# Social Fairness in Semi-Supervised Toxicity Text Classification

by

Shahriar Shayesteh

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for the
Master of Computer Science Concentration on Applied AI
in
School of Electrical Engineering and Computer Science

**Abstract**

The rapid growth of user-generated content on social media platforms in the form of text caused moderating toxic language manually to become an increasingly challenging task. Consequently, researchers have turned to artificial intelligence (AI) and machine learning (ML) models to detect and classify toxic comments automatically. However, these models often exhibit unintended bias against comments containing sensitive terms related to demographic groups, such as race and gender, which leads to unfair classifications of samples. In addition, most existing research on this topic focuses on fully supervised learning frameworks. Therefore, there is a growing need to explore fairness in semi-supervised toxicity detection due to the difficulty of annotating large amounts of data. In this thesis, we aim to address this gap by developing a fair generative-based semi-supervised framework for mitigating social bias in toxicity text classification. This framework consists of two parts, first, we trained a semi-supervised generative-based text classification model on the benchmark toxicity datasets. Then, in the second step, we mitigated social bias in the trained classifier in step 1 using adversarial debiasing, to improve fairness. In this work, we use two different semi-supervised generative-based text classification models, NDAGAN and GANBERT (the difference between them is that the former adds negative data augmentation to address some of the problems in GANBERT), to propose two fair semi-supervised models called FairNDAGAN and FairGANBERT. Finally, we compare the performance of the proposed fair semi-supervised models in terms of accuracy and fairness (equalized odds difference) against baselines to clarify the challenges of social fairness in semi-supervised toxicity text classification for the first time.

Based on the experimental results, the key contributions of this research are: first, we propose a novel fair semi-supervised generative-based framework for fair toxicity text classification for the first time. Second, we show that we can achieve fairness in semi-supervised toxicity text classification without considerable loss of accuracy. Third, we demonstrate that achieving fairness at the coarse-grained level improves fairness at the fine-grained level but does not always guarantee it. Fourth, we justify the impact of the labeled and unlabeled data in terms of fairness and accuracy in the studied semi-supervised framework. Finally, we demonstrate the susceptibility of the supervised and semi-supervised models against data imbalance in terms of accuracy and fairness.

**Acknowledgements**

I would like to express my heartfelt gratitude to Professor Diana Inkpen for her unwavering support and guidance throughout my research journey. She has been more than a university mentor to me; she has had a profound impact on my life and personality. Her wisdom, encouragement, and mentorship have shaped not only my academic growth but also my personal development. I am truly grateful for the opportunities she provided me to grow intellectually, as well as the valuable life lessons she imparted along the way. Thank you, Professor Inkpen, for being an inspiring role model and for contributing to my overall growth and success.

Additionally, I would like to extend my deepest appreciation to my parents and sister for their constant support and assistance in every stage of my life. Their unwavering belief in my abilities and their unconditional love have been the driving force behind my accomplishments. I am forever grateful for their encouragement, understanding, and sacrifices they have made to help me succeed.

# Table of Contents

[1] Focusing on bias issues and fairness

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Motivation

In recent years, social media has become a part of our lives, with platforms such as Twitter, Facebook, and Gab enabling people to interact and share information through text-based platforms. While these platforms facilitate communication, they also present challenges in terms of moderating user-generated content. The massive volume of posts generated each second makes manual moderation of toxic content [1] an impossible task. This lack of proper moderation can negatively impact online communities, making them less inclusive and nurturing hostility. As a result, researchers have turned to machine learning-based approaches to identify toxic content in social media comments automatically. However, many existing machine learning models exhibit unintended bias against comments containing sensitive terms related to demographic groups, such as race and gender, leading to unfair classifications. This has encouraged the development of new approaches that address social bias in toxicity detection while maintaining the model's performance in an acceptable range (Andročec, 2020; Mathew et al., 2021).

## 1.2 Problem Statement and Challenges

The primary challenge in developing toxic language classifiers is mitigating unintended bias, which arises when models over-generalize the data distribution during training and

[1] Generally, *toxic language* can be defined as rude and disrespectful phrases, especially in the form of text or comment.

associate certain words related to demographic groups with toxicity or non-toxicity (Dixon et al., 2018a; Zhang et al., 2020a). This bias can result from imbalanced training data across labels and demographic groups, as well as intrinsic bias in the word or sentence embedding representation.

Moreover, most existing research on this topic focuses on fully supervised learning, where all the training data is annotated. However, due to the cost of annotating the vast number of comments generated every second, deploying fully supervised models is impractical. Consequently, there is a growing need to develop frameworks that work with a limited amount of annotated data and leverage the power of unlabeled data to create models that not only exhibit high classification performance but also ensure fairness in their decisions (Chen et al., 2020). [2]

Achieving fairness in semi-supervised learning (SSL) presents unique challenges compared to supervised learning (SL) but also offers advantages. Some of the previous works on fairness in SSL (Zhang et al., 2020b,c; Zhu et al., 2022; Chakraborty et al., 2021) identify some challenges and opportunities in this field. However, not all the points in the previous works hold for this study, we will report them for clarity. The challenges can be summarised as follows:

1. **Noise from pseudo labeling:** In SSL with pseudo labeling, a classifier that is trained on limited labeled data predicts pseudo labels for the unlabeled data, which may not be accurate. Therefore, pseudo labels introduce noise into the dataset and model. This noise may worsen fairness issues by creating or amplifying biases.

2. **Representation discrimination:** If unlabeled data have inherent biases or underrepresent certain groups, SSL might inherit these biases when leveraging unlabeled data, affecting fairness.

3. **Model complexity:** SSL often involves more complex models than SL, making it harder to address fairness concerns. Incorporating fairness constraints in SSL algorithms can also be more complicated.

4. **The disparate impact of SSL:** SSL may have a disparate impact on different groups, leading to unfair outcomes and perpetuating existing biases.

The advantages can be listed as follows (Zhang et al., 2020b,c; Chakraborty et al., 2021):

[2] While a fair semi-supervised toxicity text classification has yet to be discovered, there have been some investigations into the bias concerns of semi-supervised learning settings in recent years.

1. **Utilizing abundant unlabeled data:** SSL can use vast amounts of available data, potentially helping to achieve a better trade-off between accuracy and fairness by incorporating more underrepresented data points.

2. **Better generalization:** SSL models can generalize better to unseen data due to their use of labeled and unlabeled data during training, resulting in more accurate predictions, which can benefit fairness objectives as well.

The problem that this research aims to address is the development of a fair semi-supervised framework for mitigating social bias in toxicity text classification. Achieving fairness in SSL presents unique challenges, but it also offers advantages. By carefully addressing challenges and leveraging advantages, SSL could contribute to the development of fairer machine learning models in toxicity text classification.

## 1.3 Research Objective

This study aims to develop a fair semi-supervised generative-based framework that mitigates social bias in toxicity text classification. The proposed framework leverages unlabeled and labeled data to enhance classifier performance while preserving fairness across demographic groups. This research compares the performance (in classification and in fairness ability) of the proposed fair semi-supervised models with the studied baselines [3]. The primary objective of this work is to evaluate if fairness is achievable in toxicity text classification, considering the trade-off between accuracy and fairness. Moreover, we raise other concerns as research questions to better understand the challenges and opportunities associated with achieving fairness in semi-supervised toxicity text classification. Ultimately, this study aims to contribute to developing more effective and equitable machine learning (ML) models that can be used to promote social good in the toxicity text classification field.

The research hypothesis and questions are listed as follows:

$H$: We can achieve fairness in semi-supervised toxicity classification by considering the trade-off between accuracy and fairness.

$Q1$: In the proposed framework, can subtle differences in the implementation of our fair semi-supervised models greatly affect their performance in terms of fairness and accuracy?

[3]Baselines here are fully supervised models (FullBERT and FairFullBERT), limited supervised models (BERT and FairBERT), and semi-supervised generative models (GANBERT and NDAGAN)

$Q2$: Can we improve fairness for fine-grained terms related to demographic groups while debiasing our classifier on coarse-grained terms?

$Q3$: What role does labeled data play in semi-supervised toxicity text classification in terms of fairness and accuracy?

$Q4$: What role does unlabeled data play in semi-supervised toxicity text classification in terms of fairness and accuracy?

## 1.4 Scope and Limitations

This research primarily focuses on developing a fair semi-supervised generative-based framework for toxicity text classification that mitigates social bias. In this context, we concentrate on the following aspects:

- Investigating the challenges and opportunities associated with fairness in semi-supervised learning, specifically in the context of toxicity text classification.
- Proposing and implementing novel fair semi-supervised generative-based models (FairNDA-GAN and FairGANBERT) that leverage unlabeled data to enhance classification performance while preserving fairness across demographic groups.
- Evaluating the effectiveness of our proposed framework in terms of accuracy and fairness, and comparing their performance with the baselines such as fully supervised, limited supervised cases, and semi-supervised models.
- Exploring the role of labeled and unlabeled data in semi-supervised toxicity text classification with respect to fairness and accuracy.

However, certain limitations to our study that may affect the generalizability or applicability of our findings:

1. **Dataset limitations:** Our study relies on two benchmark toxicity datasets, which may not represent all types of online content or cover all demographic groups adequately. The findings may be limited in their applicability to other datasets or contexts.
2. **Bias mitigation method:** We apply adversarial debiasing as our bias mitigating method, and the effectiveness of our proposed framework may depend on this specific method's success. Other debiasing techniques may yield different results.

3. **Language and domain specificity:** Our research primarily focuses on English language data and may not be directly applicable to other languages or domain-specific content, which may require different preprocessing techniques or model architectures.

4. **Model generalizability:** While our study compares the performance of our proposed framework with several baselines, there may be other state-of-the-art models or techniques that could offer different results or trade-offs between accuracy and fairness.

Despite these limitations, our research contributes valuable insights into the challenges and opportunities of fairness in semi-supervised toxicity text classification and paves the way for future work in this important area.

## 1.5 Overview of the Proposed Framework

In this work, our proposed framework introduces two novel models, FairNDAGAN and FairGANBERT, which aim to mitigate social bias in toxicity text classification in a semi-supervised learning setting. Both models follow a two-step training process. However, since FairGANBERT is similar to FairNDAGAN, except for not employing the negative data augmentation (NDA) process. We briefly describe the two models below (for details, see Chapter 4).

### 1.5.1 FairNDAGAN

FairNDAGAN is a fair semi-supervised model that leverages adversarial debiasing to mitigate bias in the base semi-supervised model, NDAGAN (Shayesteh and Inkpen, 2022). The training procedure of FairNDAGAN consists of two main phases: pre-training and post-training.

In the pre-training stage, we train the NDAGAN and the adversarial network separately (we train the adversarial network on the last hidden layer of the discriminator in NDAGAN while keeping the NDAGAN weights fixed). In the post-training stage, we remove the generator, which does not contribute to the debiasing process. The remaining encoder and discriminator components are referred to as the "poison classifier". In this stage, the adversarial network and poison classifier are trained simultaneously in alternating epochs to minimize the influence of unintended biases related to gender and race in the poison classifier.

### 1.5.2 FairGANBERT

FairGANBERT is the second fair semi-supervised model that leverages adversarial debiasing to mitigate bias in the base semi-supervised model, GANBERT (Croce et al., 2020). The FairGANBERT follows a similar development process as FairNDAGAN, with the primary distinction being the absence of the NDA process. As a result, FairGANBERT focuses on using the GANBERT architecture for semi-supervised learning and applying the adversarial debiasing technique to mitigate social bias. The training procedure of FairGANBERT consists of similar pre-training and post-training stages as described for FairNDAGAN.

## 1.6 Contributions

In this study, we make several key contributions to fairness-aware semi-supervised learning, focusing on toxicity text classification.

**Achieving Fairness in SSL:** For the first time in toxicity text classification, we perform a comprehensive analysis to compare the performance of our proposed semi-supervised models against semi-supervised and supervised baselines in terms of both accuracy and fairness. Our models show that we can achieve fairness in semi-supervised toxicity text classification considering the trade-off between accuracy and fairness.

**Fair Semi-Supervised Learning Framework:** Based on our knowledge, we are the first to propose fair semi-supervised learning models that integrate fairness considerations into the decision process for toxicity text classification. Our models are designed to balance classification performance with fairness using the adversarial debiasing technique, and they greatly outperform their non-fair counterparts in terms of fairness metrics.

**Understanding the Role of Labeled and Unlabeled Data:** We explore the impact of the amount of labeled and unlabeled data on both fairness and accuracy in semi-supervised learning models. Our findings suggest that simply increasing the amount of unlabeled data does not necessarily improve model accuracy or fairness, highlighting the need for more strategic use of unlabeled data. Also, we found that increasing the amount of labeled data is beneficial for accuracy; however, in terms of fairness, the trends we observed do not suggest that more labeled data is beneficial.

**Investigation of Coarse-Grained vs. Fine-Grained Fairness [4]:** We demonstrate that achieving fairness at the coarse-grained level improves fairness at the fine-grained level

[4]In this work, similar to (Baldini et al., 2022), we refer to gender and race as coarse-grained terms

but does not always guarantee it. This insight underscores the need for more considerations in model design and evaluation to achieve fairness at a granular level in semi-supervised learning.

**Influence of Data Imbalance on Accuracy and Fairness:** Our study uncovers potential vulnerabilities of both supervised and semi-supervised models to imbalanced datasets. We highlight the necessity of integrating data augmentation methods and balancing techniques when dealing with such datasets.

Through these contributions, our work deepens the understanding of fairness in semi-supervised toxicity text classification and provides a foundation for future research in this important area.

## 1.7 Thesis Outline

The remainder of this thesis is organized as follows:

1. **Chapter 2: Background and Related Works** provides a comprehensive review of existing literature and research in the area of fairness in ML, with a particular focus on toxicity text classification. This chapter also analyses the related works in this field.

2. **Chapter 3: Datasets** details the data sources utilized in this research and provides an overview of the data's characteristics and distribution.

3. **Chapter 4: Methodology** outlines the semi-supervised learning framework and fairness methods implemented in this research. It explains in detail our proposed models, FairNDAGAN and FairGANBERT. Then, introduces the baselines and evaluation metrics.

4. **Chapter 5: Evaluation and Discussion** First it introduces experiments and data sampling regime, then it presents the results from the experiments conducted with our proposed models and provides a detailed analysis and discussion. Finally, it summarized our key findings and the limitations of the work.

---

and specific terms related to gender and race in each dataset such as women and black are considered as fine-grained terms. For more information about how we formulate the fairness problem in this work, please look at section 3.4.

5. **Chapter 6: Conclusion and Future Work** summarizes the findings of this research and discusses their implications. It also proposes potential directions for future research in fairness-aware toxicity text classification.

Each chapter progressively builds upon the previous one, allowing for a comprehensive understanding of the topic.

# Chapter 2

# Backgrounds and Related Works

In this chapter, we look at the background of social fairness in semi-supervised toxicity classification and analyze the related works. First, we introduce some general background related to artificial intelligence (AI) and machine learning (ML), then we focus on classification, semi-supervised learning (SSL), and fairness problems. Following this, we delve into natural language processing (NLP) and introduce text classification. Finally, we analyze related works associated with social bias issues in toxicity text classification and introduce fairness notions.

Throughout this chapter, we aim to provide a solid foundation for understanding the interplay between social fairness and text classification. Also, we describe the challenges of social fairness in semi-supervised classification [1] to pave the way for meaningful discussions and contributions in the subsequent chapters.

## 2.1 Artificial Intelligence (AI)

AI is a multidisciplinary field at the confluence of computer science and cognitive science. The purpose of AI is to simulate human-like intelligence [2]. Therefore, AI is a field that aims to develop algorithms to perform tasks with human-like efficiency, faster (Khanzode and Sarode, 2020).

[1] Based on the best of our knowledge, there is no work that specifically focuses on social fairness in semi-supervised toxicity text classification.

[2] Definition of intelligence: intelligence, in a broad sense, means the ability to learn, adapt, understand, create, and memorize (Khanzode and Sarode, 2020).

## 2.2 Machine Learning (ML)

ML is a subfield of AI focused on developing models that learn patterns and relationships from datasets to generalize their knowledge about a domain to make accurate inferences on unseen data instances (Awad and Khanna, 2015).

### 2.2.1 General Procedure in ML

According to Awad and Khanna (2015), a general procedure to deploy a model in ML involves the following steps:

1. **Collecting and preprocessing data:** This involves collecting relevant data and preparing it for analysis by cleaning, transforming, and organizing the information.
2. **Selecting a proper ML model:** Based on the problem and the nature of the data, an appropriate ML model is selected to capture the underlying patterns and relationships in the data distribution.
3. **Training the model:** The model is built on the training dataset, which consists of examples from the problem domain. The model learns from the training data using an optimization algorithm that adjusts the model's parameters to minimize the prediction error.
4. **Evaluating the model:** The model's performance is assessed using a separate set of data (the validation or test dataset) that the model did not see during training. The evaluation is helpful since it determines how well the model can generalize the problem and whether it can make accurate predictions on new data instances.
5. **Hyper-parameter tuning and optimization:** The model may be adjusted or optimized to improve its performance based on the evaluation results. This process might involve tuning hyperparameters, selecting different features, or trying a different model altogether.
6. **Deploying the model:** Once the model is fine-tuned and optimized, it can be deployed in real-world applications to make predictions and decisions or automate tasks.

### 2.2.2 Deep Learning (DL)

DL is a subfield of ML and AI and played a crucial role in the Fourth Industrial Revolution. ML focuses on developing algorithms to identify patterns in data; however, DL expands this by employing artificial neural networks (ANN) to model complex hierarchical representations, enabling it to learn from vast amounts of data. Moreover, the ability to model complicated data representations often leads to superior generalization capabilities compared to traditional ML methods (Sarker, 2021).

### 2.2.3 Potential Harms in ML Models: Social Bias

Increasingly, ML algorithms have become a fundamental part of our life, offering many beneficial applications in various fields. However, they are prone to make biased decisions that cause unfair consequences for individuals and society. In the decision-making context, if there is no favor toward individuals or groups based on their characteristics, fairness is guaranteed. Although if bias is present in the data and algorithm, then the resulting model cannot make its decisions fairly (Mehrabi et al., 2021; Baldini et al., 2022).

There are several real-world examples that demonstrate the problem of fairness in ML models deployed in the real worlds scenarios, such as Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software (Brennan and Dieterich, 2018), which displayed bias towards African-Americans in forecasting recidivism risk. Another example is AI systems with prejudice against darker-skinned beauty competitors or inaccurate facial recognition algorithms for specific ethnicities. These examples illustrate the importance of social fairness in ML models and why models need to prevent biased outcomes and ensure that the advantages of AI and ML are distributed equally throughout society (Mehrabi et al., 2021).

## 2.3 Classification

Classification, as a fundamental concept in ML, separates data into various pre-defined categories based on their shared properties. The significance of classification lies in its ability to generalize and compress information, allowing for a better understanding of data (Drummond, 2010).

The classification task can be implemented in different procedures depending on the type and structure of the data they work with and the specific goals they aim to achieve.

These procedures can be classified as supervised, unsupervised, and semi-supervised. Despite their differences, they seek to create classifiers to learn patterns, make predictions, or optimize actions based on the provided data. In the following subsections, we briefly describe the main types of automatic classifiers (Mahesh, 2020).

### 2.3.1 Supervised Learning (SL)

SL in classification involves learning a function that maps inputs to outputs using a labeled dataset consisting of input-output pairs. In SL, the algorithm is trained on the labeled dataset to make predictions on the new unseen data. However, if we cannot access enough annotated data, we should consider other tasks, such as unsupervised learning or semi-supervised learning (Sen et al., 2020).

### 2.3.2 Unsupervised Learning (UL)

UL in classification focuses on discovering hidden structures or patterns within unlabeled data to group them into different clusters based on a specific similarity measure. Unlike SL, which relies on a labeled dataset with input-output pairs to guide the learning process, UL works with datasets that do not have predefined output labels. This lack of annotation forced the unsupervised methods to rely on inherent relationships in the data or on groupings within the data based on their similarities in some features, such as color, size, or shape (Dike et al., 2018). Unsupervised classification (clustering) algorithms include, but are not limited to, K-means clustering, Hierarchical clustering, and Gaussian mixture models (GMMs). These algorithms are commonly used in areas such as image classification, document classification, and speech recognition, where no predefined set of categories or labels are available (Dike et al., 2018).

### 2.3.3 Semi-Supervised Learning (SSL)

SSL is an ML approach that lies between supervised and unsupervised learning. It utilizes both labeled and unlabeled data during the training process. This method of learning is beneficial in real-world scenarios where acquiring labeled data can be expensive and time-consuming, but unlabeled data is abundant. By leveraging both data types sufficiently, SSL can build more accurate models trained on a limited number of annotated data while reducing the need for large amounts of labeled training data (Reddy et al., 2018).

The significance of using labeled and unlabeled data in semi-supervised classification lies in its ability to overcome the limitations of purely supervised or unsupervised techniques. For example, SL requires a lot of annotated data to achieve high classification accuracy, which can be costly and time-consuming. On the other hand, UL identifies clusters or structures within the data without guidance, leading to less accurate classifications, especially for unknown data or domains (Reddy et al., 2018). However, to make SSL work, we need to make some assumptions.

#### 2.3.3.1 Semi-Supervised Learning Assumptions

According to Chapelle et al. (2009), SSL operates under certain assumptions about the structure of the data. These assumptions are crucial for effectively applying SSL algorithms, as they enable generalization from a limited training set to a larger set of unseen cases. The main assumptions in SSL are:

- **The Smoothness Assumption:** It suggests that if two input points $(x_1, x_2)$ are close in a high-density area, their respective outputs $(y_1, y_2)$ should also be close. On the other hand, if the input points are separated by a low-density area, their outputs should be far apart.
- **The Cluster Assumption:** It states that if data points are part of the same cluster, they belong to the same class. It also implies that decision boundaries should be located in low-density regions (if a decision boundary is placed in a high-density region, it could divide a cluster into different classes, which goes against this assumption).
- **The Manifold Assumption:** It states that high-dimensional data lies on a low-dimensional manifold (estimating the true data distribution in high-dimensional spaces can be difficult due to the exponential volume growth with each added dimension). However, if the input data resides on a lower-dimensional manifold, it is possible to find a low-dimensional representation using unlabeled data and subsequently use labeled data to solve the simplified task.

#### 2.3.3.2 Semi-Supervised Learning Methods

Based on Ouali et al. (2020), several SSL methods have been introduced over time, which can be categorized into four main groups:

- **Consistency Regularization (Consistency Training):** This approach assumes that applying realistic perturbations to unlabeled data points (augmented version of the data) should not significantly change the model predictions. Therefore, the model is trained to maintain consistent predictions for an unlabeled example and its perturbed version.

- **Proxy-label Methods (Pseudo Labeling):** These methods use a model trained on limited labeled data to create additional annotated training examples by labeling instances from the unlabeled set using specific heuristics. Examples include Self-training, Co-training, and Multi-View Learning.

- **Generative Models:** These models learn transferable features from one task to another, similar to supervised settings. They generate samples from the data distribution $p(x)$ and learn transferable features to a supervised task $p(y|x)$ with target labels $y$.

- **Graph-Based Methods:** These methods view labeled and unlabeled data points as nodes in a graph, aiming to propagate labels from labeled to unlabeled samples based on the similarity between two nodes $x_i$ and $x_j$, reflected by the edge strength $e_{ij}$.

Apart from these primary categories, some SSL techniques focus on entropy minimization, encouraging the model to make confident predictions by minimizing prediction entropy. Consistency training can also be considered a proxy-label method with a subtle difference in enforcing prediction consistency rather than computing the cross-entropy loss (Ouali et al., 2020).

In addition, SSL methods can be classified based on two learning paradigms: transductive learning and inductive learning. Transductive learning focuses on applying a trained classifier to the unlabeled instances observed during training and does not generalize to unseen instances. On the other hand, inductive learning aims to learn a classifier that can generalize to unobserved instances at test time. Generally, inductive learning is the most common paradigm in SSL, as it allows models to generalize to new data points and applies to most real-world scenarios. However, transductive learning can be helpful in some situations where only a limited set of labeled data is available, and it is unnecessary to generalize to new data points (Ouali et al., 2020).

## 2.4 Bias Issue in Classification

The increasing use of ML algorithms in classification tasks has led to the belief that these systems offer more objective predictions than humans. However, ML algorithms are not always unbiased in classification, as they may learn historical biases from the training data. The impact of these biases on people's lives, especially in areas like employment, loans, and criminal justice, has raised concerns about algorithmic fairness in classification tasks. In addition, recent cases have demonstrated how biased algorithms can lead to discriminatory outcomes based on race or gender. Bias issues in the classification task yielded growing concerns about improving fairness in the decision-making algorithm. As a result, in the following years, causes for unfairness in classification have been identified, including biases in datasets, missing data, biases from algorithmic objectives, and the use of proxy attributes for sensitive attributes. However, addressing fairness in ML models is complex, as there is often an inherent trade-off between accuracy and fairness. To effectively tackle this issue, we should explore various definitions and measures of fairness and fairness-enhancing mechanisms for classification tasks (Pessach and Shmueli, 2020).

### 2.4.1 Bias in Semi-Supervised Classification

The bias issue in semi-supervised classification has yet to attract much attention; some works have tried to identify the real challenges in this field. However, they could have been more successful in developing a fundamental theoretical analysis to discover the real challenges for all types of semi-supervised methods instead of focusing on pseudo-labeling and consistency regularization methods.

Zhang et al. (2020c) is one of the first works on this topic. They aim to address the lack of research on fairness in SSL and investigate how to effectively utilize unlabeled data to achieve a better trade-off between accuracy and fairness. They hypothesize that increasing the training set's size with unlabeled data can help improve the balance between fairness and accuracy. The authors propose a fairness-enhanced sampling (FS) framework in the pre-processing phase that combines pseudo labeling to first assign labels to unlabeled data, second, re-sampling to balance data based on their labels and the demographic groups they belong, and then add ensemble learning to decrease the noise and bias introduced by pseudo labeling step to achieve fair SSL. They claim that experimental results demonstrate that their FS framework can achieve fair semi-supervised learning and reach a better trade-off between accuracy and fairness than fair supervised learning.

Later, Zhang et al. (2020b) seek to solve the problem of achieving fairness in SSL

while maintaining a balance between accuracy and fairness using an in-processing method [3]. By theoretically analyzing these sources of discrimination, the authors suggest that unlabeled data can be utilized to achieve a better trade-off between accuracy and fairness. Their methodology is to utilize unlabeled data using SSL by formulating an optimization problem with objectives to optimize accuracy using supervised classifier loss, allocate labels to unlabeled data using label propagation loss, and optimize fairness levels by adding fairness constraints over both labeled and unlabeled data.

In (Chakraborty et al., 2021), the authors address ethical bias in ML models in the software engineering domain. They argue that most prior software engineering research works focused on identifying ethical bias in models rather than mitigating it. In this work, they recognize the challenges in obtaining annotated data with trustworthy ground truth labels and emphasize that ground truth can contain human bias even when available. To overcome these challenges, they propose a semi-supervised learning (SSL) framework called Fair-SSL. The framework aims to create fair classification models by first leveraging a small amount (10 %) of labeled data to generate pseudo-labels for the rest of the unlabeled data. Then, synthetic oversampling is used to balance the training data to mitigate bias and train a fairer classifier.

Zhu et al. (2022) address the problem of disparate impacts in SSL where some subpopulations, defined by demographic groups, may experience different benefits or even performance drops when using SSL. They theoretically and empirically investigate the reasons for this discrepancy and introduce a new metric called Benefit Ratio to measure the disparate impact in SSL. Furthermore, they promote the evaluation of fairness in SSL using the Equalized Benefit Ratio concept. The paper also discusses potential strategies to mitigate these disparate impacts, such as balancing the data and collecting more labeled data.

However, these works tried to identify bias challenges in the semi-supervised classification task; most of the proposed solutions to mitigate bias is based on data manipulation [4], which is not very practical in real scenarios (Zhang et al., 2020a). Moreover, Zhu et al. (2022); Zhang et al. (2020c,b) theoretically investigate the source of bias in SSL, they have focused on the pseudo-labeling methods. Therefore, a lack of research in this field, especially related to bias in semi-supervised generative models, makes it difficult to identify all the challenges in advance to propose a proper solution for fairness in semi-supervised text classification.

[3] In-processing methods are defined in section 2.7.1.2.

[4] Except (Zhang et al., 2020b)

## 2.5 Natural Language Processing (NLP)

NLP emerged in the 1950s as an interdisciplinary field, combining AI and linguistics. Then, NLP has slowly combined with other disciplines like Computer Science, Cognitive Science, and Physiology, which requires researchers and developers to expand their knowledge base. One of the first development in this field was rule-based NLP due to Chomsky's analysis of language grammar in 1956. However, NLP now encompasses a wide range of techniques and draws from diverse fields to analyze, understand, and generate human language (Nadkarni et al., 2011).

### 2.5.1 Machine Learning in NLP

Rule-based NLP faced significant challenges due to the unrestricted nature and inherent ambiguity of language. In the 1980s, a fundamental reorientation in NLP occurred, with a shift towards statistical NLP. This new view emphasized robust approximations, probabilistic ML methods, and the use of large annotated corpora. In addition, statistical NLP improved the performance of the NLP tasks by utilizing the most common pattern extracted from real data and allowing for graceful degradation [5] when faced with unfamiliar inputs (Nadkarni et al., 2011).

### 2.5.2 Deep Learning in NLP

Deep learning has revolutionized the field of NLP by leveraging the power of artificial neural networks (ANNs) with billions of trainable parameters. The availability of large datasets, combined with advances in computational power and parallelization through graphical processing units (GPUs), has enabled researchers to make significant progress in NLP tasks. Deep learning architectures, such as convolutional neural networks (CNNs), recursive neural networks (RNNs), recurrent neural networks with long short-term memory (LSTM), attention mechanisms, transformers, and residual connections with dropout, have been applied to various NLP tasks. These areas include language modeling, morphology, parsing, and semantics, which are essential for understanding the underlying structure and

[5] Based on Dymond (2021) "when machine learning models encounter data which is out of the distribution on which they were trained they have a tendency to behave poorly, most prominently over-confidence in erroneous predictions. Such behaviours will have disastrous effects on real-world machine learning systems. In this field, graceful degradation refers to the optimization of model performance as it encounters this out-of-distribution data."

meaning of human languages. As a result, deep learning has led to significant improvements in NLP applications, such as text classification, machine translation, text summarization, information extraction, and question-answering systems Otter et al. (2021).

The following section will describe the language models (LMs) in NLP. These models play a critical role in capturing the complexities of human languages, predicting word sequences, and inferring meaning from context.

### 2.5.3 Language Models (LMs)

Language models are fundamental to NLP, allowing machines to understand and predict human language by capturing the underlying structure and relationships between words. These models are essential to various applications, including machine translation, text summarization, and text classification. Their significance lies in their ability to implicitly capture syntactic and semantic relationships among words or components within a linear context (Otter et al., 2021).

Various types of LMs have been developed over the years, including statistical, neural, and transformer-based models. Statistical language models, while effective in certain scenarios, struggle with handling synonyms and out-of-vocabulary words. Neural language models, on the other hand, have significantly addressed these issues thanks to advances in artificial neural networks (ANNs) and deep learning. Some of the most prominent neural language models include recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and gated recurrent units (GRUs). Recently, transformer-based models, such as GPT and BERT, have emerged as state-of-the-art in language modeling, achieving remarkable performance across a wide range of NLP tasks. These models have pushed the boundaries of language understanding, but some challenges still remain in areas such as cross-language modeling and adapting to low-resource languages (Otter et al., 2021).

