Table 3. A Summary of Debiasing Methods Employed for Different Types of Unintended Biases as Discussed in This Survey

Bias	Debiasing Method
Sampling	Topic proximity [104], Data mixing (augmentation) [4]
Annotation	Disagreement deconvolution [39]
Lexical	Length-sensitive sampling [24], Knowledge-based term generalization [5], Data filtering [120], Regularization [49], Multi-task learning [103], Ensemble modeling [120]
Racial	Annotation priming [85], Label correction [120], Regularization [90], Adversarial training [116], Ensemble modeling [120]
Gender	Length-sensitive sampling [68], Debiased LM [10], Gender Swapping [119], Transfer learning [70]

lead to lexical bias. As a concrete example; they showed that Alphabet's Perspective API[6] gives lower toxicity scores to terms that are considered toxic in Indian context (for example, *presstitute*, a slang combining the terms *press* and *prostitute*) leading to false negative predictions. Conversely, terms such as *muslim* and *hindu* were observed to generate higher toxicity scores even in a non-toxic context, leading to false positives. A set of reproduced terms from their study is provided in Figure 4. Ghosh et al. [37] then proposed a two-step weakly supervised method to detect lexical bias for cross-geocultural harmful content. They carried out this analysis using unlabeled tweets collected from seven countries. In the first step, they identified a set of terms T that were statistically overrepresented in the tweets from a country. Their study separated T along two axes concerning the model under investigation: (i) descriptive axis: correlational vs. causal associations, and (ii) prescriptive axes: desirable vs. undesirable associations. Correlational associations refer to the terms not invoking toxicity in the model, while causal refers to the terms causing higher toxicity predictions. Similarly, desirable associations refer to the terms for which higher toxicity ratings are desirable and vice versa for undesirable. In the second step, to find the correlational associations, they built 33 hand-curated template sentences with varying degrees of toxicity (for example, "You are a <person>," "I dislike <person>"). They replaced the template with toxic entities in T and recorded the change in toxicity score. Any term in T causing the toxicity score to increase while it was not desired to is an instance of spurious correlation. They then employed various mitigation techniques such as deletion, substitution, and balancing and found that these methods did not significantly reduce bias.

4.5 Political Bias

Recently, Wich et al. [112] studied the effect of political bias on hate speech detection. Tweets were collected from political leaders who represented the left-winged, right-winged, and neutral ideologies, with each tweet heuristically marked as non-toxic. Finally, these three politically inclined corpora were used to replace non-toxic comments from an existing corpus [94, 114], resulting in politically biased datasets. It was observed that the right-wing-biased dataset produced statistically lower $F1$ lower than the other two biased datasets—*the work did not explore introducing political bias for the toxic label. We find this assumption unconvincing that politically based tweets were treated as only non-toxic.*

5 MAJOR TAKEAWAYS: A SUMMARY

This section highlights the key takeaways from surveying existing debiasing methods. The summary table of various debiasing methods for the respective bias is provided in Table 3. We also compare physical systems and bias as physical entities. This analogy highlights the drawbacks

---

[6]https://perspectiveapi.com/.

of existing debiasing methods and motivates a continuing integrated system for mitigating biases.

Inspiration from physical systems. We extend the interpretation that "*energy can neither be created nor destroyed but only transformed from one form to another*" to bias as a physical entity. One can say that as long as humans are bombarded with more information than they can process, some form of cognitive biases will keep existing and evolving as per the social structures of the zeitgeist. It also means that, despite our best efforts, our interactions (online or offline) will be riddled with biases and stereotyping—the same manifests as biases in downstream tasks of NLP. Just like the "*system at rest remains at rest unless acted upon by an external force,*" to consciously and effectively overcome biases, we need to supply our NLP pipelines with external impetus in the form of debiasing evaluations and criteria. However, one-time mitigation and evaluation of bias, as tested in the controlled setting by all the techniques discussed in this survey, will unfortunately not be an accurate (or even valid) indicator of bias reduction as the system evolves. The principle that the "*entropy of a closed system never decreases*" renders our one-time bias mitigation efforts moot. Translating entropy to the close-minded biases we harbor, our biases, once seeded, will not automatically disappear. Instead, bias, as an entity, is expected to reinforce and strengthen with time, i.e., years of systematic discrimination and marginalization leading to increased instances of hate speech (offline and online) and hate crime. Thus, we need a continuously probing setup (a regular supply of external energy) to keep the biases in the NLP pipeline in check. We can only ensure their mitigation in an ever-evolving system of human interactions by evaluating biases at every step of the workflow and analyzing it at regular intervals.

- **Categories of Bias:** As we observe in Figure 1, the bias categories based on the *sources* of harm are not necessarily exclusive. Lexical bias can be a consequence of sampling or annotation bias or both. Similarly, any category of bias based on downstream harms can be an artifact of all three categories based on the source of harm.
- **Sampling Bias:** These observations collectively suggest that the text source and the topics used for sampling have a greater influence on the bias characteristics of the dataset than the sampling strategy. Additionally, OSNs have varying tolerance towards toxic speech [65]. The difference in platform-specific policies can further affect the bias characteristics of the sampled dataset.
- **Lexical Bias:** There should be a balance of sensitive words and phrases such as explicit abuse, identity terms (especially that of minority groups), and topic words across the labels to not fabricate any spurious lexical correlations with the labels. These correlations can disrate the model's ability to capture the context of such terms. A common point of failure of models trained on such data is the conflation of identity disclosures (e.g., "I am gay") with identity attacks (e.g., "I hate all gays"), further promoting disparity among social groups.
- **Annotation Bias:** The current literature for annotating harmful content is spread on the spectrum from strict to loosely defined guidelines. Consequently, we observe that the annotation agreements also vary, with no fixed benchmark for toxic content labeling. The observations and results from the above sections highlight the importance of examining the sensitivity of the annotators towards socio-cultural factors such as dialects and race. To effectively study annotators' behavior across attributes, it is essential to have a fair number of participants.
- **Racial Bias:** Every mitigation strategy suggested to reduce racial bias in existing models assumed that both Black-oriented and White-oriented samples follow the same conditional probability $p(Y|X)$. This assumption is flawed, because using specific terms is socially acceptable for a Black person while unacceptable for someone else. Though the resources from References [8, 44, 74] have aided the investigation of racial bias in toxic speech, none of them

have access to the ground-truth labels for the dialect identities. For example, it is possible that the Blodgett LM spuriously correlates the presence of terms like "*n*gga*" and "*b*tch*" with the tweet being Black-oriented [20]. Such predictions can magnify the racial bias during toxicity detection. Moreover, the assumption that only certain social groups employ specific terms is flawed in today's era. While, on the one hand, it leads to broader adoption and acceptance of cultural terms, on the other hand, such terms are employed by extremist groups and trolls for misappropriation and mocking other cultures. Without knowing the background and intent of the online users, it is difficult to pinpoint what they wish to achieve when adopting non-native terms and dialects. The lack of awareness about dialects and their cultural importance among NLP researchers can itself be a source of bias.

