A Comprehensive Taxonomy of Bias Mitigation Methods for Hate Speech Detection

Jan Fillies
Freie Universität Berlin
Berlin, Germany
InfAI
Leipzig, Germany
fillies@infai.org

Marius Wawerek
Freie Universität Berlin
Berlin, Germany

Adrian Paschke
Freie Universität Berlin
Berlin, Germany
InfAI
Leipzig, Germany
Fraunhofer FOKUS
Berlin, Germany

Abstract

Algorithmic hate speech detection is widely used today. However, biases within these systems can lead to discrimination. This research presents an overview of bias mitigation strategies in the field of hate speech detection. The identified principles are grouped into four categories, based on their operation principles. A novel taxonomy of bias mitigation methods is proposed. The mitigation strategies are characterized based on their key concepts and analyzed in terms of their application stage and their need for knowledge of protected attributes. Additionally, the paper discusses potential combinations of these strategies. This research shifts the focus from identifying present biases to examining the similarities and differences between mitigation strategies, thereby facilitating the exchange, stacking, and ensembling of these strategies in future research.

1 Introduction

Hate speech classification plays a crucial role in moderating online discourse, yet existing machine learning (ML) models often exhibit significant bias. These biases can lead to performance degradation (Okpala et al., 2022; Ramponi and Tonelli, 2022), discrimination (Xia et al., 2020; Feldman and Peake, 2021), erosion of trust in automated systems (Geleta, 2023; Qureshi et al., 2023), and even violations of laws and regulations (Wachter et al., 2021; Kennedy et al., 2018). Despite the advancements of hate speech detection methods (Caselli et al., 2021), their uncritical application can further exacerbate harm (Dixon et al., 2018; Arango et al., 2019). Addressing these challenges requires effective bias mitigation strategies.

Current research (e.g. (Garg et al., 2022)) typically classifies mitigation techniques based on specific bias types, making it difficult to identify methods that address multiple biases simultaneously. Additionally, most studies treat bias mitigation as a single-step process (Garg et al., 2022), without considering the complexities involved in combining multiple strategies within an ML pipeline. These limitations complicate the selection and application of effective bias mitigation techniques.

This study introduces a novel perspective by shifting the focus from bias types to mitigation strategies. Instead of asking which methods can mitigate a given bias, the central question is: Which types of bias can a specific method mitigate? This research systematically evaluates mitigation strategies to assess their effectiveness against multiple bias types. This perspective is particularly relevant for ML practitioners who encounter multiple biases within a single pipeline and need holistic solutions. By mapping these techniques to different stages of an ML pipeline, this research provides a more actionable and systematic approach.

This study conducts a structured literature review to systematically analyze existing bias mitigation strategies in hate speech classification. The analysis identifies key characteristics of these strategies.

The key contributions are:

- Reframing bias mitigation by organizing strategies based on the biases they can address.
- Providing a structured taxonomy of mitigation strategies, categorized by principle of operation, requirements of protected attributes, ML pipeline stage, and targeted biases.
- Laying the foundation for analyzing compatibility, by identifying challenges in combining multiple mitigation techniques.

Through adopting a method-centered approach and systematically structuring bias mitigation strategies, this research facilitates more effective and informed bias mitigation in hate speech classification.

Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 1–16

2 Related Literature

Sources of Bias: There are different existing frameworks, some with a more technical approach (van der Wal et al., 2022) and some starting from a more philosophical point of view (Baumann et al., 2023). To bridge the two approaches, this research applies the work of Suresh and Guttag (2021). Their model depicts the entire machine learning life-cycle and divides it into six-steps and a theoretical framework involving data transformations. This process model covers stages from data generation to the final decision-making process supported by the model's classification, divided into two main sections: 'data generation' and 'model building and implementation'. Within these stages, five 'sources of harm' are identified, which correspond to three sources of bias (See Appendix A). Garg et al. (2022) expand upon these 'sources of bias' by introducing the concept of 'targets of harm,' identifying seven types of bias in total—three based on the sources of bias and four on the targets. Mehrabi et al. (2022) offer a comprehensive survey on bias and fairness in machine learning, also building on the foundations laid by Suresh and Guttag (2021). However, they also incorporate an additional framework by Olteanu et al. (2019), which emphasizes social and ethical considerations. Mehrabi et al. (2022) categorize bias into three groups: 'Data to Algorithm,' 'Algorithm to User,' and 'User to Data,' which they use to define more specific types of bias, presenting a total of 19 distinct bias descriptions.

Taxonomy of Bias Mitigation: Garg et al. (2022) offer an overview of various mitigation strategies tailored to each of the bias types they define. They identify general approaches for mitigating specific types of bias and present individual algorithms that fit within these broader strategies. While their work compiles a comprehensive range of methodologies, it does not include a taxonomy of these methods due to its focus on bias types. Kamiran and Calders (2012) provide a general overview of bias mitigation methods specifically applied before model training (preprocessing). Their aim is to reduce the model's reliance on protected attributes by transforming the underlying dataset according to the principles of the preprocessing method used. Further research has surveyed and compared various fairness-aware classification algorithms. Mehrabi et al. (2022) review definitions of fairness and fair machine learning methods across a wide range of applications and problem settings. They list algorithms previously used in specific fairness-aware learning scenarios, noting that these approaches can be categorized by their application stage into 'pre-processing,' 'in-processing,' or 'post-processing,' but they do not provide a formal taxonomy. In contrast, Jones et al. (2020) compare 28 different model pipelines across seven datasets, evaluating them based on both performance and fairness metrics.

Combining Mitigation Methods: Park et al. (2018) examine the impact of three bias mitigation methods on models that exhibit disparities in handling different gender identity terms. Their selection of mitigation algorithms allows them to assess both the individual effectiveness of each method and the combined effects when using multiple methods together. They discover that the most significant improvements in fairness metrics occur when all three mitigation methods are combined. Similarly, Feldman and Peake (2021) propose an 'end-to-end' mitigation framework that integrates three bias mitigation algorithms, each targeting a different stage of the learning pipeline. This 'fusion model' demonstrates strong performance across all test metrics, generally outperforming models that rely on a single debiasing method.

Research contribution: Previous research is focused on two tasks. Firstly, examining individual types of bias. Secondly, which methods have been applied to address bias in the hate speech detection domain. These approaches limit the ability to generalize findings across different types of bias. This research diverges from previous work by focusing on the bias mitigation methods themselves. It provides a comprehensive overview of the existing methods and strategies for mitigating bias. This research address the gaps in current research by creating a taxonomy of bias mitigation methods and further categorizing them based on underlying concepts. This framework will support the development of more effective individual methods or even a combination of methods to combat one or multiple types of bias.