### 2.5.4 Potential Harms of Large Language Models (LLMs) [6]

A growing trend in NLP is to design large language models (LLMs) trained on enormous unstructured data available on the web. These models have shown incredible performance in reasoning and understanding across textual information. The distinguished capabilities of these models compared to what we achieved in the past fascinate most humans. However,

[6] Focusing on bias issues and fairness

many studies imply that deploying language models comes with risks, especially socio-ethical harms and bias issues (Bommasani et al., 2021; Venkit et al., 2022; Narayan et al., 2022).

Fairness and bias issues in LLMs have become increasingly important, as they can lead to both intrinsic and extrinsic harm. Intrinsic bias is an inherent bias in NLP representations, such as word embeddings. In contrast, extrinsic bias relates to disparities in downstream tasks, such as variations in false positive rates (FPRs)[7] across groups defined by sensitive attributes. Although intrinsic and extrinsic biases are interconnected, and measuring intrinsic bias in language models does not necessarily reflect the behavior of models fine-tuned for specific applications (Baldini et al., 2021).

### 2.5.5 Bias and Fairness in NLP

In this section, we first introduce a high-level definition of bias and fairness, especially in NLP models, and then we provide information for different sources of bias in NLP algorithms. In NLP, fairness means that a model treats all groups of people equally, regardless of their protected attributes [8]. In addition, bias means that NLP models encode stereotypical information associated with sensitive attributes related to demographic groups in their representation or treat samples related to different demographic groups differently. As a result, to make a model fairer, we enforce a bias mitigating process to decrease disparities among different demographic groups to make the model treat samples more objectively concerning their sensitive attributes. A general trend these days is that NLP models are trained on vast amounts of text data, which inherently contain biases affecting learned representations. Barocas and Selbst (2016) categorize biases in data into:

- **Skewed sampling (feedback loop):** When an ML model's predictions become a prediction that comes true because of actions taken based on that prediction (self-fulfilling prophecy). For example, predictive policing tools are influenced by their predictions of where crime will occur, which can reinforce the model's predictions and cause biases to become stronger.
- **Tainted examples:** Social biases like racism, sexism, and homophobia can pollute datasets, and bias can get into models during training and evaluation, e.g., if a

[7] FPRs definition is given in section 2.8

[8] In the context of fairness in machine learning, protected attributes refer to certain characteristics or features of individuals or groups that should not be used as a basis for decision-making or discrimination. These attributes are often related to demographic or personal information, such as age, gender, race, ethnicity, religion, sexual orientation, or disability status (Ghassami et al., 2018).

company has a low record of hiring women, the resume screening system, identified women to be unfit for roles like Amazon's automated resume screening system.

- **Limited features:** If a feature related to certain demographic groups is less informative, then it can lead to disparities in the model performance across different groups.
- **Sample size disparities:** If the data set is imbalanced with respect to different demographic groups, it can lead to disparate model errors.
- **Proxies:** If features related to sensitive attributes are deleted, there are other features in the feature space that have a high correlation with the removed features, which causes models to consider sensitive information related to demographic groups into account.

In addition to data biases, model architectures can amplify biases due to factors such as imbalanced datasets, spurious correlations, and contextualization. Addressing fairness and bias in NLP requires understanding the biases present in data and models, and developing strategies to minimize their impact (Subramonian, n.d.). However, there is a question that can we completely remove bias in a problem? Subramonian (n.d.) mentions that removing biases totally due to their complexity in practice is not possible. Though, using the bias mitigating method can control and reduce the harm in the models.

## 2.6 Text Classification

Text classification is a crucial technique in NLP that involves assigning predefined categories to text documents. It helps a wide range of applications, including sentiment analysis, partisanship recognition, and spam detection. ML models are commonly used for this task because they extract textual patterns from labeled documents Qian et al. (2021).

### 2.6.1 Semi-Supervised Text Classification

SSL has acquired much attention in NLP, and different semi-supervised text classification frameworks have been explored in recent years. We can categorize most of the works into two distinctive categories of models: pseudo-labeling models which usually aim to iteratively predict pseudo labels for unlabeled data using models trained on limited labeled

data. Also, in order to make their pseudo-labeling process more robust and less noisy, they usually introduce data augmentation to add perturbation to the data point (usually unlabeled data) and then enforce the model to produce a similar prediction for each sample and its transformed version by adding a regularization term called consistency loss to the objective function of the model. In this way, semi-supervised text classification models can leverage unlabeled data to improve the generalization ability of the models. However, selecting a proper perturbation or data augmentation method is key in these methods. In addition, employing data augmentation techniques and consistency loss increases the time and computational complexity of the model (Zhu et al., 2022).

The most famous SSL frameworks in text classification that leverage the power of pseudo-labeling include the following. Miyato et al. (2018) use adversarial and virtual adversarial training (VAT) to the text by applying perturbations to the word embeddings. Zbiciak and Markiewicz (2023) use hierarchical structures to enable supervision from labels at higher levels to be utilized for labels at lower levels. Xie et al. (2020) utilized consistency regularization on unlabeled data by performing back translations and tf-idf word replacements. However, one of the most effective pseudo-labeling frameworks is MixText. MixText which is introduced in Chen et al. (2020) is a novel semi-supervised learning framework for text classification. It introduces a data augmentation technique similar to Mixup (Zhang et al., 2017) called TMix, which interpolates text representation in hidden space to generate a large number of augmented samples. In this framework, to leverage unlabeled data, first, the model generates low-entropy pseudo labels for the unlabeled data and the augmented version of the unlabeled data. Then, it utilizes TMix to interpolate between labeled and unlabeled (and augmented unlabeled) instances to train the Bert classifier. In this way, this framework leverages label and unlabeled data simultaneously to train the text classifier. Moreover, to make its prediction of synthetic samples generated by the TMix technique more confident and less noisy, it employs entropy minimization loss and consistency regularization. This approach results in significant performance improvements, especially when labeled training data is scarce.

On the other hand, generative-based models, such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), learn the underlying data distribution of the data by generating new samples. In SSL, these models can use labeled and unlabeled data to learn the data distribution. The discriminative part of the model is then trained using the learned distribution and the labeled data to classify the text Croce et al. (2020).

Some of the well-known frameworks that utilized VAEs are introduced in (Gururangan et al., 2019; Chen et al., 2018; Yang et al., 2017), which utilize VAEs in the form of sequence-to-sequence modeling on text classification and sequential labeling. On the other side, semi-supervised GANs (SS-GANs) based models are kernel-based SS-GANs intro-

duced in (Croce et al., 2019), GANBERT (Croce et al., 2020), and NDAGAN(Shayesteh and Inkpen, 2022) [9]. The kernel-based SS-GANs apply Kernel-based Deep Architecture (KDA) projection on the input text to transform its representation into a low-dimensional space; then, it sends the encoded representation to the discriminator for classification purposes. GANBERT, one of our semi-supervised baseline classifiers, is a framework that combines the power of BERT, a pre-trained Transformer-based architecture, with SS-GANs to improve text classification performance. The method utilizes BERT as a discriminator to produce high-quality input text representations, while a generator creates "fake" examples estimating the data distribution. This approach enables the model to exploit labeled and unlabeled data effectively, leading to better generalization capabilities in text classification tasks. In addition, GANBERT employs Feature Matching (FM) to prevent mode collapse and improve performance. However, mode collapse remains a persistent issue in most adversarial generative models, leading to decreased discriminator performance on unseen data. Therefore, NDAGAN (Generative Adversarial Learning with Negative Data Augmentation for Semi-supervised Text Classification) is introduced by Shayesteh and Inkpen (2022) to mitigate the mode collapse issue present in GANBERT, which negatively impacts their performance. NDAGAN Uses the Negative Data Augmentation (NDA) technique (Sinha et al., 2021) for the first time in text classification, NDAGAN leverages labeled and unlabeled data to fine-tune the BERT encoder and train the discriminator for better classification performance. The proposed solution generates informative NDA synthetic samples for the discriminator during training by blending the generator's output with the contextual representation of real data. This leads to improved model accuracy and enhanced out-of-distribution sample detection. This model has been used as a baseline in this work.

## 2.7 Toxicity Text Classification

Toxicity detection is a text classification task where the ML model aims to detect comments or text that contains toxic language content [10]. Toxicity benchmark datasets usually contain annotations for each sample to determine if the sample belongs to the toxic language class. Moreover, some datasets add label information related to sub-categories of

[9] This work was previously published by us and is further elaborated in this thesis, see Section 4.2.1.4 for details. However, in the published paper, we used other benchmark text classification datasets, and now we substituted the BERT encoder that is part of the model with DistilBERT to increase run-time efficiency.

[10] Based on Borkan et al. (2019) definition toxic language is "anything that is rude, disrespectful, or unreasonable that would make someone want to leave a conversation"

the toxic language, such as hate speech or offensive language Mathew et al. (2021). More information about toxicity text benchmark datasets is presented in Chapter 3.

### 2.7.1 Social Bias Issue in Toxicity Text Classification

The corpus of text sometimes contains identity term information about demographic groups. The identity terms in the text can be stated implicitly or explicitly which varies in the different corpus (Liu et al., 2021; Baldini et al., 2021). However, a fair text classifier must classify texts based on the semantic information embedded in the corpus, not identity term information related to demographic groups (Zhang et al., 2020a). It is illustrated in the body of works (Dixon et al., 2018b; Zhang et al., 2020a; Baldini et al., 2021) that text classifiers tend to make their predictions relying on information related to demographic groups, which raises an issue of bias in this NLP task.

In this section, first, we analyze different bias definitions and notions of fairness in the toxicity text classification literature. Then we discuss potential sources of bias. Afterward, we introduce previous mitigating bias approaches in this task, and finally, bias metrics and evaluation methods are discussed.

#### 2.7.1.1 Bias and Fairness Definition in Toxicity Text Classification

Dixon et al. (2018b) is one of the first works that analyze the source of biases in the text classification tasks and introduce a method to mitigate bias in toxicity detection. This work reformulates the unintended bias definition from (Hardt et al., 2016) into the text classification task. Unintended bias means when a model favours some demographic groups more than others in its prediction. On the other side, Dixon et al. (2018b) explicitly mentions that intended bias is what an ML model is designed to learn. For example, in the toxicity text detection task, a model should discriminate toxic comments from non-toxic ones based on semantic information related to toxic words (intended bias) instead of gender or other identifying information embedded in the text (unintended bias).

Unintended bias is a term that is widely used in the toxicity text classification literature, and most of the works (Dixon et al., 2018b; Zhang et al., 2020a; Baldini et al., 2021) aim to reduce it to ensure some level of fairness in this task. Also, in some of the works such as (Liu et al., 2021; Baldini et al., 2021), authors focus more on the different aspects of unintended bias presented in systems and classify them as explicit and implicit biases. Subsequently, explicit bias refers to situations when a dataset contains identity terms related to demographic groups. On the other side, implicit bias happens when identity

terms are not directly mentioned in the text while some distinct features (as proxy features) are highly correlated to demographic groups, such as the language style of the author (Liu et al., 2021).

By focusing on the concept of unintended bias in toxicity text classification, we need to ask what is the source of unintended bias in the text that makes a model biased toward certain demographic groups?

Dixon et al. (2018b) thoroughly analyzes the source of unintended bias [11], and they suggest that a high ratio of toxic to non-toxic comments for specific identity terms is the main reason a model overgeneralizes the data distribution and learns unnecessary associations, such as the word "gay" has a relation with toxicity because it mostly appears in the toxic comments. Additionally, they have noted that other reasons, such as dataset size, model training method, and the disproportionate relationship between comment length and toxicity, negatively affect the fairness and may cause unintended bias.

The works of (Zhang et al., 2020a; Park et al., 2018) also referred to (Dixon et al., 2018b) to analyze the source of bias in the toxicity text classification task. Zhang et al. (2020a) mention that the source of bias comes from imbalance (high) rates of toxic samples to non-toxic samples for specific demographic subgroups. As a result, a model trained on these datasets may capture unintended biases related to identity terms (of the demographic groups) and perform unintended biased on its decisions. Park et al. (2018) targets the gender bias problem in the abusive language detection task and points out that overgeneralizing problem in the supervised text classifier results from memorizing frequently repeated identity terms in the toxic comments. Therefore, the high frequency of identity terms in toxic comments is the source of bias in line with the findings in (Dixon et al., 2018b).

Qian et al. (2021) reword the unintended bias as unintended dataset biases. In this work, they identify the source of unintended dataset biases, either come from label bias or keyword bias. The label bias issues are referred to as document-level bias, where labels are distributed unevenly through the training samples. The consequence of the label bias issue is the potential risk of predicting the majority class by the models trained on such training sets. On the other hand, keyword bias is concerned with word-level bias or cases where a high correlation exists between some specific keywords and class labels. As a result, a model trained on datasets containing word-level bias tends to associate some terms with a specific label instead of considering semantic information of the text (Waseem and Hovy, 2016b; Liu and Avci, 2019).

[11] In this work, we use the term 'unintended bias' to refer to explicit unintended bias unless otherwise specified.

Liu et al. (2021) focus on how to mitigate the implicit unintended bias in text classification and empirically shows that even if demographic attributes are not explicitly mentioned in the text still, imbalance label distribution for each of the demographic groups is the main reason for a model to make unintended bias in their predictions.

By looking at the previous works in social fairness in toxicity text classification literature, we can point out that the primary source of unintended bias, in general, is an imbalance ratio that comes from the keyword or document level.

#### 2.7.1.2 Bias Mitigating Methods in Toxicity Text Classification

In this part, we summarize the mitigating methods introduced in the previous works and then we classify the mitigating techniques based on how they apply to the model.

Dixon et al. (2018b) is one of the first works focused on fairness in toxicity text classification on a jigsaw toxicity dataset. They have identified that unintended bias against demographic groups mostly comes from keyword bias and comment lengths [12] (toxic comments tend to be shorter in the training dataset). Therefore, they have introduced a bias mitigating method that aims to improve the model fairness by balancing the toxic to the non-toxic ratio for sensitive terms related to demographic groups in pre-defined comment length bins. Since this ratio is high for sensitive terms, they add non-toxic comments containing biased terms of different lengths from Wikipedia articles in an unsupervised manner. [13]

Zhang et al. (2020a) reframe the unintended bias in the text classification datasets introduced in (Dixon et al., 2018b). First, they assume that there exists a non-discrimination distribution free of any bias that is unknown to us; however, the training dataset as a discrimination distribution is available. It is assumed that the discrimination distribution is sampled independently from the non-discrimination distribution following discrimination rules such as social prejudice or keyword-level bias. Therefore, they define unintended bias as a form of selection bias from non-discrimination distribution to discrimination distribution. Finally, they have proposed a framework to revive the non-discrimination distribution from training datasets by applying instance weighting. Instance weighting helps to approximate the non-discrimination loss during the training using the training dataset (which comes from the discrimination loss).

[12] They use CNNs architecture for their classifier

[13] To collect labeled, non-toxic data without human supervision, they assume that Wikipedia articles cannot contain toxic information in nature. Thus they can balance the jigsaw toxicity dataset by adding non-toxic data from the Wikipedia articles dataset.

Park et al. (2018) apply three different mitigating techniques for gender bias in abusive language detection. These techniques are Gender Swaps (GS), a data augmentation method that swaps gender identity information in training data, Debiased Word Embeddings (DE) introduced originally in (Bolukbasi et al., 2016) to remove gender stereotypical information from the word representation, Bias Fine-Tuning (FT) method to fine-tune model with a larger less biased corpus (source dataset) to regularize and prevent the model overfit the smaller and more biased dataset(target dataset). They claim that these methods effectively reduce the bias in the models.

Zhang et al. (2020a) criticize two aforementioned model-agnostic debiasing frameworks that work based on data manipulation methods. They mention that data manipulation is not always practical. For example, Dixon et al. (2018b) propose to use data supplementation to balance the rate of toxic to non-toxic comments across the sensitive demographic groups. However, collecting new samples concerning sensitive demographic groups and sentence length is difficult due to the high cost of collection and annotation. Additionally, Park et al. (2018) introduce a data augmentation method that requires applying gender-swapping to sentences that contain gender identity terms. This method is impractical when we have many different demographic groups, such as racial bias mitigating the problem [14]. Moreover, gender-swapping data augmentation can generate meaningless sentences such as "He gives birth." (Sun et al., 2019).

Gencoglu (2020) analyzes fairness in the cyberbullying task and suggests adding fairness constraints to the model objective function during training to guide the model to make more unbiased decisions toward sensitive demographic groups. This work imposes fairness constraints on the objective function during the training to impose the disparity between false negative rate (FNR) and false positive rate (FPR) (for more information, see section 2.6.3.2) of all the sensitive groups and overall rates in an acceptable range. Additionally, They have mentioned that previous works in this field mainly focus on changing the dataset by balancing or oversampling, or sample weighting (Dixon et al., 2018b; Zhang et al., 2020a; Park et al., 2018), which introduce additional calibration to hyper-parameters of the models that have a huge influence on the models' performance and fairness. For example, adding examples for sensitive groups can fundamentally change the decision boundary in the feature space and impair the model's generalization performance. Also, These methods, referred to as the pre-processing methods (See classifying Mitigating Strategies in this section ), cannot keep up with the changes in view toward the concept of bias over time, which makes them infeasible to deploy in real-life applications.

[14] There are many sensitive terms related to different races in the text.

Pruksachatkun et al. (2021) formalized the fairness problem as model robustness against small perturbations. Then, they add these perturbations to be different identify terms related to demographic groups because protected attributes such as gender information are mainly irrelevant to the decision-making process in the classification task. Therefore, they suggest using certified robustness approaches to increase the fairness of the text classification tasks. In robustness literature, certified robustness approaches are deployed to guarantee those model predictions are invariant to small changes in word substitution attacks. Qian et al. (2021), as mentioned before, considers the source of unintended bias, either document-level or keyword-level bias. Then, they introduce a model-agnostic debiasing framework, CORSAIR, that employs counterfactual inference to mitigate bias in the models after training as opposed to factual inference when the debiasing happens before and during the training. Basically, CORSAIR aims to make unbiased decisions using biased observations by decreasing the influence of unintended confounders in inference time. The CORSAIR framework has three main components: first, biased learning, the classifier trained on training data without considering any bias mitigating method. Second, bias distillation on the test set, where an input sample and two counterfactual instances of the input (fully blindfolded to capture label bias and partially blindfolded to capture keyword bias) are fed into the classifier. Third, the bias removal stage aims to remove the effect of keyword bias and label bias from the final prediction of a sample in test time, and this happens by elementwise subtraction of the classifier output of two counterfactual instances found in step 2 from the original sample prediction (controlled by a factor called lambda). Finally, they show experimentally that their framework effectively imposes fairness on the model's decision while maintaining generalizability.

Adversarial debiasing is introduced in (Zhang et al., 2018b) as a model-agnostic approach that mitigates biases in models when training data contains information about demographic groups. This method aims to ensure that classifiers do not learn to make their predictions based on the samples' demographic information (unintended bias). As a result, the model becomes more robust against biases related to sensitive attributes related to demographic groups and provides fairer classification results. Compared to previous methods, adversarial debiasing offers several advantages. First, it relies on something other than data manipulation, making it more practical where data collection or augmentation may be challenging or costly. Second, unlike pre-processing methods that require additional calibration of hyperparameters and may impair model generalization, adversarial debiasing directly addresses bias within the model's training process. However, adversarial debiasing also faces some challenges. It requires careful tuning of the adversarial component's strength to ensure it does not compromise the model's performance on the primary task. Additionally, it may be computationally expensive due to the added

complexity of the adversarial training process.

Basu Roy Chowdhury et al. (2021) present a framework in text classification that tries to leverage the adversarial debiasing method introduced in (Elazar and Goldberg, 2018) to eliminate demographic information encoded in data representations during the training of the text classification model. They named their framework "Adversarial Scrubber" (ADS)" and consists of four modules: Encoder, Scrubber, Bias Discriminator, and Target Classifier. The Encoder generates contextual representations of the input text, which the Scrubber then processes to produce fair representations for the target task. The Bias Discriminator and Target Classifier predict the protected attribute and target label from the Scrubber's output respectively. The framework is trained end-to-end adversarially for text classification in a fully supervised regime.

By focusing on the related bias-mitigating methods in text classification, specifically in the toxicity text classification task, we have decided to employ the adversarial debiasing technique as a step to our semi-supervised classifiers to mitigate unintended bias in our models. We chose adversarial debiasing because it has been shown to work well in reducing bias without hurting the accuracy of the models. Moreover, it fits easily into our existing semi-supervised learning framework with minor changes, and it is particularly suited for multi-label debiasing scenarios similar to this work when we aim to de-bias the models against race and gender.

Finally, bias mitigating approaches can be classified into three categories based on how they enforce debiasing on the models (Lohia et al., 2019; Bellamy et al., 2018; Baldini et al., 2021). These categories are called pre-processing, in-processing, and post-processing (Pessach and Shmueli, 2022).

The pre-processing methods mainly focus on modifying the training data before the training process, such as methods proposed by Dixon et al. (2018b); Zhang et al. (2020a); Park et al. (2018). For example, Dixon et al. (2018b) collected more data to address the keyword bias problem in the dataset. Regarding the work (Noroozi et al., 2019), the main advantage of pre-processing methods is that does not change the ML algorithms, making it easy to deploy.

The in-processing methods modify the ML algorithm during the training to enforce algorithmic fairness. This can be done by adding a term to the model's objective function to add fairness constraint (Gencoglu, 2020) or employing adversarial debiasing to mitigate bias in the model (Zhang et al., 2018b). Pruksachatkun et al. (2021) also used certified robustness methods to improve fairness and robustness during the training. It is useful to note that Noroozi et al. (2019) considers in-processing mechanisms as the most robust methods for debasing models regarding the accuracy and fairness metrics.

Conversely, the post-processing methods include techniques that transform the model's prediction to make the decision less biased toward sensitive demographic groups. One example of a simple post-processing method is defining a specific threshold for each sensitive group to control fairness across the task. An important advantage of post-processing methods over other mentioned approaches is "the ability to avoid retraining model" (Noroozi et al., 2019).

## 2.8 Bias Notions and Evaluation Metrics in Algorithmic Fairness

This section summarizes the most common algorithmic fairness notions for group fairness [15] in classification. In the algorithmic fairness for classification literature, we have identified four main notions of fairness that a system should satisfy to make it a fairer model. However, before introducing them, we need to define some important concepts that help to understand fairness notions. According to Pessach and Shmueli (2022), we define the following concepts:

- **Unprivileged group:** A group typically disadvantaged or underrepresented in a specific domain, such as women, people of color, or people with disabilities.
- **Privileged group:** A group typically advantaged or overrepresented in a specific domain, such as men, white people, or people without disabilities.
- **False positive rate (FPR):** The ratio of false positive predictions to the total number of negative instances in a dataset.

$$FPR = \frac{FP}{N} \tag{2.1}$$

  where $FP$ represents the number of false positive predictions, and $N$ represents the total number of negative instances in the dataset.

[15] Group fairness is a concept that aims to guarantee that models do not perform biased decisions against certain demographic groups such as race or gender. In contrast, individual fairness focuses on treating identical individuals similarly, regardless of their group membership (Pessach and Shmueli, 2022).

- **True positive rate (TPR):** The ratio of true positive predictions to the total number of positive instances in a dataset.

$$TPR = \frac{TP}{P} \tag{2.2}$$

where $TP$ represents the number of true positive predictions, and $P$ represents the total number of positive instances in the dataset.

Then, according to (Pessach and Shmueli, 2022; Dixon et al., 2018b; Zhang et al., 2020a), the fairness notions are:

- **Disparate impact:** The disparate impact notion measures fairness by computing the ratio of the positive prediction rate of the unprivileged group to the privileged group. The mathematical formula for this metric is defined as follows:

$$\frac{Pr(Y_{pred} = 1|S = unprivileged)}{Pr(Y_{pred} = 1|S = privileged)} \leq (1 + \epsilon) \tag{2.3}$$

where $Y_{pred} = 1$ represents a positive prediction, and $S$ represents the identity terms related to demographic groups ($S = 1$ is for samples in the privileged group). It is common to select $\epsilon = 0.2$ to follow the "80 percent rule" notion in disparate impact law.

- **Demographic parity or statistical parity:** The demographic parity notion measures the difference between the positive prediction rates of the privileged and unprivileged groups. The mathematical formula for this metric is defined as follows:

$$|Pr(Y_{pred} = 1|S = unprivileged) - Pr(Y_{pred} = 1|S = privileged)| \leq \epsilon \tag{2.4}$$

where a lower rate indicates more fairness in the model.

There are several disadvantages identified for the models that measure fairness using disparate impact and demographic parity. First, these two notions do not consider the base rate of each group. Therefore, once the base rate (actual positive outcome rate) among the groups is highly different, a fair classifier measured by these notions is considered unfair. In addition, to satisfy fairness measures by these two notions, two similar samples from different groups may be treated differently by the system, which negatively affects the individual fairness of a system. As a result of these limitations, researchers were motivated to develop other notions of fairness, such as equalized odds and equalized opportunity (Pessach and Shmueli, 2022).

- **Equalized odds:** Equalized odds is a notion commonly used in algorithmic fairness in toxicity text classification, which measures the difference between false positive rates (FPRs) and true positive rates (TPRs)[16] of the two groups separately. Equalized odds can be defined as follows:

$$P(\hat{Y} = 1|S = unprivileged, Y = 1) - P(\hat{Y} = 1|S = privileged, Y = 1) \leq \epsilon \quad (2.5)$$

and

$$P(\hat{Y} = 0|S = unprivileged, Y = 0) - P(\hat{Y} = 0|S = privileged, Y = 0) \leq \epsilon \quad (2.6)$$

where $\hat{Y} = 1$ represents a positive prediction, $Y = 1$ indicates a positive outcome, and $S$ represents the identity terms related to demographic groups ($S = 1$ is for samples in the privileged group). In terms of TPRs and FPRs notion, we can formulate them as follows:

$$|FPR_{unprivileged} - FPR_{privileged}| \leq \epsilon \qquad \text{and} \qquad |TPR_{unprivileged} - TPR_{privileged}| \leq \epsilon \quad (2.7)$$

- **Equalized opportunity:** Equalized opportunity is a relaxed version of the equalized odds notion, where a model makes the true positive rates (TPRs) equal across various groups. Equalized opportunity can be defined as follows:

$$P(\hat{Y} = 1|S = unprivileged, Y = 1) - P(\hat{Y} = 1|S = privileged, Y = 1) \leq \epsilon \quad (2.8)$$

In terms of the true positive rates (TPRs), we can formulate equalized odds as:

$$|TPR_{unprivileged} - TPR_{privileged}| \leq \epsilon \quad (2.9)$$

where the goal is to equalize the TPRs among the groups.

Although equalized odds and equalized opportunity are developed to overcome the limitations of demographic parity and disparate impact, they also have limitations. For example, equalizing the TPRs and FPRs may result in larger differences in other error rates, such as FNRs (false negative rates) or TNRs (true negative rates), causing the overall accuracy of the classifier to degrade. Moreover, these notions may not be applicable to all classification tasks, particularly when the base rates of different groups are very different (Pessach and Shmueli, 2022).

[16] The difference between TPRs is called True Positive Parity; for FPRs are False Positive Parity (Pessach and Shmueli, 2022).

# Chapter 3

# Datasets

In this work, we evaluate our models on two benchmark toxicity datasets. The datasets were selected for their diverse sources and varied demographic representations, ensuring a comprehensive evaluation of our model. One of the essential factors in selecting these datasets is providing information about demographic groups each sample has targeted[1]. The benchmark datasets are HateXplain, and Wikipedia Comments Dataset: Personal Attack (Wiki Toxicity).

## 3.1 Hatexplain

HateXplain is a large-scale benchmark dataset for explainable automated hate speech detection (Mathew et al., 2021). It consists of over 20,000 posts from Gab and Twitter, with each post annotated from three different perspectives: a basic 3-class classification (hate, offensive, or normal), the target community (the community is the victim of hate speech or offensive speech in the post), and human-generated rationales (portions of the post that support the given label). In our work, we use HateXplain for toxicity text classification. Therefore, to classify samples as toxic or non-toxic, similar to (Baldini et al., 2022), we consider the hate and offensive labels as toxic and normal labels as non-toxic. However, HateXplain has some limitations; first, it lacks external contexts, such as profile bio, user gender, or post history, which might be useful for classification tasks. Additionally, it only

[1] Baldini et al. (2022) suggest research community add targeted demographic groups' information to datasets. This will make research on this topic more accessible.

focuses on the English language, ignoring multilingual hate speech. Despite these limitations, the dataset is a valuable resource for addressing various aspects of hate speech detection, including fairness and interpretability (Mathew et al., 2021).

### 3.1.1 Statistical Information

The HateXplain dataset consists of 20,148 posts collected from Twitter and Gab. The dataset is categorized into hateful, offensive, and normal. Table 3.1 shows the number of posts for each class on both platforms.

| **Class** | **Twitter** | **Gab** | **Total** |
|---|---|---|---|
| Hateful | 708 | 5,227 | 5,935 |
| Offensive | 2,328 | 3,152 | 5,480 |
| Normal | 5,770 | 2,044 | 7,814 |
| Undecided | 249 | 670 | 919 |
| **Total** | **9,055** | **11,093** | **20,148** |

Table 3.1: HateXplain dataset details. HateXplain includes instances where all three annotators assigned different labels to the same post. These instances are referred to as "Undecided" (Mathew et al., 2021). In this work similar to Baldini et al. (2022), since we employ this dataset in the toxicity classification task, and undecided cases have two labels related to toxicity (as we considered hate and offensive labels as toxic), we take undecided samples as toxic samples.

Table 3.1 shows that the number of hateful posts is higher on Gab compared to Twitter, while normal posts are more common on Twitter. The "Undecided" category represents cases where all three annotators assigned different labels to a post, making it difficult to determine the most appropriate class.

| **Group** | **Fine-grained annotation** |
|---|---|
| Race | African, Arab, Asian, Caucasian, Hispanic |
| Gender [2] | Men, Women, Heterosexual, Gay |

Table 3.2: The sensitive groups (coarse-grain on the left and fine-grained on the right) for the HateXplain dataset (Baldini et al., 2022; Mathew et al., 2021).

[2]Similar to Baldini et al. (2022), we chose to combine the categories of sexual orientation and gender,

Table 3.2 presents the sensitive groups considered in the HateXplain dataset and their corresponding fine-grained annotations. This information is relevant for understanding the different demographic groups targeted in the dataset. The groups include race, religion, and gender, with fine-grained annotations for specific ethnicities, religions, and gender categories. In our study, we collapse the fine-grained annotations during training into the coarse-level groups. However, we use fine-grained demographic information for evaluating the fairness of the studied models on the fine-grained terms using a test set to find an answer for research question $Q2$. Moreover, we ignore the information about religion when debiasing as it would introduce numerous additional fine-grained identity terms during evaluation, which would not contribute additional information.

| **Text** | **Label** | **Targets** |
|---|---|---|
| why is anna acting like she black | Normal | Race |
| like brazil many are descendants of west africa social constructs labels classifications have ruined a many people who falsely self identify as white or just do not even know abt their native land all b c of them bum ass caucazoids aka caucasians | Hate | Race |

Table 3.3: Samples from HateXplain dataset (Baldini et al., 2022).