6 DIRECTION FOR FUTURE RESEARCH

This section discusses the common challenges across bias mitigation in toxic speech and their possible solutions. Some of these challenges point to the general area of toxicity detection modeling and has been touched upon in other surveys on toxic speech [17, 117].

- **Cognizance towards side effects.** Owing to resource constraints, the study of bias and its mitigation has focused on reducing only one bias at a time. While bias mitigation is vital for toxic speech detection, it is important to acknowledge its ability to introduce newer biases in the pipeline. As stated earlier, the taxonomy of bias is overlapping in nature. For example, lexical bias can be a source of racial and gendered harm. Meanwhile, the lexical bias could be introduced due to spurious data collection and annotation biases. We observe that researchers often failed to acknowledge this critical aspect while evaluating their proposed mitigation methods (Appendix B). Initial work in the direction of intersectional bias has been led by Kim et al. [51], who analyzed the combined impact of gender and race (Section 4.3). However, the analysis and evaluation of the interplay of various biases on toxicity detection remains an open question.
- **Data collection and annotation.** As observed and discussed in Section 3.1, the source and topic of content can have an overarching impact on the characteristics of the dataset curated. While random sampling is closer in characteristics to the real-world distribution, they are highly skewed towards non-hate, which makes collecting toxic comments hard. Meanwhile, priming for specific topics, hashtags, or users to increase the toxic content introduces unintended biases into the dataset and the modelling pipeline. Recently Rahman et al. [77] proposed an **information retrieval (IR)**-based approach to collecting hate speech from Twitter. Their IR-inspired method increased the coverage of hate compared to existing datasets. Such cross-domain methods can help increase the relevance of the content that should be filtered for labelling. Keeping the biases in check can lead to better generalization. Meanwhile, proposing and standardizing code books for data annotation of toxicity labels can help reduce the variability in labels employed by different datasets. A case study is the large-scale harassment annotation code book [38] that helped label 35k instances of various categories of online harassment, including hate speech. When researchers want to propose a new toxicity class, a clear distinction and relation to existing labels should be made. Another exciting area of research can be employing weakly supervised or unsupervised methods of data augmentation [87] and domain adaption [57] to reduce the need for annotations and improve generalizability of toxicity detection models.
  Gap in social and computational understanding of toxicity. It is also essential to highlight the gap between the social and psychological studies of harassment faced by vulnerable groups and its computation analysis by technical researchers and industrialists. Currently,

the two systems operate in silos with little interaction between them. Such practices lead to uninformed people developing toxicity detection pipelines without a nuanced understanding of the social issues. For example, without being cognizant of the social/mental impact of objectifying women, it can be tricky to crawl public forums that capture toxicity against gendered body imaging. Therefore, we need to closely incorporate and synergize social issues and their corresponding data sampling and annotation.

- **Gender is not just binary.** Existing literature in the area of gender debiasing in NLP as well as in toxicity detection has evaluated gender as binary (male vs. female). In their recent work, Dev et al. [23] provided a general overview of how non-binary individuals are at risk of erasure and misgendering at the hands of existing language models. These harms trickle down to the task of toxicity detection as well and, unfortunately, its full extent has not been studied yet. As observed in the existing analysis of annotation and lexical biases, annotators' lack of awareness around gender fluidity can lead to inconsistent labels. More so, the toxicity models predict both "I am a homosexual" and "I hate homosexuals" as toxic due to the presence of the word "homosexual," which has been historically used to detest the LGBTQ+ community and has only recently been reclaimed.

- **Language is not static.** Extending from the previous point, the case of word reclamation is a part of the discussion around the use of static hate lexicons, offensive dictionaries, and static knowledge graphs, which cannot account for the evolving language and the evolving social-cultural aspects [93]. Recently, Qian et al. [76] proposed a prototypical learned model for hate speech classification that aims to capture the evolving hateful content as it develops. Such models need to be extended to debiasing methods as well.

- **Proactive bias mitigation.** Zhou et al. [120] showed that instead of targeting existing bias-ridden datasets, downstream harms could be better mitigated by incorporating mitigation techniques starting in the early stages of the learning pipeline, such as data sampling and annotation. While the mitigation methods proposed for annotation and sampling biases are tied to their respective steps in the machine learning pipeline, the methods for other categories are distributed throughout the pipeline. Additionally, most of the surveyed papers demonstrated bias mitigation as a single-step solution [24, 64]. However, it is essential to be "bias-aware" throughout the learning pipeline [70]. A good reference for this can be Reference [98], which formalized the complete pipeline as a sequence of data transformations and defined the potential sources of harm.

- **Out-of-domain evaluation.** Zhou et al. [120] showed that mitigation techniques displaying encouraging results on in-domain samples failed to reduce disparity when tested on out-of-domain datasets. This is an alarming finding, as most mitigation techniques are tested on in-domain data, putting their generalizability in question. This observation entails introducing diverse benchmarks to standardize existing and future work. For example, a benchmark [107] was recently developed to systematically compare gender-bias mitigation techniques in visual recognition models.

- **Collecting user feedback.** The Jigsaw team utilized user feedback as a key source of bias mitigation[7] for its Perspective API. Setting up of a feedback infrastructure, wherever possible, allows a collection of data that better represents the target population.

- **World beyond English text.** In the majority of literature around toxicity detection and debiasing, we mostly consider English datasets. English, being the most widely available language on the Internet, has most accurate preprocessing tools. This has created a knowledge gap when applying toxicity systems for other languages. Especially since what can be

---

[7]https://medium.com/jigsaw/unintended-bias-and-names-of-frequently-targeted-groups-8e0b81f80a23.

considered toxic in English-speaking geographies may not be considered toxic in other geographies. The initial study in cross-cultural bias is being led by the work of Ghosh et al. [37]. However, the extensive study of toxicity bias in non-English and code-mixed settings remains non-existent. Additionally, the workaround flagging harmful content has focused majorly on text-based features, as they are easier to collect. Meanwhile, the usage of memes and videos (short clips and long ones) spreading toxic and harmful content has been gaining momentum [50, 72, 73]. We need to study the impact of bias in multi-modal content as well. The combination of the individual as well as unified multi-modal bias is an exciting and open area of research.

7 CONCLUSION

While a reasonable amount of work has been dedicated to unintended bias in toxic speech detection, we observe that its mitigation needs further exploration. Due to the multi-faceted and dynamic nature of online toxicity and its unintended biases, conducting a systematic study helps better understand the scope of the bias mitigation strategies. To our knowledge, no survey approaches this subject, focusing on the proposed methods. Therefore, we filled the gap through this survey. We developed a taxonomy of bias based on their source and target of harm. This categorization enables us to effectively discuss these methods and their drawbacks, challenges, and future work directions. We also draw attention to the need to handle more psychographic biases. For example, toxic speech based on political interests is a known issue on OSNs. However, the study of discrimination based on political leanings in harmful speech detection is still an under-explored avenue [112]. We also conducted a case study to introduce the concept of bias shift due to knowledge-based bias mitigation methods. While certainly not exhaustive, we called attention to a list of common challenges and pitfalls of bias-handling methods for toxic speech detection.