3 Methodology

Through a structured literature research, key bias mitigation concepts were identified as foundation for the taxonomy and further research. This approach is similar to the works of Yin and Zubiaga (2021) and Garg et al. (2022). First, a set of general keywords related to the domain of hate speech and

toxic speech is collected ('task names'). Given the absence of a consistent, operationalized standard and the scientific broadness and ambiguity of existing definitions, previous research may have been categorized differently depending on the author's understanding. Thus a single all-encompassing definition would have limited the scope of the study and excluded relevant literature. Next, a set of keywords related to the specific topic of investigation is defined. For this study these keywords were synonyms for 'bias mitigation' ('mitigation names'). These two sets were then combined into several queries, with the goal of identifying existing studies of mitigation methods for the problem of hate speech detection. Afterwards a second wave of queries was created. Here the names of possible mitigation strategies ('mitigation strategies') were combined with the 'task names' to consider research that did not explicitly aim for bias mitigation. These 'mitigation strategies' keywords were sourced from existing literature, with the goal to extend the literature collection and review current developments in the area. All prepared queries were then handed to Google Scholar as the primary search engine to discover relevant research. An overview of the utilized keywords can be found in Appendix B.

The literature research for bias mitigation studies ended in October 2023, but individual searches for specific methods continued until February 2024. Starting from these sources, citations, and cross-references were utilized to extend the collection of literature. The publications were evaluated according to the journal they were published in, the year of publication, the amount of citations and the relevance of the abstract, introduction and conclusion. If the research passed these initial hurdles, further investigation was undertaken. After this process 83 publications were utilized in this research. The extracted mitigation strategies were combined into a taxonomy of bias mitigation. Additionally both the framework by Suresh and Guttag (2021) and the extracted bias mitigation principles are combined in the work.

4 Framework of Bias

To position the mitigation strategies an explicit, shared understanding of bias needs to be defined.

Suresh and Guttag (2021) define seven potential sources of harm. Theoretically, each source is aligned with a distinct kind of bias. However, in practice these types may not necessarily be mutually exclusive. The bias types proposed by Suresh and Guttag (2021) are: 1. Historical Bias, 2. Representation Bias, 3. Measurement Bias, 4. Aggregation Bias, 5. Learning Bias, 6. Evaluation Bias and 7. Deployment Bias. A brief introduction to each bias type can be found in Appendix C. The relation between bias types and the theoretical representation of the machine learning life cycle (Data Collection, Data Preparation, Model Development, Model Evaluation, Model Postprocessing, Model Deployment) can be seen in Appendix A Figure 2.

5 Strategies Principles of Bias Mitigation

To organize the bias mitigation strategies based on the bias they address, this Section outlines all different types of identified bias mitigation approaches. It differentiates them from each other, based on their conceptual approach, and delineates if the model is model-agnostic (independent of the underlying machine learning model).

A complete overview of all identified mitigation methods, along with examples from applied research, is provided in Appendix D.

5.1 Model Dependent Methods

Prediction Manipulation Prediction manipulation focuses on adjusting the class labels assigned by a model to reduce bias. Instead of directly outputting labels, a classifier typically assigns a probability vector to each sample. Instead of choosing the label with the highest associated probability, different selection algorithms can be used (Pleiss et al., 2017). Depending on the chosen approach, both individual and group fairness can be improved.

Change in Model Optimization Training an ML model is an optimization process guided by a loss function, such as cross-entropy, which penalizes incorrect predictions based on their confidence level. This assumes equal costs for false positives and negatives, which may not suited for all applications. In general, different loss functions and optimizations approaches within the model can be utilized for mitigating biases.

To address fairness, regularization terms can be added to the loss function. Agarwal et al. (2018) integrate fairness constraints, while Ravfogel et al. (2020) reduce bias by targeting word embeddings. Attention mechanisms can also be adjusted to ensure fairer treatment, as shown by Gaci et al. (2022) and Attanasio et al. (2022). These methods enhance

fairness but require re-training when adjustments are made.

Adversarial Debiasing Adversarial Debiasing is a technique used to reduce bias in machine learning models by altering the training process. It combines two tasks: classifying text as toxic or non-toxic and using an adversarial model to predict protected attributes. The goal is to train the model to accurately classify hate speech while preventing it from identifying protected attributes, thereby minimizing bias (Xia et al., 2020; Han et al., 2021).

This method requires data with feature vectors, hate speech labels, and protected attribute labels. The model architecture typically includes a shared encoder, an adversarial model, and a classifier. During training, the information collected by observing the adversarial model can be applied to disrupt any bias the classifier might learn (Xia et al., 2020; Han et al., 2021; Zhang et al., 2018). Despite being resource-intensive, adversarial debiasing is flexible, making it an effective tool for bias mitigation in machine learning models.

Ensemble Models Ensemble Models, or multiple classifier systems (MCS) (Roli et al., 2001), combine predictions from multiple classifiers to improve accuracy and reduce errors. Typically, they use the majority-vote rule, where the most common label is chosen (Kamiran et al., 2018).

Ensemble models boost robustness and fairness by leveraging various model architectures and aggregation methods. Despite their higher resource demands, their flexibility in incorporating different models and strategies makes them effective for bias mitigation and performance improvement (Kamiran et al., 2018; Nascimento et al., 2022).

Explainable AI (XAI) Understanding how machine learning models make decisions is seen as the first step in the mitigation process, therefore XAI is considered a part of the mitigation strategies and included in this research. XAI methods provide insights into decision-making processes, helping in model evaluation, regulatory compliance, and development of mitigation strategies, though they may require additional computational resources (Kuhl et al., 2023; Qureshi et al., 2023).

Attention mechanisms, such as those in models like GPT (Radford et al., 2019), offer insights into decision processes by highlighting which parts of the input are most influential (Lindsay, 2020). Mathew et al. (2021) introduce HateXplain, a dataset with human-annotated rationales for hate speech detection, allowing evaluation of model attention against human reasoning. Qureshi et al. (2023) use feature importance methods to suggest non-offensive alternatives.

5.2 Model Agnostic Methods

Word Manipulation A word significantly correlated with a class label is defined as a 'bias sensitive word' (BSW) (Badjatiya et al., 2019). A classifier can learn unintended relations between the toxicity label and benign words (Dixon et al., 2018). These relations can be effectively combated by different word manipulation strategies. Approaches range from masking via tokens or k-nearest neighbors to placing Named-entity tags (Badjatiya et al., 2019; Ramponi and Tonelli, 2022). Allowing them to mitigate bias in an early step of the machine learning pipeline.

Counterfactuals In contrast to word manipulation, counterfactuals introduce new samples. By either using template structures to insert BSW into new (un-)problematic situations (Dixon et al., 2018) or switching terms against their counterparts (Zhao et al., 2018). One advantage is the ability to create samples of protected groups that may be underrepresented otherwise. However, counterfactuals can also impact the semantics of a sample. If a token is replaced without considering the context, non-realistic samples can be generated. These can be problematic as they may introduce new sources of bias, counteracting the intended goal. Although modern approaches such as the 'Social Group Counterfactuals' developed by Davani et al. (2020) can alleviate this issue.

