3-2023

Design, Determination, and Evaluation of Gender-Based Bias Mitigation Techniques for Music Recommender Systems

Sunny Shrestha
University of Denver

Design, Determination, and Evaluation of Gender-Based Bias Mitigation Techniques for Music Recommender Systems

Abstract

The majority of smartphone users engage with a recommender system on a daily basis. Many rely on these recommendations to make their next purchase, download the next game, listen to the new music or find the next healthcare provider. Although there are plenty of evidence backed research that demonstrates presence of gender bias in Machine Learning (ML) models like recommender systems, the issue is viewed as a frivolous cause that doesn't merit much action. However, gender bias poses to effect more than half of the population as by default ML systems are designed to cater to a cisgender man. This thesis takes a closer look into gender bias discovered in different ML/AI applications and provides a holistic view of bias mitigation measures proposed in literature. Then by means of user study on 20 participants this paper analyzes gender bias in music recommender systems and the efficiency of bias mitigation methods. Instead of detailing the bias mitigation methods in technical terms, this paper takes the approach of utilizing user reviews to understand the effectiveness of bias mitigation methods for gender biases. Finally, this work aims to propose solutions that can help create equitable ML/AI systems that profits all stakeholders.

Document Type

Thesis

Degree Name

M.S.

Department

Computer Science and Engineering

First Advisor

Sanchari Das

Second Advisor

Maria Calbi

Third Advisor

Daniel Pittman

Keywords

Bias mitigation methods, Gender bias, Literature review, Machine learning, Music recommender system (MRS), User study

Subject Categories

Artificial Intelligence and Robotics | Computer Engineering | Computer Sciences | Music | Other Music

Design, Determination, and Evaluation of Gender-Based Bias Mitigation Techniques for Music Recommender Systems

---

A Thesis

Presented to

the Faculty of the Daniel Felix Ritchie School of Engineering and Computer Science

University of Denver

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

---

by

Sunny Shrestha

March 2023

Advisor: Dr. Sanchari Das

Author: Sunny Shrestha
Title: Design, Determination, and Evaluation of Gender-Based Bias Mitigation Techniques for Music Recommender Systems
Advisor: Dr. Sanchari Das
Degree Date: March 2023

Abstract

The majority of smartphone users engage with a recommender system on a daily basis. Many rely on these recommendations to make their next purchase, download the next game, listen to the new music or find the next healthcare provider. Although there are plenty of evidence backed research that demonstrates presence of gender bias in Machine Learning (ML) models like recommender systems, the issue is viewed as a frivolous cause that doesn't merit much action. However, gender bias poses to effect more than half of the population as by default ML systems are designed to cater to a cisgender man. This thesis takes a closer look into gender bias discovered in different ML/AI applications and provides a holistic view of bias mitigation measures proposed in literature. Then by means of user study on 20 participants this paper analyzes gender bias in music recommender systems and the efficiency of bias mitigation methods. Instead of detailing the bias mitigation methods in technical terms, this paper takes the approach of utilizing user reviews to understand the effectiveness of bias mitigation methods for gender biases. Finally, this work aims to propose solutions that can help create equitable ML/AI systems that profits all stakeholders.

Acknowledgements

First, I am forever grateful to my thesis advisor, and mentor, Dr. Sanchari Das for including me into her lab and providing the wonderful opportunity to work in important research projects. I want to extend my heartfelt gratitude to Dr. Das for lending me valuable insights, support and guidance throughout the process of developing this thesis. This thesis and all related research has been successful because of Dr. Das and her mentorship. When I began my thesis research, I had very little idea on my research direction and process. It is Dr. Das who has helped me shape my initial idea into a coherent and meaningful thesis. I would also like to thank my oral defence committee members Dr. Daniel Pittman, Dr. Kerstin Haring and Dr. Maria Calbi. I especially want to acknowledge Dr. Pittman for teaching me about machine learning topics and always encouraging my research pursuits. I also want to thank InSPIRIT lab, headed by Dr. Das, and all of my colleagues in this lab for supporting me in many ways than I can write. Specifically, I want to thank Faiza Tazi for always helping me review my code and study designs. Tazi has provided me immense support and valuable insights that has, for which I am forever thankful. Also my heartfelt gratitude to all my study participants, this work would not have been complete without your participation and contributions.

Finally, special thanks to my collaborators Tanjila Islam and Alisa Zezulak who have contributed to chapters of my thesis. Also, thanks to Dr. Mayukh Das who has reviewed my chapters and provided me with good suggestions.

Table of Contents

List of Tables

List of Figures

1. Introduction

The evolution of Machine Learning (ML) systems and Artificial Intelligence (AI) is widely regarded as the driving force behind of the fourth industrial revolution. AI and ML are transforming technologies and archaic (old) systems in almost all sectors of living like healthcare, education, energy, marketing, medicine and so on [171]. Unlike previous industrial revolution drivers, steam engines, electricity and internet that did not have an immediate impact of people's lifestyle (people in rural areas received facilities later than most people in privilege settings), ML/AI assisted automated decision making is quickly integrating with lives of people from all social and financial statuses. Due to rapid development and implementations, ML/AI technology is both intimate and beyond user knowledge. Thus, many users who use ML/AI assisted technology are not aware of its presence, mechanisms, or impact on their own decision-making process.

The lack of digital education in user population becomes a significant issue when the ML/AI assisted technology are flawed and not properly vetted. Especially, when these ML/AI assisted devices are shaping up users' everyday lives and their decisions. One of the major flaws in these systems is the presence of unintended and harmful algorithmic biases. In 2016, researchers discovered the presence of racial bias in COMPAS [59], an ML/AI assisted application which was used to predict risk score of an offender's likelihood of re-offending crime. Upon evaluation of the risk assessments by applications like COMPAS, researchers found that Black offenders where 77 times more likely to be assessed as higher risk of re-offending than their white peers. In reality, these

predictions were often false and misguided by racial bias against Black offenders, thus making these federally backed applications unfair towards a section of population. Unfortunately, this was just the beginning of the discovery of algorithmic biases in ML/AI systems. Since then, more ML/AI assisted systems and models have been found with algorithmic biases.

Still the discovery of these biased ML/AI assisted systems, fails to make a significant impact on users, the ML/AI system creators, industry leaders and government policies. However, the lack of alarm in users and government action is understandable, because of the nature of ML/AI systems. For a lay user it is hard to make the connection between a criminal recidivism application and the app they have installed on their phone. Flaws in one application shouldn't really affect the quality of another application, but it does. Due to the lack of proper laws and policies surrounding the design, development and deployment of ML/AI systems, and vast gaps of knowledge of these systems, users who are constantly using the ML/AI applications do not know the biases embedded in these apps and are unable to see any impact such biases have on their own decision making.

This thesis aims to de-mystify the ML/AI assisted decision making systems, the algorithmic biases associated with such systems and the impact of these biases from users' perspective. Additionally, the goal of this thesis is also demonstrating the algorithmic biases that readers can most relate to and to a create a roadmap of research which readers can follow along. Hence, in this work I detail the research on algorithmic biases with the aid of literature reviews, case studies, and a user study, and I focus my work on gender biases in music recommender systems.

1.1. Motivations

The demand for Machine Learning(ML) and Artificial Intelligence(AI) systems rose by 36.1% from year 2019 to 2020. The market for ML and AI integrated solutions is projected to amount to 209.91 billion USD by the year 2029 [129]. Recommender systems account for a substantial portion of ML/AI solutions that contributes to the increasing demand of ML/AI market. Recommender systems are commonly used to recommend items to users through various popular applications like Netflix, iTunes, YouTube, Amazon and many more. Whether it is trying to listen to music, podcasts or installing a new app or looking for news, most smartphone users are bound to come in contact with a recommender system every day. Hence, in this thesis I am taking a closer look into the implementation of a recommender system and investigating gender biases in such systems.

1.2. Problem Statement

Gender bias is deeply rooted into the fabric of our society. From the design and testing of advanced vehicles to the design of cooking stove (*chulah*) in a rural village [188], majority of innovations in our world is made to fit a cis-gendered man. This greatly disadvantages half of the population in the world, as proven by much research. Still gender bias is not seen as an important enough issue that merits attention. As our societies are being automated with rapid adoption of ML and AI systems, there is a possibility that the gender biased system will soon become the default. Thus, eventually negatively affecting users who do not belong to cis-gender man category.

Additionally, the gender discrimination is so well hidden into all of our thought process that it is difficult to definitively recognize gender bias and evaluate the impact of the bias. This is preventing the discussion and implementation of gender-bias mitigation methods from taking a center stage.

1.3. Thesis Statement

Gender bias in ML and AI systems have a critical impact on users and societal development. User study in a gender biased music recommender model shows the efficiency of bias mitigation methods and the need for more user-driven solutions.

1.4. Key Contributions

To this end, this work aims to aid the ongoing research into gender bias in ML/AI systems by providing following key contribution:

- **Provide a holistic view of ML/AI gender bias research**: Gender bias in ML/AI models has been very well documented and researched in academia. This paper provides a detailed view of this research gleaned from 120 scholarly published papers. These papers provide interesting insights on gender bias and its effect on users. Additionally, these papers also propose and present innovative solutions to mitigate and prevent gender bias in ML technologies. This thesis provides a holistic view of all of these solutions and also discusses the overall impact of gender biases in ML/AI community as well as broader society.

- **Exploration and Analysis ML/AI systems**: With the aid of literature review, case studies, and model implementation of recommender systems, this paper provides a detailed view of design and implementation of different ML models. The literature reviews give the broad and diverse application of the ML models whereas the case studies provide an in-depth look into the mechanisms of ML models. These chapters help to simplify the ML/AI technologies thus encouraging readers to further explore such technologies. The final chapter of this paper provides detailed codes to create music recommender systems which I hope will inspire readers to not be intimidated by these seemingly complex looking systems.

- **Highlight User Focus**: With constant buzz words and noise surrounding new-edge ML/AI technologies, it is easy to miss the primary goal of these systems: serve users. Ultimately, all of these technologies are created to automate, enhance, and simplify user processes. However, users are often brought too late into the conversation of design, development, and deployment of these systems. Although this seems like logistically sound decision as first, it ends up distancing users from the product and prevents active participant from user-side. User focus is important and one of the goals of this thesis is to highlight that informed user feedback will only serve to push the technology further in ways ML/AI creators might not have envisioned.

- **Gender Bias Resolution**: In this thesis, I have implemented three gender bias mitigation methods to create a control versus experimental model testing with the help of user reviews. This testing will allow readers to understand how effective such bias mitigation methods are. Through the user reviews of these models, I aim to showcase that gender bias is a societal and cultural issue which manifests in human-computer interaction. Thus, gender bias mitigation might require a more nuanced approach than just technical resolution. Additionally, this experiment imparts the rare insight that a gender biased system might not be working well even for the targeted audience, cis-gendered man. Further confirming that gender bias prevention and mitigation is crucial to create a profitable and equitable product that benefits all. The paper provides a detailed overview of such solutions because the technical bias mitigation along with these approaches will help to elevate gender-bias mitigation measures.