APPENDICES
A EVALUATION METRICS
A.1 Background on Bias Evaluation

This section presents about two popular bias evaluation metrics outside the task of classification. This section briefly discusses two popular bias evaluation metrics outside the classification task.
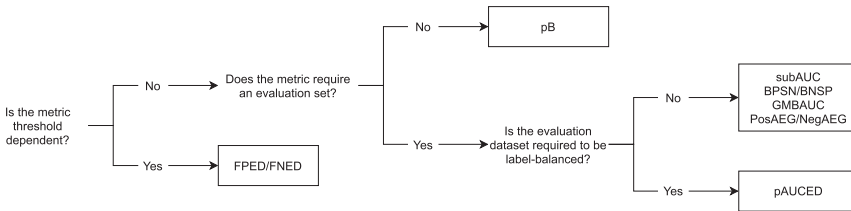
Fig. 5. A taxonomy of bias evaluation metrics popular in toxic speech detection.

Bias evaluation based on psychological tests: Back in 1988 Greenwald et al. [41] developed the **implicit association test (IAT)**[8] as a measure to capture the subconscious biases in humans. It is based on observing the difference in reaction times and accuracy in categorizing two words settings. The first setting relates male with science and female with liberal arts. The second setting reverses the relations. In both settings, the test subjects are tasked to categorize a word towards

---

[8]https://implicit.harvard.edu/implicit/.

either of these relations. It was observed that most test subjects could categorize words faster and accurately in the first relation (male = science) compared to the second (female = science). *Note, the reliability of IAT is questionable [88], and it can at best be considered only a weak indicator of bias.*[9]

Caliskan et al. [15] extended IAT to develop the **Word Embedding Association Test (WEAT)**. It measures social biases through difference in association strengths in non-contextual word embeddings (e.g., word2vec [61], GloVe [71]). It should be noted that the effectiveness of WEAT as a metric heavily depends on the terms used to seed the algorithm. Later, May et al. [60] extended WEAT for sentence embeddings, naming it the **Sentence Encoding Association Test (SEAT)**. SEAT used templates ("This is a[n] <word>") to insert individual words into the slot "<word>." These templates were designed to convey minimal focus towards the context of the inserted words, helping measure the associations a sentence encoder makes with these inserted terms. Next, it applied WEAT to these synthetic sentences' embeddings instead of word embeddings. In other words, SEAT is a generalization of WEAT on multi-word sequences.

Bias evaluation based on finding bias subspace: Bolukbasi et al. [10] defined the gender bias of a word as the correlation between the projection of the word embedding onto the gender sub-space and the manually annotated bias rating of the word [97]. This gender subspace itself was found by applying **principal component analysis (PCA)** on a set of vectors obtained by taking a difference of words that vary only in the gender context (e.g., $\overrightarrow{king} - \overrightarrow{queen}$) and retaining the top component(s). Liang et al. [55] extended this idea to contextual sentence embeddings. Unlike Bolukbasi et al. [10], to calculate the gender sub-space, they used a large number of sentence pairs, with each sentence in a pair varying just in the gender-specific word. Both the bias evaluation categories described are yet to find a direct utility in the field of bias evaluation in toxic speech classification. One rudimentary exploration for the toxic speech classification task can be the analysis of social biases in fine-tuned word embeddings or sentence encoders after training them for the downstream task dataset.

A.2 Unintended Bias Metrics

This subsection discusses metrics developed especially for bias evaluation of classification models. As described in Figure 5, we create a taxonomy of bias evaluation metrics standard in toxic speech detection literature. We list the metric abbreviations along with other commonly used abbreviations in the article in Table 2. In the following definitions, we assume T as a set of terms $t_1, t_2, \ldots, t_{|T|}$. The metrics defined for a term t can also be calculated for a designated group g among a host of vulnerable groups G.

For a test dataset D, the following definitions apply: (i) let D_t^+ be the positive (toxic) examples containing term t, (ii) let D_t^- be the negative (non-toxic) examples containing term t, (iii) let $D_{\setminus t}^+$ be the positive (toxic) examples not containing the term t, and (iv) let $D_{\setminus t}^-$ be the negative (non-toxic) examples not containing the term t.

- **Error rate equality differences (FPED, FNED):** Dixon et al. [24] introduced the metrics **False Positive Equality Difference (FPED), False Negative Equality Difference (FNED)**, and **pinned AUC Equality Difference (pAUCED)**. FPED and FNED quantify a variation of equality of opportunity. Mathematically, the authors calculated these variations of term-wise error rates (FPR_t and FNR_t) around the error rates (FPR and FNR) of the complete evaluation set:

$$FPED_T = \sum_{t \in T} |FPR - FPR_t|, \tag{5}$$

$$FNED_T = \sum_{t \in T} |FNR - FNR_t|. \tag{6}$$

Note that in an ideal case, $FPED_T = FNED_T = 0$. Both FPED and FNED are threshold-dependent and require a classifier that produces binary labels. However, many models produce probability distributions. Dixon et al. [24] and Borkan et al. [12] applied multiple threshold-agnostic metrics for such scenarios.

Usage: In general, FNR and FPR capture the upper bound on misclassification that a model can currently achieve. When employing FNR_t and FPR_t on different identity terms, if the model is biased towards the terms, then there will be higher variance in FNR_t and FPR_t, leading to higher FPED and FPED.

- **Sub-group AUC (subAUC$_t$):** It calculates AUC on $(D_t^+ \cup D_t^-)$, i.e., the examples containing the identity term t.

  Usage: $subAUC_t$ based on specific terms measures the model's ability to correctly predict the toxic and non-toxic classes based on t. A lower value means the model is not able to distinguish the toxic samples from the non-toxic ones containing term t. A higher value, however, does not mean the model is debiased. It can still give non-toxic mentions of term t a high toxicity score, but as long as it gives a toxic mention of t a relatively higher toxicity score than its non-toxic counter part, $subAUC_t$ will be high.