Synthetic Data Generator models can be used to create new artificial samples that contain both toxic and benign statements about a wide range of protected groups (Yang et al., 2020; Ng et al., 2020; Fanton et al., 2021). Different approaches range from human-in-the-loop to GPT-based Generative Adversary Networks (GAN). The key differentiation from Counterfactuals is that synthetic data can be acquired without relying on, or sampling from, the real data distribution. Therefore, it can also include unseen phenomena not present in the original dataset.

Sampling Based Sampling-based approaches (under- and over-sampling) are used to reduce class imbalance (Elrahman and Abraham, 2014). For the problem of hate speech detection, a possible usage might be equalizing the amount of toxic and non-toxic data samples by either removing or duplicating data points. Different ways of selecting

these samples exist. One possibility are clustering algorithms (Yong, 2012)). Often this class balancing is extended to the distribution of other labels e.g., protected attributes.

Sample Reweighing Reweighing algorithms are used to balance a dataset with regard to a given metric or grouping. Their underlying principle is similar to Sampling-based approaches. Sample Reweighing is based on the idea of converging the observed probability in the data set with the expected probability distribution at application time. However unlike sampling approaches, reweighing does not need to explicitly duplicate or remove any samples from the dataset. Instead, each sample is concatenated with an assigned weight. This weight determines the strength at which (mis-)classifications for this sample are factored into the loss function (penalty) of the classifier. For example Zhao et al. (2023) propose an adversarial reweighting method guided by the Wasserstein distance.

Annotation Manipulation Hate speech detection often relies on supervised learning, where datasets include feature vectors and annotated class labels. The annotation process can introduce biases, but various strategies exist to mitigate them. One approach, proposed by Li et al. (2023), is Decoupled Confident Learning (DeCoLe), which prunes samples with potentially inaccurate labels. Another method involves relabeling, also known as "massaging", which corrects erroneous labels rather than removing them, as suggested by Kamiran and Calders (2012). In addition, the general format of the annotation can be changed. Garg et al. (2022) argue that there is a need to incorporate disagreement into the labeling process, e.g. via multiple labels for the same sample. These techniques focus on improving data quality, though altering labels can influence which models are suitable for the task.

Addition of External Information Datasets provide a snapshot of data at a specific time, which can become outdated as language and platforms evolve (Garg et al., 2022). This can affect hate speech detection models, leading to performance issues when applied in real-world contexts (Ramponi and Tonelli, 2022). To improve robustness, integrating additional or diverse data is beneficial. Antypas and Camacho-Collados (2023) found that combining multiple datasets enhances model generalization. Dixon et al. (2018) used Wikipedia data for balancing the datasets, and Park et al. (2018) employed transfer learning to reduce biases. Sheth et al. (2023b) developed the PEACE framework, using sentiment and aggression cues to enhance model performance across platforms.

Overall, adding diverse data or meta-information can improve hate speech detection, while managing privacy and regulatory concerns through data sheets and statements (Mehrabi et al., 2022; Ramponi and Tonelli, 2022).

6 Taxonomy of Bias Mitigation

6.1 Proposed Taxonomy of Bias Mitigation Methods

As the second research contribution and based on current literature, all identified strategies for bias mitigation displayed in Section 5, can be categorized by four general aspects: 1. The underlying principle of operation, 2. the application stage, 3. the requirements of protected attributes and 4. the bias type mitigated. This categorization is visualized in Figure 1. The proposed taxonomy divides bias mitigation methods into four *conceptual groups* of operation principles: text-oriented, sample-oriented, model-oriented, and meta information-oriented.

Each category plays a unique role in bias mitigation, offering a new 'perspective' due to a different focus point. An overview containing all in literature identified methods for each group can be found in Appendix E.

Text-oriented methods focus on modifying the content of text samples in the dataset and are model-agnostic. They include text manipulation, counterfactuals, and synthetic data methods. Word manipulation changes individual words, counterfactuals replace identity terms with alternative tokens, and synthetic data methods generate new samples to enhance dataset diversity.

Sample-oriented methods abstract from text content and focus on dataset attributes such as sample distribution. These include sampling-based approaches, sample reweighing, and annotation manipulation. Sampling-based methods adjust the dataset by over- or under-sampling, sample reweighing alters the impact of individual samples, and annotation manipulation changes sample labels to reduce bias.

Model-oriented methods target the classifier itself, including changes in model optimization, adversarial debiasing, ensemble models, and prediction manipulation. These methods are model-

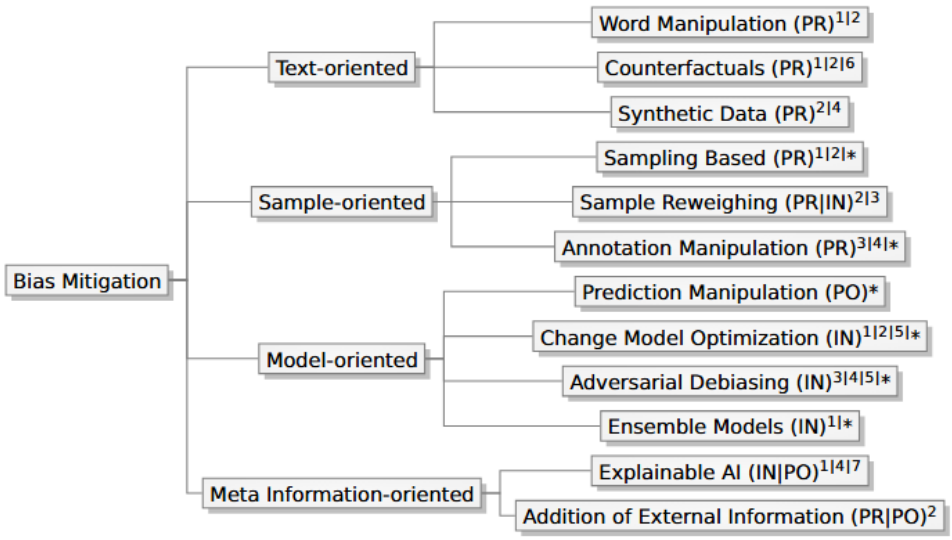

Figure 1: Taxonomy of bias mitigation methods based on their principle of operation. Each symbol marks a class of processing PR = Pre-processing, IN = In-Processing, PO = Post-Processing, Historical Bias = [1], Representation Bias = [2], Measurement Bias = [3], Aggregation Bias = [4], Learning Bias = [5], Evaluation Bias = [6], Deployment Bias = [7], Requires knowledge about the protected attribute = *

specific and can affect both training and application times, depending on the classifier's architecture and the nature of the adjustments.

Meta information-oriented methods are the most abstract, focusing on information about the ML pipeline and its development. This category includes explainable AI (XAI) and the addition of external information. XAI analyzes decision influences, while additional information aims to diversify datasets by incorporating external sources or metadata.