1.5. Thesis Organization

The thesis is divided into five chapters barring the introduction, research background, conclusion, limitation, and future work sections. First, I begin with providing related research background in section 2. This includes all the work that has been done by the research community in relation to algorithmic bias and gender bias in ML and AI applications. Then, in Chapter 3, I detail the literature review I have conducted, with help from co-authors, on algorithmic bias in ML and AI applications. In this chapter, I talk about all the different types of biases researchers have found in ML/AI systems. This chapter leads me to Chapter 4, where I conduct another literature review on specifically gender biases in ML/AI applications. In this chapter, I introduce all the bias mitigation methods that have been proposed in the literature. I also identify how gender bias differs from other biases found in ML/AI systems. For Chapter 5, I provide case studies of most commonly used ML implementations. In this chapter, with the help of ML models like Word Embeddings and Machine Translation, I provide a detailed view of the implementation of these ML applications that most users interact with daily. Thereafter, in Chapter 6, I discuss music Recommender Systems (mRS) and introduce the music recommender system that I have created. I also create three different models of the mRS each with different mitigation methods implemented. In Chapter 7, I discuss the user study I have conducted using the models I presented in earlier chapter. This chapter sheds light into user perception of gender bias in mRS and users also evaluate the mitigation methods implemented in different models. Finally, I outline the limitations of this work in Chapter 8 and the future direction of the thesis in Chapter 9. I have also included the Appendix section at the end.

2. Research Background

This section of the thesis discusses all the relevant research in the field of ML and AI technologies. Alisa Zezulak has helped me on writing and editing the related works mentioned in this section.

This thesis has been inspired from the prior research work in the field of AI, ML and Human-Computer Interaction (HCI). Especially the body of research that has provided an account of extensive use of ML/AI models in different fields of society and evidence of biases stemming from ML/AI model assisted decision making. I also noticed several indications of existing and harmful biases in these popular algorithms which has motivated my work.

2.1. Automated Decision Making

Since the industrial evolution, we have been making our way to devise engines to make our life easier, faster, and more efficient. Automated decision making is only a natural course path in this continuous evolution process. Of course, machines do not have a conscious thought as such to make decision, hence it needs to be trained to take decisions. This can be done by either defining rules that machines should abide by or by training the machine to recognized certain inputs that result on specific outcomes.

2.2. Use of ML/AI Systems

As mentioned earlier ML/AI has infiltrated everyday lives. This has motivated our work as given the exposure of humans to ML/AI the existence of biases can have harmful impacts. In this section, we detail some critical aspects of everyday life where ML/AI has already been implemented.

2.2.1. ML/AI in Policing

Machine Learning in policing technologies has become increasingly popular in recent years. As described by Fussey et al., Automated Facial Recognition (AFR) is an extremely controversial policing innovation because it centers around real-time biometric processing of captured video images. The images are analyzed by a facial recognition algorithm which matches facial features in a database, makes identifications, and then provides details such as ethnicity and warrant information [75]. By increasing efficiency in police work, providing insights from big data, and the ability to quickly identify suspects, ML is a significant factor in modern policing. However, the use of ML and AI in a field that has the ability to arrest, detain, and use deadly force against individuals raises questions about if there is enough accountability for how police departments use this software [100].

2.2.2. ML/AI in Marketing

While ML is not a new field, its application in certain disciplines is still nascent and evolving. From one field that studies and emulates human behavior to another – marketing has many similarities to ML and AI. Siau and Yang describe how AI, robotics, and ML are beginning to replace sales and marketing professionals, particularly in online stores [180]. The researchers describe how in a virtual environment, users don't have great interest in who (or what) is fulfilling their requests, as long as the requests get

fulfilled. Further research on the connection between ML and marketing reveals that some ML algorithms may perform better at analyzing user data than the current best-practice marketing strategies used [192]. In a study performed by Sundsoy et al., the researchers used metadata and social network analysis to create metrics for customers that were likely to convert into mobile internet users. This data was analyzed by an ML algorithm, where it was found that the ML algorithm had a 13 percent higher rate of customer conversion than the current best-practice marketing strategies [192]. This raises interesting questions about the future of these automated models in marketing if better performing models can gain trust from users than marketing professionals.

2.2.3. ML/AI in the Physical Sciences

In addition to the business sector, ML has had one of the greatest impacts in the science disciplines, especially the physical/material sciences. Because Machine Learning surrounds such a broad range of algorithms and modeling tools for data collection, Carleo et al. describe how its connection to scientific disciplines is unparalleled. ML methods have applications in particle physics and cosmology, quantum many-body physics, quantum computing, and chemical and material physics [30]. Many researchers have also found connections between ML and biology, citing early ML research that explored neuronal behavior, E.coli start sites, and artificial neural network architectures [194]. There is also a complex relationship between ML and the chemical sciences, where researchers in the field envision a future where the design, synthesis, characterization, and application of molecules and materials are accelerated by AI [28].

2.2.4. Advancement of ML/AI in Computer Science

Perhaps the most fundamental relationship between ML and the science disciplines comes from Computer Science. Alzubi et al. explain how Machine Learning is a

way in which AI, networked processes, and big data are brought together. This complex relationship between many different aspects of Computer Science has generated a vast amount of data that needs to be collected, stored, and analyzed, giving way to Machine Learning solutions and practices. As a result, each interaction with a system and each action performed by a user becomes a way for the system to learn and emulate these behaviors [4], a fundamental aspect of Machine Learning and Computer Science.

2.2.5. ML/AI biases and Human-Centered Computing Component of ML

Recent works to ensure the fairness, transparency, and equality of ML/AI models have gained a lot of attention from the researchers. ML/AI models are not only applied to Computer Science (CS) fields but also in other significant areas to minimize the cost as well as to speed up the process [76]. But these models are susceptible to the "garbage in, garbage out" syndrome like any system [169]. For example, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), an ML/AI assisted decision making software, which was used by the US courts for the prediction of Crime and Recidivism was biased against black offenders. In their paper, Larson et al. report on the biased nature of this software. During their analysis, they found out that this system was much more likely to mistakenly label the black defendants as a higher risk of re-offending, while wrongly flagging the white defendants as low risk [98]. By integrating personal, social, and cultural aspects while designing effective computer systems, Human-Centered Computing (HCC) can minimize the gap between end-users and computing technologies [97, 108].

2.2.6. AI and HCI Component

Though a notable number of studies have focused on the intersection of Artificial Intelligence and Human-Computer Interaction in recent years, the intertwining about

HCI and AI started almost a decade ago. Back in 2019, Grudin marked AI as a long-term vision that required expensive workstations, whereas HCI focused on a short-term goal by improving existing algorithms. He also described both fields as a rival of each other experiencing alternating periods of booming while the other suffered lack of researcher's interest as well as resources. As a result, they both competed for resources and funding within the same period of time [85]. Winograd examined the rationalistic and design approaches in terms of both HCI and AI to highlight the relationship and relevant differences to work effectively in solving real-world problems [207]. Nielsen et al. pointed out that AI systems are even capable of violating orthodox usability guidelines of the user interface (UI) design [142]. AI adopted systems may behave differently from one user to another, as they change learning over time (e.g., search engines/recommender systems returning different sets of results due to personalization and preference) . Therefore, inconsistent, and unpredictable behavior can cause confusion among the users that may lead to the abandonment of AI technology [5]. Even AI pioneer Yoshua Bengio has also predicted the over-hyped abilities of AI might be starting to cool off [177].

Such studies are critical as ML and AI is often viewed as a technical aspect of CS and other related fields without us realizing the impact of ML and AI assisted systems in the life of everyone. Inspired from prior works on AI and HCI, we wanted to explore on the user side of this research to further understand the emergence of biases in these algorithms. To this aid, we first conducted a systematic literature review to understand further on the biases research and then performed detailed analysis of two popular algorithms to explore further on the user side and the origin of this biases with human-centered data training. We explain the detailed study design of the two-part study in the next section.

2.3. Observed Biases in ML/AI Systems

2.3.1. ML Biases

Machine Learning methods and algorithms are used in nearly aspect of our society where mass data collection occurs, which means that unbiased and fair behavior is of the utmost importance when designing, implementing, and analyzing these tools. Mehrabi et al. describe how the widespread use of AI systems and applications makes it crucial to study and mitigate any biases or discriminatory behavior from these machines and algorithms [133]. The technical biases created by ML algorithms not only need a technical solution, but a social one as well. Biases have a social aspect that affect how people think and behave, creating lasting impacts surrounding discriminatory behavior, inequalities, and unfair treatment [154].

2.3.2. Healthcare Biases

The use of Machine Learning in healthcare raises many ethical concerns, especially because it can exacerbate existing biases and unfairness that already exist in the healthcare field [34, 2]. There is a robust collection of literature discussing the ML biases found in healthcare systems, including why they are so harmful and how they perpetuate existing inequalities. For example, Chen at al. describe how state-of-the-art clinical prediction models do not perform equally for women, ethnic and racial minorities, and people with public insurance [33, 34]. Researchers suggest putting patient safety and quality-of-care improvement at the forefront of solutions to mitigating ML biases in healthcare [132]. This means minimizing harm and encouraging accountability, justice, and transparency, as described by McCradden et al. [132]. Additionally, there are many ML algorithms that use electronic health record data in an attempt to avoid bias in diagnosis and treatment, however researchers have many concerns about

how these algorithms may cause an overreliance on automation, they may be based on biased data, and they may not provide clinically meaningful data [79].

2.3.3. Racial Biases

It is impossible to discuss racial biases in ML without first addressing the social factors that help create these biases. Researchers describe how there is a social and psychological complexity that comes with understanding how racial identity is embedded into our social experiences as humans [18]. Some proposed solutions include the need for more workplace diversity in high-tech industries and public policies that can detect or reduce biases in algorithmic design and execution [117]. Additionally, Benthall and Haynes explain how disadvantages associated with racial categories are created and perpetuated by segregation in housing, education, employment, and civic life, which then carry through ML algorithms and training datasets [18]. Designing fairness in Machine Learning must be adaptive, flexible, and privy to social change. As discussed above in Related Work 2.2.1, there are many racial biases that exist as a result of ML/AI in facial recognition software that is used by police departments to identify suspects. Williams describes how systems like facial recognition, predictive policing, and biometrics are developed with human prejudicial biases in the datasets, including assumptions and stereotypes that should be named, interrogated, and addressed before making any further innovations in the facial recognition field [206].