Table 3.3 provides examples of posts from the dataset and their associated labels and target communities. These examples illustrate the different types of samples language targeted communities from race demographic groups.

| **Split** | **Total** | **Race** | **Gender** |
|---|---|---|---|
| Train | 15383 | 4765 | 2952 |
| Test | 1924 | 828 | 579 |

Table 3.4: HateXplain dataset statistics: sample counts per dataset split and sensitive group.

Table 3.4 presents the sample counts of the HateXplain dataset, categorized by dataset split and sensitive group, including Race, and Gender.

despite being aware that Heterosexual and Gay are not mutually exclusive with Men and Women. We believe that this approach does not compromise the integrity of our methodology and including them in the Gender category is necessary for promoting fairness and equity in our analysis.

## 3.2 WikiPedia Comments Dataset: Personal Attack (Wiki Toxicity)

Wiki Toxicity is a large-scale dataset for studying online personal attacks and toxicity in user-generated content. It includes over 100,000 high-quality human-labeled comments and 63 million machine-labeled ones from English Wikipedia, enabling researchers to analyze personal attacks. Each comment is annotated based on whether it is a personal attack or not. In addition, the dataset helps researchers to develop and evaluate ML toxicity text classifiers for identifying and mitigating toxic behavior in online discussions. The dataset was created through a combination of crowdsourcing and ML labeling to enable a quantitative, large-scale analysis of a vast corpus of online comments (Wulczyn et al., 2017).

Despite its comprehensive nature, the Wiki Toxicity dataset has some limitations. For example, it is solely focused on the English language, which may not address the challenges of multilingual toxicity detection. Furthermore, it does not include external context information, such as user profile data, post history, or targeted community [3], which could be useful for classification tasks. Nonetheless, the Wiki Toxicity dataset is valuable for understanding and addressing online personal attacks and toxicity (Wulczyn et al., 2017).

### 3.2.1 Statistical Information

The dataset used in this study consists of 115,864 comments, each labeled as either "attack" or "non-attack". The maximum length of the comments is 400 words, and they may include emojis and special characters. The dataset is derived from discussion comments in English from Wikipedia, with each comment annotated by approximately 10 annotators regarding whether it is a personal attack or not (Wulczyn et al., 2017).

Table 3.5 presents the statistics of the Wiki Toxicity Dataset, which is divided into two categories: Attacking Comments (AC) and Non-Attacking Comments (NC). Out of a total of 115,864 comments, 14,032 comments (12.11%) are labeled as personal attacks, while the remaining 101,832 comments (87.89%) are classified as non-attacks.

Table 3.6 presents the sensitive groups considered in this dataset and their corresponding fine-grained annotations. This information is relevant for understanding the different

[3]However, it does not include any information about targeted groups; some of the works that employed this dataset for mitigating bias came up with the same targeted groups as the jigsaw toxicity dataset to analyze the fairness in the classifiers (Dixon et al., 2018a; Zhang et al., 2020a).

| Set Name | Size | Fraction |
|---|---|---|
| Attacking Comments (AC) | 14,032 | 12.11% |
| Non-Attacking Comments (NC) | 101,832 | 87.89% |
| Global Comments (GC) | 115,864 | 100% |

Table 3.5: WikiPedia Comments Dataset Statistics (Wulczyn et al., 2017)

demographic groups targeted in the dataset. The groups include race and gender, with fine-grained annotations for specific ethnicities and gender categories.

| Group | Fine-grained annotation |
|---|---|
| Race | African, African American, Black, White, European<br>Hispanic, Latino, Latina, Latinx, Mexican, Canadian, American,<br>Asian, Indian, Middle Eastern, Chinese, Japanese |
| Gender | Lesbian, Gay, Bisexual, Transgender,Trans,<br>Queer, LGBT, LGBTQ, Homosexual,<br>Straight, Heterosexual, Male, Female, Nonbinary |

Table 3.6: The sensitive groups for the given race and gender terms with their corresponding fine-grained annotations Baldini et al. (2022).

Table 3.7 shows examples of posts from the wiki toxicity dataset and their associated labels and target communities. These examples illustrate the different types of samples (Non-attacking or Attacking) language targeting communities from gender demographic groups.

| Text | Label | Targets |
|---|---|---|
| In other words, you made it up that Gamaliel identifies as Male | Non-Attacking | Gender |
| I'm not homophobic or anything, but you are fucking gay. | Attacking | Gender |

Table 3.7: Examples from Wiki Toxicity dataset

Table 3.8 presents the sample counts of the Wiki Toxicity dataset, categorized by dataset split and sensitive group, including Race, and Gender.

| Split | Total | Race | Gender |
|---|---|---|---|
| Train | 95906 | 1667 | 2952 |
| Test | 31866 | 501 | 1011 |

Table 3.8: Wiki Toxicity dataset statistics: sample counts per dataset split and sensitive group.

## 3.3 Other Potential Datasets not Used In This Work

### 3.3.1 Jigsaw Toxicity

Jigsaw Toxicity [4] is a dataset created to study and mitigate negative online behaviors, such as toxic comments. The Conversation AI team developed this dataset, a research initiative founded by Jigsaw and Google (both part of Alphabet), to improve online conversation quality. The dataset contains comments from Wikipedia's talk page edits, which have been annotated by Jigsaw for toxicity, as well as various toxicity subtypes, including severe toxicity, obscenity, threatening language, insulting language, and identity-based hate. This dataset replicates the data released for the Jigsaw Toxic Comment Classification Challenge and Jigsaw Multilingual Toxic Comment Classification competition on Kaggle. In addition, the test dataset has been merged with the test labels released after the competition's conclusion (the original test dataset does not have any information about the target groups for each sample).

### 3.3.2 Reasons for not Using Jigsaw Toxicity Dataset

Although the Jigsaw Toxicity dataset is a valuable resource for studying and mitigating negative online behaviors, we decided not to use it in our study for the following reasons. First, the Jigsaw Toxicity dataset is derived from the same source as the Wiki Toxicity dataset, i.e., Wikipedia's talk page edits. As a result, both datasets share similarities in terms of the content and targeted groups. Using two similar datasets would not provide new insights or substantially improve our model's performance. Instead, we focused on the Wiki Toxicity dataset, which is already well-suited for our study of personal attacks and toxicity in online discussions. Second, the Jigsaw Toxicity dataset is significantly larger (around ten times larger) than the Wiki Toxicity dataset, which may introduce additional computational and memory requirements for our study. Third, the original test dataset

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

in the Jigsaw Toxicity dataset does not provide information about the targeted groups for each sample. This limitation restricts our ability to evaluate our model's performance on specific targeted groups, a crucial aspect of algorithmic fairness in toxicity detection. Given these three reasons, we decided not to use the Jigsaw Toxicity dataset in our study. Instead, we focused on the Wiki Toxicity dataset, which provides a comprehensive and manageable resource for understanding and mitigating bias in online personal attacks and toxicity detection.

### 3.3.3 GAP dataset

One of the popular datasets for hate speech and toxicity detection is the GAP (Generalized Abusive and Problematic) dataset. Gap aims to provide a large-scale, diverse, and representative sample of abusive and problematic content for researchers to develop state-of-the-art works in hate speech and toxicity detection. This dataset includes data from various online platforms, annotated by multiple annotators for different types of abusive language, including hate speech, offensive language, and targeted harassment (Kennedy et al., 2022).

### 3.3.4 Reasons for not Using GAP Dataset

The GAP dataset provides annotations for targeted communities, including race/ethnicity, religious identity, sexual orientation, gender identity, ideology, nationality, political identity, and mental/physical health (Kennedy et al., 2022). However, in our study, we chose not to use the GAP dataset for two reasons. First, the dataset describes which groups are targeted for around 11 coarse grain groups where gender identity and race are included in 0.9% and 3.4% of the total samples. Therefore, it makes it challenging to draw meaningful conclusions or build effective models to address hate speech against specific races or gender identities. Second, the dataset lacks fine-grained information about the targeted groups. For instance, it does not provide details about which specific race or gender has been targeted in the hate speech. The absence of such detailed information hinders the applicability of the dataset for our study, as we aim to analyze and address hate speech on a more granular level. Given these limitations, we decided not to use the GAP dataset in our study, as it does not provide the necessary depth and representation of data required for our research objectives (**?**).

### 3.3.5 Sexist Tweets

The Sexist Tweets dataset consists of 136,052 tweets collected over a period of 2 months. From all the collected tweets, 16,914 tweets were annotated, with 3,383 identified as sexist content sent by 613 users, 1,972 as racist content sent by nine users, and 11,559 labeled as neither sexist nor racist, sent by 614 users. The dataset aims to provide a realistic representation of hate speech by not balancing the data, as hate speech is a real but limited phenomenon (Waseem and Hovy, 2016a).

### 3.3.6 Reasons for not Using Sexist Tweets Dataset

However, we decided not to use the Sexist Tweets dataset for two reasons. First, unfortunately, due to Twitter's rules, some tweet IDs in the dataset have expired, making it impossible for us to collect the exact same dataset as the original ones used in (Park et al., 2018). This limitation reduces the amount of data accessible to us and makes it difficult to truly test our model's performance since we are not sure if the training and testing data would really represent the distribution we are looking for.

Previous works in toxicity classification have faced a significant challenge due to the lack of information about targeted communities in most available datasets. This limitation makes it difficult to study and address social biases in text classification tasks effectively. To the best of our knowledge, the datasets mentioned earlier are the primary candidates for this specific task. However, we believe that more comprehensive information is needed to advance the study of social bias in toxicity text classification. As a result, we encourage the research community to enrich existing datasets with additional information regarding targeted communities. In our future work, we plan to explore and develop reliable unsupervised methods to identify targeted communities in toxicity detection tasks, which will further contribute to addressing social biases in this field.

## 3.4 Problem Reformulation and Demographic Group Identification

In this work, we reformulate the fairness problem in toxicity classification differently from most previous works, inspired by recent work related to social fairness in toxicity text classification (Baldini et al., 2022). This new view of the problem consists of two main aspects:

1. **No privileged and non-privileged groups:** Unlike common fairness scenarios (for example, hiring), where decisions must be fair to groups based on characteristics like race or gender, we do not explicitly have privileged or unprivileged groups. For example, in terms of race in the hiring scenario, men are the privileged group who benefit more from the classifier while females are an unprivileged (under-represented) group where the system discriminates against them. However, in this work, we are doing something different than focusing on privileged or non-privileged groups. Instead, our goal is to make a model that can determine if a comment is toxic or not, without considering sensitive terms related to demographic groups in the decision-making process. To do this, we introduce two binary features for each sample, indicating whether a sample contains gender (1) or not (0), and if samples contain sensitive information related to race (1) information or not (0). In this way, we are focusing on the presence or absence of sensitive demographic terms, as the definition of unintended bias in toxicity text classification operates differently than in other hiring scenarios.

2. **Coarse demographic group identifications for adversarial debiasing process:** In the training phase, we debias the last hidden layer representation of the classifier to remove gender and race information based on coarse-grained information. For example, if a sample targeted women, the coarse-grained binary gender label for this sample is 1 and race is 0, and the adversarial network aims to predict if the last hidden layer representation of the classifier for this sample has gender and race information. Then based on the adversarial network loss on the prediction for this sample, we modify the weights of the classifier to not contain gender information. In this way, we debias the model based on the coarse-grained information related to gender or race, rather than fine-grained ones. This means that the adversarial debiasing network predicts if each sample contains gender and race information instead of predicting specific fine-grained terms like "women" or "white" (if we debias the classifier on fine-grained terms, the adversarial debiasing network must predict a binary label for each fine-grained terms available in each dataset). This approach has two main advantages:

    - Enough training data for all the fine-grained terms may not be available in the training data which would make it difficult for the model to avoid overgeneralizing (overfitting) on these terms (also, test sets may not contain enough samples for each sensitive attribute, which make it difficult to evaluate fairness for all the fine-grained terms in the benchmark datasets).
    - Fine-grained terms referring to particular demographic groups are diverse and

subject to change over time. Adapting the model to new cultural and societal changes would be infeasible if we debiased the model based on every single fine-grained term.

Therefore, by reformulating the problem in this manner, we can develop an approach that better addresses fairness concerns in toxicity classification without explicitly focusing on privileged or non-privileged groups while leveraging coarse demographic group information to guide the adversarial process. In the next chapter, we outline our proposed models and introduce baselines in detail.

# Chapter 4

# Methodology

## 4.1 Overview

In this chapter, we present the methodology and the step-by-step process to develop the proposed fair semi-supervised generative-based framework. This framework aims to enforce social fairness in semi-supervised toxicity text classification to mitigate bias without significant loss in classification performance. For clarity and completeness, we also outline the baselines to which we compare our models and discuss the technical aspects underlying our experimental setup. Finally, we introduce the evaluation metrics employed in this work at the end of this chapter.

## 4.2 Proposed Framework

In this section, we present our proposed fair semi-supervised framework and briefly introduce the semi-supervised and supervised baselines. Our proposed framework consists of two stages. First, we train a semi-supervised generative-based text classifier (either NDA-GAN or GANBERT). Second, we employ an adversarial debiasing approach to mitigate bias and enhance fairness in the trained classifier.

It is essential to mention that due to the high similarity in design between FairNDAGAN and FairGANBERT (the only distinction being the use of the NDA process in the first stage for FairNDAGAN), we will only present FairNDAGAN to introduce the proposed framework in detail, to avoid redundancy.

### 4.2.1 FairNDAGAN

In this section, we introduce all the components of FairNDAGAN in detail, and at the end, we put all the components together to propose the final fair semi-supervised framework. First, we provide some background information for the models we build on top of.

#### 4.2.1.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) is a powerful transformer-based pre-trained model introduced by Devlin et al. (2018) for a wide range of NLP tasks. BERT is pre-trained on a large corpus of text data, learning contextual representations of words in a sentence from both left-to-right and right-to-left directions, which allows it to capture the full context of words more effectively. BERT has achieved state-of-the-art results in various NLP tasks, including sentiment analysis, named entity recognition, and question-answering. In addition, it has become a popular choice for fine-tuning downstream tasks due to its ability to transfer knowledge from the pre-trained to the task-specific models, reducing the training data and time required for the target task (Devlin et al., 2018).

However, BERT is computationally expensive and has many parameters, making it challenging to deploy in real-world scenarios with limited computational resources. In order to address these challenges, we employed a distilled version of BERT called DistilBERT. Therefore, we introduce DistilBERT first and then identify how it can effectively be used in the text classification task (Sanh et al., 2019).

#### 4.2.1.2 DistilBERT

DistilBERT is a smaller and faster version of BERT, with only 66 million parameters compared to BERT's 110 million parameters, introduced by Sanh et al. (2019). It has been designed to maintain most of the original model's performance while significantly reducing its size and computational complexity. The distillation process involves training a smaller model, DistilBERT, to simulate the behaviour of the larger, more complex BERT model. This is achieved by transferring the knowledge from the teacher model (BERT) to the student model (DistilBERT) using knowledge distillation (Sanh et al., 2019). DistilBERT has approximately cut down the number of parameters in BERT in half. Therefore, it makes it more efficient in terms of memory and computational requirements. In addition, the reduced size and complexity of DistilBERT allow it to be more easily deployed when limited computational resources in real-world scenarios are accessible (Sanh et al., 2019).

On the other side, the shift from BERT to DistilBERT causes a slight performance drop [1], posing a trade-off that may be acceptable for some applications but possibly unsatisfactory for those that need the highest performance.

Finally, BERT is preferred when you need top accuracy, and when computational resources and time are not a primary concern. Conversely, DistilBERT is employed in scenarios where computational resources are limited. It's also a better choice when you need quick results, as it is faster [2] than BERT (Sanh et al., 2019).

In our version of the NDAGAN model, unlike the original one introduced in our previous work[3], we make a key modification: we use DistilBERT as the encoder. This change is also applied to the original GANBERT architecture as outlined by Croce et al. (2020)[4]. By utilizing DistilBERT, we can accelerate the training process and lower computational demands, while ensuring high performance for the task of toxicity text classification.

In order to use DistilBERT for various NLP tasks such as text classification, we need to tokenize [5] the input text using the DistilBERT tokenizer. According to (Su et al., 2021), tokenization in DistilBERT involves the following steps:

1. **Text normalization**: The input text is lowered case and converted into Unicode format.

2. **WordPiece tokenization**: The text is tokenized into subword units using the WordPiece tokenizer. This process helps the model handle out-of-vocabulary words by breaking them into smaller, known subword units.

3. **Adding special tokens**: The [CLS] token is added to the beginning of the tokenized sequence, and the [SEP] token is added to the end. These tokens help the model distinguish between different input sequence types and identify a sentence's start and end.

[1] In fact, it scores 97% of what BERT does on the GLUE language understanding benchmark (Sanh et al., 2019)

[2] DistilBERT achieves a 60% speed improvement compared to BERT. In (Sanh et al., 2019), the evaluation was conducted on the STSB development set, employing a batch size of 1, and executed on a CPU (Intel Xeon E5-2690 v3 Haswell @2.9GHz).

[3] Refer to (Shayesteh and Inkpen, 2022) for the original NDAGAN model.

[4] In the original GANBERT architecture, a BERT encoder was used.

[5] Tokenization is the process of breaking down a text into smaller units called tokens, a common preprocessing step in NLP tasks to convert raw text data into a format that ML models can easily use (Webster and Kit, 1992).

4. **Padding and truncation**: The input sequences are either padded with [PAD] tokens or truncated to a fixed length, depending on the maximum sequence length the model allows. This step ensures that all input sequences have the same length, which is required for efficient batch processing.

#### DistilBERT for Classification Tasks

DistilBERT can be fine-tuned for various classification tasks by adding a fully connected layer on top of the pre-trained model. First, the input sequence is tokenized and fed into the DistilBERT model in a classification task. The unique token [CLS] is added at the beginning of the input sequence, and the token [SEP] is inserted at the end. The [CLS] token's output representation is used as a pooled representation of the entire input sequence (Su et al., 2021).

Finally, by leveraging the [CLS] token's representation, the DistilBERT model can effectively capture the global semantic information in the input text, enabling accurate toxicity classification. This approach allows for better generalization and performance compared to traditional text classification methods that rely on features like bag-of-words or n-grams, which do not capture the contextual information of the input text (Su et al., 2021).

#### 4.2.1.3 SS-GANs with Feature Matching

Semi-Supervised Generative Adversarial Networks (SS-GANs) (Salimans et al., 2016) have emerged as a powerful model for classification tasks, particularly when labeled data is limited. SS-GANs consist of two components: a generator and a discriminator. The generator learns to generate fake data samples. In contrast, the discriminator learns to classify the real and fake data into their respective classes and to distinguish between real and generated data.

In the SS-GANs, the primary purpose of the generator is not to learn the true data distribution but rather to generate samples that help the discriminator better classify real data into their respective classes. This insight was highlighted in the Dai et al. (2017). According to the authors, a "bad" generator that does not perfectly learn the true data distribution can be useful for semi-supervised learning tasks, as it can produce samples near the decision boundary. Therefore, it helps the discriminator to have a better knowledge of out-of-distribution areas in a given task which leads to a better generalization performance of a classifier (Salimans et al., 2016).

One of the challenges encountered in training GANs is mode collapse (Salimans et al., 2016), where the generator produces samples from only a few modes of the true data distribution, resulting in a limited diversity of generated samples. To address the mentioned issue, feature matching is introduced as a regularization technique in the SS-GAN (Salimans et al., 2016). The main idea behind feature matching is to encourage the generator to produce samples that match the real data's feature statistics, making it more challenging for the discriminator to distinguish between real and generated samples. Also, it is important to note that in the "bad" generator paradigm for SS-GANs, feature matching helps the generator produces more diverse out-of-distribution samples, which is helpful for the discriminator to generalize the problem more effectively (Dai et al., 2017).

In the SS-GAN with feature matching introduced by Salimans et al. (2016), the objective function can be expressed as:

$$\min_G \max_D L(D, G) \tag{4.1}$$

where $L(D, G)$ comprises three loss components:

$$L(D, G) = L_{D_{\text{sup}}} + L_{D_{\text{unsup}}} + L_G. \tag{4.2}$$

The three loss components are defined as follows:

- Discriminator supervised loss ($L_{D_{\text{sup}}}$) for labeled data, which corresponds to the correct classification of real data samples by the discriminator ($D$):

$$L_{D_{\text{sup}}} = -E_{x,y\sim p_{data}(x,y)}[\log D(y|x, y < K + 1)], \tag{4.3}$$

  where $x$ represents the samples, $K$ is the number of labels, $y$ is the classification label with $y \in [1, K + 1)$, and $p_{data}(x, y)$ denotes the true data distribution.

- Discriminator unsupervised loss ($L_{D_{\text{unsup}}}$), corresponding to the correct classification of real unlabeled data samples as not being fake by the discriminator ($D$):

$$L_{D_{\text{unsup}}} = -E_{x\sim p_{data}(x)}[\log(1 - D(y = K + 1|x))] - E_{x\sim G}[1 - \log(1 - D(y = K + 1|\hat{x}))] \tag{4.4}$$

  In this case, $x$ represents the samples, $\hat{x}$ represents the fake samples, $K + 1$ denotes the labels for data being fake (1) or real (0).

- The generator loss ($L_G$), includes two terms, the generator unsupervised loss $L_{G_{\text{unsup}}}$, which is the probability of the discriminator assigning the "fake" label $(K+1)$ to the generated samples, and the feature matching loss $\mathcal{L}_{FM}(G)$, which encourages the generator to produce samples that match the true data distribution more closely:

$$L_G = L_{G_{\text{unsup}}} + \alpha \mathcal{L}_{FM}(G), \tag{4.5}$$

where $L_{G_{\text{unsup}}}$ and $\mathcal{L}_{FM}(G)$ are defined as:

$$L_{G_{\text{unsup}}} = -E_{x\sim G}[\log D(y = K+1|x)] \tag{4.6}$$

$$\mathcal{L}FM(G) = ||E_{x\sim pdata(x)}(D_h(x)) - E_{\hat{x}\sim G}(D_h(\hat{x}))||_2^2. \tag{4.7}$$

In this context, $x$ denotes the real sample, $\hat{x}$ represents the fake sample, and $D_h(.)$ refers to the last hidden layer of the discriminator.

In SS-GANs with feature matching, the generator, and discriminator are trained alternatively to optimize their respective objectives. The generator learns to produce diverse and informative fake samples, while the discriminator learns to classify real data into their respective classes and to distinguish between real and generated data. This process improves classification performance, especially when the labeled data is limited (Salimans et al., 2016).

The training process for SS-GANs consists of two main steps, training the discriminator and training the generator. These steps are alternated during the training process, with each step aiming to improve the performance of the respective model (discriminator or generator) while the other remains fixed (Salimans et al., 2016). we can summarize the training process of the discriminator and the generator as follows:

1. **Training the Discriminator:**
   (a) Sample a batch of real data $(x)$ and noise samples $(z)$.
   (b) Generate fake data samples $(\hat{x} = G(z))$ using the generator $(G)$ and the input noise samples.
   (c) Train the discriminator $(D)$ to classify the real data into their respective classes and to distinguish between real and generated data. Update the discriminator's weights accordingly.

2. **Training the Generator:**

   (a) Sample a new batch of noise samples ($z$).

   (b) Generate fake data samples ($\hat{x}$) using the generator and the input noise samples.

   (c) Compute the feature activations of the real data ($D_h(x)$) and the generated data ($D_h(\hat{x})$) in the discriminator's intermediate layers.

   (d) Calculate the L2 distance between the mean feature activations of the real data and the mean feature activations of the generated data using the feature matching loss $\mathcal{L}_{FM}(G)$.

   (e) Compute the $L_{G_{\text{unsup}}}$ on discriminator's predictions on fake samples to distinguish between real and generated data.

   (f) Update the generator's weights to minimize $L_G$.

Finally, in the text classification task, Croce et al. (2020) introduce GANBERT that strictly follows SS-GANs with feature matching architecture, however, they employ BERT as the encoder to convert input text into contextual representation, GANBERT architecture is shown in 4.1. In GANBERT, the main idea behind feature matching is to encourage the generator to produce diverse out-of-distribution samples, making it more challenging for the discriminator to distinguish between real and generated samples (Croce et al., 2020). This leads to more accurate and robust training, as the generated samples help the discriminator learn better decision boundaries between different classes (Salimans et al., 2016). However, it has been identified that GANBERT can still suffer from mode collapse (Dai et al., 2017). To further mitigate mode collapse and its negative effect on classification performance, the NDAGAN (Shayesteh and Inkpen, 2022) introduced the NDA process, which will be discussed in detail in the next section.

#### 4.2.1.4 NDAGAN

NDAGAN (Shayesteh and Inkpen, 2022) is a novel semi-supervised generative adversarial learning model for text classification tasks. NDAGAN leverages the negative data augmentation (NDA) technique (Sinha et al., 2021) to improve the performance of existing semi-supervised generative adversarial networks (SS-GANs) in the text classification task. Specifically, NDAGAN is designed to address the mode collapse issues encountered in SS-GANs with feature matching such as GANBERT.

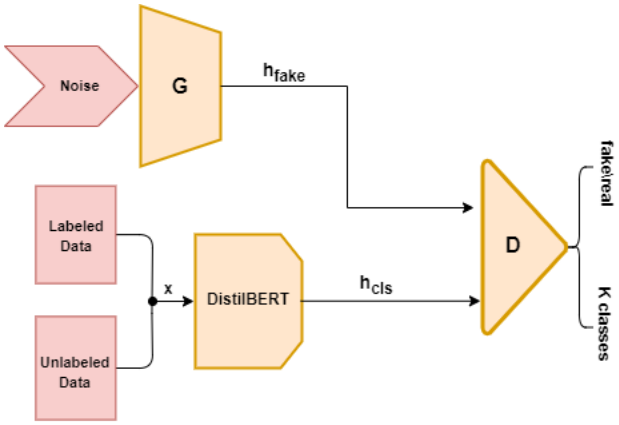



Figure 4.1: GANBERT architecture (Croce et al., 2020) consists of three components, DistilBERT, Generator, Discriminators.

The main objective of NDAGAN is to learn to classify limited labeled data through supervised loss while generalizing the data distribution by distinguishing between unlabeled and NDA synthetic data using unsupervised loss. In the following section, before introducing FairNDAGAN, we first reviewed each part of the NDAGAN, consisting of DistilBERT and SS-GAN architecture and NDA technique, then we will introduce FairNDAGAN as a whole.

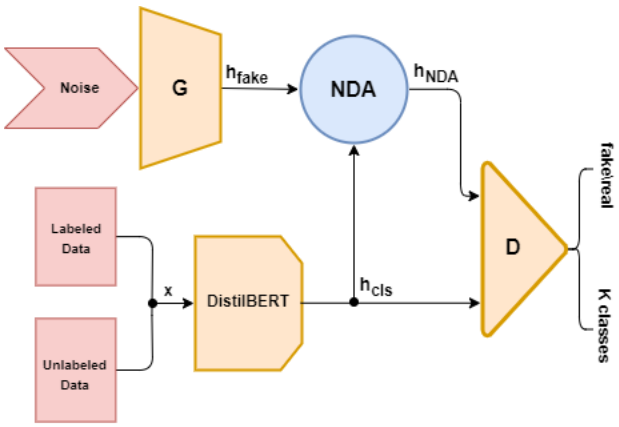



Figure 4.2: NDAGAN architecture (Shayesteh and Inkpen, 2022) consists of four components, DistilBERT, Generator, Discriminator, and NDA process.

The NDAGAN architecture (Shayesteh and Inkpen, 2022) illustrated in Figure 4.2 consists of four main components: a generator and a discriminator similar to SS-GANs, an NDA process, and a text encoder (DistilBERT). The model leverages DistilBERT as the text encoder and SS-GANs with feature matching as the underlying semi-supervised learning architecture. In addition, the NDA process is added to generate NDA synthetic samples. In this section, we describe the overall architecture of NDAGAN, its advantages over GANBERT, and how these components are integrated.

The main components of NDAGAN are as follows:

1. **Encoder:** The DistilBERT encoder is utilized as the text encoder in the NDA-GAN architecture to capture the contextual information of the input text. When processing input text, DistilBERT generates contextualized token embeddings for each input token, with the first token in every input sequence being the special CLS token (classification token). After sending the input text into the DistilBERT encoder, the contextualized embeddings corresponding to their respective CLS tokens are extracted and serve as inputs for the discriminator and NDA generator.

2. **Generator:** The generator in NDAGAN is a neural network that takes 100-dimensional noise samples as input and generates synthetic data samples as output. Its architecture includes two fully connected layers, each with a hidden size of 512 and followed by a LeakyReLU activation function and a dropout layer to learn complex patterns and prevent overfitting. Finally, a linear output layer generates a fake contextual representation of the data distribution for the NDA process.

   The (NDA) process in NDAGAN is designed to enrich the synthetic sample diversity by adding local feature information of data distribution during the Mixup method to guide the discriminator in learning areas close to the real data distribution. The operation receives as input the generator's output and the CLS token from the encoder's output, both being 768-dimensional vectors. These inputs are processed using the Mixup technique (Sun et al., 2020), creating synthetic samples that are linear interpolations of the generator's output and the encoder's CLS token. Mathematically, the Mixup process can be formulated as:

$$h_{NDA} = h_{fake} + (1 - \lambda)h_{cls} \tag{4.8}$$

   where $h_{fake}$ and $h_{cls}$ are pairs of the generator's output and the encoder's CLS token, respectively, and $\lambda$ is a hyper-parameter set to 0.85 in our experience.

This mixing makes the NDA samples less likely to reside in high-density regions [6], and due to blending with real data representations, these samples retain a reasonable closeness to the data manifold. Consequently, the introduction of NDA samples helps the discriminator's capability to identify the optimal decision boundary, thereby improving the overall classification performance (Shayesteh and Inkpen, 2022).

3. **Discriminator:** The discriminator is a neural network that takes both real and NDA synthetic data samples as input and classifies real label data into their respective classes, as well as distinguishes between real unlabeled and NDA synthetic data. In NDAGAN, the discriminator consists of two layers, each with a hidden size of 512, and a LeakyReLU activation function and dropout layer for regularization follow these layers. Finally, the model ends with a softmax layer that maps the 512-dimensional hidden space to the desired number of classes.

In the NDAGAN similar to SS-GANs, the objective function can be expressed as:

$$\min_{G'} \max_{D} L(D, G') \tag{4.9}$$

where $G'$ describes the NDA process and we can formulate it as:

$$G' = \lambda G + (1 - \lambda)\text{enc(x)}. \tag{4.10}$$

In this context, $G$ is the generator network, and enc(x) represents the contextualized representation of the [CLS] token in DistilBERT. It is important to note that $G'$ emphasizes that both the generator and encoder are being optimized during the training as a part of the NDA process.

The $L(D, G')$ comprises three loss components:

$$L(D, G') = L_{D_{\text{sup}}} + L_{D_{\text{unsup}}} + L_{G'}. \tag{4.11}$$

The three loss components are defined as follows:

- Discriminator supervised loss ($L_{D_{\text{sup}}}$) for labeled data, which corresponds to the correct classification of real data samples by the discriminator ($D$):

[6] Dai et al. (2017) claims this can happen as feature matching process imposes the generator to follow the real data distribution while in SS-GANs generator should generate complementary synthetic data.

$$L_{D_{\text{sup}}} = -E_{x,y\sim p_{data}(x,y)}[\log D(y|\text{enc(x)}, y < K+1)], \tag{4.12}$$

where $K$ is the number of labels for the main text classification task, $y$ is the classification label with $y \in [1, K+1)$, and $p_{data}(x, y)$ denotes the true data distribution for label data.