6.2 Application Stages of Bias Mitigation

Bias mitigation strategies need to be applied at a certain point in the machine learning lifecycle. Suresh and Guttag (2021) utilize a six-stage model to represent all details of this lifecycle. While useful for a fine-grained analysis this model is not well-suited for an intuitive understanding, especially from non-domain experts. Thus for this Section the focus will be on the common three-stage model categorizing approaches into either 'pre-, in- or post-processing'. In Appendix F a mapping of the mitigation strategies onto the six-stage framework by Suresh and Guttag (2021) can be found.

Figure 1 illustrates the distribution of mitigation methods across 'pre-, in- or post-processing'. Each stage has multiple methods associated with it, indicating that all parts of the ML pipeline can be useful to combat bias. Nonetheless debiasing methods are not uniformly distributed across stages. The least amount of mitigation strategies can be viewed as post-processing, with only three belonging into this category. Both pre-processing and in-processing are more prevalent with seven and five members respectively. Referencing this distribution with the conceptual groups of operation principles, introduced in Section 6.1, a trend can be seen. Each group has a majority category that almost all members belong to. Both 'Text-oriented' and 'Sample-oriented' methods concentrate on pre-processing. This aligns with their focus on the dataset, either its content or the abstracted samples. Sample reweighing is the only concept that deviates from this trend, by belonging to both pre- and in-processing.

The 'Model-oriented' group as the name suggests focuses on the model. The main concepts are either the training procedure or the underlying model architecture. This places them into the in-processing stage, as they either directly or indirectly change the way how model training occurs. An exemption in the 'Model-oriented' group is prediction manipulation. As the idea is to relabel the predictions after they were made by the model, prediction manipulation is applied during the post-processing stage. The last group consists of 'Meta information-oriented' methods, which includes explainable AI and the addition of external information. Both consist of applications that are part of post-processing, however in addition they also belong to another stage.

In total, it can be observed that, while unevenly split, mitigation methods can be utilized at all stages of the ML pipeline. Additionally, the conceptual groups provided by the proposed taxonomy from Section 6.1 offers a good intuition at which stage a method may intercept. This can reduce the workload when designing multi-stage bias mitigation interventions. As instead of analyzing the details of a mitigation strategy knowing the rough focus of a method allows placing them within the pipeline. 'Text-oriented' and 'Sample-oriented' methods intervene during pre-processing. 'Model-oriented' methods mainly target model development. 'Meta information-oriented' strategies cover at least post-processing and one additional stage, reflecting their need for broader interaction.

6.3 Requirements for Protected Attributes

Protected attributes are all attributes that should not be utilized for the prediction of a sample (Morse et al., 2022). They are considered potentially sensitive, in the sense of personal identification, discrimination and data protection. Examples for protected attributes are age, skin color, or religious orientation (Morse et al., 2022). Not all datasets include protected attributes, and recording them in general raises privacy issues (Wachter et al., 2021). There are also concerns about 'reverse discrimination' (Kamiran and Calders, 2012; Kamiran et al., 2018) and legal challenges, as regulations may hold classifiers accountable for using protected attributes (Margot E. Kaminski, 2021; Wachter et al., 2021). It has to be noted that all strategies work with the protected attributes provided. But only six strategies are applicable without information about the protected attributes. The six strategies are: Text Manipulation, Counterfactuals, Synthetic Data, Sample Reweighing, Explainable AI and Addition of Information. The strategies that require knowledge about the protected attribute are marked with "*" in Figure 1. Overall, half of the mitigation principles require sensitive information during training. Despite this, there are viable strategies to address bias without such data.

6.4 Biases targeted by Mitigation Strategies

A comprehensive overview of the existing research applications for each mitigation strategy and which biases from Suresh and Guttag (2021) they have addressed can be found in Table 1. They refer to biases as distinct sources of harm in an ML system. It can be seen that there is a difference in the attention that different bias types have received from researchers. While representation bias has been the focus of seven existing hate speech publications, both evaluation and deployment bias have only been combated once in this research field. These insights suggest that current research may be narrow in focus.

7 On Combining Mitigation Methods

Bias mitigation is often treated as a single-step intervention, where methods are applied in isolation (Garg et al., 2022). While this simplifies implementation, it limits effectiveness, as no single method optimally balances fairness and performance across all bias types. Research suggests that combining strategies can improve fairness (Feldman and Peake, 2021; Park et al., 2018), yet there is little systematic guidance on how to structure such combinations. The challenge lies in understanding interaction effects: some methods reinforce each other, while others may act independently or even neutralize each other. The order of application plays a crucial role. For example, assume that counterfactuals are first created by replacing words with their opposite-gender counterparts. If additional word manipulation methods are then applied to replace all gendered words with a token word, the resulting dataset will contain duplicate sentences, one from the original sample after tokenization and another from its counterfactual version, which is tokenized into the exact same output.

It is important to understand, how different bias mitigation techniques interact, both within and across ML pipeline stages, to identify effective multi-stage interventions. Another challenge is evaluating combined approaches: many fairness metrics only assess individual methods rather than their collective impact (Park et al., 2018; Feldman and Peake, 2021). Developing benchmarks that quantify trade-offs between performance and fairness in multi-method settings is a crucial step forward. Additionally, practical concerns, such as computational efficiency and deployment constraints, remain underexplored.

To address these challenges, the introduced conceptual groups help organize bias mitigation by clarifying their roles within the ML pipeline. This structured approach enables more effective multi-method interventions, reducing conflicts and improving scalability.

Bias Type	Mitigation Method
Historical Bias	Text Manipulation (Badjatiya et al., 2019; Ramponi and Tonelli, 2022); Counterfactual (Davani et al., 2020); Sampling Based (Ball-Burack et al., 2021); Model Optimization (Park et al., 2018; Kennedy et al., 2020a; Gaci et al., 2022; Cai et al., 2022); Ensemble Models (Nascimento et al., 2022); Explainable AI (Attanasio et al., 2022; Mathew et al., 2021; Pereira-Kohatsu et al., 2019; Qureshi et al., 2023)
Representation Bias	Text Manipulation (Badjatiya et al., 2019); Counterfactual (Davani et al., 2020; Park et al., 2018; Dixon et al., 2018); Synthetic Data (Hartvigsen et al., 2022; Ocampo et al., 2023); Sampling Based (Ball-Burack et al., 2021); Sample Reweighing (Mozafari et al., 2020); Model Optimization (Cai et al., 2022); Addition of external Information (Dixon et al., 2018; Park et al., 2018)
Measurement Bias	Sample Reweighing (Mozafari et al., 2020); Annotation Manipulation (Li et al., 2023);Adversarial debiasing (Okpala et al., 2022)
Aggregation Bias	Synthetic Data (Yang et al., 2020); Annotation Manipulation (Li et al., 2023); Adversarial debiasing (Okpala et al., 2022; Xia et al., 2020); Explainable AI (Attanasio et al., 2022; Mathew et al., 2021; Pereira-Kohatsu et al., 2019; Qureshi et al., 2023; Geleta, 2023)
Learning Bias	Model Optimization (Chen et al., 2023); Adversarial debiasing (Okpala et al., 2022)
Evaluation Bias	Counterfactual (Dixon et al., 2018)
Deployment Bias	Explainable AI (Geleta, 2023; Qureshi et al., 2023; Attanasio et al., 2022; Pereira-Kohatsu et al., 2019)

Table 1: Bias mitigation strategies categorized by targeted bias type in historical applications. This research mapped the applied strategies from cited resources to specific bias types.