2.3.4. Gender Biases

Much like racial biases, in order to understand gender biases in ML/AI models, we must understand the historical and social context of gender and technology. Moreover, systematic gender biases not only affect cisgender users, but transgender and non-binary users as well [29]. Cao et al. describe how ML designers must acknowledge

the historical and social complexity of gender, otherwise they risk building systems that perpetuate stereotyping, over and under-representation of certain groups, and fail to provide quality of service to all genders [29]. An example of gender bias in Machine Learning is provided by DeBrusk, who explains how ML algorithms that scan resumes and college applications can screen out female applicants if the training datasets reflect few women being hired in the past or few women being admitted to a college [51]. Additionally, we found several examples of gender biases in Word Embedding algorithms, which is the basis of our research in this paper. Word Embedding algorithms are text-based algorithms and natural language processing tools [73]. Researchers such as Font and Costa-jussa explore debiasing techniques for Word Embedding algorithms in their paper, providing examples of how a test list of occupations can be generated in order to study and mitigate gender biases in their proposed system [73]. This related research lays the groundwork for our study of gender biases in ML algorithms and how we can begin to analyze different solutions to the problem.

Gender bias in mRS

Gender bias is a significant issue in mRS because these systems constantly try to identify, classify, and pass decisions on people based on the archaic concept of gender identity. An automated system is gender-biased if it performs differently or poorly for a specific group or population based on gender identity. In case of mRS, due to the lack of representation of the women and non-binary in the music dataset the mRS might not work as efficiently as it does for listeners who confirm with the majority population.

A study by Epps-Darling et al. shows that listeners stream much less female or non-binary artists than male artists [63]. This shows that it is much difficult for a person to discover women or non-binary artist than popular or men artists. Consequently, because there are less listeners, the non-binary and women artist are less likely to the

recommended than men or popular artists. Thus, creating a feedback loop which is unfavorable to women or non-binary artists, or listeners who would otherwise want to listen to such music. Another study by Lebefinger demonstrates that mRS are gender-biased against women and non-binary artists and this can be improved by incorporating a balanced dataset [116]. In this study, however we wanted to take a user-centric approach to explore the issue of gender bias from a listener's perspective.

3. Algorithmic Fairness: Literature Review

The following chapter focuses on the existing algorithmic biases in ML/AI models as discovered in academic research. The chapter has benefited from the work of many authors. Author Tanjila Islam has done extensive work on topic ideation and initial data collection, and Dr. Mayukh Das has helped with the initial editing of the paper. Dr. Sanchari Das has also helped with writing and editing many versions of this chapter.

3.1. Introduction

Digital advancements and increased computational power has evolved the field of Artificial Intelligence (AI) and Machine Learning (ML) [55]. As ML/AI models are now capable of handling complex data efficiently, these models are now changing the way we work and interact with the existing technologies [140, 123]. However, due to its black-box mechanism, albeit of varying degrees [3], even ML developers are sometimes unable to understand how different algorithmic variables are weighted and combined together to estimate the functional relationship between the input features and the target variables [162]. Therefore, many AI systems are not explainable in a way that deployers can claim that their systems are free from unintended biases even with careful review of algorithm and data set [161].

Algorithmic biases used in recommender systems are often considered as static code component, however, prior research notes that such ML/AI agent biases stems

from iterative learning driven from more deep-rooted causes. Along these lines, Sun et al. highlights that recommender systems use "online learning" where dataset used to train these systems are based on human actions [190]. When users can discern the existence of algorithmic biases in the ML model their reactions may exacerbate the cyclic interaction between the recommendation of the algorithm and people's preferences. Zou and Schiebinger mentioned that when Google Translate [1] converts news articles written in Spanish into English, it occasionally rephrases certain phrases into 'he said' or 'he wrote' which were originally written to refer to women [219]. In this way, these models often introduce prejudices that have a detrimental impact on individuals or a certain groups of people [77].

These biases can stem from: A. The design of the translation model, or B. The algorithm used to train the translation model, or C. Societal bias in the data that was used to train the model, or even D. Due to the secondary models that personalize tools for the users.

To understand such biases, we conducted a detailed systematic literature review while addressing the impact of machine learning and algorithmic discrimination. For our research, we adapted and extended the study methodology from Stowell et al. and Das et al. [187, 49]. Starting from keyword search we collected $38,206$ papers and filtered $N = 192$ papers published over the course of last five years. We explored and analyzed the current trend in ML/AI research. Thereafter, we conducted a detailed analysis of $n = 30$ papers published in various disciplinary venues focusing on the user component. The key contributions of this work are:

- Identifying the most sensitive areas of application for tailored ML/AI systems where the final decision made by these systems have a crucial impact on the end-users.

[1] https://translate.google.com/

- Pinning down the importance of considering the AI systems for legal and ethical review aiming towards building accountable systems.
- Analysis of the previous research works including their methodology and solutions as well as guidance for future research presented in a consolidated way to resolve ML and algorithmic discrimination.

3.2. Key Contribution

Our findings suggest that majority of the previous work details the importance of considering the social and cultural factors while providing technical solutions and dataset for the ML implementation [31, 25]. We also identified that women and black people are mostly affected by automated decision making due to the discriminatory behavior of machine learning adopted systems [18, 3].

In the following sections, the methodology for data collection and SOK is described in section 3.3. Then the major themes discovered in the publications is discussed in section 3.4, and the implications of these themes is discussed in the section 3.5. Thereafter, the chapter by providing a summary in section 3.6.Finally, the limitation of this work and the future extension of this chapter is outlined in section 3.7 and section 3.8.

3.3. Methods

We started our review process by going through 11 systematic literature review articles to better understand the procedure [53, 135, 68, 38]. Thereafter, we adapted the study methodology for the literature review from Stowell et al. and Das et al.'s work. Their research focused on mHealth intervention for vulnerable population, and phishing and authentication respectively [187, 49]. Our systematic review methodology consisted of four major phases: (1) Keyword Extraction, (2) Database Search, (3) Data Screening and Quality Control including: Title Screening, Abstract Screening, and Full

Paper Screening, and (4) Thematic Analysis as demonstrated by the figure 3.1.

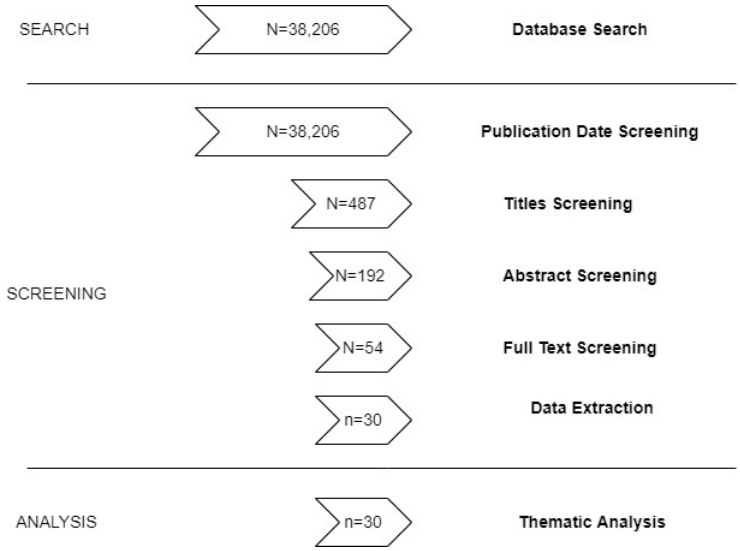

Figure 3.1.: Literature Review Method Process

3.3.1. Keyword Extraction

In the beginning, we started with 90 keywords collected from 89 papers related to algorithmic biases. After three interactions, we finalized 23 keywords which resulted in papers related to AI/ML biases. The final list of keywords included: *Search Bias, Algorithmic Bias in Machine Learning, Machine learning Fairness, Artificial Intelligence Bias, Fair Machine Learning, Gender Bias in Machine Learning, Algorithmic Discrimination, Targeted Social Programs, Disparate Impact and Algorithmic Bias, Bayesian Improved Surname Geocoding, Fairness-aware Machine Learning, Fairness in Algorithmic Decision Making, Racial Discrimination in Machine Learning, Homo Egualis in Machine Learning, Personalized News Recommendation, Algorithmic Impact Assessment, Algorithmic Accountability, Algorithmic Decision-making, Algorithm Experience, Protected Group in Machine Learning, Algorithm Fairness, Algorithmic Decision, Gender Bias in Artificial Intelligence.*

3.3.2. Database Search

We conducted the keyword-based search in several major digital databases including, Google Scholar, ACM digital library, IEEE Xplore, Science Direct, DBLP, Microsoft Academic Scholar, SSRN, and Springer. This step resulted in 38,206 papers in total. We used the Publish or Perish software[2] to collect the papers. During the time of our initial search (August 2020), searching in Google Scholar required no prior registration in the Publish or Perish software.

We implemented two search filters including keyword-based and publication years to cover the latest research, we collected the publications of the last five years from 2015 to 2020. The final 23 keywords resulted in a total of 487 papers. Papers were excluded if: (1) the primary language of the paper was not English, (2) non peer-reviewed publication, (3) a work in progress paper, or (4) the full text was not available. We contacted the authors of the papers to get access to the papers where the full text was missing, which led us to obtain 21 missing papers. After the filtering mechanism, we identified 257 papers in total. In the next phase, duplicate papers or very similar papers (at least 70% similarities with each other) were removed from the list of papers. After the removal of 65 duplicate papers, our list contained a corpus of 192 papers.

3.3.3. Data Screening and Quality Control

Title and Abstract Screening: We carefully reviewed each paper by reading the title and abstract to analyze the relevance of each paper considering our research goal. We coded the list of 192 papers as 'irrelevant ' or 'relevant ' or 'undecided' based on the content of the paper. The papers coded as 'undecided' were carefully reviewed by the authors of the paper and after three rounds of detailed discussions, the discrepancies were resolved. Finally, a total of 54 papers were selected at the end of this stage.

[2]https://harzing.com/resources/publish-or-perish

Full Paper Screening: The full text of the remaining 54 studies was examined to delve further. We focused on the study methodology, data collection and evaluation process applied in the papers, solutions proposed by the authors, and so on. 24 articles were discarded because they did not meet the criteria mentioned above and a total of 30 articles were selected after the review.

3.3.4. Thematic Analysis

During the full paper screening, we structured this process into five steps. First, we identified the most sensitive application areas discussed in the previous literature where ML/AI biases can have dire consequences but were most neglected. Secondly, we documented if previous researches focused on any special user group to understand the affected community as a result of ML/AI biases. Thereafter, we identified any existing user studies which were used to understand the people's perceptions. Fourthly, we noted data sets and methodology used by the researchers for their ML and AI-focused studies. Finally, we reviewed the proposed solutions and implications recommended by the scholars.