- Discriminator unsupervised loss ($L_{D_{\text{unsup}}}$) for real unlabeled data, corresponding to the correct classification of real data samples as not being fake by the discriminator ($D$):

$$\begin{aligned} L_{D_{\text{unsup}}} = &-E_{x\sim p_{data}(x)}[\log(1 - D(y = K+1|\text{enc(x)}))] \\ &-E_{h_{NDA}\sim G'}[1 - \log(1 - D(y = K+1|h_{NDA}))] \end{aligned} \tag{4.13}$$

  In this case, $h_{NDA}$ represent the NDA synthetic samples, $K+1$ denotes the labels for data being fake (1) or real (0), $p_{data}(x)$ represents the true data distribution for unlabeled data, and $D$ denotes the discriminator.

- The total NDA loss ($L_G{}'$) includes two terms: the NDA process loss $L_{G'_{\text{unsup}}}$, which is the probability of the discriminator assigning the "fake" label ($K+1$) to the NDA fake samples, and the feature matching loss $\mathcal{L}_{FM}(G')$, which encourages the generator to produce samples that match the true data distribution more closely:

$$L_{G'} = L_{G'_{\text{unsup}}} + \alpha\mathcal{L}_{FM}(G'), \tag{4.14}$$

  where $L_{G'_{\text{unsup}}}$ is defined as:

$$L_{G'_{\text{unsup}}} = -E_{h_{NDA}\sim G'}[\log D(y = K+1|h_{NDA})] \tag{4.15}$$

  The feature matching loss $\mathcal{L}_{FM}(G')$ is given by:

$$\mathcal{L}_{FM}(G') = ||E_{x\sim pdata(x)}(D_h(x)) - E_{h_{NDA}\sim G'}(D_h(h_{NDA}))||_2^2. \tag{4.16}$$

In this context, $D_h(.)$ refers to the last hidden layer of the discriminator.

The NDAGAN objective function separates the discriminator's and the NDA process's (G') goals, providing a model that includes the NDA mixing process and the feature matching loss to improve the model's performance.

The key contributions and advantages of NDAGAN over GANBERT (the other semi-supervised baseline) in the text classification task that was previously studied in (Shayesteh and Inkpen, 2022), include the following:

- The first adaptation of the NDA technique in the text classification task.
- Leveraging the NDA technique to enhance the performance of GANBERT by providing more informative synthetic samples that capture the local data distribution structure.
- A departure from the original NDA technique introduced in (Sinha et al., 2021), NDAGAN does not apply a non-label preserving augmentation method to the real data before mixing them with the generator samples. This decision ensures the discriminator is trained on informative NDA samples that contain useful local data distribution structures, allowing the model to learn a more optimal boundary between low-density regions and data manifolds.

In summary, NDAGAN is an innovative semi-supervised generative adversarial learning model used for text classification. This model leverages the NDA technique to enhance the performance of existing SS-GAN models while addressing their limitations. In this work, we leverage this model in the toxicity text classification as a part of our proposed model FairNDAGAN to analyze the performance of its accuracy and fairness performance in toxicity text detection tasks.

#### 4.2.1.5 FairNDAGAN: adversarial debiasing for mitigating bias in NDAGAN

The goal of adversarial debiasing is to reduce the influence of certain unintended biases or sensitive attributes in the decision-making process of a model while maintaining accuracy in an acceptable range Zhang et al. (2018a). To achieve this, an additional adversarial component, called the adversarial network, is introduced to the NDAGAN architecture. The adversarial network attempts to predict the sensitive attribute based on the last hidden layer of the discriminator, while the discriminator attempts to minimize the adversary's ability to predict the sensitive attribute.

Given this information, the optimization formula for FairNDAGAN can be represented as a min-max optimization process. We consider an optimization problem with the goal of reducing unintended biases in the model by minimizing the classification loss and fooling

the adversarial network to detect demographic information in the data representation. The initial min-max objective function can be represented as:

$$\min_{\theta_{\text{NDAGAN}}} \max_{\theta_A} L_{\text{total}}(\theta_{\text{NDAGAN}}, \theta_A) \tag{4.17}$$

where

$$L_{\text{total}}(\theta_{\text{NDAGAN}}, \theta_A) = L_{\text{NDAGAN}}(D, G^{'}) - \lambda \cdot L_{\text{adv}}(O, \hat{O}; \theta_A) \tag{4.18}$$

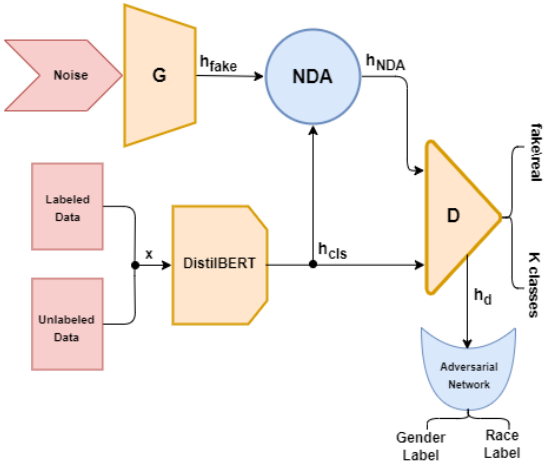
Here, $L_{\text{NDAGAN}}(D, G^{'})$ is the NDAGAN loss function given in Formula (4.11). For the adversarial training loss, $\lambda$ (the adversarial decay factor) is a hyperparameter that controls the trade-off between the NDA loss and the adversarial loss (trade-off between fairness and accuracy). In addition, $O$ and $\hat{O}$ are sets of multi-label related to gender and race for the multi-label binary classification task in the adversarial network. In this problem, the adversarial network's aim is to predict a set of labels $\hat{O}$ to predict if the input contains gender and race information. Therefore, given a set of ground truth label $O$, $L_{\text{adv}}$ compute the cross-entropy binary loss of each gender and race for the adversarial network. In this way, we can formulate $O$ ($\hat{O}$ is similar to $O$) as follows:

$$O = o_{\text{gender}} \times o_{\text{race}} \tag{4.19}$$

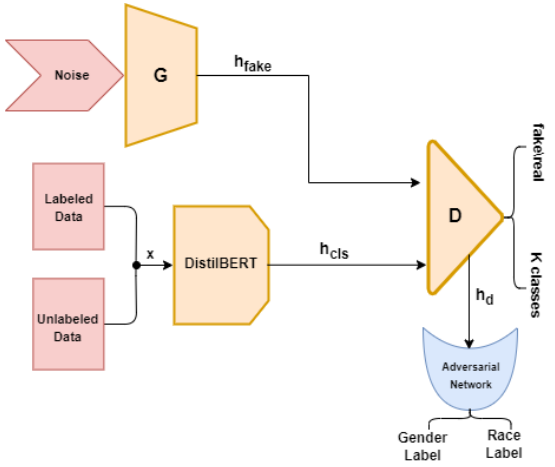
where,

$$o_{\text{gender}} \in [0, 1], o_{\text{race}} \in [0, 1] \tag{4.20}$$

However, to understand the FairNDAGAN, we must define how we incorporate the adversarial process in NDAGAN architecture and train both networks in an in-processing bias mitigating process. Based on our knowledge, this is the first work that aims to use adversarial debiasing in SS-GANs architecture to provide a fair semi-supervised framework for toxicity text classification.




(a) Pre-training representation of FairNDA-GAN




(b) Pre-training representation of FairGAN-BERT.

Figure 4.3: Pre-training representation of FairNDAGAN and FairGANBERT. The only difference between these two models is the presence of the NDA process (NDA block in 4.3 (a)).

The training procedure of FairNDAGAN consists of two main phases, pre-training, and post-training. In the pre-training stage (Figure 4.3), in each epoch, we train the NDAGAN first, and then the adversarial network is trained on the last hidden layer of the discriminator ($h_d$) while the NDAGAN weights are fixed. In the post-training stage (Figure 4.4), first, we remove the generator as it does not help the NDAGAN in the debiasing process. Then, we call the remaining poison classifier (encoder + discriminator). In this stage, the adversarial network and poison classifier are trained simultaneously in alternating epochs to minimize the influence of unintended biases related to gender and race in the poison classifier. Also, it is worth noting that to select the fair model in the post-training process; we designed a fair model selection criteria by considering accuracy and fairness trade-offs. Here, we introduce the fair model selection criterion and then detail the training process after.

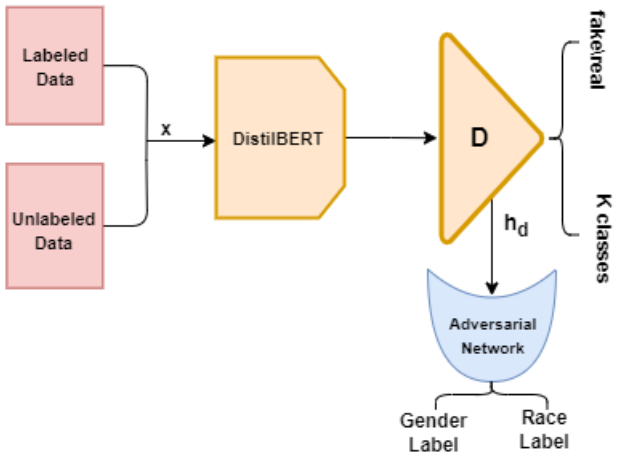



Figure 4.4: The post-training representation of FairNDAGAN and FairGANBERT. In the debiasing post-training stage, we debias the $h_d$ representation of the poison classifier on the total loss ($L_{\text{total}}(\theta_{\text{NDAGAN}}, \theta_A)$) to improve the fairness in the model.

In order to select the fairest model in the post-training stage that consider the trade-off between accuracy and fairness, we introduce a fair model selection criterion. This criterion aims to select a model with the lowest equalized odds difference (EOD)[7] for gender and race demographic groups while considering a maximum loss of accuracy as a trade-off to gain fairness while losing accuracy performance.

Before delving into details, we first need to define the concept of scaled maximum loss of accuracy to ensure the trade-off between accuracy and fairness is acceptable in the SSL task. The scaled maximum loss of accuracy is defined as a scaled-down of the maximum acceptable loss of accuracy for a fully supervised scenario where the classifier can achieve near-perfect accuracy. The idea behind scaled maximum loss of accuracy is that models' accuracy in semi-supervised learning is usually lower than in fully supervised cases due to the lack of a large number of label data. Therefore, we need to scale down the maximum acceptable loss for supervised cases to keep SSL's accuracy and fairness trade-off in an acceptable range. Therefore, the scaled maximum loss of accuracy is defined as:

$$L_{\text{scaled}} = 0.01 \times L_{\text{max}} \times A_{\text{SS-Best}} \tag{4.21}$$

where:

[7]EOD is introduced in detail in section 4.5.3

- $L_{\text{scaled}}$: Scaled maximum loss of accuracy
- $L_{\text{max}}$: Maximum loss of accuracy [8]
- $A_{\text{SS-Best}}$: The maximum accuracy achieved by the NDAGAN (GANBERT) during the pre-training stage in the semi-supervised training.

Next, we convert the EOD for race and gender into percentages to create a more intuitive unified selection criterion. Given that the minimum equalized odds difference is 0 when all subgroups are treated equally and the maximum is 2 when the absolute sum of true positive and false positive parity is 1 each, then we can define the percentage of equalized odds difference as follows:

$$EOD\% = 100 - 100 \times \frac{EOD}{2} \tag{4.22}$$

where $EOD\%$ and $EOD$ represent the percentage of equalized odds difference and equalized odds difference respectively. It is important to note that a smaller EOD means a fairer classifier while a larger percentage means a fairer classifier for $EOD\%$.

After calculating the $EOD\%$ for race and gender for each iteration in the post-training stage, we define the selection criterion as follows [9] :

$$\begin{aligned} &1. \quad |A_{\text{SS-Best}} - A_{\text{epoch}}| \leq L_{\text{scaled}} \\ &2. \quad EOD\%_{\text{Race}} + EOD\%_{\text{Gender}} \ < EOD\%^{\text{epoch}}_{\text{Race}} + EOD\%^{\text{epoch}}_{\text{Gender}} \end{aligned} \tag{4.23}$$

where:

- $A_{\text{SS-Best}}$: The maximum accuracy achieved by the NDAGAN during the pre-training stage.
- $A_{\text{epoch}}$: The accuracy of the current epoch in the post-training stage.
- $L_{\text{scaled}}$: Scaled Maximum Loss of Accuracy.
- $EOD\%_{\text{Race}}$: The best percentage of equalized odds difference for race.

[8] which is a hyper-parameter set by us to 7 percent and it can be set based on accuracy and fairness trade-off needed for a specific application.

[9] The selection criterion is satisfied when both conditions (1) and (2) are met. Also, $EOD\%_{\text{Gender}}$ and $EOD\%_{\text{Race}}$ are calculated based on the model performance on the same test set.

- $EOD\%_{\text{Gender}}$: The best percentage of equalized odds difference for gender.
- $EOD\%_{\text{Race}}^{\text{epoch}}$: The percentage of equalized odds difference for race at the current epoch.
- $EOD\%_{\text{Gender}}^{\text{epoch}}$: The percentage of equalized odds difference for gender at the current epoch.

This criterion ensures that the selected model considers an optimal balance between accuracy and fairness by including both conditions in the selection process. Finally, to have a clear view of the training process of FairNDAGAN (adversarial debiasing along with NDAGAN), we can separate the pre-training (Figure 4.3) and post-training (Figure 4.4) processes.

The pre-training phase is as follows:

1. Train the semi-supervised classifier (NDAGAN) on the labeled and unlabeled data using $L_{\text{NDAGAN}}(D, G')$.
2. Fix the weights of the NDAGAN and feed the output of the last hidden layer of the discriminator ($h_D$) to the adversarial network.
3. Train the adversarial network to predict if the $h_D$ contains information related to gender and race using binary multi-label classification. The training is performed on the adversarial network prediction loss ($L_A$) [10], with the weights of the NDAGAN fixed.
4. Continues this process for several epochs and selects the adversarial network and best NDAGAN at the epoch where the semi-supervised classifier has the best accuracy.

Also, the post-training phase can be categorized as:

1. Separate the generator and NDA process from the architecture and use the encoder and discriminator as a poison classifier for binary classification of the toxic text. However, since we use unlabeled data in this process, the poison classifier utilized unsupervised loss, the same as Formula 4.12 for unlabeled data.
2. Perform T iterations of simultaneous training for the adversarial and classifier networks:

[10] $L_A$ is the sum of binary cross-entropy losses for gender and race labels.

a) For each iteration, train the adversarial network for a single epoch while keeping the poison classifier fixed.

b) Train the poison classifier on a randomly sampled mini-batch while keeping the adversarial network fixed. However, in this stage, we consider the total loss ($L_{\text{total}}(\theta_{\text{NDAGAN}}, \theta_A)$) to adjust the weights of the poison classifier in order to improve fairness.

3. Select the fairest classifier based on the defined fair model selection criteria.

### 4.2.2 FairGANBERT

FairGANBERT is a proposed model that shares the same architecture and training steps as FairNDAGAN, with the exception that it does not utilize the NDA process. In all other respects, the models are identical to FairNDAGAN.

## 4.3 Baselines

In order to evaluate the performance of FairNDAGAN and FairGANBERT, we compare them against two primary baselines, GANBERT and NDABERT. Both baselines are semi-supervised learning classifiers for text classification that share the same architectures with respect to each other but differ in specific aspects. We discussed the detail of NDABERT and its difference from GANBERT as part of our proposed models in previous sections; however, as our semi-supervised baselines, we deployed them as below:

**GANBERT:** GANBERT serves as one of our primary semi-supervised baselines. It has the same architecture as NDAGAN but does not employ the NDA process during the training (Figure 4.1). In GANBERT, the generator and discriminator follow the same structure as in NDAGAN, but the generator's fake samples are not transformed using the NDA technique. This results in a more straightforward generator-discriminator interaction, which may be less effective in capturing the local data distribution structure. The details architecture of GANBERT is the same as NDAGAN which is outlined in section 4.2.1.4.

**NDAGAN:** The architecture of NDAGAN is described as part of our proposed fair model, FairNDAGAN in section 4.2.1.4. Therefore, no additional information is provided here to avoid duplication of content.

In addition to the semi-supervised baselines, we also compare the performance of FairNDAGAN and FairGANBERT against supervised learning baselines to evaluate their

accuracy and fairness in the context of the toxicity text classification task. We consider two main supervised baselines: DistilBERT trained on limited labeled data (BERT) and DistilBERT trained on full labeled data (FullBERT), along with their fair versions (FairBERT and FairFullBERT, respectively).

All supervised cases in our study utilize the DistilBERT for sequence classification architecture, which involves adding a simple linear classification layer to the final layer of the pre-trained transformer. This layer takes the contextualized representation (CLS token) of the input sequence generated by the transformer as input and produces a probability distribution over the possible classes for the given task.

| | **Hatexplain** | | | **Wiki Toxicity** | | |
|---|---|---|---|---|---|---|
| **Models** | $LR_{\text{Adversary}}$ | $LR_{\text{Classifier}}$ | $\lambda$ | $LR_{\text{Adversary}}$ | $LR_{\text{Classifier}}$ | $\lambda$ |
| BERT | - | 2e-5 | - | - | 5e-5 | - |
| FairBERT | 1e-4 | 2e-5 | [5,7] | 1e-4 | 2e-5 | [10,12] |
| FullBERT | - | 2e-5 | - | - | 5e-5 | - |
| FairFullBERT | 1e-4 | 2e-5 | [15,15] | 1e-4 | 2e-5 | [20,25] |

Table 4.1: Supervised models' Hyper-parameters for each dataset are set as described in this table.

**BERT:** This baseline represents a supervised learning model trained on limited labeled data. It uses the DistilBERT architecture for text classification, which is a lighter and faster version of the original BERT model, optimized for efficiency while maintaining strong performance. The goal of comparing our proposed framework and semi-supervised baselines against this baseline is to evaluate the effectiveness of semi-supervised learning in scenarios where labeled data is scarce.

**FairBERT:** This is the fair version of the BERT, which incorporates adversarial training to minimize the influence of sensitive attributes in the classification decisions. This allows us to assess the impact of adversarial debiasing on fairness and the overall performance of the model in a supervised setting with limited labeled data. The adversarial debiasing network in this model follows the same architecture as FairNDAGAN and FairGANBERT.

**FullBERT:** This baseline represents a supervised learning model trained on the fully labeled dataset. By comparing our proposed framework and semi-supervised baselines against FullBERT, we can evaluate the advantages and disadvantages of semi-supervised learning in comparison with supervised learning settings with completely labeled datasets.

**FairFullBERT:** This is the fair version of the FullBERT baseline, which includes adversarial training to achieve unbiased classification decisions. The adversarial debiasing network in this model follows the same architecture as FairNDAGAN and FairGANBERT.

It is important to mention that the architecture of the BERT and FullBERT classifiers incorporates a DistilBERT encoder, a dropout layer, and a Softmax layer. The Softmax layer inputs the CLS token output of the encoder, resulting in a probability distribution suitable for binary classification tasks. Additionally, the adversarial network ensures fairness by mitigating bias on the CLS token representation of the encoder (DistilBERT) for both FairBERT and FairFullBERT.

We trained all non-fairness-aware models on a total of 10 epochs [11]. Moreover, the fair baseline models and our proposed fair models (FairNDAGAN and FairGANBERT) have a pre-training phase of 10 epochs, followed by a debiasing post-processing stage of 20 epochs [12]. The batch size is 32 and all model training procedures are executed on a single Tesla T4 GPU, featuring a maximum of 27 GB RAM, utilizing the Google Colab platform.

## 4.4 Technical Setting

This section initially outlines the hyperparameter tuning process, then introduces technical settings related to our proposed fair framework, and the baselines. It is crucial to mention that for NDAGAN and GANBERT, we have employed the same architecture described in (Shayesteh and Inkpen, 2022; Croce et al., 2020), although we replaced BERT with DistilBERT. Furthermore, for the supervised baselines, we utilized DistilBERT for sequence classification as introduced by Huggingface [13]. In this study, we used PyTorch version 2.0.0+cu118 as our primary deep learning framework and make the code available [14] online.

In this research, we conducted a Bayesian hyperparameter tuning using the Hyperopt package to identify an optimal set of learning rates and adversarial decay factor ($\lambda$) for each model. Although we could have chosen a distinct pair of mentioned hyper-parameters for each model trained on different ratios of labeled and unlabeled data, we selected one fixed pair for each model on each dataset to gain more general insight into our models'

[11] We have tried to go over 10 epochs however, we could not find more epochs that can improve the accuracy for these architectures.

[12] We need to consider that 20 epochs of debiasing are not the same as 10 epoch of pre-training phase in terms of the model to optimize the objective function since in debasing we only use a mini-batch of data to enforce fairness as opposed to pre-training that we use whole data in each epoch.

[13] https://huggingface.co/docs/transformers/model_doc/distilbert

[14] https://github.com/shahriarshayesteh/Master_thesis.git

robustness and performance. Table 4.2 shows the hyper-parameters selected for each model for different datasets.

| Models | Hatexplain | | | | Wiki Toxicity | | | |
|---|---|---|---|---|---|---|---|---|
| | $LR_{\text{Adversary}}$ | $LR_{\text{Discriminator}}$ | $LR_{\text{Generator}}$ | $\lambda$ | $LR_{\text{Adversary}}$ | $LR_{\text{Discriminator}}$ | $LR_{\text{Generator}}$ | $\lambda$ |
| NDAGAN | - | 5e-5 | 5e-4 | - | - | 5e-5 | 3e-4 | - |
| FairNDAGAN | 1e-4 | 5e-5 | - | [9,11] | 1e-4 | 5e-5 | - | [18,22] |
| GANBERT | - | 5e-5 | 5e-4 | - | - | 5e-4 | 3e-4 | - |
| FairGANBERT | 1e-4 | 5e-5 | - | [9,11] | 1e-4 | 5e-5 | - | [28,32] |

Table 4.2: Semi-supervised models' Hyper-parameters for each dataset are set as described in this table. Please note that $LR_{\text{Discriminator}}$ is the learning rate for both the discriminator and the encoder.

## 4.5 Evaluation Metrics

In this section, we will introduce and discuss the different evaluation metrics utilized in this study to assess the performance of our models. These metrics are accuracy, balanced accuracy, and equalized odds difference (EOD), which function for different purposes and provide different insights into the model's performance. The choice of these metrics is based on the nature of our problem (including datasets and proposed models' architecture). Furthermore, given the imbalanced nature of our data, it is crucial to consider metrics that reflect the performance across all classes and demographic groups, not just the majority ones. In the following, we present the evaluation metrics in more detail.

### 4.5.1 Accuracy

We provide the standard accuracy metric to compare the overall model performance on the classification task. This metric calculates the proportion of correctly classified instances over the total number of instances. However, it is important to note that accuracy may not be an appropriate measure for imbalanced datasets as we have in this work, and it may provide a misleading impression of the model's performance. Therefore, we also measure the balanced accuracy metric to have a better understanding of our models' performance.

The accuracy can be defined as follows for binary classification:

$$\text{accuracy (Acc)} = \frac{\text{The proportion of correctly classified instances}}{\text{The total number of instances}} \tag{4.24}$$

### 4.5.2 Balanced Accuracy

In this work, we have decided to also report the balanced accuracy metric for evaluating the performance of our models, as our datasets are imbalanced. Balanced accuracy is particularly well-suited for imbalanced datasets, as it provides a more robust performance measure than standard accuracy, which can be misleading when dealing with imbalanced classes. We can define the balanced accuracy as the average of recall (also known as sensitivity or true positive rate) obtained in each class. In the case of binary classification, the balanced accuracy can be calculated as follows:

$$\text{balanced accuracy (BAcc)} = \frac{1}{2} \times \left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}}\right) \tag{4.25}$$

Here, True Positives (TP) and True Negatives (TN) represent the correctly classified positive and negative instances, respectively, while False Positives (FP) and False Negatives (FN) represent the misclassified instances.

The advantage of balanced accuracy over standard accuracy in the context of imbalanced datasets lies in its ability to give equal importance to each class's performance. By averaging the recall values for each class, balanced accuracy ensures that underrepresented classes contribute equally to the overall performance metric. This allows for a more meaningful evaluation of the model's performance, as it considers not only the overall accuracy but also the ability of the model to classify instances from minority classes correctly. Therefore, it is especially important in the context of our study to evaluate balanced accuracy along with accuracy and fairness in binary classification tasks involving imbalanced datasets appropriately. While AUC-ROC and F1-score are commonly used metrics, they were not chosen for our study due to the specific requirements of our task. AUC-ROC, although valuable for binary classification problems, can give misleading results with our imbalanced datasets as it incorporates both true positive rates (TPR) and false positive rates (FPR). F1-score, on the other hand, is about balancing precision and recall. It seeks to ensure the model is both good at predicting positive instances correctly (precision) and not missing actual positive instances (recall). While balanced accuracy is more about ensuring the model performs equally well in all classes, whether they are majority or minority classes, and It is about balancing performance across classes. Therefore, we report balanced accuracy similar to Baldini et al. (2022), to measure the accuracy performance of our models considering the imbalanced nature of datasets.

### 4.5.3 Equalized Odds Difference

We measure fairness with equalized odds difference (EOD) for race and gender. EOD is commonly used to measure the equalized odds notion of fairness. It measures the difference between false positive rates (FPRs) and true positive rates (TPRs) of two separate groups. More information about TPRs, FPRs and equalized odds is discussed in 2.8.

Here, we define EOD by focusing on race and gender and fine-grained terms related to demographic groups in each dataset. For EOD of gender:

$$\text{EOD}_{\text{Gender}} = |FPR_{\text{gender}} - FPR_{\text{non-gender}}| + |TPR_{\text{gender}} - TPR_{\text{non-gender}}|. \tag{4.26}$$

Here, gender and non-gender refer to samples that contain sensitive terms related to gender or not.

For EOD of race, we have:

$$\text{EOD}_{\text{race}} = |FPR_{\text{race}} - FPR_{\text{non-race}}| + |TPR_{\text{race}} - TPR_{\text{non-race}}|, \tag{4.27}$$

race and non-race refer to samples that contain sensitive terms related to gender or not.

For EOD of fine-grained terms, we can compute the equalized odds for each sensitive and non-sensitive-term sample, which accordingly means samples with studied sensitive terms and not having sensitive terms:

$$\text{EOD}_{\text{fine-grained-term}} = |FPR_{\text{sensitive}} - FPR_{\text{non-sensitive}}| + |TPR_{\text{sensitive}} - TPR_{\text{non-sensitive}}| \tag{4.28}$$

Therefore, using the formula 4.28, we can evaluate the fairness of the classifier across fine-grained demographic terms.

The range of the EOD is an essential factor to consider when evaluating the fairness of an ML model. The minimum value for this metric is 0, meaning that the True Positive Rate (TPR) and False Positive Rate (FPR) are equal across different demographic groups. In such a scenario, the model's performance is considered fair and unbiased for all groups. The maximum value for the Difference of Equalized Odds is 2, which occurs when there is a maximum discrepancy between the demographic groups in both TPR and FPR. Specifically, this would happen when one group has a TPR and FPR of 1 while the

other group has a TPR and FPR of 0. In this case, the model's performance is extremely biased, leading to unfair treatment and discrimination.

In practice, achieving an EOD value of exactly 0 might be challenging due to various factors such as imbalanced data or inherent biases in the training process. However, aiming for a value as close to 0 as possible is crucial to ensure fairness and equal treatment for all demographic groups.

#### 4.5.3.1 Justification for the Selection of Equalized Odds Difference as a Fairness Metric

While other fairness metrics exist, such as demographic parity or individual fairness, we have focused on the EOD for several reasons. Firstly, EOD is particularly relevant to our work as it is well suited for the binary classification task, offering insights into false positive (FPR) and true positive rates (TPR) for different demographic groups. This is vital as it directly addresses the potential biases in misclassification rates that can disproportionately impact certain groups, a critical concern in toxicity text classification tasks.

Secondly, the adversarial debiasing method we applied in our fair-enhanced models is particularly aligned with the objective of equalized odds. Adversarial debiasing aims to ensure that the predictions are independent of the sensitive attributes when the true outcome is given, which is the core idea behind equalized odds.

Lastly, while metrics like demographic parity ensure equal positive rates across different groups, they do not account for the true underlying distribution of positive outcomes in these groups and may enforce over- or under-prediction for certain groups. Individual fairness, on the other hand, is more suited for scenarios where a similarity metric between individuals can be defined, which is challenging in the context of text classification tasks.

Therefore, we believe using EOD provides a more robust and meaningful measure of fairness for our study. In addition, it aligns well with our project's goals and the nature of the data and task while effectively mitigating bias with the adversarial debiasing technique. However, we acknowledge that no single fairness metric can capture all aspects of fairness, and it has its limitations.

# Chapter 5

# Experiment and Discussion

As mentioned, this study investigates social fairness in semi-supervised toxicity text classification. In other words, we seek to understand the impact of using biased training data (imbalanced in terms of labels and keywords) on the fairness and accuracy performance of the models under study. Additionally, we aim to evaluate the performance of our proposed fair framework and baselines concerning the accuracy and fairness metrics related to coarse-grained and fine-grained terms associated with race and gender.

In this chapter, we first present the data distribution and then mention each experiment in detail. Next, we dive into the evaluation of the models and discussion of the results.

## 5.1 Data Distribution and Sampling Process for Experiments

In this section, we first present how each dataset is processed for this project. Then, we discuss the distribution of train and test sets with respect to labels, coarse-grained, and fine-grained terms related to demographic groups. Following that, we outline each experiment and describe the data sampling process.

### 5.1.1 Processed Datasets

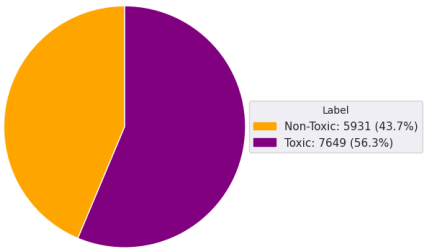
For this study, we utilize DistilBERT, which does not require manual pre-processing [1]of the input text in the datasets. To adapt the datasets for training and testing, we have developed a unified representation to demonstrate coarse-grained term labels (for gender and race), toxicity classification labels, and fine-grained term communities for each sample. Furthermore, we examine the distribution of the train and test sets to highlight the imbalanced nature of the datasets.

After extracting useful information for this task, our datasets consist of five columns: Text, Race, Gender, Label, and Fine-terms (fine-grained terms). The datasets are structured as follows:

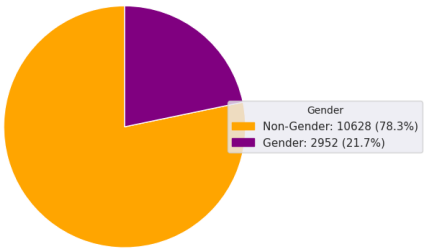
- **Text**: This column contains the comments that our model will receive as input for classification.
- **Race**: A binary feature indicating whether a sample in our dataset contains sensitive terms related to race (i.e., whether each sample targets race or not).
- **Gender**: A binary feature indicating whether a sample in our dataset contains sensitive terms related to gender (i.e., whether each sample targets gender or not).
- **Label**: A binary feature that indicates whether each sample contains toxic language or not.
- **Fine-terms**: This column contains string information about which fine-grained communities related to gender or race have been targeted in our dataset (a sample can contain no to as many fine-grained communities targeted in the text).