7.1 Example: Structuring Multi-Method Mitigation with the Taxonomy.

As a concrete example, Park et al. (2018) employed three different methods to successfully mitigate gender bias in toxic comment classification: counterfactual data augmentation, adversarial debiasing, and a change in model optimization through fine-tuning. While these methods proved effective, they were selected heuristically—without the benefit of structured guidance for expanding or systematically organizing the mitigation pipeline.

Using the proposed taxonomy, additional strategies could be integrated systematically. For example, a sample-oriented method such as sample reweighing could complement counterfactual data without altering textual content. Similarly, a meta-information-oriented technique like fairness-aware methods could be added post-hoc to audit residual bias, without modifying the model architecture or training regime.

This illustrates how the taxonomy enables modular, non-disruptive extensions to bias mitigation workflows by clarifying method roles and interactions across the ML pipeline.

8 Discussion

This research categorized the identified bias mitigation methods based on their intended target within the ML pipeline. It defined four principle groups: text-oriented, sample-oriented, model-oriented, and meta information-oriented methods. The classification is based on the operational similarities of the methods within the ML workflow. While it is possible that other classes of mitigation strategies cannot be represented within this grouping, all identified strategies presented in the research could be positioned here. It would be possible to introduce different groupings based on other aspects, but especially with the goal of making the strategies exchangeable or stackable, this grouping provides the best identified option.

This classification also highlights an uneven distribution of methods across ML pipeline stages, with more methods focused on pre-processing and in-processing, and fewer applied during post-processing. This might be due to the field of research being relatively new and pre-processing being the first step, as it is often a necessity to pre-process the datasets before training an algorithm. Therefore, pre-processing might also be the first step to be scrutinized by research.

9 Conclusion and Future Work

This research investigates bias mitigation methods, specifically for hate speech detection. All identified mitigation strategies were presented and organized based on their principles of operation into a newly developed taxonomy for bias mitigation, categorizing them into four 'conceptual groups' of operation principles: 'text-oriented,' 'sample-oriented,' 'model-oriented,' and 'meta-information-oriented.'

This research shifts the focus away from the individual biases present in a system and onto the available mitigation strategies, creating a comprehensive overview of existing strategies, introducing a novel grouping and taxonomy, and highlighting which biases and stages these strategies cover in current research. It identifies methods that require protected attributes to function and discusses factors influencing the combination of various mitigation methods. Therefore, making it easier for future research to understand which methods exist, function similarly and which could bear potential for replacing, stacking, or combining.

In the next step, exploring how to best combine these different methods is a promising avenue for further research. Future research could also consider adaptive bias mitigation frameworks that dynamically adjust mitigation strategies based on dataset properties. Additionally, integrating new mitigation strategies from adjacent fields could enhance the effectiveness of bias reduction techniques.

10 Limitations

As keyword-based searches on Google Scholar were utilized, publications not linked on the search engine or publications with wrong keyword tags were potentially excluded from the sources of information. Additionally, no complete list of all possible bias mitigation methods and strategies can exist. Especially, as research on bias mitigation is steadily growing. Another limitation of any modern fairness-aware research is the missing existence of definitions for fairness, hate speech and bias types that are shared across fields and researchers. While certain definitions have been utilized by multiple publications, no singular understanding exists. As a consequence, previous research may have been misinterpreted in this research. Similarly, no reproducibility experiments were done as part of this research. Results and findings by other researchers were assumed to be obtained in scientifically sound and valid ways. In addition, the findings proposed by this research were not experimentally confirmed. Finally, the research presented here is only applicable to the field of hate speech detection, which is itself a subarea in the field of natural language processing, which could contain more mitigation approaches not covered here.

11 Ethical Considerations

The research centers on societal interests, with a focus on the public good. The mitigation of bias in algorithmic detection of hate speech is essential to foster a harm-free environment, especially for minority groups requiring protection. Mitigating biases within datasets, labels, algorithms, trained classifiers, and predictions will aid in achieving this goal in the future. The research aims to contribute to a more diverse understanding of what constitutes mitigation of bias in hate speech detection. Potential limitations are outlined in Section 10. The research advocates for more bias mitigation in the machine learning lifecycle of hate speech classification in a real world setting. This will not just protect the target of hate more efficiently, but also minimize unjustified restrictions on freedom of speech.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. *arXiv preprint*.

Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust Hate Speech Detection in Social Media: A Cross-Dataset Empirical Evaluation.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate Speech Detection is Not as Easy as You May Think. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA. ACM.

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based Attention Regularization Frees Unintended Bias Mitigation from Lists.

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical Bias Removal for Hate Speech Detection Task using Knowledge-based Generalizations. In *The World Wide Web Conference*, pages 49–59, New York, NY, USA. ACM.

Ari Ball-Burack, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2021. Differential Tweetment:

Mitigating Racial Dialect Bias in Harmful Tweet Detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 116–128, New York, NY, USA. ACM.

Joachim Baumann, Alessandro Castelnovo, Riccardo Crupi, Nicole Inverardi, and Daniele Regoli. 2023. Bias on Demand: A Modelling Framework That Generates Synthetic Data With Bias. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1002–1013, New York, NY, USA. ACM.

Yi Cai, Arthur Zimek, Gerhard Wunder, and Eirini Ntoutsi. 2022. Power of Explanations: Towards automatic debiasing in hate speech detection.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. *HateBERT: Retraining BERT for Abusive Language Detection in English*. Association for Computational Linguistics.

Shijing Chen, Usman Naseem, and Imran Razzak. 2023. Debunking Biases in Attention. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 141–150, Toronto, Canada. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Ali Omrani, Brendan Kennedy, Mohammad Atari, Xiang Ren, and Morteza Dehghani. 2020. Fair Hate Speech Detection through Evaluation of Social Group Counterfactuals.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, New York, NY, USA. ACM.

Shaza Elrahman and Ajith Abraham. 2014. A review of class imbalance problem. *Journal of Network and Innovative Computing*, 1:332–340.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech.

Tal Feldman and Ashley Peake. 2021. End-To-End Bias Mitigation: Removing Gender Bias in Deep Learning. *arXiv preprint*.