- **Pinned AUC (pAUCED)**: **Area under the receiver operating characteristic curve (AUC-ROC or AUC)** on the complete evaluation set can be insufficient at diagnosing bias, as a low AUC does not help in identifying the bias-ridden terms in T. Similar limitation has been noticed for $subAUC_t$. To overcome these limitations, Dixon et al. [24] developed **pinned AUC ($pAUCED_t$)** for a term t or a subgroup g, which is the AUC measure on an auxiliary dataset pD_t such that $pD_t = s(D_t) \cup s(D)$ and $|s(D_t)| = |s(D)|$, where D_t is the set of comments containing term t in the evaluation set, D is the complete evaluation set, and $s(.)$ is a sampling function. Here, the auxiliary data allows examples from $s(D)$ to be selected at random while "pinning" down the underlying distribution for the samples containing term t, i.e., the distribution of t in pD_t should resemble the overall distribution of t in D as closely as possible. In layman's terms, one can also assume the "pinned" samples to be the "representative" samples capturing the overall distribution of the control variable under consideration. In the toxicity study, these control variables are the identity terms t or the vulnerable group g. However, Borkan et al. [11] showed that the ability of $pAUCED$ to reveal bias is highly dependent on the distribution of labels between the identity terms. Dixon et al. [24] avoided this drawback by generating a **synthetic bias evaluation set (SBET)** with balanced label distribution. However, generating a SBET can be a tedious task [5].

  Usage: When we have multiple subgroups $g \in G$, we can employ pinned AUC to quantify and rank the level of unintended bias of the model w.r.t. various subgroups. The group with lowest pinned AUC is carrying the highest bias.

- **Background Positive Subgroup Negative AUC (BPSN), and Background Negative Subgroup Positive AUC (BNSP)**: $BPSN$ calculates AUC on test set where non-toxic samples contain the bias term t, while toxic samples that do not, i.e., $(D_t^- \cup D_{\backslash t}^+)$. Meanwhile, in BNSP AUC, we do the reverse, selecting the toxic samples that mention a term t and non-toxic ones that do not, i.e., $(D_t^+ \cup D_{\backslash t}^-)$. The former leads to reduction in FPR, while the latter leads to reducing in FNR.

  Usage: A lower BPSN means the model confuses non-toxic mentions of the identity term (t) with toxicity examples that do not contain the term. In other words, the model assumes the non-toxic samples containing t to be close to the toxicity class in general. For example,

when the term "nigga" is used in a post, "I feel for you, my nigga," the model focuses on the high correlation between "nigga" and toxicity instead of the neutral context.

However, a lower BNSP means the model confuses toxic samples that contain t with non-toxic samples that do not contain the term. Here, even the toxic usage of t is assumed to be closer to the non-toxicity class. For example, the sentence "Whites are superior to the rest" may not receive high toxicity as the identity term "White" usually carries low negative association without context. One can blame sampling and annotation (Sections 3.1 and 3.2) biases for this.

- **Generalized Means of Bias AUC (GMBAUC)**: To obtain a mean bias score across different bias AUC metrics, Jigsaw[10] introduced a generalized mean of the bias AUC (or GMBAUC) as:

$$GMBAUC_{p,T} = (0.25 \times AUC) + \sum_{a=1}^{3} 0.25 \times \left(\frac{1}{|T|} \sum_{t=1}^{|T|} m_{a,t}^p \right)^{\frac{1}{p}}, \tag{7}$$

where AUC is the overall ROC AUC score, $m_{a,t}$ is the ath AUC-based metric calculated for term t, and p is the power of the mean function. Inclusion of overall AUC helps in capturing the downstream performance.

Usage: GMBAUC is, essentially, an average of the subgroup AUCs that helps to easily compare different debiasing settings by having to follow only one metric. A lower GMB AUC like subgroup-specific AUC is a good indicator of unintended bias in the toxicity detection pipeline against one or more groups.

- **Pinned Bias (pB)**: In a work parallel to Reference [12], Badjatiya et al. [5] introduced a family of bias evaluation metrics, pinned Bias (pB):

$$pB_T = \sum_{t \in T} \frac{|p(\text{"}toxic\text{"}|t) - \phi|}{|T|}, \tag{8}$$

where (i) $p(\text{"}toxic\text{"}|t)$ is the prediction probability of the toxicity label for a sentence containing only term t, (ii) ϕ is the pinned/threshold value that differs for different members of the metric family. For example, to penalize $p(\text{"}toxic\text{"}|t)$ values above 0.5, $\phi = min(p(\text{"}toxic\text{"}|t), 0.5)$ can be used. Here, the pinning is on the threshold probability that depicts how much toxicity can be allowed for a term. The limitation of above formulation is that ϕ is constant for all terms in T, while in the real world the thresholds are dynamic.

Usage: pB can be used in tandem or even as a replacement for pinned AUC. While pAUCED requires the creation of a balanced auxiliary dataset for each term $t \in BSW$, pB utilizes the same dataset and is easier to measure.

B CASE STUDY ON KNOWLEDGE-DRIFT

Lexical debiasing. To overcome lexical bias (Section 3.3), Badjatiya et al. [5] employed a knowledge-based generalization method. It involved the replacement of the BSWs in the training dataset with an ancestor from the WordNet [62] hypernym-tree.

Knowledge-drift hypothesis. Motivated by the energy-bias analogy (refer to Section 5), we conjecture that replacing all occurrences $w \in BSW$ with its wordnet ancestor $a \in A$ will shift the bias from w to a. While generalization can be true for all the debiasing methods we discussed in this survey, for this experiment, we focus on the wordnet-based substitution method for lexical debiasing.

---

[10]https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/.

Limited evaluation. Post substitution of W with A, the authors employed the BSW on the original W to evaluate bias. There is no discussion on evaluating the bias on terms in A, which is now a substitute for W.

Experimental setup. All experiments were carried out on a Ubuntu 18.04.5 LTS system with 126 G RAM and 32 G Tesla V100. We apply debiasing method on the W&H dataset [109] to verify the hypothesis and compare by training a BERTweet [67] classifier.

Table 4. (a): BSWs w with Their Corresponding Selected Generalizations a According to the Wordnet-3 Scheme [5]. (b): pB Values for the Set of BSWs W and A for (i) the Model Obtained on Original W&H Dataset M_{bias} and (ii) the Model Obtained after Lexical Database Generalization M_{gen}

(a)

$w \in W$	$a \in A$
muslim, prophet, woman, christian, girl, terrorist, slave, man, child, driver	being
feminist, civilian, liar, comedian, god	someone
hate, slavery, hatred, want, truth, freedom	state

(b)

Metric	M_{bias}	M_{gen}
pB_W	0.027	0
pB_A	0.002	0.016

Observation. Table 4(a) lists the replaced words and their respective generalizations. The observations in Table 4(b) indicate a shift of bias from source to target words. Since the terms in the set A are more likely to be present in non-toxic comments, this shift in bias can also be detrimental to the non-toxical class. While this is a proof of concept case study, it is expected that the other mitigation techniques by Badjatiya et al. [5] based on knowledge-based generalizations suffer from a similar shift of bias instead of actually debiasing the dataset.

Resolving bias shift. Note that this shift of lexical bias is towards a more general set of terms A, where debiasing on this generalized dataset through upsampling [24] directly from an OSN (such as Twitter) seems like a feasible next step. For example, newly scraped comments containing terms such as *muslim* can not be directly added to the training dataset. It is because the number of toxic comments can be expected to be higher. However, if *muslim* is first generalized to *being*, then randomly scraping new comments with this new keyword can lead to a significantly lower toxicity ratio.