3.4. Findings

We performed in-depth analysis of prior research which discussed the various areas impacted due to ML/AI biases.

3.4.1. Biases in Diverse Application Areas

Machine learning models are being immensely adopted by large organizations for job advertisements and recruitment process [3]. Our study identified the attempts of previous researchers who have analyzed the impact of machine learning bias in hiring purposes. Three out of 30 of the papers discussed the impact of the machine learning

model for recruitment. Few studies mentioned the existence of gender and racial bias in hiring [83, 151]. For example, Goodman et al. explained that if a racist manager gives a lower rating to non-white employees compared to white employees and these data are fed into ML models, the system will produce discriminatory results towards black people [83]. Therefore, the presence of discriminatory behavior in society can influence the decision produced by automated systems and eventually impact the overall hiring process as well. Social welfare and healthcare services are sensitive areas that can immensely impact a person's well-being hence such areas require a closer examination when it comes to understanding ML/AI biases on services provided to different population groups.

Existence of Biases in Welfare Service:

Only two out of 30 discussed the impact of ML/AI biases in social welfare services. The usage of the algorithmic tool in the child welfare system has faced a lot of criticism that resulted in the termination of deploying algorithmic systems [25]. To understand the experience of end-users, Brown et al. conducted a human-centered approach by adopting a participatory design methodology in the context of the child welfare system [25]. They recruited and interviewed families, social workers, and specialists, persons of color to explore how they were affected by the child welfare systems. Participants exhibited system level concerns regarding the final decision-making process. Similarly, Noriega et al. implemented a series of prediction algorithms to estimate the poverty margin. Their findings indicated that algorithmic decision-making systems could produce discriminatory behavior based on geographical locations. Their studies showed that poverty prediction algorithm may exclude poor people living in urban area than their rural counterparts [144].

Impact of Biases in Healthcare System:

Only three papers discussed the impact of ML biases in healthcare services [31, 218, 183]. Carroll et al. discusses an important scenario where ML models are mostly trained on the "men-only" data [31]. As body functions including blood glucose, body weight, body mass index, and physical activity differ from men to women, it raises a concern in the healthcare arena. If healthcare professionals do not understand how and why decisions were made, algorithmic decisions making may have a threatening impact in this sector.

3.4.2. High Risk of Gender and Racial Biases

Twenty-one out of 30 papers acknowledged the existence of machine learning biases that affected certain groups of people. Previous studies mentioned that machine learning model can behave differently based on individuals protected characteristics, for example, race, sex, political beliefs, geographic location, ethnic origin, genetic/health status which can have a global impact [83, 25, 144, 82, 18, 211]. For example, Allhutter et al. mentioned the controversial claim against Austrian Public Employment Service (AMS) which discriminates a certain group of people based on gender, ethnicity, people with disabilities, and care obligations [3]. Zhiltsova et al. mentioned the existence of biases against French cognates in NLP (Natural Language Processing) systems [217]. However, other studies discussed several notions of biases including societal bias, iterated algorithmic biases, sampling biases, historical biases including subjective bias of individuals, and institutionalized biases [218, 9, 105, 191].

3.4.3. Automated Decision Making- Human Factors

Only three papers conducted user studies or collected responses from the users to know about their perception of algorithmic decision making. Cotter et al. showed

that algorithmic knowledge varies depending on socioeconomic advantages [40]. They mentioned that technologies like big data, algorithms, machine learning adopted systems are understood through usage and experience. Therefore, the socioeconomic background shapes the information and knowledge we learn about these technologies. On the other hand, Brown et al. conducted several workshops to understand the concerns of affected communities in the child welfare system. Their study revealed that participants fear that these systems may treat them unfairly as they are treated in society. People of color shared that majority of the systems are not made for them because these systems are usually not developed by people of color [25]. A team of researchers analyzed the impact of bias in rating platforms. They showed that users can identify algorithmic bias during their regular usage of a system, and they do want to inform others about it if the platform allows [64].

3.4.4. Social and Cultural Aspects

Thirteen papers discussed the importance of human, social, and cultural factors while addressing ML/AI biases. Goodman et al. draw upon literature from economics which brings together both social science and computer sciences [83]. They mentioned that by linking both fields, it is possible to address algorithmic discrimination. Other studies pointed out discrimination not only depends on technical causes but also on social effects [82, 105].

Fairness of Machine Learning Model:

Sixteen papers discussed various notions of fairness in the context of the machine learning model. Few studies [217, 9, 105, 210, 89, 13] focused on individual and group fairness, while others [90, 124, 201, 213, 211, 84, 183, 52] addressed disparate impact, disparate treatment, statistical parity which are different notions of fairness in

automated decision making. Noriega et al. showed the presence of unfairness in the system where poor elderly households in Mexico are more likely to be excluded from getting the financial stipend than their traditional nuclear family counterparts [144]. While addressing the fairness of automated decision-making system, Seymour et al. identified that black-box techniques are not enough to guarantee that a system is fair unless the entire problem domain is exhaustively searched [174]. Dwyer et al. showed that how the lack of algorithmic fairness can have a bigger impact on our society [62]. They analyzed the dynamic media market in Korea including algorithmic recommender systems, news portals, and pointed out that their media industry is dominated by a relatively small set of players whose primary focus was to manage the fairness of news distribution on the portals. However, algorithmic fairness is subjective and difficult to achieve as the decision making process is largely dependent on the data used for training and added into the system.

Privacy, Security, Legal, and Ethical Concern:

Though our research primarily did not focus on the privacy and security of the ML/AI adopted systems, we identified that users are concerned about how their information are being used by these systems. Nine papers discussed the privacy, security, and ethical concern of AI/ML biases. While collecting the personal data about individuals, there is a possibility of disclosing the private information of an individual to the decision-maker [89]. Additionally, the ML/AI models may violate individual privacy by collecting unnecessary data [13]. In recent years, security experts have also recognized a growing pattern of security attacks on ML systems and also identified new vulnerabilities associated with these systems [148]. Asudeh et al. pointed out that ML models are nowadays a favorite target for the attacker due to the skewed dataset [9]. They discussed that lack of coverage in the dataset opens up the possibility for the adversarial

attack. Few researchers [31, 124, 62, 61] mentioned that ML/AI models are unable to satisfy the need of underlying ethical needs of the individuals, while others [64, 211, 174, 62] discussed the need for increased transparency in the machine learning system. Besides, technical approach let alone may fail to satisfy the underlying legal and ethical needs of ML/AI models [124].

3.5. Discussion and Implication

We found several far-reaching impacts of ML/AI biases discussed throughout the years via prior research.

3.5.1. Third-Party Algorithmic Auditing

ML/AI models are now widely used in various sectors to make decisions. We identified that critical areas could have crucial impact on users due to existence of the biases. Our findings suggest that further research is needed to build more tailored machine learning models in critical sectors, such as healthcare, public welfare services, etc. [25, 31]. To ensure transparent and reliable systems, organizations should focus on algorithmic inspection by third-party authority. Additionally, previous researchers suggested using the automated system as technical support but the final decision, especially for recruitment should be made by the employer themselves [3].

3.5.2. Bias-Aware-Design

Our study shows the existence of critical biases in AI/ML data sets discussed in prior literature. Among the existence of various forms of biases mentioned in our findings, we identified that racial and gender biases are prominent in the literature. To address these biases, five out of 30 papers adopted data pre/post-processing or both techniques [9, 105, 89, 195, 90] to generate unbiased data sets. Eslami et al. [64]

emphasizes on building "bias-aware-design" so that users can comprehend algorithmic biases.

Federated Learning

Another method proposed to mitigate biases from automated systems is the use of Federated Learning (FL). Generally, a prior set of data is used to train an ML/AI algorithm. This results in the integration of biases into the algorithm due to a lack of representation of certain populations based on the locality of the data. In FL, a decentralized data set from across different services and devices is used to train the ML model. FL is proposed to address the issues of privacy, ownership, and bias rising from silos and centralized datasets [214, 21].

FL utilizes data available from multiple clients like smartphones, servers, data centers, and so on. The learning happens in the following way: each device supported by the clients downloads a generic model for local training, then the downloaded model will learn and improve with the local data. This improved model or related gradient information is uploaded to the cloud in an encrypted mode. Thereafter, these updated models from different clients are used to create a new general model. This process is repeated until an optimum performance is reached [121]. Noticeably, no personal data from individual devices are shared across platforms, the initial training occurs within the devices, and only the trained model is shared. This is why FL is considered a privacy preserving learning mechanism. Although FL promises a new way to balance bias-free AI development while preserving privacy, it is still in a nascent stage of development and more research is required in this field to understand FL better [122, 110].

3.5.3. Policy Reforms Addressing Privacy Concerns

ML models require the use of large training datasets to provide accurate predictions. The mass collection of large sets of personal, location-based, and behavioral information on private users presents very serious risks of data misuse, breach, or even loss [141]. In such scenarios, it is imperative to build safeguards against these risky situations. Currently, ML model and dataset handling is fully controlled and managed by individual entities with little to no oversight. Data protection laws are so minimal that even big companies such as LinkedIn have no legal recourse to prevent data scraping from other companies[3].

Furthermore, there is a growing body of research that clearly demonstrates bias in automated systems, especially against minority, marginalized, or under-represented populations [143]. Regardless of this research, there is very wide and rapid implementation of ML and AI systems across the globe, even at the federal level. According to a recent report published by Shaheen and Kasi, out of 12 institutes that were surveyed, 45% of them had implemented some form of AI and ML tools. ML systems are already in use to enhance predictions in fields like healthcare, policing, community welfare and so on [175].

3.5.4. Ethical and Legal Reviews

Our research shows that majority of ML models collect unnecessary private data from individuals. Though researchers have proposed a framework that collects minimal sets of information from the users, their framework is unable to ensure the uniform sets of privacy for the individuals [13]. Hence, we suggest that developers need prior research before collecting private information from the users. Furthermore, when additional sets of information are collected from a specific group of users, they should be notified about

[3]Bloomberg article

it with proper justification. Similarly, previous researchers have also identified a lack of ethical considerations for algorithmic decision-making systems. Especially, more focus is needed to ensure that AI agents can behave morally in the field of connected health. Even so, AI can itself be held as accountable or not is also a matter of discussion [31]. In a recent article, evidence shows that if algorithms indirectly discriminate end-users, these systems are subject to judicial review [77]. Therefore, we suggest introduction/revision to the policies for addressing algorithmic biases.