Yacine Gaci, Boualem Benatallah, Fabio Casati, and Khalid Benabdeslem. 2022. Debiasing Pretrained Text Encoders by Paying Attention to Paying Attention. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9582–9602, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2022. Handling Bias in Toxic Speech Detection: A Survey.

Raisa Romanov Geleta. 2023. *Exploring the Role of AI and XAI in Hate Speech Detection on Social Media: A Study on User Trust*. Master Thesis, Johannes Kepler University Linz, Austria.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021. Diverse Adversaries for Mitigating Bias in Training.

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.

Gareth P. Jones, James M. Hickey, Pietro G. Di Stefano, Charanpal Dhanjal, Laura C. Stoddart, and Vlasios Vasileiou. 2020. *Metrics and methods for a systematic comparison of fairness-aware machine learning algorithms*. arXiv.

Przemyslaw Joniak and Akiko Aizawa. 2022. Gender Biases and Where to Find Them: Exploring Gender Bias in Pre-Trained Transformer-based Language Models Using Movement Pruning.

Ratnesh Kumar Joshi, Arindam Chatterjee, and Asif Ekbal. 2023. Saliency Guided Debiasing: Detecting and mitigating biases in LMs using feature attribution. *Neurocomputing*, page 126851.

Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33.

Faisal Kamiran, Sameen Mansha, Asim Karim, and Xiangliang Zhang. 2018. Exploiting reject option in classification for social discrimination control. *Information Sciences*, 425:18–33.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joseph Hoover, Aida Azatian, Alyzeh Hussain, Austin Lara, gabriel olmos, Adam Radwan Omary, Christina Park, Clarisa Wijaya, Xin Wang, Yong Zhang, and Morteza Dehghani. 2018. *Introducing the Gab Hate Corpus: Defining and applying hate-based rhetoric to social media posts at scale*. Language Resources and Evaluation.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020a. *Contextualizing Hate Speech Classifiers with Post-hoc Explanation*. Association for Computational Linguistics.

Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020b. *Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application*. arXiv.

Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-aware Classification. In

Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18, pages 853–862, New York, New York, USA. ACM Press.

Ulrike Kuhl, André Artelt, and Barbara Hammer. 2023. *For Better or Worse: The Impact of Counterfactual Explanations' Directionality on User Behavior in xAI*, volume 1903. Springer, Cham.

Yunyi Li, Maria De-Arteaga, and Maytal Saar-Tsechansky. 2023. Mitigating Label Bias via Decoupled Confident Learning.

Grace W. Lindsay. 2020. Attention in Psychology, Neuroscience, and Machine Learning. *Frontiers in Computational Neuroscience*, 14:29.

Pranay K. Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R. Varshney, and Ruchir Puri. 2019. Bias Mitigation Post-processing for Individual and Group Fairness. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2847–2851. IEEE.

Margot E. Kaminski. 2021. The right to explanation, explained. In Sharon K. Sandeen, Christoph W. Rademacher, and Ansgar Ohly, editors, *Research handbook on information law and governance*, Research handbooks in information law series, pages 278–299. Edward Elgar Publishing Limited, Cheltenham, UK and Northampton, Massachusetts.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):1–35.

Lily Morse, Mike Horia M. Teodorescu, Yazeed Awwad, and Gerald C. Kane. 2022. Do the ends justify the means? variation in the distributive and procedural fairness of machine learning algorithms. *Journal of Business Ethics*, 181(4):1083–1095.

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):e0237861.

Francimaria R.S. Nascimento, George D.C. Cavalcanti, and Márjory Da Costa-Abreu. 2022. Unintended bias evaluation: An analysis of hate speech detection and gender bias mitigation on social media using ensemble learning. *Expert Systems with Applications*, 201:117032.

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. 2020. SSMBA: Self-Supervised Manifold Based Data Augmentation for Improving Out-of-Domain Robustness. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1268–1283.

Nicolas Ocampo, Elena Cabrio, and Serena Villata. 2023. Playing the Part of the Sharp Bully: Generating Adversarial Examples for Implicit Hate Speech Detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2758–2772, Toronto, Canada. Association for Computational Linguistics.

Ebuka Okpala, Long Cheng, Nicodemus Mbwambo, and Feng Luo. 2022. AAEBERT: Debiasing BERT-based Hate Speech Detection Models via Adversarial Learning. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, 2:13.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. *Reducing Gender Bias in Abusive Language Detection*. Association for Computational Linguistics.

Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. 2019. Detecting and Monitoring Hate Speech in Twitter. *Sensors (Basel, Switzerland)*, 19(21).

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On Fairness and Calibration. *arXiv preprint*.

Muhammad Deedahwar Mazhar Qureshi, M. Atif Qureshi, and Wael Rashwan. 2023. Toward Inclusive Online Environments: Counterfactual-Inspired XAI for Detecting and Interpreting Hateful and Offensive Tweets. In *Explainable Artificial Intelligence*, Communications in Computer and Information Science, pages 97–119, Cham. Springer Nature Switzerland and Imprint Springer.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. *Language Models are Unsupervised Multitask Learners*. OpenAI.

Alan Ramponi and Sara Tonelli. 2022. Features or Spurious Artifacts? Data-centric Baselines for Fair and Robust Hate Speech Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3027–3040. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. *arXiv preprint*.

Fabio Roli, Giorgio Giacinto, and Gianni Vernazza. 2001. Methods for Designing Multiple Classifier Systems. In *Multiple classifier systems*, Lecture Notes in Computer Science, pages 78–87, Berlin. Springer.

Paras Sheth, Tharindu Kumarage, Raha Moraffah, Aman Chadha, and Huan Liu. 2023a. Causality Guided Disentanglement for Cross-Platform Hate Speech Detection.

Paras Sheth, Tharindu Kumarage, Raha Moraffah, Aman Chadha, and Huan Liu. 2023b. PEACE: Cross-Platform Hate Speech Detection- A Causality-guided Framework.

Harini Suresh and John V. Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *EAAMO '21: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 32:1–9.

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. *arXiv preprint*.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law. *SSRN Electronic Journal*.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting Racial Bias in Hate Speech Detection.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative Data Augmentation for Commonsense Reasoning. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1008–1025.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions.

Yang Yong. 2012. The Research of Imbalanced Data Set of Sample Sampling Method Based on K-Means Cluster and Genetic Algorithm. *Energy Procedia*, 17:164–170.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA. ACM.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2*, pages 15–20.

Xuan Zhao, Simone Fabbrizzi, Paula Reyero Lobo, Siamak Ghodsi, Klaus Broelemann, Steffen Staab, and Gjergji Kasneci. 2023. *Adversarial Reweighting Guided by Wasserstein Distance for Bias Mitigation*. arXiv.

A Framework of Bias

Figure 2 displays the framework of bias from Suresh and Guttag (2021).

B Keywords

The 'task names' keywords were: {'hate speech detection', 'abusive language detection', 'offensive language detection', 'toxic speech detection'}.