C   REPRODUCIBILITY OF DEBIASING METHODS

Motivation. In this section, we reproduce some of the bias mitigation methods described in the respective sections of bias. We mainly extend upon the results reported by Zhou et al. [120] and Park et al. [70] to incorporate W&H [109] and Davidson [21] datasets along with Founta [34]. By reproducing these methods, we can comment not just on the theoretical discussion of the debiasing techniques but also on their engineering feasibility. We hope to give the readers a glimpse of various settings (dataset, embedding, evaluation combination) for the debiasing methods presented in this survey.

Experimental Setup. All experiments were carried out on a Ubuntu 18.04.5 LTS system with 126 G RAM and 32 G Tesla V100. The hyper-parameters and epochs were kept intact as provided in the respective methodology.

C.1   Racial Bias

For studying racial bias, we reproduce the results reported in Zhou et al. [120]. As observed in Table 5(c) for the large-scale dataset (like Founta [34]), data filtering methods like Aflite that

Table 5. Reproduced Results for Racial Debiasing as Experimented by Zhou et al. [120]

(a)

Method/Result	W FPR	B FPR	W %	B %
Random	0.072	0.133	0.267	0.481
Original	0.025	0.024	0.162	0.380
LearnedMixIn	0.028	0.024	0.178	0.405

(b)

Method/Result	W FPR	B FPR	W %	B %
Random	0.029	0.022	0.045	0.033
Original	0.032	0.009	0.063	0.015
LearnedMixIn	0.193	0.308	0.194	0.313

(c)

Method/Result	W FPR	B FPR	W %	B %
Random	0.017	0.058	0.303	0.719
Original	0.040	0.115	0.300	0.725
Datamap-Easy	0.018	0.053	0.302	0.717
Datamap-Ambi	0.020	0.077	0.307	0.725
Datamap-Hard	0.020	0.073	0.287	0.716
Aflite	0.017	0.058	0.303	0.719
LearnedMixIn	0.0275	0.075	0.281	0.694

(a) Results on W&H [109], (b) Results on Davidson [21], (c) Results on Founta [34].

eventually operate on only 33% of the dataset provide a comparable drop (less is more) in *FPR*. Meanwhile, the LearnedMixIn method [18] that adopts model regularization provides the smallest percentage of Black tweets samples wrongly labelled as toxic. For W&H and Davidson datasets, applying data filtering and utilizing only 33% of the dataset would leave us only $\approx 5k$ and $\approx 8k$ tweets, respectively. Hence, we did not test data filtering on small-scale datasets and only reported unexpected and LearnMixIn results. Debiasing in both datasets via LearnMixIn (Tables 5(a) and 5(b)) is not better than the original dataset composition. This we assume could be happening due to the small volume of samples and the presence of other biases (lexical, topical) in the datasets.

C.2 Gender Bias

Here, we reproduce the results from Park et al. [70] by testing the combination of embedding and model against an exhaustive set of gender debiasing techniques. For the transfer learning, we test on Waseem and Davidson using Founta for training. In line with existing literature, we observe in Table 6 that pre-trained word embedding, on average, gives higher FPED and FNED due to their intrinsic gender bias. Similarly, as observed from Table 7, transfer learning leads to a higher F1 score, but it comes at the cost of increasing FPED and FNED. A combination of debiased word2vec, gender-swapping, and transfer learning, while a more complex set, seems to be the best tradeoff between reducing F1 and increasing FPED-FNED values. However, no one setting is best across datasets.

Table 6. Reproduced Results for Gender Debiasing Employing CNN, GRU, and α-GRU on Three Different Embeddings: *R*: Randomly Initialized, *F*: Fasttext, *W*: word2vec

(a)

Method	Emd.	FPR	FNR	FPED	FNED	F1
CNN	R	0.039	0.331	0.067	0.586	0.885
	F	0.060	0.213	0.097	0.384	0.902
	W	0.053	0.247	0.088	0.445	0.899
GRU	R	0.076	0.258	0.134	0.439	0.879
	F	0.071	0.277	0.120	0.477	0.877
	W	0.082	0.227	0.141	0.400	0.882
α-GRU	R	0.043	0.243	0.073	0.426	0.907
	F	0.080	0.228	0.142	0.414	0.885
	W	0.080	0.202	0.142	0.365	0.892

(b)

Method	Emd.	FPR	FNR	FPED	FNED	F1
CNN	R	0.010	0.842	0.019	1.561	0.917
	F	0.013	0.762	0.025	1.415	0.926
	W	0.008	0.768	0.016	1.439	0.928
GRU	R	0.043	0.654	0.082	1.222	0.916
	F	0.059	0.624	0.108	1.160	0.908
	W	0.035	0.735	0.065	1.370	0.914
α-GRU	R	0.037	0.750	0.069	1.372	0.912
	F	0.041	0.695	0.074	1.274	0.914
	W	0.047	0.683	0.087	1.256	0.910

(c)

Method	Emd.	FPR	FNR	FPED	FNED	F1
CNN	R	0.04	0.327	0.071	0.586	0.883
	F	0.048	0.247	0.077	0.449	0.902
	W	0.044	0.262	0.069	0.487	0.901
GRU	R	0.143	0.325	0.254	0.560	0.786
	F	0.070	0.125	0.125	0.216	0.909
	W	0.097	0.119	0.174	0.203	0.89
α-GRU	R	0.081	0.232	0.142	0.422	0.883
	F	0.098	0.183	0.170	0.327	0.884
	W	0.081	0.202	0.138	0.357	0.891

(a) Results on W&H [109], (b) Results on Davidson [21], (c) Results on Founta [34].

Table 7. Reproduced Results for Gender Debiasing Employing CNN, GRU, and α-GRU on a Combination of Three Debiasing Techniques: *DE*: Debiased word2vec, *GS*: Gender Swapping, and *FT*: Transfer Learning by Training on Founta [34]

(a)

Method	DE	GS	FT	FPR	FNR	FPED	FNED	F1
	✓	-	-	0.0294	0.316	0.045	0.567	0.897
	-	✓	-	0.047	0.297	0.080	0.555	0.890
	-	-	✓	0.054	0.331	0.090	0.600	0.875
CNN	✓	✓	-	0.054	0.267	0.087	0.491	0.893
	-	✓	✓	0.118	0.909	0.218	1.422	0.644
	✓	-	✓	0.124	0.871	0.223	1.365	0.656
	✓	✓	✓	0.118	0.909	0.218	1.422	0.644
	✓	-	-	0.070	0.297	0.123	0.515	0.872
	-	✓	-	0.076	0.247	0.130	0.427	0.882
	-	-	✓	0.095	0.427	0.166	0.723	0.818
GRU	✓	✓	-	0.048	0.342	0.084	0.581	0.874
	-	✓	✓	0.134	0.816	0.252	1.27	0.668
	✓	-	✓	0.064	0.712	0.118	1.185	0.746
	✓	✓	✓	0.065	0.708	0.119	1.177	0.746
	✓	-	-	0.065	0.297	0.115	0.525	0.877
	-	✓	-	0.075	0.255	0.130	0.456	0.881
	-	-	✓	0.061	0.278	0.107	0.487	0.885
α-GRU	✓	✓	-	0.097	0.232	0.169	0.400	0.872
	-	✓	✓	0.071	0.247	0.132	0.426	0.886
	✓	-	✓	0.085	0.243	0.148	0.445	0.877
	✓	✓	✓	0.092	0.247	0.155	0.437	0.871