3.6. Chapter Summary

Machine Learning and Artificial Intelligence algorithmic biases produced by automated systems have a large impact on end-users [190]. Thus, to understand the current research in this area, a detailed systematic literature review ($n = 30$) following the methodology of other studies was conducted. Here, the research shows that algorithmic biases have detrimental impact on users in different sectors like for Healthcare System, Public Welfare Service, Recruitment sectors, etc. The findings also document the prevalence of racial and gender biases in the automated systems affecting particularly women and black people. Thus, the need to integrate cultural and human factors to combat unconscious biases is highlighted in this chapter. The chapter concludes by focusing on considering human-in-the-loop while developing automated decision-making systems and suggest researchers incorporate an interdisciplinary research approach for better suggestions from scholars of different expertise. In the following chapter we will see the literature on gender bias specifically and discuss the bias reduction and mitigation methods proposed in the research.

3.7. Limitation

This chapter is vital in providing a consolidated overview of prior studies published on the topic of ML biases, gender biases, and bias mitigation methods. The research conducted in this chapter is limited to publication between year 2015 and for the year 2021. The analysis of the algorithmic biases can also be expanded on and more visual aids can be added to enhance readability.

3.8. Future Work

In the future direction, case studies and user studies by first creating a prototype addressing these bias concerns and next testing the prototype with the set of users can be conducted to explore the themes explored in this chapter. An expansion of this chapter could also be an in-depth study of one of the biases with better visuals.

4. Gender Biases: Literature Review

As seen in previous chapter, the research shows that ML/AI models have many existing algorithmic biases. Due to wide and varied implementation of these models, the impact of the algorithmic biases are often differing. The following chapter focuses on algorithmic biases based on a user's gender identification. The goal of the chapter is to pull readers' focus into the manifestations of gender bias in ML/AI models, the difference of gender bias from other algorithmic biases and shed light on the impact of these biases. Dr. Sanchari Das has assisted in the review and edit of the chapter. In the previous chapter, I explored the topic of algorithmic biases in automated systems. Once looking into the different types of biases and its potential impact on the minority population, in this chapter I have narrowed my focus to gender-specific biases.

4.1. Introduction

Gender bias are the algorithmic biases that disproportionately disadvantage a population based on their gender identity. Prior to 1990's, gender was defined as a binary concept based on a person's biological sex. However, in current times the definition of gender and gender identity is evolving away from the restrictive biological sexual identities (men and women) [7]. In the context of this paper, gender is used as an inclusive term comprising both binary and non-binary sexual identities. Hence, the term "gender bias" here refers to any kind of discrimination in either accuracy or performance in the automated system depending on the gender identification of a population.

Gender bias in Automated Systems: Gender bias has been a topic of interest for many researchers in the past decade. Due to widespread application of Machine Learning (ML) and Artificial Intelligence (AI) assisted systems and lack of an oversight in the deployment of these models, the fairness aspect of these models have come under focus. Highly computation-heavy and often challenging to interpret, the ML and AI components of computer-assisted decision-making make it difficult to understand and implement. Furthermore, due to the opaque (*black-box models*) nature of these systems, it is not easy to follow the validity and accuracy of the decisions made by these systems from an ethical and legal perspective [170]. Human history is full of examples of unfair treatment of minority groups, and it is well documented into the data I have generated over time. Gender bias is one such issue which is rooted in human social culture and development. ML and AI models rely on these flawed data to learn human decision making hence, it is inevitable that these systems inherit the historical biases present. As stated by Bender et. al. these models exacerbate existing biases and further perpetuate stereotypes, causing significant impact on marginalized population [17].

Additionally, most of these models are designed and created by the most privileged people of the society so their perspective on the fairness of the outcomes of these models could be skewed and uninformed. Broadly in the context of ML and AI implementation, a model is gender biased if the model's performance and/or output is biased against a faction of population based on their gender. For example, in the research conducted by Buolamwini & Gebru I see that the facial recognition systems were highly inaccurate (more than 60%) when it comes to classifying the faces of women of color. In this ground-breaking paper, the authors further demonstrated that the model was most accurate for people who identified as male and of white skin tone [27]. Gender bias has been studied extensively in Natural Language Processing (NLP) systems because this is the most visible form of gender bias [189]. Especially, in widely used language transla-

tion systems pronouns are assigned to profession confirming to the gender stereotypes, for example doctors and pilots are automatically translated to *he/him* pronouns and nurse and flight attendants are assigned to *she/her* pronouns [36].

However, gender bias is most harmful when the bias is not as visible. Especially, in the systems of social programs, national defense, justice systems and policing, which implements ML/AI algorithm, when the decision made by the automated systems might be gender biased but there is no definite way to confirm. Additionally, the ML/AI systems usually use the binary concept of gender which does not reflect the real world. There is a bigger concern for LGBTQIA+ community, when people are entirely misgendered and wrongly classified because the models are not equipped to manage this category.

These revelations has led to even more research into gender bias detection and mitigation methods that could help the ML and AI models to prevent gender bias in automated decision making. The fairness debate although nascent is a widely accepted concept and there is continuous effort to mitigate these biases from both academia and the industry. The industry leaders like IBM has a dedicated code repository, AIF360 [1] that encourages ML and AI developers to learn, utilize and normalize the use of bias detection and mitigation methods within their models.

Still more effort and work is needed to prevent gender biases and make these applications more inclusive and fairer across all users. Although every scientific publication and research lends immense insight into ML and AI systems and pervasiveness of gender biases in the implementations, there is a need for a holistic account of all the work done in this area. A comprehensive review of all the research will give us the benefit of looking into the areas that are well researched and the areas that might need more research. It also helps us understand the trend of gender biases within ML/AI systems.

[1] https://aif360.mybluemix.net/

4.1.1. Key Contributions

To this end, in this paper I provide a comprehensive view of all the research conducted in the field of gender biases propagated by ML and AI systems in the many implementations. The Systematization of Knowledge (SOK) is an effective format to provide a unified view of several aspects of gender biases in different forms and stages of ML and AI application. Furthermore, a detailed analysis of academic publications will help researchers and ML/AI engineers to understand the gaps in research that needs more attention. In short, this chapter aims to aid the ongoing research into gender bias in ML/AI systems by providing following key contribution:

- **Provide an overview of different themes and topics explored in the ML/AI gender bias research papers**: There are many interesting and innovative concepts presented by different authors published in the gender bias research field. These ideas bring novel solutions and perspective to the issue of gender fairness within ML and AI communities. Hence, it is valuable to create a holistic account of these ideas as they can inspire further conversation and actions to prevent gender bias.

- **Discuss different bias detection and mitigation methods proposed in these papers for different ML/AI algorithms**: There are many gender-bias detection and mitigation methods proposed in the literature but there is little wide-spread application of these methods. Since many of these methods are provided within the context of specific type of algorithm, these solutions could remain hidden from researchers working on a separate set of algorithms. So, it is important to gather these concepts and recount these ideas as they can result in creating more solutions to a similar problem in a different ML/AI application.

- **Shed light into the less studied aspects of ML/AI gender bias research and provide an argument for the need of more attention to these less explored topics**: The ethical and legal aspect of these gender bias issues are seldom discussed. Similarly, I need more studies that collaborate with user community and field experts to get a good grasp on their perspective of gender bias in ML/AI systems.

In the following sections, the methodology for data collection and SOK is described in section 4.2. Then the major themes discovered in the publications is discussed in section 4.3, and the implications of these themes is discussed in the section 4.4. Thereafter, the chapter is summarized in section 4.5 . Finally, the limitation of this work and the future extension of this chapter is outlined in section 4.6 and section 4.7.

4.2. Methodology

I began this study by first looking at similar prior SOKs published in the field to better understand the methodologies of conducting a thorough literature review. I reviewed papers by Stowell et al. and Das et al. to understand the methodology of conducting a systematic literature review. Their research focused on mHealth intervention for vulnerable population, and phishing and authentication respectively [187, 49]. Drawing inspiration from these papers I have implemented following methods in this study: (1) keyword-based database search, (2) data screening and quality control: content screening based on paper's title, abstract and full text, and (3) data analysis.

This literature review is guided primarily by the following research questions:

- What is the current research landscape on gender specific biases present in ML systems and models?
- What technical solutions are proposed to detect, mitigate, or eliminate gender specific biases in prior research?

- What area of research needs more attention from research community or requires further investigation?

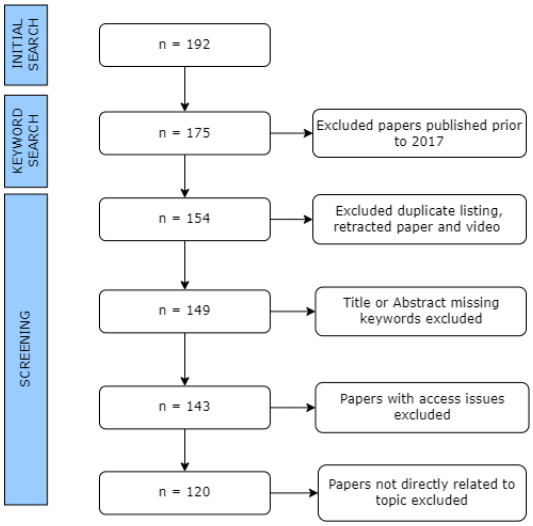

Figure 4.1.: Paper Collection Methodology Diagram

4.2.1. Database Search

First, I conducted a brief overview of different research published under gender bias in ML/AI systems. This helped me gather keywords that would be fruitful in searching publications under the topic of gender bias in automated systems. Furthermore, the research goals also motivated the keywords I finally used to gather publication for the following SOK. I wanted to make by search precise and move away from using ambiguous terms like "automated decision making systems". Hence, I finalized following keywords to search the databases for publication on gender bias topics in ML/AI systems: "*gender*" , "*algorithmic bias*" , "*gender bias*", "*gender bias in automated system*" and "*gender bias in machine learning*". I also utilized arious combinations of these words using "+ " or logical connectors "AND" and "OR" to gather as many publications as I could.

In this study, I have used the Publish or Perish software to collect the papers. This is because this software allowed me to conduct search into multiple digital libraries at once and also allowed me to filter results based on publication year, titles, and other criteria. With the help of this software I was able to conduct keyword-based search across digital libraries including, Google Scholar, ACM Digital Library and IEEEXplore. At the time of the initial search (December 2021), searching in Google Scholar required no prior registration. I limited the search to the publications from year 2010 up to 2021. I added this year restrictions because I reviewed publications prior to 2010 and found that any research prior to 2010, will not reflect the current developments in the field. From the search I gathered 192 papers for the review.