The keywords for 'mitigation names' were: {'bias mitigation', 'debiasing', 'combat bias', 'fair classification', 'fairness-aware classification', 'harm mitigation', 'removing bias', 'prevent bias', 'handling bias'}.

The keywords for 'mitigation strategies' were: {'Text Removal', 'Masking','Word Generalization','Filtering','Word Replacements','Token Generalization','Counterfactuals','Template Test Set','Data Augmentation','Data Creation','Artificial Data', 'Synthetic Data','Synthetic Samples','Preferential Sampling', 'Sampling', 'Sample Reweighing', 'Sample Pruning', 'Relabeling', 'Annotation Uncertainty Modeling', 'Label Uncertainty', 'Prediction Manipulation', 'Fair Training Metric', 'Debiasing Word Embeddings','Attention Regularization', 'Model Pruning', 'Transfer Learning', 'Model Pretraining', 'Transformer-based Debiasing', 'Adversarial Debiasing', 'Ensemble Models', 'Explainable AI', 'Understandable AI', 'Explainable Machine Learning', 'Counterfactual Explanations'}.

C Introduction to the different Bias Types

The bias types proposed by Suresh and Guttag (2021) are:

Historical Bias: This bias arises from pre-existing societal and historical inequalities that shape data before collection begins. Even with accurate measurements, marginalized groups may face disadvantages due to systemic disparities.

Representation Bias: This occurs when the sampled population fails to accurately reflect the real-world application population. Underrepresentation of certain groups can lead to reduced model robustness and fairness.

Measurement Bias: This bias stems from inconsistencies in feature and label definitions during

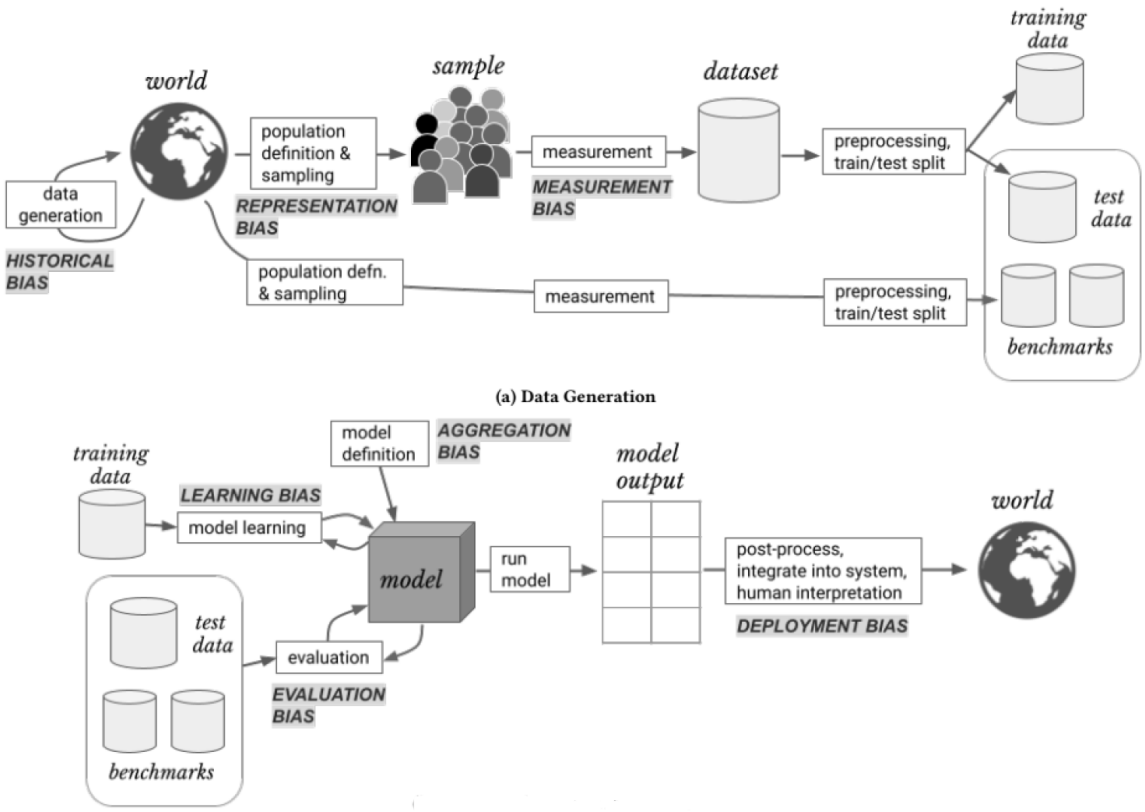

Figure 2: Sources of Harms and their related Types of Bias. This diagram is copied from Suresh and Guttag (2021).

data collection. Variability in human judgment (e.g. annotator bias), guidelines, or external factors can lead to systematic errors affecting model generalization.

Aggregation Bias: This bias arises when a model assumes that all data points follow the same input-label relationship, disregarding different data origins and subgroup-specific differences. Variations in language, context, or cultural background can lead to systematic misrepresentations in the model.

Learning Bias: Arising from choices made during model training, this bias reflects disparities in performance across different groups. The optimization of one metric may unintentionally compromise fairness, privacy, or other critical objectives.

Evaluation Bias: This occurs when the benchmark data or metrics used for assessing a model fail to capture real-world variations. A model may perform well in testing but struggle with new data.

Deployment Bias: This bias emerges when a model is used in real-world environments in ways that differ from its intended design. Human and institutional interactions can distort its application, leading to unintended consequences.

D Sources for Mitigation Strategies

For each identified bias mitigation strategy the concrete methods with their sources in literature are displayed.

Text Manipulation: Removal (Ramponi and Tonelli, 2022)], Masking (Ramponi and Tonelli, 2022), Word Generalization (Badjatiya et al., 2019), Filtering (Ramponi and Tonelli, 2022).

Counterfactuals: Word Replacements (Park et al., 2018; Davani et al., 2020; Joshi et al., 2023), Template Test Set (Dixon et al., 2018).

Synthetic Data: Data Augmentation (Ng et al., 2020; Yang et al., 2020), Artificial Data Creation (Hartvigsen et al., 2022; Ocampo et al., 2023; Fanton et al., 2021).

Sampling Based: Preferential Sampling (Kamiran and Calders, 2012; Ball-Burack et al., 2021).

Sample Reweighing: Uniform Weights (Kamiran and Calders, 2012; Krasanakis et al., 2018), Individual Weights(Zhao et al., 2023).

Annotation Manipulation: Pruning Inaccurate Samples(Li et al., 2023), Relabeling(Kamiran and Calders, 2012), Uncertainty Modeling(Garg et al., 2022; Kennedy et al., 2020b; Sheth et al., 2023a).

Prediction Manipulation: Group Fairness

(Pleiss et al., 2017), Individual Fairness (Lohia et al., 2019).