(b)

Method	DE	GS	FT	FPR	FNR	FPED	FNED	F1
	✓	-	-	0.446	0.537	0.791	1.0	0.658
	-	✓	-	0.008	0.793	0.016	1.476	0.925
	-	-	✓	0.037	0.098	0.056	0.149	0.940
CNN	✓	✓	-	0.008	0.860	0.015	1.567	0.917
	-	✓	✓	0.085	0.927	0.148	1.707	0.867
	✓	-	✓	0.235	0.774	0.424	1.446	0.792
	✓	✓	✓	0.129	0.860	0.224	1.580	0.849
	✓	-	-	0.0415	0.685	0.075	1.278	0.914
	-	✓	-	0.035	0.735	0.065	1.370	0.914
	-	-	✓	0.491	0.501	0.712	1.0	0.679
GRU	✓	✓	-	0.0468	0.716	0.083	1.29	0.908
	-	✓	✓	0.009	1.0	0.014	1.821	0.898
	✓	-	✓	0.003	1.0	0.004	1.846	0.901
	✓	✓	✓	0.006	0.932	0.011	1.698	0.909
	✓	-	-	0.034	0.762	0.0618	1.409	0.912
	-	✓	-	0.031	0.738	0.057	1.366	0.916
	-	-	✓	0.079	0.120	0.123	0.182	0.906
α-GRU	✓	✓	-	0.035	0.781	0.065	1.402	0.909
	-	✓	✓	0.041	0.762	0.075	1.378	0.907
	✓	-	✓	0.038	0.787	0.068	1.439	0.907
	✓	✓	✓	0.014	0.860	0.024	1.555	0.914

✓marks the combination of debiasing used. (a) Results on W&H [109], (b) Results on Davidson [21].

REFERENCES

[1] Osman Aka, Ken Burke, Alex Bauerle, Christina Greer, and Margaret Mitchell. 2021. Measuring model biases in the absence of ground truth. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES'21)*. Association for Computing Machinery, New York, NY, 327–335. DOI:https://doi.org/10.1145/3461702.3462557

[2] Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the 4th Workshop on Online Abuse and Harms*. 184–190.

[3] Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? Analysis and detection of religious hate speech in the Arabic Twittersphere. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'18)*. IEEE, 69–76.

[4] Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 45–54.

[5] Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *Proceedings of the World Wide Web Conference*. 49–59.

[6] Ricardo Baeza-Yates. 2020. Bias in search and recommender systems. In *Proceedings of the 14th ACM Conference on Recommender Systems (RecSys'20)*. Association for Computing Machinery, New York, NY, 2. DOI:https://doi.org/10.1145/3383313.3418435

[7] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5454–5476.

[8] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1119–1130. DOI:https://doi.org/10.18653/v1/D16-1120

[9] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *J. Statist. Mechan.: Theor. Experim.* 2008, 10 (2008), P10008.

[10] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., Red Hook, NY, 4356–4364.

[11] Daniel Borkan, Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Limitations of pinned AUC for measuring unintended bias. *arXiv preprint arXiv:1903.02088* (2019).

[12] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Proceedings of the World Wide Web Conference.* 491–500.

[13] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[14] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research,* Vol. 81), Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. Retrieved from: https://proceedings.mlr.press/v81/buolamwini18a.html.

[15] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[16] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186. DOI : https://doi.org/10.1126/science.aal4230

[17] Tanmoy Chakraborty and Sarah Masud. 2022. Nipping in the bud: Detection, diffusion and mitigation of hate speech on social media. arXiv:2201.00961 [cs.SI].

[18] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19).* Association for Computational Linguistics, 4069–4082. DOI : https://doi.org/10.18653/v1/D19-1418

[19] Thomas Davidson and Debasmita Bhattacharya. 2020. Examining racial bias in an online abuse corpus with structural topic modeling. *arXiv preprint arXiv:2005.13041* (2020).

[20] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the 3rd Workshop on Abusive Language Online.* Association for Computational Linguistics, 25–35. DOI : https://doi.org/10.18653/v1/W19-3504

[21] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media.*

[22] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444* (2018).

[23] Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, 1968–1994. DOI : https://doi.org/10.18653/v1/2021.emnlp-main.150

[24] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* 67–73.

[25] Jacob Eisenstein. 2017. *Identifying Regional Dialects in On-line Social Media.* John Wiley & Sons, Ltd, 368–383. DOI : https://doi.org/10.1002/9781118827628.ch21

[26] Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. 2011. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics, 1365–1374. Retrieved from: https://aclanthology.org/P11-1137.

[27] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, 345–363. DOI : https://doi.org/10.18653/v1/2021.emnlp-main.29

[28] Christian Ezeibe. 2020. Hate speech and election violence in Nigeria. *J. Asian Afric. Stud.* 56, 4 (2020), 0021909620951208.

[29] Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the Evalita 2018 task on automatic misogyny identification (AMI). *EVALITA Eval. NLP Speech Tools Ital.* 12 (2018), 59.

[30] Agneta Fischer, Eran Halperin, Daphna Canetti, and Alba Jasini. 2018. Why we hate. *Emot. Rev.* 10, 4 (2018), 309–320.

[31] Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes et al. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the 3rd Workshop on Abusive Language Online.* 94–104.

[32] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.* 51, 4 (2018), 1–30.

[33] Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of hate speech datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference.* 6786–6794.

[34] Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of Twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media.*

[35] Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. 2019. On making reading comprehension more comprehensive. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering.* 105–112.

[36] Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L. Alfonso Ureña-López. 2021. A survey on bias in deep NLP. *Appl. Sci.* 11, 7 (Apr. 2021), 3184. DOI: https://doi.org/10.3390/app11073184

[37] Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. 2021. Detecting cross-geographic biases in toxicity modeling on social media. In *Proceedings of the 7th Workshop on Noisy User-generated Text (W-NUT’21).* Association for Computational Linguistics, 313–328. DOI: https://doi.org/10.18653/v1/2021.wnut-1.35

[38] Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. 2017. A large labeled corpus for online harassment research. In *Proceedings of the ACM Web Science Conference (WebSci’17).* Association for Computing Machinery, New York, NY, 229–233. DOI: https://doi.org/10.1145/3091478.3091509

[39] Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI’21).* Association for Computing Machinery, New York, NY. DOI: https://doi.org/10.1145/3411764.3445423

[40] Leon Graumas, Roy David, and Tommaso Caselli. 2019. Twitter-based polarised embeddings for abusive language detection. In *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW’19).* IEEE, 1–7.