In order to illustrate the imbalances in the datasets, several pie charts have been prepared, each highlighting different aspects of the data distribution. For example, Figure 5.1 shows how HateXplain training sets are distributed based on the label (Figure 5.1(a)), gender (Figure 5.1(b)), race (Figure 5.1(c)). In HateXplain, there is an imbalance in the label and keyword levels (gender and race) where almost 56% of data are toxic comments, and 76% and 85% of data contain no gender and no race information, respectively.

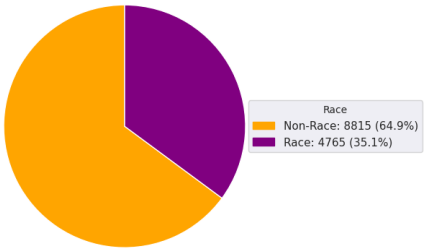
[1]In this work, we apply the DistilBERT tokenizer which tokenized the text in the proper format for the DistilBERT encoder.




(a) Distribution of training data based on label.



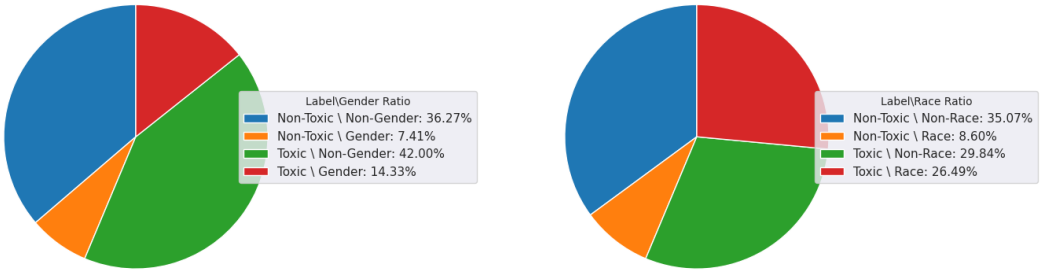

(b) Distribution of training data based on gender.




(c) Distribution of training data based on race.

Figure 5.1: Pie charts illustrate how the HateXplain training set is distributed based on the label, gender, and race binary features.

Figure 5.2 shows how labels and coarse-grained keywords such as race and gender are jointly distributed throughout the training set of HateXplain. The pie chart 5.2(a), and 5.2(b) show that HateXplain has more toxic labels for gender and race than non-toxic which may cause the model to capture unintended bias where it may assume a non-toxic sentence contain a sensitive word related to demographic groups is toxic (which can cause discrimination).

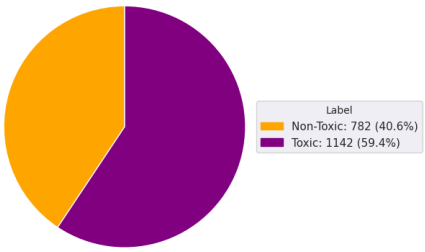



(a) Distribution of training data based on label and gender.

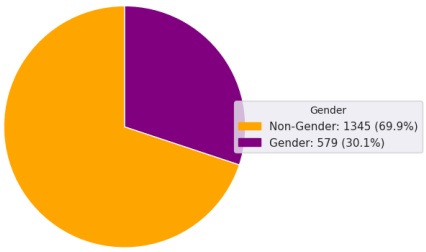
(b) Distribution of training data based on label and race.

Figure 5.2: Pie charts illustrate how the HateXplain training set is jointly distributed based on the label, gender, and race binary features.

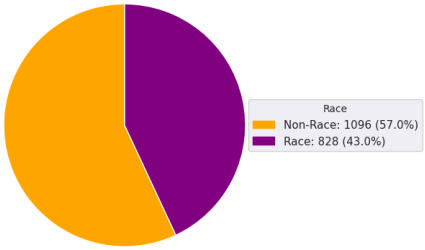
As illustrated in Figures 5.3 and 5.4, the HateXplain test set closely reflects the distribution observed in the training set. Evaluating our models on the HateXplain test set, which follows real-world keyword-level distribution patterns in social media (Mathew et al., 2021), shows a higher proportion of toxic comments for those containing demographic information. This approach allows for a more precise evaluation of the models by simulating the models' performance in real-world scenarios.




(a) Distribution of test data based on label.

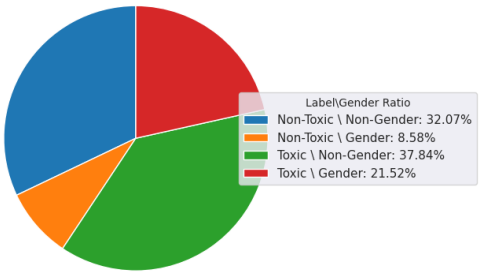



(b) Distribution of test data based on gender.

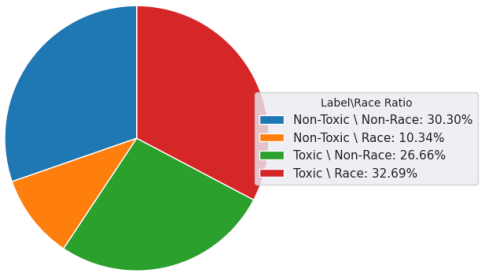



(c) Distribution of test data based on race.

Figure 5.3: Pie charts illustrate how the HateXplain test set is distributed based on the label, gender, and race binary features.



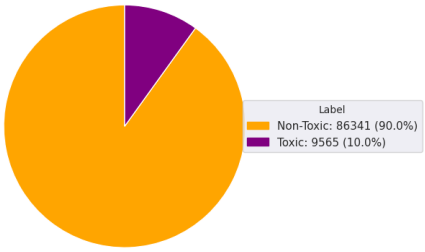

(a) Distribution of test set based on label and gender.



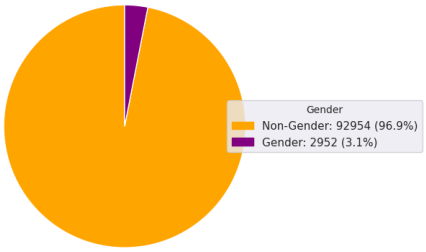

(b) Distribution of test set based on label and race.

Figure 5.4: Pie charts illustrate how the HateXplain test set is jointly distributed based on the label, gender, and race binary features.

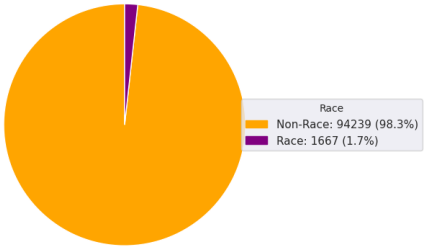
On the other side, the Wiki Toxicity dataset is a highly imbalanced dataset. By looking at pie charts in Figure 5.5, we can see that most of the comments are non-toxic (around 90%) and only 2% and 3% of the comments contain information about gender and race.




(a) Distribution of training data based on label.



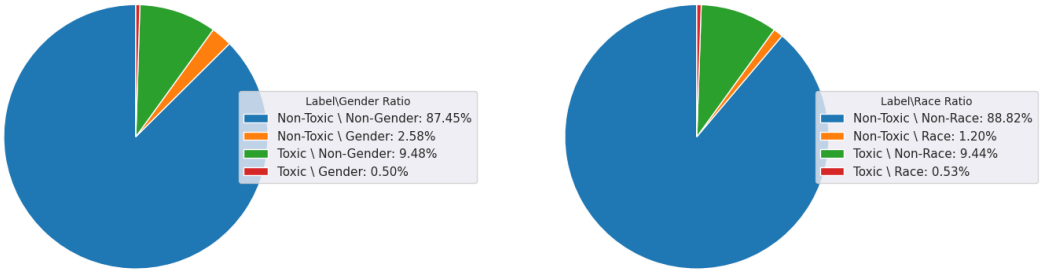

(b) Distribution of training data based on gender.




(c) Distribution of training data based on race.

Figure 5.5: Pie charts illustrate how the Wiki Toxicity training set is distributed based on the label, gender, and race binary features.

Figure 5.6 shows how labels and demographic keywords such as gender and race are jointly distributed throughout the training set of Wiki Toxicity. The pie chart 5.6(a), and 5.6(b) show that the Wiki Toxicity has more non-toxic labels for gender and race than toxic which may cause the model to capture unintended bias where it may assume a toxic sentence contain a sensitive word related to demographic groups is non-toxic.

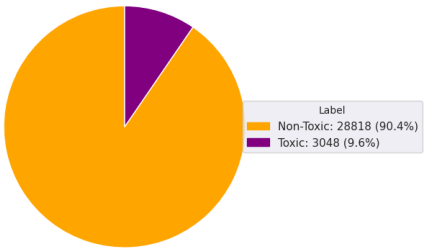



(a) Distribution of training data based on label and gender.

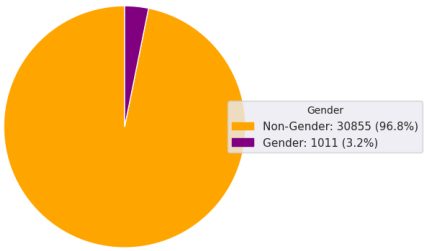
(b) Distribution of training data based on label and race.

Figure 5.6: Pie charts illustrate how the Wiki Toxicity training set is jointly distributed based on the label, gender, and race binary features.

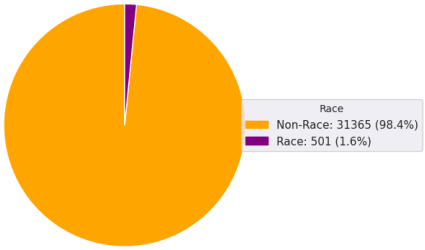
As illustrated in Figures 5.7 and 5.8, the Wiki toxicity test has a similar distribution to the training set. Therefore, it helps us to have a precise evaluation of our models trained on the training set.




(a) Distribution of test set based on label.

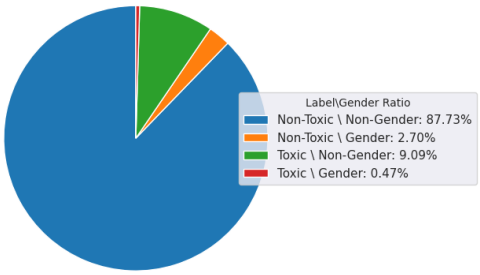



(b) Distribution of test set based on gender.

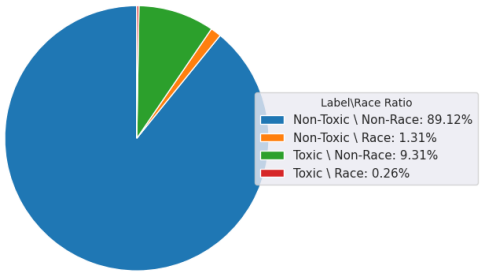



(c) Distribution of test set based on race.

Figure 5.7: Pie charts illustrate how the Wiki Toxicity test set is distributed based on the label, gender, and race binary features.




(a) Distribution of test data based on label and gender.




(b) Distribution of test data based on label and race.

Figure 5.8: Pie charts illustrate how the Wiki Toxicity test set is jointly distributed based on the label, gender, and race binary features.

Finally, two datasets with different natures and distributions can make the classification and evaluation of the results more challenging but more inclusive in toxicity text classification. For example, the HateXplain dataset is a benchmark dataset for hate speech detection and contains samples from social media such as Twitter and Gab, which follow the real-world distribution of social media where most of the samples contain demographic information that is toxic (Mathew et al., 2021). On the other hand, a highly imbalanced Wiki Toxicity dataset is collected from the English Wikipedia page comments, which is considered to have more non-toxic content (Zhang et al., 2018a).

### 5.1.2 Significance of Fine-grained Terms in Evaluation

The use case of fine-grained terms in the datasets is in the evaluation process, where we aim to answer research question $Q2$ and assess the performance of the models in terms of fairness on these specific demographic groups in each dataset. By doing so, we can determine the effectiveness of our debiasing approach on a more granular level. However, to rely on the model's performance concerning the fine-grained terms, we must assess whether we have sufficient data in the testing sets for each demographic group related to gender and race in each dataset. In (Baldini et al., 2022), the results for the fine-grained terms with more than 100 samples in the test set were reported. However, this study presents the results for those fine-grained communities with over 50 samples in the test set. Nonetheless, we should be cautious that the results for fine-grained terms with smaller test samples might be less reliable.

Table 5.1 and 5.2 illustrates the number of samples of fine-grained terms in each dataset.

| Fine-Grained Term | Count | Fine-Grained Term | Count |
|---|---|---|---|
| trans | 1622 | lesbian | 39 |
| white | 728 | african | 28 |
| black | 632 | bisexual | 26 |
| gay | 491 | heterosexual | 25 |
| straight | 385 | transgender | 23 |
| male | 359 | mexican | 20 |
| female | 192 | japanese | 18 |
| american | 191 | hispanic | 14 |
| homosexual | 179 | middle eastern | 14 |
| asian | 95 | canadian | 12 |
| indian | 81 | african american | 6 |
| european | 73 | latina | 5 |
| queer | 53 | latino | 3 |
| chinese | 47 | lgbt | 1 |

Table 5.1: Fine-grained terms and their counts for Wiki Toxicity test set.

| Fine-Grained Term | Count |
|---|---|
| African | 2341 |
| Women | 1697 |
| Homosexual | 1494 |
| Refugee | 986 |
| Arab | 978 |
| Caucasian | 816 |
| Men | 526 |
| Asian | 435 |
| Hispanic | 391 |

Table 5.2: Fine-grained terms and their counts for HateXplain test set.

### 5.1.3 Experiments and Data Sampling Regime

To address the hypotheses and research questions posed in the introduction, we have designed three experiments to provide insights into these aspects. This section presents

an overview of each experiment, followed by an analysis of the data sampling strategies employed for each. Finally, we aim to analyze the result of these experiments in the following sections to effectively evaluate the performance of our models in terms of accuracy and fairness and ultimately shed light on the key concerns outlined earlier in the study.

#### 5.1.3.1 Experiment 1: Overall Performance Regarding Fairness and Accuracy

In Experiment 1, the data sampling process is designed to create a training set that contains labeled and unlabeled training data for semi-supervised models and labeled training data for the limited supervised cases. The labeled training set is sampled with a specific distribution for coarse-term communities (gender, race) based on the specific proportions for each class. The proportions are defined as follows:

- **Toxic Label Distirbution**: [10%, 45%, 45%] - This list represents the percentage of labeled data related to each coarse-grained term in toxic samples in the training set. The first element represents the percentage of samples with no race and gender indication (equal to non-protected samples), the second element represents the percentage of samples related to gender, and the third element represents the percentage of samples related to race.
- **Non-Toxic Label Distirbution**: [80%, 10%, 10%] - This list represents the percentage of labeled data related to each coarse-grained term in non-toxic samples in the training set. The first element represents the percentage of samples with no race and gender indication, the second element represents the percentage of samples related to gender, and the third element represents the percentage of samples related to race.

| | HateXplain | | | | Wiki Toxicity | | | |
|---|---|---|---|---|---|---|---|---|
| Label Ratio | 0.005 | 0.01 | 0.05 | 0.1 | 0.0008 | 0.0016 | 0.0080 | 0.0161 |
| Label Count | 77 | 154 | 770 | 1539 | 77 | 154 | 766 | 1541 |

Table 5.3: The label ratio and label count per class for each dataset is given. For instance, r = 0.005 for HateXpalin means that we have 77 labeled data for toxic texts and 77 labeled data for non-toxic texts in the training set.

In this experiment, we first sample a number of labeled data based on the given set of label ratios per class in each dataset, as described in Table 5.3 (the distribution of the

labeled data in terms of demographic groups is set based on the given label distribution for each class). Then, we utilize the remaining data as unlabeled data for semi-supervised models. Experiment 1 serves as our primary experiment, intending to evaluate our model's performance in terms of both accuracy and fairness. After that, we will analyze the hypothesis $H$ and research questions $Q1$ introduced in the introduction to comprehensively assess our model's performance and address the critical concerns outlined earlier in the study. In addition, in this experiment, we evaluate the effect of coarse-grained term debiasing on fine-grained terms fairness to answer research question $Q2$.

#### 5.1.3.2 Experiment 2: Effect of labeled data on Accuracy and Fairness

In Experiment 2, the data sampling process is designed to create a training set that contains labeled and unlabeled training data for semi-supervised models and labeled training data for the limited supervised cases. The labeled training set is sampled with a specific distribution for coarse-term communities (gender, race) based on the specific proportions for each class. The proportions remain as defined in Experiment 1:

- **Toxic Label Ratio**: [10%, 45%, 45%]
- **Non-Toxic Label Ratio**: [80%, 10%, 10%]

Additionally, in this experiment, we randomly select a subset of 5,000 and 50,000 unlabeled data points from the HateXplain and the Wiki Toxicity datasets subsequently. Within this subset, 60 percent of the samples are from the toxic label category, and 40 percent are from the non-toxic label category. We further sample these unlabeled data points according to the following distributions:

- **Toxic Labeled / Unlabeled Data Ratio**: [10%, 45%, 45%] - This list represents the percentage of unlabeled data related to each coarse-term community in toxic labeled samples in the training set. The first element represents the percentage of samples with no race and gender indication, the second element represents the percentage of samples related to gender, and the third element represents the percentage of samples related to race.
- **Non-Toxic Labeled / Unlabeled Data Ratio**: [80%, 10%, 10%] - This list represents the percentage of unlabeled data related to each coarse-term community in non-toxic labeled samples in the training set. The first element represents the percentage of samples with no race and gender indication, the second element represents

the percentage of samples related to gender, and the third element represents the percentage of samples related to race.

In Experiment 2, we aim to investigate the impact of labeled data on the model's performance in terms of accuracy and fairness while fixing the number of unlabeled data to eliminate the possible effect of unlabeled data in our study. This will enable us to analyze further the research question $Q3$, introduced in Section 1.3.

#### 5.1.3.3 Experiment 3: Effect of Unlabeled Data on Accuracy and Fairness

In Experiment 3, we aim to investigate the impact of varying amounts of unlabeled data on the model's performance in terms of accuracy and fairness. To achieve this, we select two fixed ratios of labeled data for each dataset [2] and then vary the number of unlabeled data points (four different values depending on each dataset's size). The reason we have selected two different label ratios is to make sure that the results of this experiment are not affected by the ratios of labeled data. The sampling distribution for labeled and unlabeled data remains consistent with those used in Experiment 2.

| | HateXplain | | | | Wiki Toxicity | | | |
|---|---|---|---|---|---|---|---|---|
| Unlabeled Count | 1000 | 2500 | 5000 | 10000 | 10000 | 25000 | 50000 | 95000 |

Table 5.4: The number of unlabeled data selected for each dataset in experiment3.

In this experiment, we aim to explore how the quantity of unlabeled data affects the model's performance in terms of accuracy and fairness. We aim to achieve this by systematically varying the amount of unlabeled data to further analyze the research question $Q4$ introduced in the introduction.

## 5.2 Evaluation and Discussion

This section will detail our experimental results, systematically examining the patterns, trends, and implications that emerge from our datasets. We aim to provide insight into the validity of our hypotheses and answer the research questions that have been proposed. Each experiment will be individually investigated, with the intent to draw meaningful

[2] We selected 0.05 and 0.1 for HateXplain, and 0.0016, 0.0161 for Wiki Toxicity

conclusions that align with our research objectives. Ultimately, we will summarize the key trends and limitations to offer a comprehensive view of our findings, providing an enriched perspective on this topic.

It is important to note that we do not conduct any statistical analysis tests in this study. This is because there are no fair semi-supervised frameworks for comparison with our model results. The introduced baselines merely serve to demonstrate how enforcing fairness can enhance fairness criteria considering the balance between accuracy and fairness. All the results are reported over five runs with associated standard deviation.

### 5.2.1 Experiment 1: Overall Performance Regarding Fairness and Accuracy

This section analyzes the classification and fairness performance metrics for different models on both datasets. This assessment allows us to provide more comprehensive insights into the performance of various models, including FairNDAGAN and FairGANBERT from our proposed framework, as well as several baseline models. Finally, it helps us to provide a good view of the hypothesis and research questions Q1, and Q2.

#### 5.2.1.1 HateXplain Dataset Results

In this section, we initially refer to Table 5.5 and 5.6 to analyze our models' performance on the HateXplain dataset in terms of accuracy scores (accuracy, and balanced accuracy).

Let us first identify trends related to $H$. All the fair models have comparable accuracy with non-fair counterparts in terms of accuracy. Nevertheless, we must remember that selection criteria are based on accuracy; therefore, fair models' accuracy is within a range of non-fair models. However, this is not the case for balanced accuracy, where FairBERT balanced accuracy at ratio 0.005 is almost 11% less than BERT. For other models, such as our proposed fair semi-supervised models (FairNDAGAN, FairGANBERT), balanced accuracy is in the same range as NDAGAN and GANBERT. Therefore, generally, we can identify that there is a reasonable trade-off between the fair models and non-fair counterparts in terms of accuracy scores. However, to fully confirm $H$, we need to look at the fairness scores as well.

In line with $Q1$, we observe some interesting trends in the context of fair semi-supervised models. Looking at balanced accuracy (BAcc), FairGANBERT slightly outperforms FairNDAGAN across all label ratios, except for r=0.1. On the other hand, when considering accuracy, FairNDAGAN tends to outperform FairGANBERT across all label ratios, except for

r=0.005. These opposing trends highlight a view toward $Q1$ that there is no total winner between our proposed models regarding the accuracy scores on HateXplain. For example, based on a specific application, we may select to go with FairNDAGAN if accuracy is important or FairGANBERT if balanced accuracy is critical.

| Model | BAcc | | | |
|---|---|---|---|---|
| | (r = 0.005) | (r = 0.01) | (r = 0.05) | (r = 0.1) |
| NDAGAN | 64.57 ± 2.35 | 69.90 ± 1.62 | 73.61 ± 1.41 | 76.06 ± 0.82 |
| FairNDAGAN | 62.42 ± 4.32 | 67.05 ± 3.08 | 72.33 ± 2.01 | 74.79 ± 0.46 |
| GANBERT | 65.76 ± 3.70 | 69.89 ± 1.74 | 73.74 ± 1.02 | 75.52 ± 1.54 |
| FairGANBERT | 64.85 ± 3.79 | 68.52 ± 2.19 | 72.70 ± 1.41 | 74.70 ± 0.66 |
| BERT | 64.44 ± 2.15 | 68.31 ± 1.51 | 73.07 ± 1.08 | 75.16 ± 1.01 |
| FairBERT | 53.34 ± 2.68 | 66.37 ± 2.59 | 73.03 ± 0.68 | 73.94 ± 1.42 |
| FullBERT | 78.14 ± 0.80 | 78.14 ± 0.80 | 78.14 ± 0.80 | 78.14 ± 0.80 |
| FairFullBERT | 76.13 ± 3.54 | 76.13 ± 3.54 | 76.13 ± 3.54 | 76.13 ± 3.54 |

Table 5.5: This table presents the results of Experiment 1. Each column corresponds to a different label ratio and displays the balanced accuracy of the respective models on the HateXplain dataset. The results are averaged over five runs on the test data, and the standard deviation for each model is also provided. For more information about the number of labeled data per class in each ratio, please refer to Table 5.3. It is worth noting that FulllBERT and FairFullBERT are trained on the fully labeled training set and different ratios do not apply to them, however, to provide an easier comparison, we add their results for all ratios.

There are some other interesting trends that we can identify. For instance, accuracy generally decreases when transitioning from SSL models to limited supervised learning (SL) models. Also, NDAGAN and GANBERT can achieve around 76% accuracy at r=0.1, close to the fully supervised FullBERT model that achieves 79.20% accuracy. Furthermore, regarding balanced accuracy, FairGANBERT and FairNDAGAN often outperform FairBERT, especially at label ratio r=0.005. These trends indicate that SSL models can maintain competitive accuracy and balanced accuracy with Fully supervised cases and outperform limited supervised models. This could further add insight into how SSL models leverage unlabeled data to generalize the data distribution.

It is worth noting that FullBERT, trained on all labeled data, does not surpass 79.20% in terms of accuracy or 76.13% in terms of balanced accuracy, which underlines the complexity

of the HateXplain dataset. This provides important context for the performance of the other models, highlighting the necessity of advanced modeling techniques to handle such challenging datasets.

| Model | Acc | | | |
|---|---|---|---|---|
| | (r = 0.005) | (r = 0.01) | (r = 0.05) | (r = 0.1) |
| NDAGAN | 68.24 ± 1.73 | 71.34 ± 0.67 | 75.02 ± 0.79 | 76.51 ± 0.77 |
| FairNDAGAN | 65.76 ± 1.36 | 69.28 ± 2.11 | 71.53 ± 1.46 | 73.51 ± 1.78 |
| GANBERT | 68.62 ± 2.42 | 71.01 ± 0.56 | 74.67 ± 0.84 | 76.61 ± 0.99 |
| FairGANBERT | 66.01 ± 2.91 | 68.50 ± 1.92 | 71.53 ± 1.68 | 73.17 ± 1.52 |
| BERT | 65.69 ± 2.07 | 70.14 ± 1.43 | 74.37 ± 0.43 | 76.22 ± 0.67 |
| FairBERT | 61.55 ± 1.63 | 67.62 ± 2.43 | 71.36 ± 1.44 | 73.15 ± 0.91 |
| FullBERT | 79.20 ± 0.39 | 79.20 ± 0.39 | 79.20 ± 0.39 | 79.20 ± 0.39 |
| FairFullBERT | 77.19 ± 1.81 | 77.19 ± 1.81 | 77.19 ± 1.81 | 77.19 ± 1.81 |

Table 5.6: This table presents the results of Experiment 1, showing the accuracy of our proposed models (FairNDAGAN and FairGANBERT) and baseline models on the HateXplain dataset. Each column corresponds to a different label ratio and displays the accuracy of the respective models, averaged over five runs and the standard deviation for each model is also provided. It is worth noting that FulllBERT and FairFullBERT are trained on the fully labeled training set and different ratios do not apply to them, however, to provide an easier comparison, we add their results for all ratios.

We now analyze the fairness metric (EOD). The information about EOD for gender and race demographic groups in HateXpalin is illustrated in Table 5.7, and 5.8.

In the context of fair semi-supervised models, some interesting trends emerge. First, for both gender and race EOD, FairNDAGAN, and FairGANBERT consistently perform better than their non-fair counterparts (NDAGAN and GANBERT) across all label ratios. This finding aligns with our hypothesis $H$ that introducing fairness to our models can improve fairness metrics. However, regarding $Q1$, there is no clear dominant model for either gender or race EOD between FairNDAGAN and FairGANBERT.

When comparing the non-fair semi-supervised models (GANBERT and NDAGAN), we do not observe a clear trend for either gender or race EOD. However, both models generally perform better than the BERT and FullBERT models regarding race EOD. This highlights the potential of semi-supervised models to maintain better fairness even with limited labeled data. However, we need to consider that the low labeled data regime in the

SSL models such as r = 0.005 have a very lower accuracy performance compared to the fully supervised cases; a lower EOD individually does not guarantee better generalization in terms of accuracy and fairness. Therefore, for a comprehensive analysis, we need to consider both measures to compare the performance of the models.

| Model | $\text{EOD}_{\text{Gender}}$ | | | |
|---|---|---|---|---|
| | (r = 0.005) | (r = 0.01) | (r = 0.05) | (r = 0.1) |
| NDAGAN | 0.12 ± 0.01 | 0.16 ± 0.06 | 0.12 ± 0.07 | 0.10 ± 0.07 |
| FairNDAGAN | 0.07 ± 0.04 | 0.06 ± 0.02 | 0.05 ± 0.02 | 0.04 ± 0.03 |
| GANBERT | 0.10 ± 0.04 | 0.11 ± 0.04 | 0.12 ± 0.04 | 0.10 ± 0.03 |
| FairGANBERT | 0.10 ± 0.03 | 0.05 ± 0.02 | 0.06 ± 0.02 | 0.02 ± 0.01 |
| BERT | 0.12 ± 0.04 | 0.13 ± 0.05 | 0.10 ± 0.04 | 0.15 ± 0.03 |
| FairBERT | 0.01 ± 0.01 | 0.11 ± 0.03 | 0.07 ± 0.02 | 0.10 ± 0.02 |
| FullBERT | 0.07 ± 0.03 | 0.07 ± 0.03 | 0.07 ± 0.03 | 0.07 ± 0.03 |
| FairFullBERT | 0.05 ± 0.03 | 0.05 ± 0.03 | 0.05 ± 0.03 | 0.05 ± 0.03 |

Table 5.7: This table presents the results of Experiment 1, showing the equalized odds difference for the gender demographic group as a measure of fairness. Each column corresponds to a different label ratio and displays the $EOD_{\text{Gender}}$ of the respective models, averaged over five runs, and the standard deviation for each model is also provided. It is worth noting that FulllBERT and FairFullBERT are trained on the fully labeled training set and different ratios do not apply to them, however, to provide an easier comparison, we add their results for all ratios.

As we transition from SL models to the SSL models from our proposed framework, we noticed that the fairness metrics generally improve by introducing fairness considerations. Notably, FairNDAGAN and FairGANBERT consistently outperform FairBERT and FairFullBERT regarding $EOD_{\text{Race}}$ across all label ratios. For $EOD_{\text{Gender}}$, while FairNDAGAN and FairGANBERT perform comparably to FairFullBERT, they outperform it when the ratio of labels is 0.1 per class.

Additionally, when comparing the performance across different demographic groups, it is noticeable that the models tend to perform better regarding $EOD_{\text{Race}}$ than $EOD_{\text{Gender}}$. This observation might point towards the potential challenges in achieving fairness across different demographic groups, calling for tailored strategies for each group.

| Model | $\text{EOD}_{\text{Race}}$ | | | |
|---|---|---|---|---|
| | (r = 0.005) | (r = 0.01) | (r = 0.05) | (r = 0.1) |
| NDAGAN | 0.12 ± 0.07 | 0.07 ± 0.06 | 0.13 ± 0.04 | 0.14 ± 0.03 |
| FairNDAGAN | 0.03 ± 0.01 | 0.06 ± 0.05 | 0.06 ± 0.05 | 0.07 ± 0.03 |
| GANBERT | 0.11 ± 0.08 | 0.10 ± 0.09 | 0.12 ± 0.05 | 0.16 ± 0.08 |
| FairGANBERT | 0.04 ± 0.01 | 0.05 ± 0.05 | 0.04 ± 0.02 | 0.07 ± 0.04 |
| BERT | 0.10 ± 0.10 | 0.09 ± 0.08 | 0.14 ± 0.05 | 0.12 ± 0.07 |
| FairBERT | 0.02 ± 0.02 | 0.05 ± 0.04 | 0.06 ± 0.03 | 0.11 ± 0.02 |
| FullBERT | 0.15 ± 0.02 | 0.15 ± 0.02 | 0.15 ± 0.02 | 0.15 ± 0.02 |
| FairFullBERT | 0.11 ± 0.06 | 0.11 ± 0.06 | 0.11 ± 0.06 | 0.11 ± 0.06 |

Table 5.8: This table presents the results of Experiment 1, showing the equalized odds difference for the race demographic group as a measure of fairness. Each column corresponds to a different label ratio and displays the $EOD_{\text{Race}}$ of the respective models, averaged over five runs, and the standard deviation for each model is also provided. It is worth noting that FulllBERT and FairFullBERT are trained on the fully labeled training set and different ratios do not apply to them, however, to provide an easier comparison, we add their results for all ratios.