4.2.2. Data Screening and Quality Control

Once I had all the papers, I manually went through the papers to further refine the corpus. I created following exclusion criteria to refine the papers I had collected:

- I excluded the paper if the paper was not written in English given the primary evaluation done in the same language.
- I removed the paper if the full text of the paper was inaccessible, behind a paywall or had loading errors. I contacted the authors in that case, and I kept the paper in the list if I obtained those papers.
- I excluded the paper if the paper was incomplete or retracted, or not published on peer-reviewed journals and/or conferences.
- I put a time constraint in the inclusion of the papers and analyzed those papers published on or after 2017. I intentionally put this criterion to evaluate the recently published work in the last six years.

When I looked into the papers during this excluding exercise, I found that most of the papers published prior to 2017 were not relevant to the topic at hand: gender bias in automated decision-making systems. Also, few papers that were published prior to 2017 and had relevant information had additional or updated work published after 2017, hence I arrived at the final exclusion criteria listed above.

I implemented the exclusion criteria in three phases of screening. First, I reviewed just the title, keywords, and abstracts of the paper. Then, I reviewed the full text of the papers and created a codebook based on the paper's focus. Finally, I analyzed the methods and implications of the papers in a detailed manner to arrive at the final review corpus. I removed a total of 19 papers from the papers based on the criteria mentioned, thus resulting in 173 papers. Next, I excluded papers that showed up twice which further reduced the number of papers in the corpus to 154.

Title and Abstract Screening: I further screened the remainder of 154 papers based on their titles and abstracts. In this step, I carefully reviewed the collection of papers to makes sure they had relevant keywords which includes words like "gender", "bias/biases", "machine learning" and/or "artificial intelligence", within the titles and abstracts. Based on the presence of these keywords, I classified papers into two categories: 'relevant', 'some relevancy' and 'irrelevant'. During this processing I excluded more papers that fell into irrelevant category, resulting in total corpus count of 149.

Full Paper Screening: In this phase, I conducted a quick review of the full text of the 149 papers mainly focusing on the study methodology, implications, algorithms explored, and solutions proposed. In this step I removed 6 papers from the corpus because the full-text version of these papers were inaccessible or behind a paywall. I also further excluded 23 papers because they did not directly discuss about gender biases in the automated decision-making systems. Thus, I ended up with a total of 120 papers as summarized by the diagram 4.1.

4.2.3. Analysis

After the screening, I conducted a detailed review of 120 publications focusing on the algorithms explored in the paper, methodology followed, and solutions proposed to detect and/or mitigate gender biases in the automated decision-making systems. I also further analyzed the papers based on the publication year to trace the research trend on this topic over the years. As demonstrated in the figure 4.2, there has been drastic increase in research into gender bias in automated decision-making systems over the years. I also created a codebook to categorize papers into different groups based on the focus of the paper. Details of this codebook can be viewed on this table 4.1.

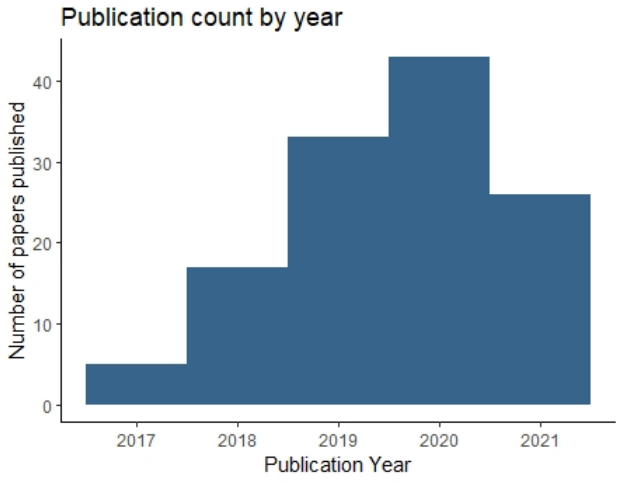

Figure 4.2.: Paper Publication Timeline Over the Years

4.3. Results

4.3.1. Gender Bias in Literature

Forty-eight out of 120 papers collected for this study discussed about the presence of gender bias in a variety of ML and AI applications. The diverse fields studied in these papers show the varied applications automated decision-making systems have

and, it helps us realize the severity of the gender fairness issue. It also demonstrates that unintentional bias can have drastic effects on minority populations.

Fourteen papers discussed **Natural Language Processing (NLP)** systems and the presence of gender biases in these systems. Researchers studied NLP algorithms like Word Embeddings, Coreference Resolution and Global Vector of Word Representation (GloVe). In these papers authors discuss the presence of inherent bias in the human languages which is then codified into the ML and AI NLP operations through the data. Here, NLP operations refers to functions like embeddings, coreference resolution, dialogue generation, machine translation, text parsing, sentiment analysis, hate speech detection, and so on [20, 189]. Authors Chen et al., look across nine human languages, including English, Spanish, Arabic, German, French and Urdu, and find gender bias in gendered nouns for profession words [35] in Word Embeddings. In another paper authors Guo & Caliskan, explore the intersectional bias present in English static Word Embeddings [87]. They find that women of African American and Mexican descent were most biased against because of their racial and gender identity. Similarly, authors also look into machine learning models and study the effects of gender biases in these applications [186, 155]. Gender bias has also been detected and studied in application that predict a person's profession [8, 165] and gender [113]. Other NLP applications discussed in these papers that are affected by the gender bias are sentiment analysis [72], emotion identification [130], and customer review analysis [136]. Two papers discuss the representation of women in audio-visual medium, these papers discuss how a biased system can affect the gender-equality movement by presenting women in gender normative fashion. Author Miren Gutierrez points out that the Google image search results for powerful profession like CEO, news reporters or movie directors lack women representation [88]. Similar paper by Singh et al. also highlights the issue of lack of women representation in occupations images in various digital platforms [182].

The **Automated Facial Analysis, image classification and recognition algorithms** is the second most studied ML application in these papers. The research by Buolamwini et al. is a pioneering paper that shed light into algorithm fairness in automated systems was also conducted on Facial Recognition Technologies (FRT) [27]. Authors Srinivas et al. also conduct gender bias analysis on off-the shelf and government prescribed FRTs and found these systems to have biases even though the creators claim otherwise [185]. Biases affects FRTs by affecting its performance accuracy creating mis-labelled or mis-classified faces for minority population [130, 14, 112, 158].

Automated decision-support systems are prolific in the field of **advertisement, marketing, and recruitment** systems. Authors Howcroft and Rubery discuss the effects of gender bias in the labor markets in disrupting social order and point out the need to tackle these biases from outside-in (*fixing the issue in the society before fixing the algorithm.* They discuss how implicit biases of the users, rooted in our social norms and habits, feed into these biased systems to create a regressive loop [96]. Another paper by Shekawat et al. discuss the presence of gender bias in ad-personalization applications that expose users to biased advertisements continuously through their devices [179]. In similar vein, Raghavan et al. present the legal implication of of recruitment systems that are gender biased [157].

Five papers talk about the presence of gender bias in **recommender systems** and, **search and ranking** algorithms. In these papers too, I see the authors point out biases in the systems and how it affects our lives on the ground-level. Lambrecht and Tucker study the tailored job listings, which based on applicants' gender, present different job opportunities to different applicants. The jobs shown to women were discriminatory in nature as they were shown far fewer STEM ads than men [114]. Another paper by Tang et al. replicate this study but provide an interesting insight into the implicit biases held by the applicants. In this paper, they demonstrate that job applicants

are also affected by and affirm to the gender stereotypes i.e., men tend to apply for more technical and ambitious jobs as compared to women [193]. This paper further confirms theory presented by Howcraft et al. that the issue of gender bias is rooted in society and requires outside-in approach. Furthermore Wang et al. goes deeper and demonstrates how the implicit bias in the users interact with the biased systems creating a regressive loop, for example a biased system shows a gender-stereotypical job listing to applicants and applicants perpetuates this by selecting from these jobs instead of searching for non-stereotypic listing [202]. Finally, Shakespeare et al.'s paper study the presence of gender bias in music recommender systems and shows the effect of these biases [176].

Some papers also delve into the presence of gender biases in **AI & robotics** technologies. For example, some papers look into different specific incidences of gender bias in justice systems, medical robots, and self-driving cars [95, 24, 5]. Lopez et al. present the existence of implicit gender bias in virtual reality where the users' bias affect the virtual avatars, they tend to choose [127]. Righetti et al. analyze the significant consequences of a biased model and argue for the importance of proper legislation and multidisciplinary mitigation approach to prevent such biases in AI and robotics [160].

An interesting paper by Crockett et al. explored the gender bias effect on deception detection systems that uses Non-Verbal Behavioral cues exhibited by people and predicts if the subject is deceptive. Although they didn't find any significant effect of subject's gender on the prediction accuracy, they argued that the classifiers used to detect deception should be trained separately for different genders because that tends to work better for either gender than a one-size fits all approach [46].

There were only two papers that studied the gender bias in automated decision-making systems used specifically for **governing and policymaking** purposes. The paper by Ester Shein with the help of poverty attorney highlights the ground reality of AI decision-makings in human social programs. The paper points out that although AI

Focus of papers	Paper Count (Percent)	Sub-Themes
Gender Bias Analysis	48(39.7%)	AI[4], NLP[14], Facial Data Analysis[8], Legal & Ethical Implication[9], Recommender Systems[3], Healthcare & Medicine [2], Policy & Government [2], Search & Ranking[2], Marketing[1], Automated Systems[1], Automated Recruitment[1]
Mitigation Methods	34(28.1%)	NLP[14], Facial Data Analysis[8], Recommender Systems[3], Classification[2], Legal & Ethical Implication[1], Marketing[1], NA[1]
Detection Methods	19(15.7%)	NLP[11], Facial Data Analysis[3], Automated Recruitment[2], Individual Fairness[1], Unwanted Associations[1]
User Studies	8(6.6%)	NLP[2], Facial Data Analysis[1] ,Legal & Ethical Implication[1], Search & Ranking[1], Recommender Systems[1], Others[2]
Case Studies	5(4.1%)	NLP[2] ,Legal & Ethical Implication[1], Classification[1], Recommender Systems[1]
Literature Reviews	7(5.8%)	NLP[3], Facial Data Analysis[1], Healthcare & Medicine [1],Search & Ranking[1], Bias Mitigation Frameworks[1]

Table 4.1.: Distribution of Papers Collected for this Review Based on the Focus of Paper

automates the systems faster and efficient, it might not necessarily be accurate. Due to the nature of human-services programs the fairness of these systems is crucial in these systems and the cases that comes across these systems require a nuanced solution which AI systems are not capable of [178]. Likewise, Hicks demonstrate the effect of gender bias in government identification card issuing algorithms. It highlights the lack of representation of the non-binary and queer community in automated decision-making systems [91]. This is one of the few papers that advocates the need of representation of LGBTQ+ population in automated decision-making models.