Change Model Optimization: Training Metric (Agarwal et al., 2018; Garg et al., 2022; Kennedy et al., 2020a), Word Embeddings (Ravfogel et al., 2020; Park et al., 2018), Attention Regularization (Gaci et al., 2022; Attanasio et al., 2022; Cai et al., 2022), Movement Pruning (Joniak and Aizawa, 2022), Transfer Learning (Park et al., 2018).

Adversarial Debiasing: Internal (Xia et al., 2020), External (Okpala et al., 2022), Multiple (Han et al., 2021).

Ensemble Models: Reject Option Classifaction (Kamiran et al., 2018), Various Experts (Nascimento et al., 2022).

Explainable AI: Attention Highlighting (Attanasio et al., 2022; Mathew et al., 2021), Monitoring (Pereira-Kohatsu et al., 2019), Counterfactual Explanations (Qureshi et al., 2023; Kuhl et al., 2023).

Addition of External Information: Data Augmentation (Dixon et al., 2018; Park et al., 2018), Training Datasets (Antypas and Camacho-Collados, 2023), Related Cues (Sheth et al., 2023b).

E Taxonomy with Approaches and Sources

Figure 3 displays the taxonomy further enriched with the approaches from the literature. An easier accessible version can be found on GitHub[1]

F Mitigation Methods in the Six Stage Bias Model

Table 2 displays the mitigation approaches and their location within the six stage bias model from (Suresh and Guttag, 2021).

[1]https://github.com/fillies/BiasMitigationTaxonomy

Mitigation method	Da.Col.	Da.Proc.	Mo.Dev.	Mo.Eval.	Mo.Post.	Mo.Dep.
Text Manipulation		x				
Counterfactuals		x		x		
Synthetic Data	x	x				
Sampling based		x				
Sample Reweighing		x	x			
Annotation Manipulation	x	x			x	
Prediction Manipulation					x	
Change Model Optimization			x	x		
Adversarial Debiasing			x	x		
Ensemble Models			x		x	
Explainable AI			x	x		x
Addition of Information	x	x		x		x
Total:	3	7	5	5	3	2

Table 2: Bias Mitigation Principles categorized into the Six Stage Model by (Suresh and Guttag, 2021). Each column represents a stage. Da.Col = Data Collection. Da.Proc. = Data Processing. Mo.Dev. = Model Development. Mo.Eval. = Model Evaluation. Mo.Post. = Model Post-Processing. Mo.Dep. = Model Deployment. A horizontal line delineates the four different conceptual groups introduced in this research.

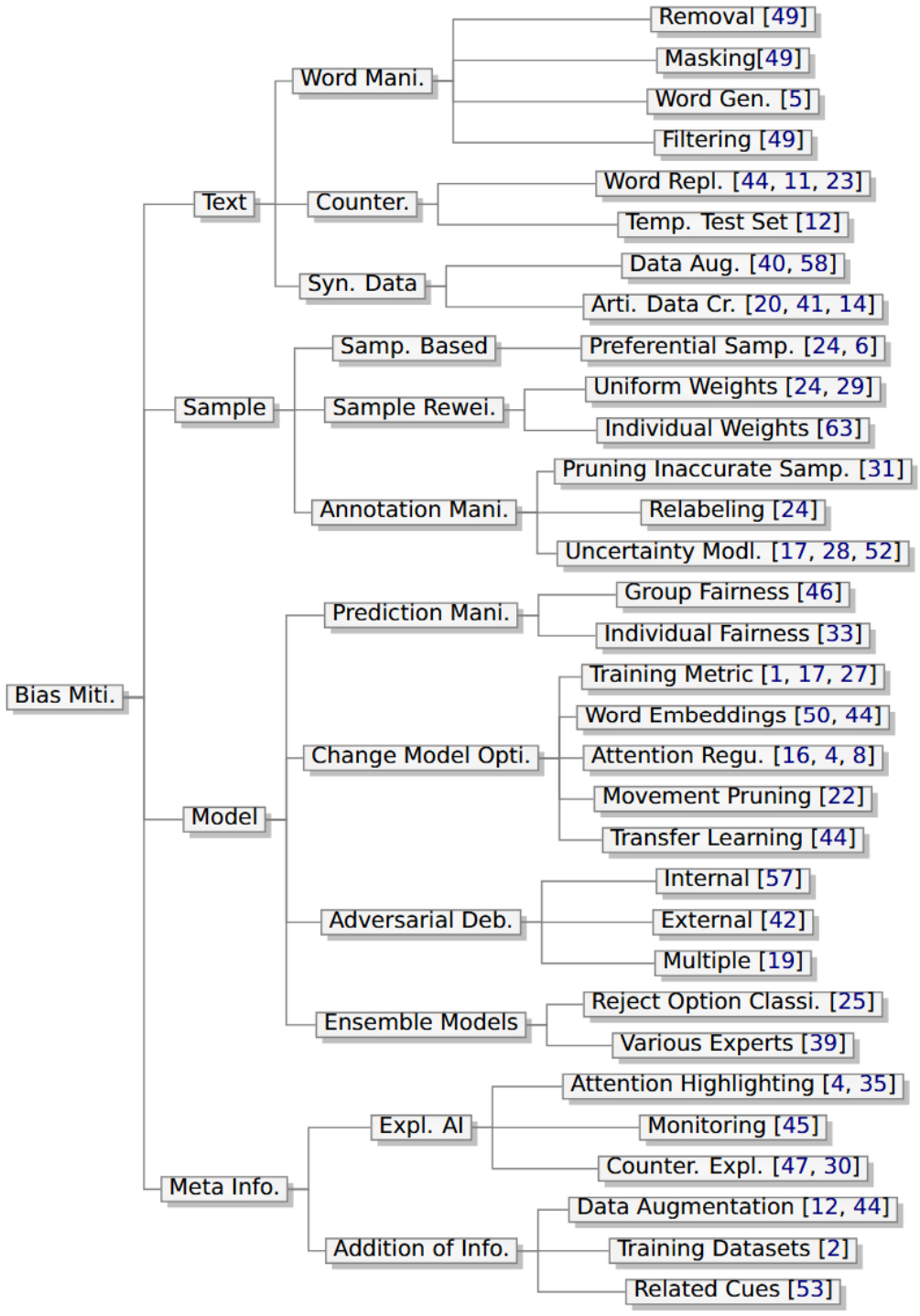

Figure 3: Classes broken down in concrete methods for bias mitigation with the corresponding citations. Miti. = Mitigation, Mani. = Manipulation, Gen. = Generalization, Counter. = Counterfactuals, Repl. = Replacements, Temp. = Template, Syn. = Synthetic, Aug. = Augmentation, Arti. = Artificial, Cr. = Creation, Samp. = Sampling, Rewei. = Reweighing, Opti. = Optimization, Deb = Debiasing, Model. = Modeling, Expl. = Explainable, Regu. = Regularization, Classi. = Classifaction