We now combine the analysis of accuracy scores (accuracy, balanced accuracy) and fairness metrics (gender and race EOD) to better understand the effectiveness of our proposed models, FairNDAGAN and FairGANBERT, on the HateXplain dataset.

A key question that motivated our research is understanding the trade-off between accuracy performance and fairness in semi-supervised learning (SSL) models. When considering fairness alongside accuracy performance, we find that our fair semi-supervised models (FairNDAGAN and FairGANBERT) can achieve comparable accuracy performance to the semi-supervised models (NDAGAN and GANBERT) while greatly improving fairness metrics. This aligns with our initial hypothesis $H$, suggesting that it is possible to maintain good performance while improving fairness through semi-supervised learning approaches. Therefore, it validates that introducing fairness considerations into SSL models can improve the fairness metrics without a considerable drop in performance.

In addition, the advantage of SSL models becomes more pronounced at higher label ratios such as r = 0.1. Despite having limited labeled data compared to fully supervised models, FairNDAGAN and FairGANBERT can reach comparable accuracy scores performance with the FairFullBERT and completely outperform the FairFullBERT regarding EOD for race and gender.

In line with $Q1$, While FairNDAGAN and FairGANBERT perform well, they have no clear winner across all metrics. The choice between the two would depend on the specifics of the task and the prioritized evaluation metric. Therefore, for research $Q1$, we cannot find a dominant winner between our proposed fair semi-supervised models, at least on the HateXplain dataset.

Our analysis also uncovers some challenges in achieving fairness. While our fair models improve fairness metrics, they still show some discrepancies between gender and race EOD, with $EOD_{\text{Race}}$ generally being lower on HateXpalin. This suggests that achieving fairness across all demographic groups remains challenging and may require more tailored strategies for different demographic groups.

In conclusion, our comprehensive evaluation demonstrates that FairNDAGAN and FairGANBERT effectively address the challenges posed by the HateXplain dataset. They maintain performance while improving fairness, particularly in scenarios with a higher level of limited labeled data. Therefore, our fair proposed models show that they could be effective in the context of toxic language detection. However, we must still analyze the result on other datasets and resources. In the next section, we analyze the results of the models on the Wiki Toxicity dataset and then try to discuss the hypothesis and research questions in more detail.

#### 5.2.1.2 Wiki Toxicity Dataset

In this part, first, we will analyze Table 5.9 and 5.10 to find out how the models perform on the Wiki Toxicity dataset in terms of accuracy scores (balanced accuracy and accuracy). Before delving into details, we must consider that the Wiki Toxicity dataset is a highly imbalanced dataset, and balanced accuracy is a more reliable metric to measure the classification performance of our models than accuracy.

Let us first identify trends related to $H$. All the fair models have comparable accuracy with non-fair counterparts. Nevertheless, we must remember that selection criteria are based on accuracy; therefore, fair models' accuracy is within a range of non-fair models. Also, in terms of balanced accuracy, FairNDABERT and FairBERT, and FairFullBERT, have an acceptable loss of accuracy compared to their non-fair counterparts. However, this is not the case for FairGANBERT which cannot achieve comparable performance with GANBERT regarding balanced accuracy except in r = 0.0161. Therefore, generally, we can say that fairness is achievable with minimal loss of performance in the SSL setting; however, in a highly imbalanced dataset, we must be careful about the selection criteria or

introduce techniques to decrease the negative effect of imbalance in the data on models' performance.

For trends related to $Q1$, there are interesting patterns in the context of our proposed fair semi-supervised framework. Regarding balanced accuracy (BAcc), FairNDAGAN generally outperforms FairGANBERT across different label ratios. This trend, however, does not apply to accuracy, where FairNDAGAN and FairGANBERT perform similarly with no clear winner. These opposing trends highlight a view toward $Q1$ that there is no total winner between our proposed models regarding the accuracy of Wiki toxicity. However, considering the highly imbalanced nature of Wiki Toxicity, we may select to go with FairNDAGAN for similar scenarios (For the final conclusion, we need to consider the EOD for race and gender between these two models to do a better selection.).

| Model | BAcc | | | |
|---|---|---|---|---|
| | (r = 0.0008) | (r = 0.0016) | (r = 0.008) | (r = 0.0161) |
| NDAGAN | 83.50 ± 6.61 | 82.60 ± 5.02 | 85.86 ± 4.60 | 87.49 ± 1.73 |
| FairNDAGAN | 78.70 ± 4.88 | 68.41 ± 2.99 | 81.59 ± 2.85 | 87.42 ± 4.74 |
| GANBERT | 85.82 ± 4.09 | 84.74 ± 5.34 | 85.67 ± 4.12 | 86.69 ± 2.15 |
| FairGANBERT | 68.44 ± 6.44 | 67.05 ± 5.33 | 75.09 ± 4.36 | 78.77 ± 5.36 |
| BERT | 86.74 ± 2.26 | 88.50 ± 2.06 | 90.83 ± 0.58 | 90.26 ± 0.73 |
| FairBERT | 84.33 ± 0.99 | 85.34 ± 4.15 | 88.71 ± 3.07 | 88.07 ± 3.17 |
| FullBERT | 91.23 ± 0.72 | 91.23 ± 0.72 | 91.23 ± 0.72 | 91.23 ± 0.72 |
| FairFullBERT | 91.74 ± 0.20 | 91.74 ± 0.20 | 91.74 ± 0.20 | 91.74 ± 0.20 |

Table 5.9: This table presents the results of Experiment 1, showing the balanced accuracy scores for our proposed models (FairNDAGAN and FairGANBERT) and baseline models on the Wiki Toxicity dataset. Each column corresponds to a different label ratio and displays the balanced accuracy of the respective models, averaged over five runs. The standard deviation for each model is also provided. For more information about the number of labeled data points per class in each ratio, please refer to Table 5.3.

Transitioning from SSL models to limited supervised learning (SL) models, we observe that the BERT and FairBERT generally outperform SSL models regarding balanced accuracy. However, the SSL models, particularly NDAGAN and GANBERT, demonstrate superior performance in terms of accuracy. This discrepancy could be attributed to the highly imbalanced nature of the Wiki Toxicity dataset, where accuracy might not accurately represent the model's classification ability. Therefore, our semi-supervised models

(fair and non-fair) do not perform well on highly imbalanced datasets as their balanced accuracy is less than BERT and FairBERT.

| Model | Acc | | | |
|---|---|---|---|---|
| | (r = 0.0008) | (r = 0.0016) | (r = 0.008) | (r = 0.0161) |
| NDAGAN | 90.50 ± 5.37 | 93.47 ± 3.05 | 94.84 ± 0.63 | 95.22 ± 0.40 |
| FairNDAGAN | 89.33 ± 2.27 | 89.61 ± 3.59 | 91.71 ± 2.52 | 94.78 ± 0.67 |
| GANBERT | 90.38 ± 6.66 | 93.04 ± 2.34 | 95.12 ± 0.57 | 95.18 ± 0.34 |
| FairGANBERT | 87.73 ± 3.31 | 90.13 ± 3.80 | 92.51 ± 2.45 | 94.81 ± 1.16 |
| BERT | 86.74 ± 2.26 | 88.50 ± 2.06 | 90.83 ± 0.58 | 90.26 ± 0.73 |
| FairBERT | 84.33 ± 0.99 | 85.34 ± 4.15 | 88.71 ± 3.07 | 88.07 ± 3.17 |
| FullBERT | 91.23 ± 0.72 | 91.23 ± 0.72 | 91.23 ± 0.72 | 91.23 ± 0.72 |
| FairFullBERT | 91.74 ± 0.20 | 91.74 ± 0.20 | 91.74 ± 0.20 | 91.74 ± 0.20 |

Table 5.10: This table presents the results of Experiment 1, showing the accuracy scores for our proposed models (FairNDAGAN and FairGANBERT) and baseline models on the Wiki Toxicity dataset. Each column corresponds to a different label ratio and displays the accuracy of the respective models, averaged over five runs. The standard deviation for each model is also provided.

In the context of EOD for gender and race, the closer the EOD value is to 0, the fairer the model is considered. Therefore, based on the $EOD_{Gender}$ in Table 5.12, we observe that FairNDAGAN and FairGANBERT models have shown a considerable improvement in fairness compared to their non-fair counterparts NDAGAN and GANBERT across all ratios. This indicates that the adversarial debiasing introduced in the training of FairNDAGAN and FairGANBERT is an effective bias mitigating technique. In particular, FairGANBERT appears to have the lowest $EOD_{Gender}$ scores, suggesting it is the fairest among the semi-supervised models in terms of gender demographic. On the other hand, FairBERT and FairFullBERT also show lower $EOD_{Gender}$ scores compare to fair semi-supervised models.

| Model | $EOD_{\text{Gender}}$ | | | |
|---|---|---|---|---|
| | (r = 0.0008) | (r = 0.0016) | (r = 0.008) | (r = 0.0161) |
| NDAGAN | 0.16 ± 0.05 | 0.20 ± 0.05 | 0.22 ± 0.06 | 0.18 ± 0.06 |
| FairNDAGAN | 0.09 ± 0.03 | 0.09 ± 0.03 | 0.11 ± 0.05 | 0.08 ± 0.03 |
| GANBERT | 0.13 ± 0.03 | 0.17 ± 0.06 | 0.23 ± 0.05 | 0.22 ± 0.09 |
| FairGANBERT | 0.03 ± 0.01 | 0.07 ± 0.03 | 0.06 ± 0.02 | 0.07 ± 0.01 |
| BERT | 0.14 ± 0.05 | 0.19 ± 0.06 | 0.17 ± 0.05 | 0.17 ± 0.06 |
| FairBERT | 0.04 ± 0.02 | 0.09 ± 0.03 | 0.05 ± 0.03 | 0.09 ± 0.03 |
| FullBERT | 0.05 ± 0.03 | 0.05 ± 0.03 | 0.05 ± 0.03 | 0.05 ± 0.03 |
| FairFullBERT | 0.04 ± 0.03 | 0.04 ± 0.03 | 0.04 ± 0.03 | 0.04 ± 0.03 |

Table 5.11: This table presents the results of Experiment 1, showing the equalized odds difference for the gender demographic group as a measure of fairness. Each column corresponds to a different label ratio and displays the $EOD_{\text{Gender}}$ of the respective models, averaged over five runs. The standard deviation for each model is also provided.

Similarly, the $EOD_{Race}$ in Table 5.12 shows that FairGANBERT and FairNDAGAN outperform their non-fair counterparts regarding fairness across different demographic groups. The $EOD_{Race}$ for these models is closer to 0, indicating a more balanced prediction for different racial groups. Conversely, comparing supervised and semi-supervised models, FairBERT and FairFullBERT generally demonstrate better fairness than FairNDAGAN and FairGANBERT.

As we mentioned that FairGANBERT is the only model that has an unreasonable loss of balanced accuracy except in a ratio of 0.0161 for Wiki Toxicity, However, we can see that it can outperform FairNDAGAN in terms of fairness. Again, this shows that lower accuracy can bring more fairness and to select a fair model that has a good classification performance, we need to consider both measures to avoid underfitting the data distribution.

Both EOD for gender and race shows that FairFullBERT outperforms fair semi-supervised models. However, in the case of FairBERT, they have almost the same performance in terms of fairness. Therefore, generally, we think that our semi-supervised models (our proposed framework as well) cannot handle highly imbalanced datasets very well.

| Model | $EOD_{\text{Race}}$ | | | |
|---|---|---|---|---|
| | (r = 0.0008) | (r = 0.0016) | (r = 0.008) | (r = 0.0161) |
| NDAGAN | 0.16 ± 0.04 | 0.18 ± 0.06 | 0.15 ± 0.06 | 0.10 ± 0.02 |
| FairNDAGAN | 0.13 ± 0.03 | 0.13 ± 0.03 | 0.13 ± 0.02 | 0.09 ± 0.02 |
| GANBERT | 0.17 ± 0.05 | 0.15 ± 0.07 | 0.18 ± 0.05 | 0.14 ± 0.04 |
| FairGANBERT | 0.03 ± 0.01 | 0.10 ± 0.02 | 0.06 ± 0.03 | 0.07 ± 0.02 |
| BERT | 0.15 ± 0.03 | 0.31 ± 0.07 | 0.17 ± 0.06 | 0.13 ± 0.05 |
| FairBERT | 0.07 ± 0.02 | 0.25 ± 0.09 | 0.10 ± 0.03 | 0.05 ± 0.02 |
| FullBERT | 0.10 ± 0.02 | 0.10 ± 0.02 | 0.10 ± 0.02 | 0.10 ± 0.02 |
| FairFullBERT | 0.08 ± 0.02 | 0.08 ± 0.02 | 0.08 ± 0.02 | 0.08 ± 0.02 |

Table 5.12: This table presents the results of Experiment 1, showing the equalized odds difference for the race demographic group as a measure of fairness. Each column corresponds to a different label ratio and displays the $EOD_{\text{Race}}$ of the respective models, averaged over five runs. The standard deviation for each model is also provided.

Finally, we aim to provide a unified analysis of accuracy scores and fairness metrics on both datasets to reach a conclusion for hypothesis $H$ and research question $Q1$, repeated here for convenience.

**H: We can achieve fairness in semi-supervised toxicity classification by considering the trade-off between classification performance and fairness.**

The semi-supervised learning models FairNDAGAN and FairGANBERT show considerable improvements in fairness compared to their non-fair counterparts NDAGAN and GANBERT while maintaining a relatively high level of accuracy. This is evident from the EOD scores for both gender and race demographics, which are generally lower for the fair models. This finding aligns with the hypothesis that integrating fairness considerations into semi-supervised learning models can improve fairness without greatly sacrificing accuracy.

However, we need to consider that our semi-supervised models are sensitive to highly imbalanced datasets compared to supervised cases. Therefore, to use them in such circumstances, we must find a way to reduce the negative effect of data imbalance on the models' classification and fairness performance.

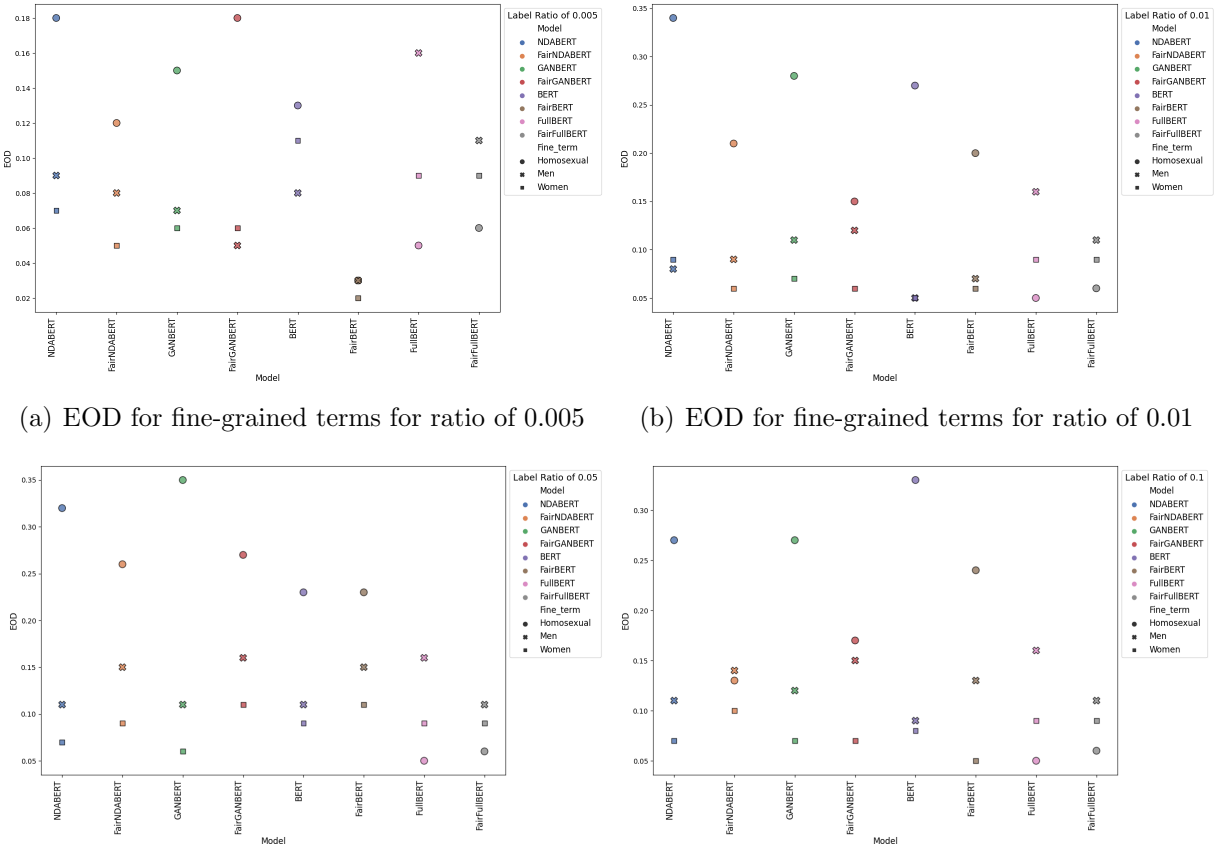
**Q1: Can FairNDAGAN greatly outperform FairGANBERT in terms of fairness and accuracy?**

There is no conclusive evidence to indicate that either FairNDAGAN or FairGANBERT can outperform the others in both datasets. While both models exhibited a slight advantage over the other in different scenarios, such differences were not noteworthy enough to draw a definitive conclusion. However, it is worth noting that both models have similar architectures and employ identical post-processing (debiasing) stages.

#### 5.2.1.3 Fairness in Fine-Grained Terms

In this section, we seek to analyze how the coarse-grained debiasing of the non-fair classifiers can affect the fairness of the fine-grained terms to identify if we can achieve fairness across all fine-grained demographic groups using adversarial debiasing.

Please notice that in this analysis, we aim to provide an answer to the research question $Q2$. Therefore, we do not look to analyze all the details, as fine-grained term analysis is strictly affected by the number of samples related to each term in the test set, and comparing detailed trends may not be accurate as each fine-grained term has a different number of samples presented in the test set as reported in Tables 5.1 and 5.2.




(a) EOD for fine-grained terms for ratio of 0.005

(b) EOD for fine-grained terms for ratio of 0.01

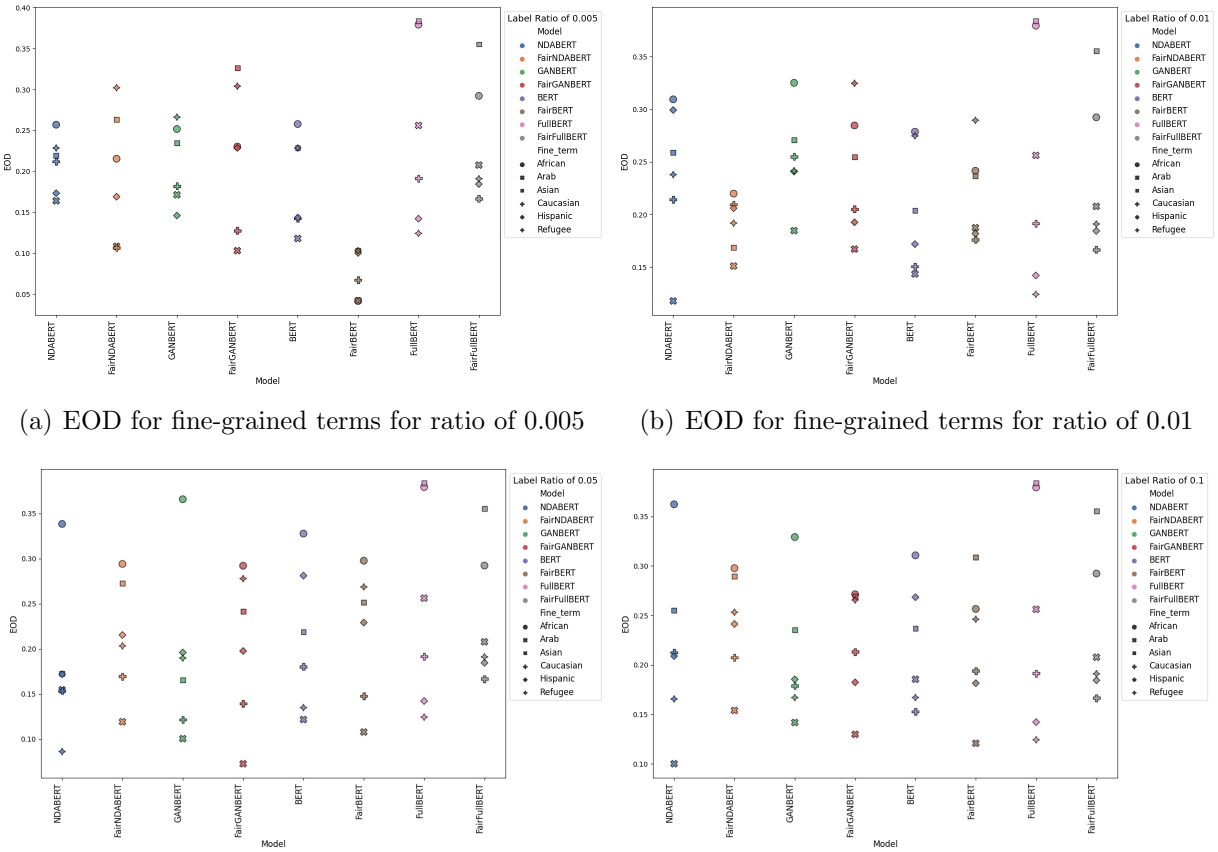
(c) EOD for fine-grained terms for ratio of 0.05

(d) EOD for fine-grained terms for ratio of 0.1

Figure 5.9: This Figure reports the EOD for fine-grained terms related to gender in the HateXplain dataset over different label ratios for models. It is important to note that FullBERT and FairFullBERT are trained on fully-labeled training sets and different ratios of labeled data do not apply to these models.

Looking at Figure 5.9, we can identify two general trends showing that the fair version of each model has a lower maximum EOD of fine-grained terms and, additionally, the differences between maximum and minimum EOD values of fine-grained terms related to gender in the HateXpalin dataset in the fair models compared to their non-fair counterparts declined. For example, in Figure 5.9(a), fine-grained terms for NDABERT range from 0.07 for women to 0.18 for Homosexual, and after applying the coarse-grained debiasing, FairNDAGAN's EOD for Homosexual decreased to 0.12. Moreover, the difference between the EOD of fine-grained terms in FairNDAGAN declined to 0.07 from 0.11 in NDABERT.

This trend applies almost to all models in all ratios. However, the only exception is in the ratio of 0.01, when FairGANBERT's EOD difference between its gender fine-grained terms is around 0.13 while it is 0.09 for GANBERT. We investigated more to insure why this exception happens and noticed that in Table 5.7, $EOD_{gender}$ for both models are the same. However, again we need to point out that we cannot rely on exceptions and details trends because of the aforementioned reasons for the inaccuracy issue in the fine-grained terms fairness analysis.




(a) EOD for fine-grained terms for ratio of 0.005

(b) EOD for fine-grained terms for ratio of 0.01

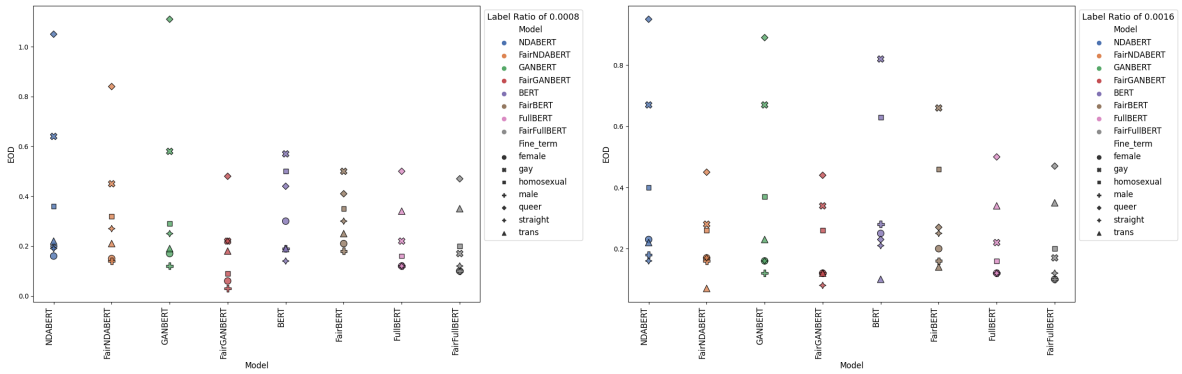
(c) EOD for fine-grained terms for ratio of 0.05

(d) EOD for fine-grained terms for ratio of 0.1

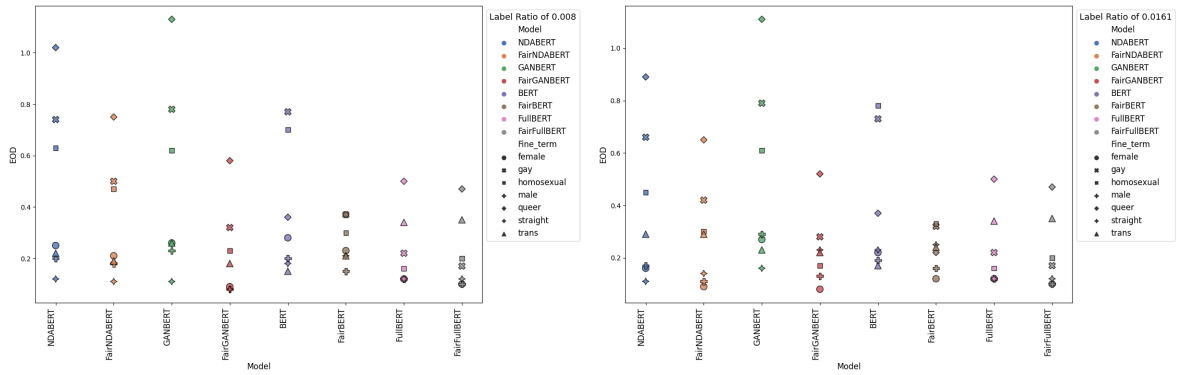
Figure 5.10: This Figure reports the EOD for fine-grained terms related to race in the HateXplain dataset over different label ratios for models. It is important to note that FullBERT and FairFullBERT are trained on fully-labeled training sets and different ratios of labeled data do not apply to these models.

In Figure 5.10, we can identify the same pattern as the differences between maximum and minimum EOD values of fine-grained terms in fair models is less than their non-fair counterparts; also, in the majority of the cases, the fair version of each model decreases the maximum EOD of fine-grained terms in respects to non-fair models.

On the other hand, the ratio of 0.005 does not follow these trends at all. Although by looking at 5.8, we identify a decline in overall $EOD_{Race}$ of fair models to their non-fair version in this ratio. Fair models do not decrease the differences between maximum and minimum EOD values of fine-grained terms; instead, they radically decline the EOD of several terms while increasing the others. This could indicate that always achieving fairness in coarse-grained terms does not guarantee fair treatment of fine-grained demographic groups. However, we think this happens due to low accuracy models achieved in lower labeled data ratios such as 0.005, where models cannot fully capture the underlying data distribution to generalize the problem well. Therefore, the low EOD for gender and race, in this case, is not a sign of fairness, but it is a sign of underfitting the data distribution.



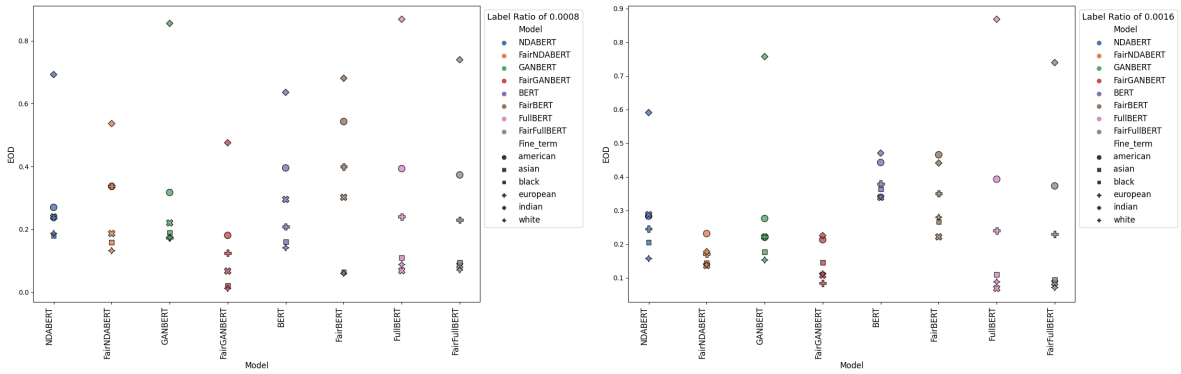

(a) EOD for fine-grained terms for ratio of 0.0008 (b) EOD for fine-grained terms for ratio of 0.0016



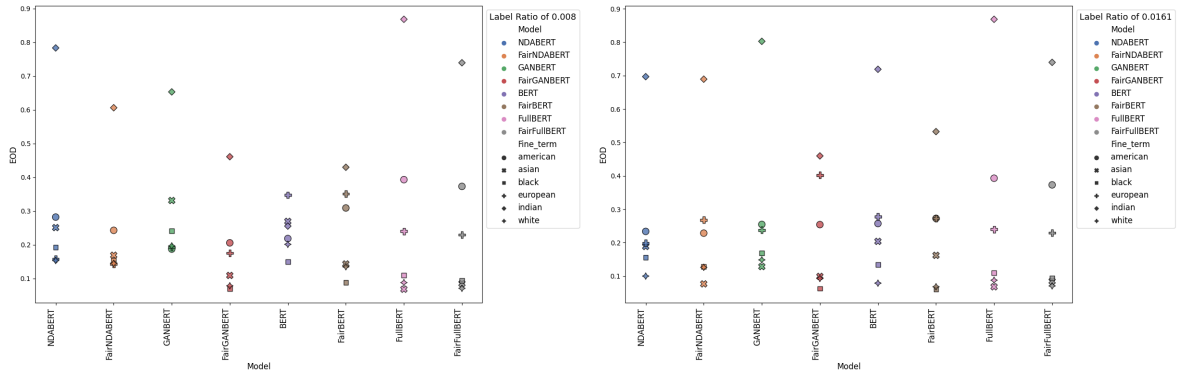

(c) EOD for fine-grained terms for ratio of 0.008 (d) EOD for fine-grained terms for ratio of 0.0161

Figure 5.11: This Figure reports the EOD for fine-grained terms related to gender in the Wiki Toxicity dataset over different label ratios for models. It is important to note that FullBERT and FairFullBERT are trained on fully-labeled training sets and different ratios of labeled data do not apply to these models.

Figure 5.11, and 5.12 follow precisely the same trends mentioned before, not only the maximum EOD of fine-grained terms in fair models with respect to non-fair counterparts decrease, but the difference between max and min EOD of fine-grained terms for each fair model in all ration decreased in compared to their non-fair models.




(a) EOD for fine-grained terms for ratio of 0.0008 (b) EOD for fine-grained terms for ratio of 0.0016




(c) EOD for fine-grained terms for ratio of 0.008 (d) EOD for fine-grained terms for ratio of 0.0161

Figure 5.12: This Figure reports the EOD for fine-grained terms related to race in the Wiki Toxicity dataset over different label ratios for models. It is important to note that FullBERT and FairFullBERT are trained on fully-labeled training sets and different ratios of labeled data do not apply to these models.

Finally, we aim to provide a unified answer to the research question $Q2$.