Some papers study the existence and effects of gender bias in automated decision-making systems used in **medicine and healthcare** operations. Narla et al. touch upon the need to prevent gender biases in skin cancer detection algorithm [139]. In a similar vein, Paviglianiti et al. focus on medicinal devices specializing in Vital-ECG for predicting cardiac diseases [149].

Surprisingly, I found nine papers that expanded on the **legal and ethical implication** of gender biases in automated systems [104, 71]. This includes a paper by Koene et al., which is a work-in-progress paper regarding an IEEE industry standard to prevent algorithmic biases [109]. Similarly, a paper by Karimi et al. discusses the presence of gender bias in criminal recidivism and highlights how a biased system affect female prisoners [102]. Raghavan et al. looks into the legal implication of recruitment using a biased automated decision-making system [157].

Some papers describe the implication of gender biases from an ethical perspective. In their papers, Bird et al. and Glymour & Herington provide a comprehensive view of different types of biases based on the scope of the errors [19]. Glymour & Herington also measure the severity of these biases and lay out the implication of these biases [81]. Authors Gilbert & Mintz demonstrate the relationship between machines, humans, and data and show the impact of human cognitive bias in machine learning

pipeline [80]. Moreover, the paper by Donnelly & Stapleton shows how a gender-biased system reinforces gender bias and can harm the marginalized population. This is an interesting paper as it focuses on the importance of creating a fair system that does not inflict discrimination against minorities [58]. Finally, Fleisher et al. and D'Ignazio present the concept of individual fairness [71] and participatory design [47] to remediate gender bias from algorithmic systems.

A total of nine papers reviewed the gender bias issue in **Recommender and Search Engine Optimization(SEO)** algorithms. Authors studied bias in music recommender systems [176, 134], career recommendation systems [200] and both SEO and Recommender systems in general [11, 202, 147, 145, 23, 78]. One interesting paper by Howard et al. looked into specific incidences of gender biases in AI and robotics. For example, incidents of gender biases in justice systems, medical robot, and self-driving car [95].

4.3.2. Bias Mitigation Methods & Frameworks

Thirty-four out of 120 papers propose different gender bias mitigation methods and frameworks. Algorithmic bias are usually prevented or mitigated by manipulating the source of the bias. In most cases the source is either the training corpus or the algorithm itself. Based on the phase of the training when the model designers introduce the intervention there are three different types of algorithmic bias mitigation [67]:

- Pre-Processing: In these methods, the intervention is introduced before the training starts. For example, data manipulation/augmentation, creating a checklist to vet the algorithm/data, targeted data collection are some of the tasks that can be done to prevent any unintentional bias ahead of time.

- In-Processing: In these methods, the intervention occurs during the model training phase. Adversarial learning is popular debiasing method that falls under the

Algorithm Family	Sub-Category	Paper	Mitigation Method
NLP	Image Processing	[184]	Corpus Level Constraints
	Dialogue Systems	[125]	Adversarial Learning
	Voice Processing	[44]	Checklists + Representative Data
	NA	[41]	Algorithm Auditing
	Language Processing	[99]	Corpus Level Constraints + Posterior Regularization
	Language Processing	[17]	Data Statements
	Word Embeddings	[212]	Ridge Regression
	Word Embeddings	[156]	Scrubbing, Debaising and Strong Debiasing
	Word Embeddings	[204]	Double-Hard Debiasing
	Word Embeddings	[128]	Counterfactual Data Augmentation
	Word Embeddings	[131]	Counterfactual Data Augmentation
	Word Embeddings	[215]	Gender Neutral Word Embedding
	Word Embeddings	[205]	Representative Data
	Abusive Language Detection	[181]	Equalized Odds processing
	NA	[41]	Auditing Algorithms
	NA	[92]	Gender Bias Taxonomy
	Advertising	[66]	Greedy Algorithm
Automated Facial Analysis	Facial Recognition Task	[203]	Adversarial Debiasing
	Face Attribute Recognition	[103]	Representative Data
	Gender Classification	[208]	Representative Data (Racial + LGBTQIA+)
	Facial Processing Technology	[50]	Representative Data + Human Annotated Data
	Automated Face Analysis	[48]	Multi-task Convolution Neural Network
	Facial Classification Task	[137]	Data Augmentation
	Facial Recognition Task	[138]	Adversarial Regularizer
	Facial Recognition Task	[54]	Adversarial Debiasing
	Automated Face Analysis	[104]	Algorithmic Equity Toolkit
Recommender System	NA	[11]	Explore and Exploit Paradigm
	Music Recommender	[134]	Resampling and Rebalancing Data
	Job Recommender	[78]	Greedy Algorithm
	Job Recommender	[16]	Greedy Set Cover and Linear Programming
	Job Recommender	[199]	Annotated Data + LiFT Framework
Object Classification	NA	[32]	Removing and Relabelling Data + FAIR_FLASH
	Collected Inference	[216]	Langarian Relaxation
	Classification	[67]	Disparate Impact Remover + Adversarial Debiasing + Calibrated Equalized Odds

Table 4.2.: Different Bias Mitigation Methodologies Proposed by the Papers Reviewed in this Study.

in-processing methods as the correction or debiasing occurs while the model is training. Applying corpus level constraints, relabeling the data are also some examples of this type of debiasing.

- Post-Processing: These mitigation methods are applied post training and are the most easily applicable methods among all three types of mitigation methods. Posterior regularization and Calibrated Equalized Odds fall under this category. This technique attempts to rectify the outcomes while minimizing errors.

Additionally, researchers Bender & Friedman present three broad categories of algorithmic biases based on their origin [17],

- Pre-existing biases: these stem from biased social norms and practices. These biases get introduced into the ML systems through data.
- Technical biases: these are of technical nature and thus are introduced into ML and AI systems when the creators implement certain technical constraints and decision.
- Emergent biases: these are biases caused when ML system trained for a specific purpose is implemented for a different goal. For example, a FRT trained for Caucasian population when implemented on Asian population will tend to perform poorly thus creating bias.

In the following paragraphs, I discuss the mitigation methods that address specific methods that target pre-existing biases or data bias and technical biases or algorithmic bias.

Data Bias Mitigation

Authors Bender & Friedman suggest creating and maintaining data statements as a professional practice can reduce unwanted bias in the ML modeling. Here, by data

statements they mean providing pertinent information on the dataset that is going to be used to train the ML models. By understanding the type and characteristics of the dataset, ML model creators and users will be able to gauge the prediction quality of the ML model and its appropriate application [17]. For example, if the data collected is not representative of the general population (minorities missing) or unbalanced (over or under representation of certain population), data statements provide such information on the data so that algorithm designers could use this information to proactively implement bias mitigation methods. Cramer et al. recognize these biases and thus propose a quite simple, yet effective method to tackle gender biases. They introduce the idea of using a checklists and ML/AI engineers getting acquainted with the world in which the model is going to be implemented. This helps engineers pause and think of the outcomes they would want to see instead of getting down to coding with little thought [44] of the eventual impact of the system on the users. Baeza-Yates and Courtland also emphasize the idea of understanding the context of model implementation and the data used for model training to prevent pre-existing/data biases [11, 41].

Authors also propose a balanced dataset for ML model training as a solution to dealing with data bias [203]. Here, a balanced dataset means a dataset that is representative of all demographics and comprised of both minority and majority population in equal proportion. There are different ways to achieve a balanced dataset like collecting more data from minority population, creating augmented data for the minority population [184], or removing the majority population data for the model training. The Facial Recognition Technology (FRT) suffers from unbalanced dataset because the publicly available face datasets have comparatively less faces of minority population than majority population. To resolve this author Karkkienen & Joo propose a dataset comprised of 7 race groups: White, Black, Indian, East Asian, Southeast Asian, Middle East, and Latino [103]. Dass et al. apply a similar approach to create a comprehensive and

representative dataset by using mugshots data for mixed race groups: Black Hispanic, White Hispanic, Black non-Hispanic and White non-Hispanic [50]. Similarly, authors Wu et al. bring the issue of lack of representation of non-binary population in the ML facial classification models. To rectify this, they propose two new databases; a racially balanced dataset with a subset of LGBTQIA+ population and a dataset that consists of a gender-inclusive faces for binary and non-binary population [208].

Another interesting method proposed in these papers is Counterfactual Data Augmentation (CDA) to mitigate gender biases in Natural Language Processing (NLP) models [131, 128]. CDA is a method of data manipulation in which alternative version of the present data is added into the corpus to overall balance the gender representation in the corpus. For example; if the corpus has overwhelmingly high proportion of statements associating male gender with the profession *doctor*, like: "***He** is a doctor*" or "*The doctor provided **his** expert advice*" , then the counterfactual data will be created and added to the corpus like:"***She** is a doctor*" or "*The doctor provided **her** expert advice*" . This way in the overall corpus the gender representation is balanced hence, making it less likely that the resulting model will have gender bias stemming from the training data. In their paper authors Maudslay et al. present the concept of direct and indirect bias present in NLP and argue that CDA or a version of CDA can tackle both of such biases. In their solution they propose substitution of augmented data instead of blind addition, to maintain the grammar and discourse coherence, and bipartite graph matching of names as a better CDA approach [131].

Algorithmic Bias Mitigation

Algorithmic bias mitigation measures involve updating the algorithm and the conditions on the algorithm in order to arrive at an optimal prediction. Unlike the data

bias mitigation methods, these methods are very varied. However, there seems to be many post-processing debiasing methods.

Many of the authors demonstrate the use of adversarial debiasing techniques in various forms to get rid of gender bias from the ML/AI models. Adversarial learning is recognized as an effective measure to remove biases from ML models. Adversarial debiasing is an in-process debiasing technique in which the goal is to increase prediction accuracy while simultaneously reducing adversary's ability to predict protected values from the output prediction[2]. For example, in a credit-worthiness algorithm with gender as a protected attribute, the prediction of a person's worthiness should be highly accurate while also being ambiguous on the person's gender. In this type of learning the goal is to ..."minimize the information extracted by the *encoder* that can be maximally recovered by a parameterized model, *discriminator*" [94]. Case in point, in their paper Morales et al. utilize adversarial technique to remove sensitive information from the learning process which results into a fair and privacy-preserving facial analysis model. This learning strategy named as SensitiveNets removes sensitive information such as gender and ethnicity from the images while still being able to recognize and classify facial gestures or multi-modal learning [138]. In similar vein, there are other paper that have used similar adversarial learning approach to debias visual recognition algorithm [203], dialogue systems [125] and facial recognition system [54].