[41] Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *J. Personal. Soc. Psychol.* 74, 6 (1998), 1464–1480. DOI: https://doi.org/10.1037/0022-3514.74.6.1464

[42] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, 3309–3326. DOI: https://doi.org/10.18653/v1/2022.acl-long.234

[43] Martie G. Haselton, Daniel Nettle, and Paul W. Andrews. 2015. The evolution of cognitive bias. In *The Handbook of Evolutionary Psychology.* John Wiley & Sons, Inc., 724–746. DOI: https://doi.org/10.1002/9780470939376.ch25

[44] Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. 2020. Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. *arXiv preprint arXiv:2002.10361* (2020).

[45] Ferenc Huszár, Sofia Ira Ktena, Conor O’Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. 2021. Algorithmic amplification of politics on Twitter. *Proc. Nat. Acad. Sci.* 119, 1 (Dec. 2021), e2025334119. DOI: https://doi.org/10.1073/pnas.2025334119

[46] Òscar Garibo i Orts. 2019. Multilingual detection of hate speech against immigrants and women in Twitter at SemEval-2019 task 5: Frequency analysis interpolation for hate in speech detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation.* 460–463.

[47] Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. 2020. Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. In *Proceedings of the International Conference on Learning Representations.* Retrieved from: https://openreview.net/forum?id=BkxRRkSKwr.

[48] Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez et al. 2018. TheGab Hate Corpus: A collection of 27k posts annotated for hate speech. *PsyArXiv* (2018). DOI: https://doi.org/doi:10.31234/osf.io/hqjxn

[49] Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. *arXiv preprint arXiv:2005.02439* (2020).

[50] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790* (2020).

[51] Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional Bias in Hate Speech and Abusive Language Datasets. arXiv:2005.05921 [cs.CL].

[52] Senthil Kumar, Aravindan Chandrabose, and Bharathi Raja Chakravarthi. 2021. An overview of fairness in data—Illuminating the bias in data pipeline. In *Proceedings of the 1st Workshop on Language Technology for Equality, Diversity*

and Inclusion. Association for Computational Linguistics, 34–45. Retrieved from https://aclanthology.org/2021.ltedi-1.5.

[53] Sachin Kumar, Shuly Wintner, Noah A. Smith, and Yulia Tsvetkov. 2019. Topics to avoid: Demoting latent confounds in text classification. *arXiv preprint arXiv:1909.00453* (2019).

[54] Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *Proceedings of the International Conference on Machine Learning.* PMLR, 1078–1088.

[55] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, 5502–5515. DOI : https://doi.org/10.18653/v1/2020.acl-main.488

[56] Haochen Liu, Wei Jin, Hamid Karimi, Zitao Liu, and Jiliang Tang. 2021. The authors matter: Understanding and mitigating implicit bias in deep text classification. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021.* Association for Computational Linguistics, 74–85. DOI : https://doi.org/10.18653/v1/2021.findings-acl.7

[57] Florian Ludwig, Klara Dolos, Torsten Zesch, and Eleanor Hobley. 2022. Improving generalization of hate speech detection systems to novel target groups via domain adaptation. In *Proceedings of the 6th Workshop on Online Abuse and Harms (WOAH'22).* Association for Computational Linguistics, 29–39. DOI : https://doi.org/10.18653/v1/2022.woah-1.4

[58] Raghvendra Mall, Mridul Nagpal, Joni Salminen, Hind Almerekhi, Soon-Gyo Jung, and Bernard J. Jansen. 2020. Four types of toxic people: Characterizing online users' toxicity over time. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society (NordiCHI'20).* Association for Computing Machinery, New York, NY. DOI : https://doi.org/10.1145/3419249.3420142

[59] Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science.* 173–182.

[60] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics, 622–628. DOI : https://doi.org/10.18653/v1/N19-1063

[61] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations.* Retrieved from: http://arxiv.org/abs/1301.3781.

[62] George A. Miller. 1995. WordNet: A lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[63] Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Author profiling for abuse detection. In *Proceedings of the 27th International Conference on Computational Linguistics.* 1088–1098.

[64] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PloS One* 15, 8 (2020), e0237861.

[65] Luke Munn. 2020. Angry by design: Toxic communication and technical architectures. *Human. Soc. Sci. Commun.* 7, 1 (July 2020). DOI : https://doi.org/10.1057/s41599-020-00550-7

[66] Jose I. Navarro. 2013. The psychology of hatred. *Open Criminol. J.* 6, 1 (Apr. 2013), 10–17. DOI : https://doi.org/10.2174/1874917801306010010

[67] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations.* Association for Computational Linguistics, 9–14. DOI : https://doi.org/10.18653/v1/2020.emnlp-demos.2

[68] Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence.* 149–155.

[69] Nedjma Ousidhoum, Yangqiu Song, and Dit-Yan Yeung. 2020. Comparative evaluation of label agnostic selection bias in multilingual hate speech datasets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'20).* 2532–2542.

[70] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, 2799–2804. DOI : https://doi.org/10.18653/v1/D18-1302

[71] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14).* 1532–1543.

[72] Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Detecting harmful memes and their targets. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021.* Association for Computational Linguistics, 2783–2796. DOI : https://doi.org/10.18653/v1/2021.findings-acl.246

[73] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, 4439–4455. Retrieved from: https://aclanthology.org/2021.findings-emnlp.379.

[74] Daniel Preoţiuc-Pietro and Lyle Ungar. 2018. User-level race and ethnicity predictors from Twitter text. In *Proceedings of the 27th International Conference on Computational Linguistics*. 1534–1545.

[75] Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2018. Leveraging intra-user and inter-user representation learning for automated hate speech detection. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 118–123. DOI: https://doi.org/10.18653/v1/N18-2019

[76] Jing Qian, Hong Wang, Mai ElSherief, and Xifeng Yan. 2021. Lifelong learning of hate speech classification on social media. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2304–2314. DOI: https://doi.org/10.18653/v1/2021.naacl-main.183

[77] Md Mustafizur Rahman, Dinesh Balakrishnan, Dhiraj Murthy, Mucahid Kutlu, and Matthew Lease. 2021. An information retrieval approach to building datasets for hate speech detection. In *Proceedings of the 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. Retrieved from: https://openreview.net/forum?id=jI_BbL-qjJN.

[78] Dante Razo and Sandra Kübler. 2020. Investigating sampling bias in abusive language detection. In *Proceedings of the 4th Workshop on Online Abuse and Harms*. 70–78.

[79] Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. Structural topic models for open-ended survey responses. *Amer. J. Polit. Sci.* 58, 4 (2014), 1064–1082.