**Q2: Can we improve fairness for fine-grained terms related to demographic groups in fair models while debiasing our non-fair models on coarse-grained terms?**

The general trend illustrates that debiasing non-fair models on coarse-grained terms using adversarial debiasing can indeed improve the fairness in fine-grained terms. Therefore, for cases such as social fairness in online toxicity text classification where the fine-grained terms related to demographic groups change over time, we can apply fairness on coarse-

grained terms to enforce fairness on fine-grained terms. However, the quality of adversarial debiasing in decreasing the difference between EOD of fine-grained terms in the fair model is not comparable with Baldini et al. (2022), where they achieve almost identical EOD for fine-grained terms by debiasing the coarse-grained terms.

### 5.2.2 Experiment 2: Effect of labeled data on Accuracy and Fairness

In this experiment, we aim to see the effect of labeled data on the accuracy scores and fairness of the proposed fair semi-supervised and baseline semi-supervised models. Therefore, we look to identify whether adding more labeled data could affect accuracy and fairness directly. At the same time, we keep the unlabeled data fixed and aim to draw a meaningful conclusion related to research question $Q3$.

#### 5.2.2.1 HateXplain

The general trends in Table 5.13, and 5.14 for accuracy scores in HateXplain show that as the label ratio increases and both the balanced accuracy (BAcc) and the accuracy (Acc) scores increase for all models tested (NDAGAN, FairNDAGAN, GANBERT, FairGANBERT). This trend suggests that the presence of more labeled data in the training set improves the models' overall accuracy performance. For instance, in the case of FairNDAGAN, the balanced accuracy increases from 62.42% at label ratio 0.005 to 74.79% at label ratio 0.1, and the accuracy increases from 65.76% at label ratio 0.005 to 73.51% at label ratio 0.1. Similar trends are observed for the other models as well.

| Model | BAcc | | | |
|---|---|---|---|---|
| | (r = 0.005) | (r = 0.01) | (r = 0.05) | (r = 0.1) |
| NDAGAN | 64.57 ± 2.35 | 69.90 ± 1.62 | 73.61 ± 1.41 | 76.06 ± 0.82 |
| FairNDAGAN | 62.42 ± 4.32 | 67.05 ± 3.08 | 72.33 ± 2.01 | 74.79 ± 0.46 |
| GANBERT | 65.76 ± 3.70 | 69.89 ± 1.74 | 73.74 ± 1.02 | 75.52 ± 1.54 |
| FairGANBERT | 64.85 ± 3.79 | 68.52 ± 2.19 | 72.70 ± 1.41 | 74.70 ± 0.66 |

Table 5.13: This table presents the results of Experiment 2, showing the balanced accuracy scores for our proposed models (FairNDAGAN and FairGANBERT) and semi-supervised baseline models on the HateXplain dataset. Each column corresponds to a different label ratio and displays the balanced accuracy of the respective models, averaged over five runs. The standard deviation for each model is also provided. For more information about the number of labeled data points per class in each ratio, please refer to Table 5.3.

We can identify some interesting trends from Table 5.13, and 5.14. For instance, the FairNDAGAN and FairGANBERT models, designed to promote fairness, generally show slightly lower accuracy and balanced accuracy than their non-fair counterparts (NDAGAN and GANBERT). This decrease in performance is associated with the fairness and accuracy trade-off, where an increase in fairness might lead to a slight compromise in accuracy. However, the difference between the fairness-integrated and non-fair models is moderate, indicating that promoting fairness does not seriously impair the models' performance. Although, in this experiment, we do not analyze $H$, this observation emphasizes on our previous finding that creating more fair semi-supervised models without substantial sacrifice on accuracy is possible.

Moreover, although this experiment is designed to detect the effect of labeled data, we can still provide some insights about the overall performance of models when we have fewer unlabeled data compared to experiment 1. In this case, with 5,000 unlabeled data in HateXplain, the NDAGAN and FairNDAGAN models slightly outperformed the GANBERT and FairGANBERT models in terms of accuracy. This suggests that the NDA process might be more effective when the number of unlabeled data is low.

| Model | Acc | | | |
|---|---|---|---|---|
| | (r = 0.005) | (r = 0.01) | (r = 0.05) | (r = 0.1) |
| NDAGAN | 68.24 ± 1.73 | 71.34 ± 0.67 | 75.02 ± 0.79 | 76.51 ± 0.77 |
| FairNDAGAN | 65.76 ± 1.36 | 69.28 ± 2.11 | 71.53 ± 1.46 | 73.51 ± 1.78 |
| GANBERT | 68.62 ± 2.42 | 71.01 ± 0.56 | 74.67 ± 0.84 | 76.61 ± 0.99 |
| FairGANBERT | 66.01 ± 2.91 | 68.50 ± 1.92 | 71.53 ± 1.68 | 73.17 ± 1.52 |

Table 5.14: This table presents the results of Experiment 2, showing the accuracy scores for our proposed models (FairNDAGAN and FairGANBERT) and semi-supervised baseline models on the HateXplain dataset. Each column corresponds to a different label ratio and displays the accuracy of the respective models, averaged over five runs. The standard deviation for each model is also provided.

By looking at Table 5.15 and 5.16, we can compare the EOD values for the different label ratios to investigate the general relationship between more labeled data and its influence on fairness metric (EOD). For example, for $EOD_{Gender}$, NDAGAN and GANBERT do not show a consistent trend. While NDAGAN's $EOD_{Gender}$ initially increases with more data (from r = 0.005 to r = 0.01), it then decreases (from r = 0.01 to r = 0.05) and then drops even further (from r = 0.05 to r = 0.1). GANBERT, on the other hand, initially increases, then remains the same, and finally decreases. FairNDAGAN and FairGANBERT both show a clear downward trend in $EOD_{Gender}$ with more labeled data, which can indicate that more labeled data leads to better fairness outcomes for these models.

| Model | $EOD_{\text{Gender}}$ | | | |
|---|---|---|---|---|
| | (r = 0.005) | (r = 0.01) | (r = 0.05) | (r = 0.1) |
| NDAGAN | 0.12 ± 0.04 | 0.16 ± 0.05 | 0.12 ± 0.04 | 0.10 ± 0.04 |
| FairNDAGAN | 0.07 ± 0.02 | 0.06 ± 0.02 | 0.05 ± 0.02 | 0.04 ± 0.02 |
| GANBERT | 0.10 ± 0.02 | 0.11 ± 0.03 | 0.12 ± 0.05 | 0.10 ± 0.05 |
| FairGANBERT | 0.10 ± 0.04 | 0.05 ± 0.03 | 0.06 ± 0.02 | 0.02 ± 0.01 |

Table 5.15: This table presents the results of Experiment 2, showing the equalized odds difference for the gender demographic group as a measure of fairness. Each column corresponds to a different label ratio and displays the $EOD_{\text{Gender}}$ of the respective models, averaged over five runs. The standard deviation for each model is also provided.

For $EOD_{Race}$, NDAGAN and FairNDAGAN do not show a clear trend. While NDA-

GAN's EOD initially decreases, it finally increases again. FairNDAGAN, on the other hand, initially decreases, then remains the same, and finally slightly increases. GANBERT shows an increasing trend in EOD, indicating that more labeled data might lead to worse fairness outcomes. FairGANBERT, on the other hand, shows a slightly increasing trend.

| Model | $EOD_{\text{Race}}$ | | | |
|---|---|---|---|---|
| | (r = 0.005) | (r = 0.01) | (r = 0.05) | (r = 0.1) |
| NDAGAN | 0.12 ± 0.03 | 0.07 ± 0.02 | 0.13 ± 0.04 | 0.14 ± 0.03 |
| FairNDAGAN | 0.03 ± 0.01 | 0.06 ± 0.02 | 0.06 ± 0.02 | 0.07 ± 0.03 |
| GANBERT | 0.11 ± 0.03 | 0.10 ± 0.03 | 0.12 ± 0.02 | 0.16 ± 0.03 |
| FairGANBERT | 0.04 ± 0.01 | 0.05 ± 0.02 | 0.04 ± 0.02 | 0.07 ± 0.03 |

Table 5.16: This table presents the results of Experiment 2, showing the equalized odds difference for the race demographic group as a measure of fairness. Each column corresponds to a different label ratio and displays the $EOD_{\text{Race}}$ of the respective models, averaged over five runs. The standard deviation for each model is also provided.

The results of Experiment 2 on HateXplain show that the trend between more labeled data and fairness is dependent on the model and the demographic group.

For gender, FairNDAGAN and FairGANBERT, the models integrating fairness measures, consistently show a decrease in $EOD_{Gender}$ as the amount of labeled data increases, indicating an improvement in fairness for the gender demographic. However, this trend is less apparent for the non-fair models, NDAGAN and GANBERT. These observations imply that for gender fairness, more labeled data leads to better fairness outcomes in HateXplain, specifically for the fair models.

On the other hand, the relationship between more labeled data and race fairness needs to be clarified. While FairGANBERT shows a slightly increasing trend in $EOD_{Race}$, FairNDAGAN does not show a clear trend. NDAGAN and GANBERT also do not show consistent trends for race fairness. This suggests that the effect of adding more labeled data on fairness is more complex and may depend on other factors, such as the specific model used or the characteristics of the data.

However, the only trend that we identify is that in higher ratios of labeled data, fair models are able to improve the fairness more than their non-fair counterparts (by decreasing EOD). For example, $EOD_{Gender}$ in ratio 0.001 decrease around 0.05 between FairNDAGAN and NDAGAN, and no decrease between, FairGANBERT and GANBERT while these

values for ratio 0.1 is 0.06 and 0.08 respectively. It is important to note that we are still not sure if this trend is because of more labeled data or generally more data because, in the debasing process, we have used unlabeled and labeled data to improve fairness (both labeled and unlabeled data have coarse-grained labels related to gender and race.).

Overall, these results suggest that the impact of more labeled data on fairness depends on the model and the demographic group. It also suggests that incorporating fairness measures into models can improve fairness outcomes, especially when more labeled data is available. Despite this, the relationship between more labeled data and fairness is not straightforward, and more research may be needed to understand this relationship better.

#### 5.2.2.2 Wiki Toxicity

To analyze the results fully for Wiki Toxicity on accuracy scores, we consider Table 5.17, and 5.18.

| Model | BAcc | | | |
|---|---|---|---|---|
| | (r = 0.0008) | (r = 0.0016) | (r = 0.008) | (r = 0.0161) |
| NDAGAN | 81.94 ± 2.98 | 84.85 ± 2.72 | 85.92 ± 2.91 | 85.62 ± 3.46 |
| FairNDAGAN | 84.44 ± 4.28 | 86.85 ± 3.78 | 88.03 ± 4.64 | 87.45 ± 3.28 |
| GANBERT | 76.08 ± 4.15 | 84.36 ± 1.27 | 85.58 ± 4.93 | 86.32 ± 2.28 |
| FairGANBERT | 84.23 ± 3.74 | 82.89 ± 4.98 | 85.96 ± 4.50 | 88.25 ± 3.16 |

Table 5.17: This table presents the results of Experiment 2, showing the balanced accuracy scores for our proposed models (FairNDAGAN and FairGANBERT) and semi-supervised baseline models on the Wiki Toxicity dataset. Each column corresponds to a different label ratio and displays the balanced accuracy of the respective models, averaged over five runs. The standard deviation for each model is also provided. For more information about the number of labeled data points per class in each ratio, please refer to Table 5.3.

From the Tables, we can identify that NDAGAN and GANBERT accuracy scores consistently increase as we add more labeled data. Interestingly, the balanced accuracy of FairNDAGAN and FairGANBERT tends to increase as the label ratio increases, suggesting that having more labeled data can improve the model's performance in terms of balanced accuracy. However, the accuracy performance is somewhat mixed for all models. For instance, FairNDAGAN performed best at the label ratio of 0.0161, while FairGANBERT performed best at the label ratio of 0.0016. Therefore, the general trend illustrates the

availability of more labeled data improves the balanced accuracy of the models, particularly for FairNDAGAN and FairGANBERT. However, the relationship between the amount of labeled data and accuracy is more complex, and other factors, such as the number of unlabeled data in the training data, can influence it.

In addition, we observed some noteworthy trends in analyzing our models' performance on the Wiki Toxicity dataset. First, FairNDAGAN and FairGANBERT showed superior performance compared to their non-fair counterparts regarding balanced accuracy. This enhancement could be attributed to the semi-supervised baselines' inability to fully fit the data distribution in a highly imbalanced dataset like Wiki Toxicity. The adversarial debiasing stage allowed the models to train further on the data, improving their performance. It's worth noting that when the number of unlabeled data was larger in Experiment 1, we did not observe this trend, indicating that we need more evidence for comprehensive justification of this trend.

This contradiction between higher balanced accuracy for fair models with respect to non-fair counterparts highlights the complexity of model evaluation in imbalanced scenarios. Because the training set is relatively balanced (due to the sampling regime for Experiment 2), but the test set is highly imbalanced, non-fair models may overfit the balanced characteristics of the training set, thereby underperforming on the imbalanced test set. In contrast, fair models like FairNDAGAN and FairGANBERT are designed to minimize disparities in performance across different classes or demographic groups. They might be more robust to the shift from a balanced training set to an imbalanced test set due to their focus on minimizing disparities in performance across classes. Therefore, the higher balanced accuracy of the fair models, in this case, could be attributed to their ability to handle better the shift in class distribution from the training set to the test set. This highlights the potential benefits of the adversarial debiasing technique when dealing with imbalanced data, even when the imbalance is only present in the test set.

In conclusion, the results of our experiments demonstrate the complicated dynamics between model fairness, accuracy, and the distribution of labeled data in semi-supervised learning. Furthermore, they highlight the potential of fair models in handling class imbalances and the role of labeled data in model performance. Future research could delve deeper into these by providing further insights into the trade-offs and interaction between fairness and accuracy scores in semi-supervised learning in the case of highly imbalanced datasets.

| Model | Acc | | | |
|---|---|---|---|---|
| | (r = 0.0008) | (r = 0.0016) | (r = 0.008) | (r = 0.0161) |
| NDAGAN | 94.08 ± 0.82 | 94.76 ± 1.15 | 94.80 ± 0.73 | 95.39 ± 0.36 |
| FairNDAGAN | 92.21 ± 2.89 | 89.77 ± 3.56 | 93.66 ± 1.44 | 95.31 ± 0.61 |
| GANBERT | 80.78 ± 4.50 | 94.95 ± 0.92 | 94.88 ± 0.90 | 95.40 ± 0.12 |
| FairGANBERT | 92.52 ± 1.92 | 93.57 ± 1.91 | 94.30 ± 2.08 | 92.88 ± 3.81 |

Table 5.18: This table presents the results of Experiment 2, showing the accuracy of our proposed models (FairNDAGAN and FairGANBERT) and semi-supervised baseline models on the Wiki Toxicity dataset. Each column corresponds to a different label ratio and displays the balanced accuracy of the respective models, averaged over five runs. The standard deviation for each model is also provided. For more information about the number of labeled data points per class in each ratio, please refer to Table 5.3.

We can compare the EOD values for the different label ratios to investigate the general relationship between more labeled data and its influence on fairness metric (EOD) in the Wiki Toxicity dataset.

| Model | $EOD_{\text{Gender}}$ | | | |
|---|---|---|---|---|
| | (r = 0.0008) | (r = 0.0016) | (r = 0.008) | (r = 0.0161) |
| NDAGAN | 0.13 ± 0.05 | 0.19 ± 0.05 | 0.24 ± 0.04 | 0.21 ± 0.07 |
| FairNDAGAN | 0.06 ± 0.02 | 0.05 ± 0.03 | 0.11 ± 0.04 | 0.08 ± 0.04 |
| GANBERT | 0.12 ± 0.05 | 0.21 ± 0.08 | 0.24 ± 0.07 | 0.22 ± 0.04 |
| FairGANBERT | 0.07 ± 0.03 | 0.03 ± 0.01 | 0.10 ± 0.03 | 0.11 ± 0.05 |

Table 5.19: This table presents the results of Experiment 2, showing the equalized odds difference for the gender demographic group as a measure of fairness. Each column corresponds to a different label ratio and displays the $EOD_{\text{Gender}}$ of the respective models, averaged over five runs. The standard deviation for each model is also provided.

From Tables 5.19 and 5.20, we can compare the EOD values for the different label ratios to investigate the general relationship between more labeled data and its influence on fairness metric (EOD) in the Wiki Toxicity dataset.

For both gender and race categories, there appears to be no straightforward relationship between an increase in labeled data and the EOD values for the models. Furthermore, the

EOD does not consistently decrease or increase with more labeled data, suggesting that the fairness of the models (as measured by EOD) is not directly proportional to the amount of labeled data available. Also, When fairness measures are incorporated (FairNDAGAN and FairGANBERT), both models perform similarly, with FairGANBERT showing slightly better EOD at lower label ratios for the gender category. The FairNDAGAN and FairGANBERT models generally have lower EOD values than their non-fair counterparts. However, the relationship between EOD and the amount of labeled data is complex, and Other factors might influence it.

| Model | $\text{EOD}_{\text{Race}}$ | | | |
|---|---|---|---|---|
| | (r = 0.0008) | (r = 0.0016) | (r = 0.008) | (r = 0.0161) |
| NDAGAN | 0.10 ± 0.04 | 0.15 ± 0.04 | 0.17 ± 0.06 | 0.12 ± 0.06 |
| FairNDAGAN | 0.08 ± 0.03 | 0.13 ± 0.04 | 0.11 ± 0.03 | 0.09 ± 0.04 |
| GANBERT | 0.08 ± 0.04 | 0.16 ± 0.04 | 0.16 ± 0.06 | 0.13 ± 0.04 |
| FairGANBERT | 0.06 ± 0.02 | 0.10 ± 0.04 | 0.11 ± 0.05 | 0.12 ± 0.04 |

Table 5.20: This table presents the results of Experiment 2, showing the equalized odds difference for the gender demographic group as a measure of fairness. Each column corresponds to a different label ratio and displays the $EOD_{\text{Race}}$ of the respective models, averaged over five runs. The standard deviation for each model is also provided.

**Q3: What is the role of the labeled data in semi-supervised toxicity text classification in terms of fairness and accuracy?**

Generally speaking, the availability of more labeled data improves the classification performance of our semi-supervised models in terms of accuracy and balanced accuracy. However, the amount of labeled data does not appear to have a straightforward relationship with the fairness of the models, as measured by the EOD for both race and gender demographic groups. Neither an increase nor a decrease in the EOD values is consistently observed with increasing amounts of labeled data. This suggests that the fairness of the model, in terms of both race and gender, is not directly proportional to the amount of labeled data, at least in our proposed and baseline semi-supervised models.

### 5.2.3 Experiment 3: Effect of Unlabeled Data on Accuracy and Fairness

In this experiment, we aim to see the effect of unlabeled data on the accuracy scores and fairness of the proposed fair semi-supervised and baseline semi-supervised models. In other words, we wanted to see if increasing the number of unlabeled data would positively affect these two criteria or not.

This section first looks at the trends in the results across each dataset. Then, we draw conclusions related to research question $Q4$. It is important to note that we have conducted the experiments for each dataset with two different labeled data ratios (one low, one high) to make sure the amount of labeled data does not impact the results. Here, we only present results from the lower labeled data ratio; however, since the outcomes of higher ratios do not contribute any additional information to our analysis, we report them in Appendix A for clarity.

#### 5.2.3.1 HateXplain

Regarding Balanced Accuracy (BAcc) (Table 5.21), we can see that as the quantity of unlabeled data increases, the BAcc scores for NDAGAN and GANBERT models show a slight increase, peaking at n=5000 and then decreasing slightly. FairNDAGAN sees a slight decline in performance up to n=5000 and then a 4% jump at n=10000. FairGANBERT's performance is somewhat unpredictable, dropping at n=5000, then increasing again at n=10000.

| Model | BAcc | | | |
|---|---|---|---|---|
| | (n = 1000) | (n = 2500) | (n = 5000) | (n = 10000) |
| NDAGAN | 64.79 ± 1.34 | 64.89 ± 1.96 | 66.10 ± 1.78 | 65.23 ± 3.53 |
| FairNDAGAN | 64.48 ± 4.90 | 63.92 ± 4.63 | 63.93 ± 4.10 | 67.36 ± 3.95 |
| GANBERT | 64.02 ± 4.72 | 64.61 ± 1.56 | 65.57 ± 3.03 | 65.21 ± 2.61 |
| FairGANBERT | 61.11 ± 4.13 | 63.98 ± 4.76 | 60.70 ± 3.00 | 64.41 ± 3.60 |

Table 5.21: This table presents the balanced accuracy results of models in Experiment 3 for the HateXplain dataset. Each column corresponds to a different number of unlabeled data in the training set, and the labeled data ratio is 0.005. Results are computed by averaging over five runs, and the standard deviation for each model is also provided.

Looking at accuracy (Acc) in Table 5.22, all models maintain a relatively stable performance as unlabeled data increases. The highest accuracy score is seen with NDAGAN, while the lowest is seen with FairGANBERT.

By looking at the accuracy scores results for HateXplain, It is evident that the accuracy and balanced accuracy does not necessarily improve if we increase the number of unlabeled data.

| Model | Acc | | | |
|---|---|---|---|---|
| | (n = 1000) | (n = 2500) | (n = 5000) | (n = 10000) |
| NDAGAN | 67.85 ± 2.38 | 67.79 ± 2.33 | 68.14 ± 1.87 | 68.17 ± 3.01 |
| FairNDAGAN | 66.83 ± 3.31 | 67.02 ± 3.15 | 65.96 ± 1.75 | 66.75 ± 3.70 |
| GANBERT | 67.99 ± 2.93 | 67.79 ± 2.61 | 67.54 ± 2.91 | 67.94 ± 2.18 |
| FairGANBERT | 65.88 ± 3.00 | 67.03 ± 3.57 | 65.51 ± 2.85 | 66.80 ± 3.41 |

Table 5.22: This table presents the accuracy results of models in Experiment 3 for the HateXplain dataset. Each column corresponds to a different number of unlabeled data in the training set, and the labeled data ratio is 0.005. Results are computed by averaging over five runs, and the standard deviation for each model is also provided.

From Table 5.23, and 5.24, by considering fairness measured by $EOD_{\text{Gender}}$, FairNDAGAN and FairGANBERT models seem to maintain relatively stable and maintain lower score than semi-supervised baselines. NDAGAN and GANBERT, on the other hand, have slightly higher $EOD_{\text{Gender}}$ scores, and no trend that indicates increasing the unlabeled data improve $EOD_{\text{Gender}}$ can be identified.

For $EOD_{\text{Race}}$, the scores for all models seem relatively stable, with a slight increase or decrease as n increases. FairNDAGAN and FairGANBERT models maintain lower scores, indicating better fairness related to race.

| Model | $\text{EOD}_{\text{Gender}}$ | | | |
|---|---|---|---|---|
| | (n = 1000) | (n = 2500) | (n = 5000) | (n = 10000) |
| NDAGAN | 0.10 ± 0.04 | 0.08 ± 0.03 | 0.08 ± 0.03 | 0.09 ± 0.02 |
| FairNDAGAN | 0.05 ± 0.02 | 0.02 ± 0.01 | 0.03 ± 0.01 | 0.02 ± 0.01 |
| GANBERT | 0.06 ± 0.03 | 0.08 ± 0.03 | 0.11 ± 0.04 | 0.07 ± 0.02 |
| FairGANBERT | 0.03 ± 0.02 | 0.05 ± 0.02 | 0.05 ± 0.02 | 0.02 ± 0.01 |

Table 5.23: This table presents the $EOD_{\text{Gender}}$ results of models in Experiment 3 for the HateXplain dataset. Each column corresponds to a different number of unlabeled data in the training set, and the labeled data ratio is 0.005. Results are computed by averaging over five runs, and the standard deviation for each model is also provided.

In general, In terms of fairness, for both $EOD_{\text{Gender}}$, and $EOD_{\text{Race}}$ we cannot identify a meaningful trend that guarantees fairness improvement by increasing in the number of unlabeled data. However, there are some decreases in EOD for both race and gender for some models, but it is not consistent and considerable.

| Model | $\text{EOD}_{\text{Race}}$ | | | |
|---|---|---|---|---|
| | (n = 1000) | (n = 2500) | (n = 5000) | (n = 10000) |
| NDAGAN | 0.15 ± 0.05 | 0.15 ± 0.05 | 0.14 ± 0.04 | 0.09 ± 0.04 |
| FairNDAGAN | 0.04 ± 0.03 | 0.06 ± 0.02 | 0.05 ± 0.03 | 0.07 ± 0.02 |
| GANBERT | 0.16 ± 0.06 | 0.15 ± 0.05 | 0.13 ± 0.05 | 0.13 ± 0.04 |
| FairGANBERT | 0.05 ± 0.02 | 0.05 ± 0.03 | 0.07 ± 0.03 | 0.06 ± 0.02 |

Table 5.24: This table presents the $EOD_{\text{Race}}$ results of models in Experiment 3 for the HateXplain dataset. Each column corresponds to a different number of unlabeled data in the training set, and the labeled data ratio is 0.005. Results are computed by averaging over five runs, and the standard deviation for each model is also provided.

In summary, the addition of unlabeled data in this context does not seriously affect the accuracy and fairness of the models. However, these findings suggest that the role of unlabeled data in semi-supervised toxicity text classification is complex, and we need to select high-quality unlabeled data where the selection criteria are set based on our needs and it is in line with the problem we aim to address.

#### 5.2.3.2 Wiki Toxicity

In terms of Balanced Accuracy (BAcc) shown in Table 5.25, the results indicate a slight decrease in BAcc as the number of unlabeled instances (n) increases for NDAGAN, FairNDAGAN, and GANBERT. However, FairGANBERT shows a slight increase in performance up to n=25000 and then a decrease.

| Model | BAcc | | | |
|---|---|---|---|---|
| | (n = 10000) | (n = 25000) | (n = 50000) | (n = 95000) |
| NDAGAN | 83.85 ± 2.81 | 83.02 ± 3.04 | 81.94 ± 2.98 | 82.84 ± 4.83 |
| FairNDAGAN | 83.02 ± 5.56 | 77.51 ± 6.68 | 84.44 ± 4.28 | 81.28 ± 4.61 |
| GANBERT | 83.05 ± 2.46 | 81.47 ± 1.41 | 76.08 ± 3.15 | 81.38 ± 2.04 |
| FairGANBERT | 84.59 ± 3.34 | 85.93 ± 3.88 | 84.23 ± 3.74 | 80.60 ± 2.46 |

Table 5.25: This table presents the accuracy results of models in Experiment 3 for the Wiki Toxicity dataset. Each column corresponds to a different number of unlabeled data in the training set, and the labeled data ratio is 0.0008. Results are computed by averaging over five runs, and the standard deviation for each model is also provided.

Regarding Accuracy (Acc) illustrated in Table 5.26, NDAGAN and GANBERT maintain relatively stable performance as unlabeled instances increase. FairNDAGAN shows a small decrease in performance as n increases, while FairGANBERT's performance fluctuates. Generally, no consistent pattern shows that adding more unlabeled data can increase the accuracy scores in this context.

| Model | Acc | | | |
|---|---|---|---|---|
| | (n = 10000) | (n = 25000) | (n = 50000) | (n = 95000) |
| NDAGAN | 93.66 ± 1.81 | 93.84 ± 1.85 | 94.08 ± 0.82 | 93.57 ± 1.96 |
| FairNDAGAN | 91.81 ± 2.38 | 91.06 ± 3.04 | 92.21 ± 2.89 | 90.84 ± 3.35 |
| GANBERT | 93.74 ± 1.98 | 94.46 ± 0.83 | 80.78 ± 4.50 | 94.44 ± 0.74 |
| FairGANBERT | 91.03 ± 3.21 | 89.94 ± 2.67 | 92.52 ± 1.92 | 91.84 ± 2.67 |

Table 5.26: This table presents the accuracy results of models in Experiment 3 for the Wiki Toxicity dataset. Each column corresponds to a different number of unlabeled data in the training set, and the labeled data ratio is 0.0008. Results are computed by averaging over five runs, and the standard deviation for each model is also provided.

Considering fairness for gender In Table 5.27, measured by $EOD_{\text{Gender}}$, FairNDAGAN and FairGANBERT models seem to maintain relatively stable and lower EOD scores as the number of unlabeled instances increases, indicating consistent fairness related to gender. NDAGAN and GANBERT, on the other hand, have slightly higher $EOD_{\text{Gender}}$ scores. However, we can not identify any consistent trend that suggests adding more unable data necessarily improves fairness in terms of EOD.

| Model | $\text{EOD}_{\text{Gender}}$ | | | |
|---|---|---|---|---|
| | (n = 10000) | (n = 25000) | (n = 50000) | (n = 95000) |
| NDAGAN | 0.17 ± 0.01 | 0.23 ± 0.08 | 0.21 ± 0.07 | 0.24 ± 0.08 |
| FairNDAGAN | 0.09 ± 0.04 | 0.14 ± 0.05 | 0.08 ± 0.04 | 0.13 ± 0.05 |
| GANBERT | 0.22 ± 0.09 | 0.20 ± 0.09 | 0.22 ± 0.04 | 0.21 ± 0.09 |
| FairGANBERT | 0.13 ± 0.04 | 0.09 ± 0.02 | 0.11 ± 0.04 | 0.06 ± 0.02 |

Table 5.27: This table presents the $EOD_{\text{Gender}}$ results of models in Experiment 3 for the Wiki Toxicity dataset. Each column corresponds to a different number of unlabeled data in the training set, and the labeled data ratio is 0.0008. Results are computed by averaging over five runs, and the standard deviation for each model is also provided.

For $EOD_{\text{Race}}$, the scores for all models seem stable, with a slight increase or decrease as n increases. For example, FairNDAGAN and FairGANBERT models maintain lower scores, indicating better fairness related to race, however, their $EOD_{\text{Race}}$ does not decrease as n increases. The only model that shows some fairness improvement as the number of unlabeled data increases is GANBERT, although the improvement is not considerable. Generally, the addition of unlabeled data in this context does not seem to have a considerable positive effect on the accuracy and fairness of the models.

| Model | $EOD_{Race}$ | | | |
|---|---|---|---|---|
| | (n = 10000) | (n = 25000) | (n = 50000) | (n = 95000) |
| NDAGAN | 0.12 ± 0.04 | 0.10 ± 0.03 | 0.12 ± 0.03 | 0.12 ± 0.02 |
| FairNDAGAN | 0.08 ± 0.04 | 0.11 ± 0.03 | 0.09 ± 0.02 | 0.08 ± 0.03 |
| GANBERT | 0.16 ± 0.04 | 0.11 ± 0.05 | 0.13 ± 0.04 | 0.12 ± 0.06 |
| FairGANBERT | 0.09 ± 0.04 | 0.09 ± 0.04 | 0.12 ± 0.04 | 0.10 ± 0.03 |

Table 5.28: This table presents the $EOD_{Race}$ results of models in Experiment 3 for the Wiki Toxicity dataset. Each column corresponds to a different number of unlabeled data in the training set, and the labeled data ratio is 0.0008. Results are computed by averaging over five runs, and the standard deviation for each model is also provided.

**Q4: What is the role of unlabeled data in semi-supervised toxicity text classification in terms of fairness and accuracy?**

Overall, as the amount of unlabeled data (n) increases, the accuracy (Acc) and Balanced Accuracy (BAcc) of the models fluctuate without a clear trend towards increasing or decreasing. For instance, the FairGANBERT model on the HateXplain dataset shows a slightly increasing trend in BAcc with increasing unlabeled data for the r= 0.005, but the trend is inconsistent for the r = 0.1. Similarly, for the Wiki Toxicity dataset, the FairNDAGAN model shows a rise in BAcc with increasing unlabeled data from 10000 to 50000 but then a fall when it reaches 95000. This implies that simply increasing the amount of unlabeled data does not necessarily improve model accuracy, and there may be other factors at play, such as the quality of the unlabeled data, the model's ability to leverage this data or the inherent complexity of the task.

In terms of fairness, measured by EOD for both gender and race, increasing the amount of unlabeled data does not greatly changes the EOD scores for all the models. This suggests that incorporating fairness into the model's design may be more influential than the amount of unlabeled data when achieving fairness in predictions.

In conclusion, the role of unlabeled data in semi-supervised toxicity text classification is complex and not entirely predictable. While it contributes to the models' performance in terms of accuracy and fairness, other factors, including the models' design and the quality of the unlabeled data, also play significant roles (Tian et al., 2004). Some previous studies have shown that the value of unlabeled data can greatly enhance learning performance under certain conditions, suggesting that the specific usage of unlabeled data can be more critical than a large number of unlabeled samples (Singh et al., 2008). Furthermore, not all

unlabeled data contribute equally to these improvements. Ren et al. (2020) have demonstrated that assigning unique weights to each unlabeled example, rather than treating all unlabeled data as equal, could cause better performance in semi-supervised learning tasks. Therefore, while simply adding unlabeled data does not guarantee improvements in fairness and accuracy, we hypothesize that, for future work, carefully sampling high-quality unlabeled data (where quality is defined based on context and our specific requirements) and intelligently incorporating it into the model could effectively improve performance in terms of both accuracy and fairness.

### 5.2.4 Key Findings and Limitations

In this section, we summarize the key findings as follows:

- Our study indicates that considering the fairness accuracy trade-off, the proposed fair semi-supervised toxicity classification framework can improve fairness without considerable loss of accuracy. Both FairNDAGAN and FairGANBERT models demonstrated this, showing considerable improvements in fairness as measured by lower EOD scores for both gender and race demographics while having comparable accuracy scores with their non-fair counterparts.

- The comparative performance of FairNDAGAN and FairGANBERT was inconclusive. Both models demonstrated slight advantages in different scenarios, but these differences were not noteworthy enough to declare a clear winner.

- The availability of labeled data generally improved the classification performance of our models in terms of accuracy and balanced accuracy. However, the amount of labeled data does not have a straightforward relationship with the fairness of the models, indicating the complexity of the relationship between labeled data and model fairness.

- The role of unlabeled data in semi-supervised toxicity text classification is complex and unpredictable. While increasing the amount of unlabeled data did not necessarily improve model accuracy or fairness.

- Our results indicate that debiasing non-fair models on coarse-grained terms using adversarial debiasing can improve fairness in fine-grained terms. This finding could be particularly useful in domains, such as online toxicity text classification, where fine-grained demographic terms change over time. However, the efficacy of adversarial

debiasing in reducing the difference between EOD scores for fine-grained terms in fair models is not as high as other techniques introduced in (Baldini et al., 2022).

Limitations of this work are identified as follows:

- Our semi-supervised models appear to be sensitive to highly imbalanced datasets, which is a common characteristic of real-world data. This could limit the utility of these models in practical applications and indicate the need for additional strategies like data augmentation and balancing techniques in such scenarios.
- We have not yet investigated the quality of the unlabeled data, which could be a key aspect in enhancing the performance of semi-supervised models. Future research should consider exploring this aspect further.
- The fairness of our models was evaluated only on gender and race demographics. The models' fairness for other social groups or demographic categories remains an open question and a potential limitation of our study.
- In the context of fine-grained demographic groups, achieving fairness in coarse-grained terms does not always guarantee fair treatment for these groups. This issue seems to be more apparent when models are unable to fully capture the underlying data distribution due to low accuracy, potentially leading to underfitting.
- Our models tend to perform differently regarding EOD for different demographic groups such as gender and race. This points to potential challenges in achieving fairness across different demographic groups and underscores the need for tailored strategies for each group or using a bias-mitigating method that can handle different demographic groups more equally.

# Chapter 6

# Conclusion and Future Work

This chapter initially reflects on the contributions and findings of this work. Then, it comprehensively summarizes the importance of understanding and achieving fairness in semi-supervised toxicity text classification. This chapter further highlights potential directions for future research. Finally, it concludes with an ethical statement, reaffirming our commitment to fairness, inclusivity, and integrity in our research.

## 6.1 Summary of Contributions

Through our research, we have made several contributions in the domain of fairness-aware semi-supervised learning, particularly focusing on toxicity text classification.

**Achieving Fairness in Semi-Supervised Learning:** We performed a comprehensive analysis to compare the performance of our proposed semi-supervised models against semi-supervised and supervised baselines in terms of both accuracy and fairness for the first time in toxicity text classification. Our models show that we can achieve fairness in semi-supervised toxicity text classification considering the trade-off between accuracy and fairness.

**Fair Semi-Supervised Learning Framework:** For the first time, we designed and developed two fair semi-supervised learning models, FairNDAGAN and FairGANBERT, to analyze social fairness in semi-supervised text classification.

**Understanding the Role of Labeled and Unlabeled Data:** We have explained the influence of both labeled and unlabeled data on the fairness and accuracy aspects of semi-supervised learning models.

**Investigation of Coarse-Grained vs. Fine-Grained Fairness:** We showed that reaching fairness at the coarse-grained level improves fairness at the fine-grained level but does not always guarantee it.

**Influence of Data Imbalance on Accuracy and Fairness:** Our study revealed susceptibilities of both supervised and semi-supervised models to imbalanced datasets in terms of accuracy and fairness.

The findings and insights from our work provide a solid foundation for future research, pushing advancements in understanding and achieving fairness in semi-supervised learning. Finally, we hope that our effort in this field will encourage other researchers to analyze social fairness in semi-supervised toxicity text classification.

## 6.2 Conclusion

As the digital landscape continues to evolve and social media platforms are taking their way more into our daily life, the issue of toxic language in user-generated comments has emerged as a significant challenge. Although ML models have been developed to detect and moderate such content automatically, they often exhibit unintended biases, particularly against comments containing sensitive demographic terms. Moreover, most of the developed frameworks for fair toxicity detection rely heavily on annotated datasets, and acquiring labels for different scenarios and platforms is expensive and time-consuming. Therefore, there is a need to develop state-of-the-art fair frameworks to detect toxic language with a limited number of label data.

In this research, we have initially introduced a fair semi-supervised framework to investigate if fairness is possible in a semi-supervised setting considering the trade-off between classification performance and fairness. Our findings suggest that incorporating fairness considerations into semi-supervised learning models can greatly improve fairness metrics without seriously sacrificing accuracy. Furthermore, our study highlighted the importance of fairness at the fine-grained level. Although our fair framework using adversarial debiasing achieved some level of fairness at the fine-grained level, we identify a need for a tailored way to enforce fairness more at the fine-grained level between demographic groups.

Moreover, we investigated the role of labeled and unlabeled data in our proposed semi-supervised framework regarding fairness and accuracy. Our research revealed that increasing the quantity of unlabeled data does not directly correlate with improved model performance or fairness. Instead, we think that the quality of unlabeled data and how

it is selected and incorporated plays an important role into the fairness and classification performance of semi-supervised models.

In our exploration of the role of labeled data in semi-supervised learning, we identified that the number of labeled data greatly influenced the accuracy of our models. However, fairness particularly was not affected by the number of labeled data. We think that in terms of fairness, the quality of data plays a more important role than the quantity of the data. However, this suggestion needs to be investigated in detail.

It is important to note that the fair models, FairNDAGAN and FairGANBERT, from our proposed fair semi-supervised framework consistently outperformed their non-fair semi-supervised counterparts in terms of fairness, demonstrating the potential of semi-supervised learning in fairness-aware ML. However, none of the fair models greatly outperformed the other, suggesting that the choice between these models might depend on the specific nature of the data, such as the degree of data imbalance. Furthermore, our study revealed the potential vulnerabilities of semi-supervised models to imbalanced datasets, suggesting the necessity of integrating data augmentation methods and balancing techniques in such cases.

In conclusion, our study has shed light on the potential and challenges of semi-supervised learning for fairness-aware toxicity text classification for the first time. We believe that our findings will provide a foundation for future research in this important area.

## 6.3 Future Work

In terms of future work, we have identified several bottlenecks and various directions that can be pursued to advance our understanding of fairness in semi-supervised toxicity text classification:

**Targeted groups detection**: In this study, we presented a range of benchmark datasets for toxicity text classification, highlighting the absence of information regarding targeted groups. This makes a significant challenge for fairness-related research within this domain. Consequently, developing an automated tool that detects targeted groups within these benchmarks can significantly advance research in this field.

**Improved metrics:** Given the importance of considering both accuracy and fairness, future research could focus on developing more sophisticated metrics that capture both aspects simultaneously.

**Improved data handling:** Future research could also focus on more effective ways to handle imbalanced data, such as advanced data augmentation and balancing techniques.

**Better use of unlabeled data:** Our findings suggest that simply increasing the amount of unlabeled data does not necessarily improve model accuracy or fairness. Therefore, future work could investigate ways to select and use high-quality unlabeled data more effectively.

**Fairness at the fine-grained level:** As our study revealed, achieving fairness at the coarse-grained level does not always translate into fairness at the fine-grained level. Thus, future work could focus on methods to ensure fairness at more granular levels.

**Non-binary gender identities** : In this work, such as most of the works, we considered genders are binary. However, there is a need to investigate the complexities of non-binary gender representations in toxicity text classification. Specifically, we suggest focusing on the role of non-binary gender identities and various gender pronouns in a fair toxicity text classification. This would be a significant step towards more inclusive and fair models. However, to consider all gender identities, there is a need to collect more diverse datasets, including samples that represent diverse gender identities.

**Exploring newer architectures and methodologies:** In this study, we employed GANs as our primary architecture. However, the rapidly evolving landscape of ML provides many newer alternative architectures and methods that can be employed for toxicity text classification, particularly when labeled data is limited. Specifically, future work could delve into the application of prompt engineering methods and employ the abilities of large language models (LLMs) for this task. These approaches offer promising avenues for improving both model performance and efficiency and may help to further unravel the trade-off between accuracy, fairness, and data constraints in the domain of semi-supervised toxicity text classification.

## 6.4 Ethical Statement

This research strictly followed the ethical guidelines and principles outlined by the University of Ottawa.

We acknowledge and affirm the existence and validity of non-binary gender identities. While our research currently addresses gender in a binary context due to the limitations of our data and existing research frameworks, we understand that this is a simplification and does not capture the full spectrum of human gender identities. We hope that future work will continue to develop methodologies to include non-binary and underrepresented gender identities in the context of fairness in machine learning.

Despite our endeavours to promote fairness and mitigate bias in machine learning models, we acknowledge that creating an absolutely fair model remains a complex and unsolved challenge. This research represents our best effort to contribute to the ongoing discourse and development toward achieving this goal.

During the preparation of this research, AI tools such as ChatGPT and Grammarly were employed for coding assistance and text editing. However, it should be considered that all ideas, concepts, and texts are original by the author, and without human inspection, interpretation, and approval; no text was directly used from AI-generated content. These tools were used strictly as assistance and not as sources of original content or ideas.

The author affirms that this research has been conducted with the utmost integrity. We were transparent in our methods and findings, and we welcome scrutiny and further discussion to enhance the quality and impact of this work.

# Bibliography


D. Andročec. Machine learning methods for toxic comment classification: a systematic review. *Acta Universitatis Sapientiae, Informatica*, 12(2):205–216, 2020.

M. Awad and R. Khanna. *Machine Learning*, pages 1–18. Apress, Berkeley, CA, 2015. ISBN 978-1-4302-5990-9. doi: 10.1007/978-1-4302-5990-9_1. URL https://doi.org/10.1007/978-1-4302-5990-9_1.

I. Baldini, D. Wei, K. N. Ramamurthy, M. Yurochkin, and M. Singh. Your fairness may vary: Pretrained language model fairness in toxic text classification. *arXiv preprint arXiv:2108.01250*, 2021.

I. Baldini, D. Wei, K. Natesan Ramamurthy, M. Singh, and M. Yurochkin. Your fairness may vary: Pretrained language model fairness in toxic text classification. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2245–2262, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.176. URL https://aclanthology.org/2022.findings-acl.176.

S. Barocas and A. D. Selbst. Big data's disparate impact. *California Law Review*, 104: 671, 2016.

S. Basu Roy Chowdhury, S. Ghosh, Y. Li, J. Oliva, S. Srivastava, and S. Chaturvedi. Adversarial scrubbing of demographic information for text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 550–562, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.43. URL https://aclanthology.org/2021.emnlp-main.43.

R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366755. doi: 10.1145/3308560.3317593. URL https://doi.org/10.1145/3308560.3317593.

T. Brennan and W. Dieterich. Correctional offender management profiles for alternative sanctions (compas). *Handbook of recidivism risk/needs assessment tools*, pages 49–75, 2018.

J. Chakraborty, H. Tu, S. Majumder, and T. Menzies. Can we achieve fairness using semi-supervised learning? *arXiv preprint arXiv:2111.02038*, 2021.

O. Chapelle, B. Scholkopf, and A. Zien, Eds. Semi-supervised learning (chapelle, o. et al., eds.; 2006) [book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009. doi: 10.1109/TNN.2009.2015974.

J. Chen, Z. Yang, and D. Yang. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. *arXiv preprint arXiv:2004.12239*, 2020.

R. T. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

D. Croce, G. Castellucci, and R. Basili. Kernel-based generative adversarial networks for weakly supervised learning. In *International Conference of the Italian Association for Artificial Intelligence*, 2019.

D. Croce, G. Castellucci, and R. Basili. GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.191. URL https://aclanthology.org/2020.acl-main.191.

Z. Dai, Z. Yang, F. Yang, W. W. Cohen, and R. R. Salakhutdinov. Good semi-supervised learning that requires a bad gan. *Advances in neural information processing systems*, 30, 2017.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

H. U. Dike, Y. Zhou, K. K. Deveerasetty, and Q. Wu. Unsupervised learning based on artificial neural network: A review. In *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, pages 322–327, 2018. doi: 10.1109/CBS.2018.8612259.

L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, 2018a.

L. Dixon, J. Li, J. S. Sorensen, N. Thain, and L. Vasserman. Measuring and mitigating unintended bias in text classification. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018b.

C. Drummond. *Classification*, pages 171–171. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_111. URL https://doi.org/10.1007/978-0-387-30164-8_111.

J. Dymond. Graceful degradation and related fields. *arXiv preprint arXiv:2106.11119*, 2021.

Y. Elazar and Y. Goldberg. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1002. URL https://aclanthology.org/D18-1002.

O. Gencoglu. Cyberbullying detection with fairness constraints. *IEEE Internet Computing*, 25(1):20–29, 2020.

A. Ghassami, S. Khodadadian, and N. Kiyavash. Fairness in supervised learning: An information theoretic approach. In *2018 IEEE international symposium on information theory (ISIT)*, pages 176–180. IEEE, 2018.

S. Gururangan, T. Dang, D. Card, and N. A. Smith. Variational pretraining for semi-supervised text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5880–5894, Florence, Italy, July 2019.

Association for Computational Linguistics. doi: 10.18653/v1/P19-1590. URL https://aclanthology.org/P19-1590.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

B. Kennedy, M. Atari, A. M. Davani, L. Yeh, A. Omrani, Y. Kim, K. Coombs, S. Havaldar, G. Portillo-Wightman, E. Gonzalez, J. Hoover, A. Azatian, A. Hussain, A. Lara, G. Cardenas, A. Omary, C. Park, X. Wang, C. Wijaya, Y. Zhang, B. Meyerowitz, and M. Dehghani. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56:79–108, 2022. ISSN 1574-0218. doi: 10.1007/s10579-021-09569-x. URL https://doi.org/10.1007/s10579-021-09569-x.

K. C. A. Khanzode and R. D. Sarode. Advantages and disadvantages of artificial intelligence and machine learning: A literature review. *International Journal of Library & Information Science (IJLIS)*, 9(1):3, 2020.

F. Liu and B. Avci. Incorporating priors with feature attribution on text classification. *arXiv preprint arXiv:1906.08286*, 2019.

H. Liu, W. Jin, H. Karimi, Z. Liu, and J. Tang. The authors matter: Understanding and mitigating implicit bias in deep text classification. *arXiv preprint arXiv:2105.02778*, 2021.

P. K. Lohia, K. N. Ramamurthy, M. Bhide, D. Saha, K. R. Varshney, and R. Puri. 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, 2019.

B. Mahesh. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9:381–386, 2020.

B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875, 2021.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 09 2011. ISSN 1067-5027. doi: 10.1136/amiajnl-2011-000464. URL https://doi.org/10.1136/amiajnl-2011-000464.

A. Narayan, I. Chami, L. Orr, and C. Ré. Can foundation models wrangle your data? *arXiv preprint arXiv:2205.09911*, 2022.

V. Noroozi, S. Bahaadini, S. Sheikhi, N. Mojab, and S. Y. Philip. Leveraging semi-supervised learning for fairness using neural networks. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 50–55. IEEE, 2019.

D. W. Otter, J. R. Medina, and J. K. Kalita. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):604–624, 2021. doi: 10.1109/TNNLS.2020.2979670.

Y. Ouali, C. Hudelot, and M. Tami. An overview of deep semi-supervised learning. *arXiv preprint arXiv:2006.05278*, 2020.

J. H. Park, J. Shin, and P. Fung. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*, 2018.

D. Pessach and E. Shmueli. Algorithmic fairness. *arXiv preprint arXiv:2001.09784*, 2020.

D. Pessach and E. Shmueli. A review on fairness in machine learning. *ACM Comput. Surv.*, 55(3), feb 2022. ISSN 0360-0300. doi: 10.1145/3494672. URL https://doi.org/10.1145/3494672.

Y. Pruksachatkun, S. Krishna, J. Dhamala, R. Gupta, and K.-W. Chang. Does robustness improve fairness? approaching fairness with word substitution robustness methods for text classification. *arXiv preprint arXiv:2106.10826*, 2021.

C. Qian, F. Feng, L. Wen, C. Ma, and P. Xie. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5434–5445, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.422. URL https://aclanthology.org/2021.acl-long.422.

Y. C. A. P. Reddy, P. Viswanath, and B. E. Reddy. Semi-supervised learning: a brief review. *International journal of engineering and technology*, 7:81, 2018.

Z. Ren, R. Yeh, and A. Schwing. Not all unlabeled data are equal: Learning to weight data in semi-supervised learning. *Advances in Neural Information Processing Systems*, 33:21786–21797, 2020.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

I. H. Sarker. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, 2(6):420, 2021.

P. C. Sen, M. Hajra, and M. Ghosh. Supervised classification algorithms in machine learning: A survey and review. In *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018*, pages 99–111. Springer, 2020.

S. Shayesteh and D. Inkpen. Generative adversarial learning with negative data augmentation for semi-supervised text classification. *The International FLAIRS Conference Proceedings*, 35, May 2022. doi: 10.32473/flairs.v35i.130722. URL https://journals.flvc.org/FLAIRS/article/view/130722.

A. Singh, R. D. Nowak, and X. Zhu. Unlabeled data: Now it helps, now it doesn't. In *NIPS*, 2008.

A. Sinha, K. Ayush, J. Song, B. Uzkent, H. Jin, and S. Ermon. Negative data augmentation. *arXiv preprint arXiv:2102.05113*, 2021.

X. Su, T. Miller, X. Ding, M. Afshar, and D. Dligach. Classifying long clinical documents with pre-trained transformers. *arXiv preprint arXiv:2105.06752*, 2021.

A. Subramonian. Fairness and bias mitigation: A guide to natural language processing with allennlp. https://guide.allennlp.org/fairness#1, n.d. Accessed on [insert date].

L. Sun, C. Xia, W. Yin, T. Liang, P. S. Yu, and L. He. Mixup-transformer: dynamic data augmentation for nlp tasks. *arXiv preprint arXiv:2010.02394*, 2020.

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

Q. Tian, J. Yu, Q. Xue, and N. Sebe. A new analysis of the value of unlabeled data in semi-supervised learning for image retrieval. In *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*, volume 2, pages 1019–1022 Vol.2, 2004. doi: 10.1109/ICME.2004.1394376.

P. N. Venkit, M. Srinath, and S. Wilson. A study of implicit bias in pretrained language models against people with disabilities. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1324–1332, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.113.

Z. Waseem and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June 2016a. Association for Computational Linguistics. doi: 10.18653/v1/N16-2013. URL https://aclanthology.org/N16-2013.

Z. Waseem and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016b.

J. Webster and C. Kit. Tokenization as the initial phase in nlp. pages 1106–1110, 01 1992. doi: 10.3115/992424.992434.

E. Wulczyn, N. Thain, and L. Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399, 2017.

Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020.

Z. Yang, Z. Hu, R. Salakhutdinov, and T. Berg-Kirkpatrick. Improved variational autoencoders for text modeling using dilated convolutions. In *International conference on machine learning*, pages 3881–3890. PMLR, 2017.

A. Zbiciak and T. Markiewicz. A new extraordinary means of appeal in the polish criminal procedure: the basic principles of a fair trial and a complaint against a cassatory judgment. *Access to Justice in Eastern Europe*, 6(2):1–18, Mar. 2023.

B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA, 2018a. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278779. URL https://doi.org/10.1145/3278721.3278779.

B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018b.

G. Zhang, B. Bai, J. Zhang, K. Bai, C. Zhu, and T. Zhao. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. *arXiv preprint arXiv:2004.14088*, 2020a.

H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

T. Zhang, T. Zhu, M. Han, J. Li, W. Zhou, and P. S. Yu. Fairness constraints in semi-supervised learning. *arXiv preprint arXiv:2009.06190*, 2020b.

T. Zhang, T. Zhu, J. Li, M. Han, W. Zhou, and S. Y. Philip. Fairness in semi-supervised learning: Unlabeled data help to reduce discrimination. *IEEE Transactions on Knowledge and Data Engineering*, 34(4):1763–1774, 2020c.

Z. Zhu, T. Luo, and Y. Liu. The rich get richer: Disparate impact of semi-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=DXPftn5kjQK.

# APPENDICES

# Appendix A

# More Results for Experiment 3

In Appendix A, we provide the detailed results of Experiment 3 for the higher labeled data ratio scenarios for both datasets. Although the outcomes do not significantly alter our conclusions, we believe it is crucial to report the results for transparency.

## A.1 Accuracy Scores

From Table A.1, A.2,A.3, and A.3, we can observe the same pattern as identified in the main report where more unlabeled data do not necessarily improve the accuracy scores of the semi-supervised frameworks.

| Model | BAcc | | | |
|---|---|---|---|---|
| | (n = 1000) | (n = 2500) | (n = 5000) | (n = 10000) |
| NDAGAN | 75.63 ± 0.89 | 74.59 ± 1.30 | 74.84 ± 1.10 | 75.53 ± 0.86 |
| FairNDAGAN | 74.90 ± 0.91 | 74.25 ± 1.17 | 74.64 ± 1.12 | 75.18 ± 1.29 |
| GANBERT | 75.47 ± 0.93 | 74.53 ± 2.36 | 74.41 ± 1.42 | 74.35 ± 1.23 |
| FairGANBERT | 74.87 ± 0.71 | 75.02 ± 1.31 | 74.18 ± 3.79 | 74.82 ± 0.79 |

Table A.1: This table presents the balanced accuracy results of models in Experiment 3 for the HateXplain dataset. Each column corresponds to a different number of unlabeled data in the training set, and the label data ratio is 0.1. Results are computed by averaging over five runs, and the standard deviation for each model is also provided.

| Model | Acc | | | |
|---|---|---|---|---|
| | (n = 1000) | (n = 2500) | (n = 5000) | (n = 10000) |
| NDAGAN | 76.57 ± 0.95 | 75.99 ± 0.62 | 76.26 ± 0.41 | 76.40 ± 0.48 |
| FairNDAGAN | 74.01 ± 2.04 | 72.98 ± 2.16 | 73.85 ± 2.28 | 74.03 ± 2.40 |
| GANBERT | 76.55 ± 0.89 | 76.02 ± 0.71 | 76.08 ± 0.46 | 76.03 ± 1.00 |
| FairGANBERT | 73.51 ± 1.51 | 73.85 ± 2.34 | 74.57 ± 1.43 | 73.59 ± 2.21 |

Table A.2: This table presents the accuracy results of models in Experiment 3 for the HateXplain dataset. Each column corresponds to a different number of unlabeled data in the training set, and the label data ratio is 0.1. Results are computed by averaging over five runs, and the standard deviation for each model is also provided.

| Model | BAcc | | | |
|---|---|---|---|---|
| | (n = 10000) | (n = 25000) | (n = 50000) | (n = 95000) |
| NDAGAN | 86.09 ± 3.24 | 85.06 ± 2.64 | 85.62 ± 3.46 | 85.67 ± 1.41 |
| FairNDAGAN | 88.58 ± 3.52 | 88.07 ± 2.82 | 87.45 ± 3.28 | 86.95 ± 2.68 |
| GANBERT | 85.13 ± 3.80 | 85.99 ± 1.06 | 86.32 ± 2.28 | 85.41 ± 2.73 |
| FairGANBERT | 87.60 ± 3.10 | 89.14 ± 1.40 | 88.25 ± 3.16 | 90.30 ± 1.05 |

Table A.3: This table presents the accuracy results of models in Experiment 3 for the Wiki Toxicity dataset. Each column corresponds to a different number of unlabeled data in the training set, and the label data ratio is 0.0161. Results are computed by averaging over five runs, and the standard deviation for each model is also provided.

| Model | Acc | | | |
|---|---|---|---|---|
| | (n = 10000) | (n = 25000) | (n = 50000) | (n = 95000) |
| NDAGAN | 95.46 ± 0.28 | 95.46 ± 0.17 | 95.39 ± 0.36 | 95.47 ± 0.28 |
| FairNDAGAN | 91.79 ± 2.32 | 94.06 ± 2.99 | 95.31 ± 0.61 | 94.12 ± 2.72 |
| GANBERT | 95.49 ± 0.23 | 95.45 ± 0.16 | 95.40 ± 0.12 | 95.48 ± 0.37 |
| FairGANBERT | 94.65 ± 1.63 | 94.54 ± 0.84 | 92.88 ± 3.81 | 92.35 ± 3.09 |

Table A.4: This table presents the accuracy results of models in Experiment 3 for the Wiki Toxicity dataset. Each column corresponds to a different number of unlabeled data in the training set, and the label data ratio is 0.0161. Results are computed by averaging over five runs, and the standard deviation for each model is also provided.

## A.2 Fairness Scores

Regarding Table A.5, A.6,A.7, and A.8, the fairness metrics also followed a similar identified trend, suggesting that the amount of unlabeled data has a consistent effect regardless of the labeled data ratio. Therefore, these results further reinforce the findings presented in the main body of the work. In addition, the consistent patterns across different labeled data ratios emphasize the robustness of our conclusions and provide added confidence in the generalizability of our findings.

| Model | $\text{EOD}_{\text{Gender}}$ | | | |
|---|---|---|---|---|
| | (n = 1000) | (n = 2500) | (n = 5000) | (n = 10000) |
| NDAGAN | 0.09 ± 0.03 | 0.13 ± 0.04 | 0.14 ± 0.03 | 0.16 ± 0.04 |
| FairNDAGAN | 0.03 ± 0.02 | 0.08 ± 0.01 | 0.06 ± 0.03 | 0.05 ± 0.02 |
| GANBERT | 0.10 ± 0.05 | 0.12 ± 0.04 | 0.14 ± 0.03 | 0.16 ± 0.03 |
| FairGANBERT | 0.05 ± 0.03 | 0.05 ± 0.04 | 0.05 ± 0.02 | 0.06 ± 0.03 |

Table A.5: This table presents the $EOD_{\text{Gender}}$ results of models in Experiment 3 for the HateXplain dataset. Each column corresponds to a different number of unlabeled data in the training set, and the label data ratio is 0.005. Results are computed by averaging over five runs, and the standard deviation for each model is also provided.

| Model | $\text{EOD}_{\text{Race}}$ | | | |
|---|---|---|---|---|
| | (n = 1000) | (n = 2500) | (n = 5000) | (n = 10000) |
| NDAGAN | 0.19 ± 0.04 | 0.17 ± 0.06 | 0.16 ± 0.06 | 0.14 ± 0.05 |
| FairNDAGAN | 0.08 ± 0.03 | 0.08 ± 0.03 | 0.08 ± 0.03 | 0.08 ± 0.02 |
| GANBERT | 0.19 ± 0.04 | 0.15 ± 0.04 | 0.15 ± 0.05 | 0.16 ± 0.05 |
| FairGANBERT | 0.08 ± 0.03 | 0.06 ± 0.02 | 0.11 ± 0.05 | 0.06 ± 0.03 |

Table A.6: This table presents the $EOD_{\text{Race}}$ results of models in Experiment 3 for the HateXplain dataset. Each column corresponds to a different number of unlabeled data in the training set, and the label data ratio is 0.005. Results are computed by averaging over five runs, and the standard deviation for each model is also provided.

| Model | $\text{EOD}_{\text{Gender}}$ | | | |
|---|---|---|---|---|
| | (n = 10000) | (n = 25000) | (n = 50000) | (n = 95000) |
| NDAGAN | 0.17 ± 0.01 | 0.23 ± 0.08 | 0.21 ± 0.07 | 0.24 ± 0.08 |
| FairNDAGAN | 0.09 ± 0.04 | 0.14 ± 0.05 | 0.08 ± 0.03 | 0.13 ± 0.05 |
| GANBERT | 0.22 ± 0.07 | 0.20 ± 0.06 | 0.22 ± 0.04 | 0.21 ± 0.05 |
| FairGANBERT | 0.13 ± 0.04 | 0.09 ± 0.02 | 0.11 ± 0.04 | 0.06 ± 0.02 |

Table A.7: This table presents the $EOD_{\text{Gender}}$ results of models in Experiment 3 for the Wiki Toxicity dataset. Each column corresponds to a different number of unlabeled data in the training set, and the label data ratio is 0.0161. Results are computed by averaging over five runs, and the standard deviation for each model is also provided.

| Model | $\text{EOD}_{\text{Race}}$ | | | |
|---|---|---|---|---|
| | (n = 10000) | (n = 25000) | (n = 50000) | (n = 95000) |
| NDAGAN | 0.12 ± 0.03 | 0.10 ± 0.05 | 0.12 ± 0.03 | 0.12 ± 0.02 |
| FairNDAGAN | 0.08 ± 0.04 | 0.11 ± 0.04 | 0.09 ± 0.04 | 0.08 ± 0.03 |
| GANBERT | 0.16 ± 0.04 | 0.11 ± 0.05 | 0.13 ± 0.04 | 0.12 ± 0.05 |
| FairGANBERT | 0.09 ± 0.03 | 0.09 ± 0.04 | 0.12 ± 0.04 | 0.10 ± 0.03 |

Table A.8: This table presents the $EOD_{\text{Race}}$ results of models in Experiment 3 for the Wiki Toxicity dataset. Each column corresponds to a different number of unlabeled data in the training set, and the label data ratio is 0.0161. Results are computed by averaging over five runs, and the standard deviation for each model is also provided.

In conclusion, these supplementary results confirm our main findings for Experiment 3, where increasing unlabeled data do not necessarily improve the accuracy and fairness of the semi-supervised models studied in this work.