[80] Jonathan Rosa. 2019. *Looking Like a Language, Sounding Like a Race*. Oxford University Press.

[81] Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the European refugee crisis. *arXiv preprint arXiv:1701.08118* (2017).

[82] Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 175–190. DOI: https://doi.org/10.18653/v1/2022.naacl-main.13

[83] Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM Conference on Web Science (WebSci'19)*. Association for Computing Machinery, New York, NY, 255–264. DOI: https://doi.org/10.1145/3292522.3326032

[84] Joni Salminen, Sercan Sengün, Juan Corporan, Soon Gyo Jung, and Bernard J. Jansen. 2020. Topic-driven toxicity: Exploring the relationship between online toxicity and news topics. *PLoS One* 15, 2 (Feb. 2020), e0228723. DOI: https://doi.org/10.1371/journal.pone.0228723

[85] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1668–1678.

[86] Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and H. Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1146–1151.

[87] Sheikh Muhammad Sarwar and Vanessa Murdock. 2022. Unsupervised domain adaptation for hate speech detection using a data augmentation approach. *Proc. Int. AAAI Conf. We Soc. Media* 16, 1 (May 2022), 852–862. Retrieved from: https://ojs.aaai.org/index.php/ICWSM/article/view/19340.

[88] Ulrich Schimmack. 2021. Invalid claims about the validity of implicit association tests by prisoners of the implicit social-cognition paradigm. *Perspect. Psycholog. Sci.* 16, 2 (Mar. 2021), 435–442. DOI: https://doi.org/10.1177/1745691621991860

[89] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media*. 1–10.

[90] Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. Association for Computational Linguistics, 3419–3425. DOI: https://doi.org/10.18653/v1/D19-1341

[91] Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 5248–5264. DOI: https://doi.org/10.18653/v1/2020.acl-main.468

[92] Arthur K. Spears. 2021. African-American language use: Ideology and so-called obscenity. *African-American English*, Routledge, 249–276. https://doi.org/10.4324/9781003165330-9

[93] Luc Steels. 2016. Human language is a culturally evolving system. *Psychonom. Bull. Rev.* 24, 1 (July 2016), 190–193. DOI: https://doi.org/10.3758/s13423-016-1086-6

[94] Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GermEval task 2, 2019 shared task on the identification of offensive language. *Preliminary Proceedings of the 15th Conference on Natural Language Processing (KONVENS'19, October 9–11, 2019 at Friedrich-Alexander-Universität Erlangen-Nürnberg)*, German Society for Computational Linguistics & Language Technology und Friedrich-Alexander-Universität Erlangen-Nürnberg, München [u.a.], 352–363. https://nbn-resolving.org/urn:nbn:de:bsz:mh39-93197.

[95] J. Suler. 2004. The online disinhibition effect. *Cyberpsychol. Behav.: Impact Internet, Multim. Virt. Real. Behav. Societ.* 7, 3 (2004), 321–326.

[96] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, 1630–1640. DOI: https://doi.org/10.18653/v1/P19-1159

[97] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976* (2019).

[98] Harini Suresh and John V. Guttag. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002* (2019).

[99] Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'20).* Association for Computational Linguistics, 9275–9293. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.746

[100] Shawn Tan, Yikang Shen, Chin-wei Huang, and Aaron Courville. 2019. Investigating biases in textual entailment datasets. *arXiv preprint arXiv:1906.09635* (2019).

[101] Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the 1st ACL Workshop on Ethics in Natural Language Processing.* Association for Computational Linguistics, 53–59. DOI: https://doi.org/10.18653/v1/W17-1606

[102] Daniel Trottier. 2020. Denunciation and doxing: towards a conceptual model of digital vigilantism. *Global Crime* 21, 3-4 (2020), 196–212. DOI: https://doi.org/10.1080/17440572.2019.1591952

[103] Ameya Vaidya, Feng Mai, and Yue Ning. 2020. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. *Proc. Int. AAAI Conf. Web Soc. Media* 14, 1 (May 2020), 683–693. Retrieved from: https://ojs.aaai.org/index.php/ICWSM/article/view/7334.

[104] Oskar Van Der Wal, Jaap Jumelet, Katrin Schulz, and Willem Zuidema. 2022. The birth of bias: A case study on the evolution of gender bias in an English language model. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP'22).* Association for Computational Linguistics, 75–75. DOI: https://doi.org/10.18653/v1/2022.gebnlp-1.8

[105] Juliet Van Rosendaal, Tommaso Caselli, and Malvina Nissim. 2020. Lower bias, higher density abusive language datasets: A recipe. In *Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language.* 14–19.

[106] Sebastian Wachs, Manuel Gámez-Guadix, and Michelle F. Wright. 2022. Online hate speech victimization and depressive symptoms among adolescents: The protective role of resilience. *Cyberpsychol., Behav., Soc.l Netw.* 25, 7 (2022), 416–423. DOI: https://doi.org/10.1089/cyber.2022.0009

[107] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. 2020. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 8919–8928.

[108] Zeerak Waseem. 2016. Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the 1st Workshop on NLP and Computational Social Science.* 138–142.

[109] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop.* 88–93.

[110] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and Social Risks of Harm from Language Models. arXiv:2112.04359 [cs.CL].

[111] Maximilian Wich, Hala Al Kuwatly, and Georg Groh. 2020. Investigating annotator bias with a graph-based approach. In *Proceedings of the 4th Workshop on Online Abuse and Harms.* 191–199.

[112] Maximilian Wich, Jan Bauer, and Georg Groh. 2020. Impact of politically biased data on hate speech classification.
       In *Proceedings of the 4th Workshop on Online Abuse and Harms*. 54–64.
[113] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: The problem of
       biased datasets. In *Proceedings of the Conference of the North American Chapter of the Association for Computational
       Linguistics: Human Language Technologies*. 602–608.
[114] Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 shared task on the
       identification of offensive language, Austrian Academy of Sciences.
[115] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of
       the 26th International Conference on World Wide Web*. 1391–1399.
[116] Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings
       of the 8th International Workshop on Natural Language Processing for Social Media*. Association for Computational
       Linguistics, 7–14. DOI : https://doi.org/10.18653/v1/2020.socialnlp-1.2
[117] Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: A review on obstacles and
       solutions. *PeerJ Comput. Sci.* 7 (2021), e598.
[118] Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020. Demographics should not
       be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In *Proceedings of
       the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics,
       4134–4145. DOI : https://doi.org/10.18653/v1/2020.acl-main.380
[119] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference
       resolution: Evaluation and debiasing methods. In *Proceedings of the Conference of the North American Chapter of the
       Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics,
       15–20. DOI : https://doi.org/10.18653/v1/N18-2003
[120] Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021. Challenges in automated debi-
       asing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association
       for Computational Linguistics: Main Volume*. Association for Computational Linguistics, 3143–3155. Retrieved from:
       https://www.aclweb.org/anthology/2021.eacl-main.274.