Some authors look into greedy algorithm to train their models to get the desired outcome. Here the greedy algorithm delivers a fair model because the algorithm tries to maximize the fairness metric, as designed by the model creators, as it trains. In their paper, Barnabo et al. look into using three different greedy set cover methods and a linear programming method to get a representative professional team for labor match [16]. Here, the greedy algorithm is trying to maximize the diversity of the professional team the algorithm picks while still meeting the labor match. For example,

[2]AIF360-AdversarialDebiasing

the workers picked by the algorithm to form a team should be able to complete the task at hand, the total labor cost of the team should be as minimum as possible, and the team should represent workers from all classes (like men, women and non-binary or workers from different races, ages as so on). Essentially, the algorithm will keep trying to put together a team of works for a work requirement that meets all the criteria mentioned above. Geyik et al. also use a post-processing greedy algorithm approach to mitigate gender biases in LinkedIn talent search. In their paper, they find that the debiased greedy algorithm yielded a representative sample 95% of the time in comparison to a non-debiased algorithm [78]. Farnand et al. also utilize greedy algorithm to mitigate gender bias in Influence Maximization problem i.e., maximizing profits of an advertiser in a social network. The authors identify the statistical metric that denote fairness for example fair allocation of resource across groups. With the help of their greedy algorithm, they try to maximize this property of fair allocation thus delivering a fair system [66].

Other methods proposed include using Equalized Odds processing technique, which is a popular method which is also included in the IBM AIF360 library [181], using post-process regularization technique [99, 138], Multi-task Convolution Neural Network Approach (MTCNN) [48] and Langragian relaxation for collective inference [216]. One interesting method proposed by Feldman & Peake comprised of using a mix of different debiasing techniques like disparate impact remove (pre-processing method), adversarial debiasing (in-processing method) and calibrated equalized odds(post-processing method). I have summarized all of these different methods in the table 4.2.

4.3.3. Bias Detection Methods & Frameworks

Detection of unwanted gender biases in a ML/AI model is as important as the mitigation of the biases. The detection frameworks allows to create benchmarks that

model designers can use to vet these ML/AI models that will be implemented in far-reaching systems. In the paper collected for this review, nineteen out of 120 papers presented a detection mechanism or framework to assess the presence of gender bias in an algorithm.

A comprehensive detection framework proposed by Schwemmer et al. shows FRT systems like Google Cloud Vision, Amazon Rekognition, Microsoft Azure Computer Vision contain gender bias when compared against human coded dataset. All of the systems were able to accurately identify a person as women when the picture confirmed with feminine stereotype like hair length, makeup and so on. Some of the systems even labelled images with stereotypical feminine words like "kitchen" or "cake" when in fact nothing of that sort was present in the pictures. Furthermore, the authors point out that the identification of images is binary and there is no room for LGBTQ+ population in the prediction results [172]. Serna et al. present an InsideBias detection model that detects bias in deep neural network systems that classify and analyze facial data [173]. Booth et al. also review gender bias in recruitment using video interview analysis. Their paper analyzes the bias present in image processing by utilizing psychometrics and affective computing [22]. Author Pena et al. also look into bias in automated recruitment systems using FairCVtest, a gender bias detection framework that detects bias in training data [150].

The Winograd schema proposed by Levesque et al. has inspired some of the detection methods suggested in these papers. The Winograd schema operates on commonsense reasoning questions that is asked to the machine to test if the machine can distinguish the nuances of human languages as competently as most humans are able to. A Winograd schema usually involves twin ambiguous sentences that differ in one or two words, it requires a sense of the situation, reasoning, and intention of the sentence to identify the correct form of the sentence. It is used to test the commonsense reasoning

of artificial intelligence [119]. Taking up this idea author groups Rudinger et al. and Sakaguchi et al. have proposed Winograd based questionnaire framework which can be used to test the presence of gender bias in co-reference resolution systems and word association algorithms respectively [164]. In their paper Rudinger et al., the authors present the system with Winograd style sentence-pairs that use profession and differ only in pronouns. Here, the ML system has to predict the pronoun of the profession based on the sentence. For example, the paramedic performed CPR on the passenger even though he/she/they knew it was already too late. The task for the system is to predict the appropriate pronoun for the paramedic [163].

Similar to Rudinger et al. paper, which detects gender bias in co-reference resolution, the majority of detection methods discussed in these papers target NLP tasks like sentiment analysis [196, 107, 166], information retrieval [159], Word Embeddings [87, 115] and a combination of language processing tasks [10, 93, 56] for gender bias detection. Authors Rekabsaz & Schedl have also looked into bias detection methods in Information Retrieval(IR) models. Using metrics like RankBias and AverageRankBias, authors demonstrate that IR models like BERT-Base, BM25, KNRM, MatchPyramid, PACRR, ConvKNRM and BERT-Large are all male inclined [159]. The authors use the metrics mentioned above by defining a value, in this case a mathematical representation of the magnitude or occurrences of gender definitive words like *he/him/she/her* in a document and measuring the averages of this value in the rank lists generated by the metrics. While Rekabsaz & Schedl have focused on group fairness the paper by Aggrawal et al. provide a comprehensive fairness detection methods for individual fairness. In this framework they make use of test cases for the algorithm to detect any discriminatory attributes employed by the algorithm to arrive at the prediction [1].

Another interesting framework discussed by Li et al. in their paper, is the DENOUNCER (Detection of Unfairness in Classifiers), it is a bias detection framework that

takes in training dataset, a set of sensitive attributes in the dataset like race, gender, age etc. and classifiers to be used for the computation. Using these inputs DENOUNCER is able to conduct a fairness detection and present the true versus the predicted value, hence making it plain for the model creators if their models are biased. For example, if a user wants to check if race is a fair classifier for criminal recidivism prediction. They can use DENOUNCER to select a dataset, COMPAS, and elect a classifier (race of the individual) and run the prediction. The DENOUNCER would run different prediction algorithms and compare the outcome of the prediction (will the individual reoffend) to real outcome (did the individual reoffend) and conduct fairness evaluation of the outcome. Thus, the result will reflect if the classifier selected scored high in fairness evaluation or not [120]. Finally, authors Tramer et al. look into Unwarranted Associate (UA) framework that detects unwanted associations automated systems [197].

4.3.4. Users Perspective on Gender Biases

Out of 120 papers only 8 papers focus on user studies in relation to gender specific biases in ML model. These user studies provide interesting and insightful look into how the ML models are designed, deployed, and perceived. In their paper, Fosch et al. conducted a very short survey across Twitter users to understand and quantify gender-bias present in Twitter's gender assignment algorithm. Twitter like any other social media platform thrives on personal ads that are catered to users based on their race, gender, lifestyle, political leanings and so on. In most cases, when users do not volunteer their gender information Twitter's algorithm assigns a gender to their users inferred from their app activity.

In this study conducted over four days with 109 Twitter users, researchers found that for users who did not provide their gender to the platform, the Twitter algorithm misgendered straight men 8% of the time. In contrast the misgendering for gay men

and straight women was much higher 25% and 16% respectively. Not surprisingly the non-binary population were misgendered in every case. Furthermore, even if the users tried to update their gender orientation in the platform, the ads were still biased and corresponding to the gender assigned to them. Thus, the only recourse to escape from these ads was to opt out of the personalized ads entirely [74]. Although, this was a very short study and the research community needs to conduct more studies like this to get a full picture of the nature of gender bias in the Twitter platform, it is very evident that Twitter's algorithm is significantly discriminatory against non-binary community, and straight women.

A similar study conducted with search engines does a deeper dive into the complex nature of gender bias in both platform and the users of the platform. In their paper, authors Otterbacher et al. use the result of image search results and the Ambivalent Sexism Inventory (ASI) to understand the interaction of gender-bias in the results and user's perception of it. ASI is a scoring system in which participants are measured for two types of sexism: Hostile Sexism (HS) which views minority gender in negative light and Benevolent Sexism (BS) which views minority gender in less negative albeit through stereotypical lenses. In their study, the researchers show users a grid of images and ask the users to guess the query used in the search engine which might have resulted in those images. Then the researchers reveal the actual query used and ask the participants to compare their answer with the query. These questions along with the ASI score helped the researchers to arrive at the conclusion that participants who scored higher in the ASI scale i.e., displayed sexist tendencies, tended to not see any gender-bias in the biased search result images. This study conducted on 280 participants across US, UK and India reveals that a biased search engine perpetuates gender stereotypes and sexism [147], thus further exacerbating the issue.

In their paper, Hitti et al. study user perception of gender bias by conducting a survey on 44 participants. They find that about 90% of the participants understand the concept of gender bias. The participants identify gender stereotypes (100% agreement), Gender Generalization (90% agreement), and abusive language (80% agreement) as three significant sub-types of gender bias [93]. Similarly Wang et al. also look into users preference on gender-biased versus gender-fair systems by conducting an online study on 202 university students. In this study, the researchers also gauge the users perception on their role in perpetuating gender bias in the recommender systems using a career recommender. They find that participants prefer a gender-biased (recommending jobs based on gender stereotypes) system as it confirms with their own implicit biases. Through this study, researchers suggest that gender bias is a societal issue and technical mitigation methods are simply not enough to remove gender biases from automated decision-making systems [202].

On the other hand, there are also user studies that look into the creation side of gender-biased models. In a unique study, Cowgill et al. study the behavior of close to 400 AI engineers when they are tasked to design a system that predicts standardized test scores for a demographic with a differing circumstance. In this study, the authors cleverly intervene the model creation process with a gender-bias awareness module, to study if such warning or idea changes the creators' model. Here engineers were asked to study on gender-bias awareness module before continuing their model creation. Surprisingly, after this intervention most of the models tended to over-estimate the test scores for female demographic. This study reveals that bias is a very nuanced topic and more thought should be put into how to educate ML/AI creators on tackling such issue [42].

Another expert study conducted on ML practitioners reveals the complexity of addressing the issue of bias in application [37]. In their paper, authors Andrus et al.

outline different hurdles faced by ML/AI practitioners when they are trying to mitigate gender or racial biases in practice. The majority of practitioner agree that they simply do not have access to demographic data with sensitive information, unless they work in healthcare, employment (HR, recruiting) or financial institution, they cannot gain access to racially balanced datasets. They also talk about the organizational priority and legal limitation that holds them back from vetting their algorithms for biases [6].

4.3.5. Literature reviews

In this study, I found eight papers that also conducted literature reviews on gender-bias in various ML/AI application. Unlike this work, these literature reviews mainly focused on specific algorithmic group or ML/AI implementation for example, NLP. In their paper, Khalil et al. review 24 academic publications to analyze the gender-bias in Facial Analysis Technology. Through their literature review the researchers show that facial analysis systems rely heavily on stereotypes to classify ambiguous facial features, thus leading to gender biases. The authors pull readers focus into the importance of algorithmic auditing and more academic research. Authors argue that by providing more attention into this issue, I can invoke positive action to prevent and mitigate gender biases in image classification [106]. An interesting paper by O'Reilly Shah also looks into the gender bias in the field of medicine [146].

In their paper, Blodgett et al. review 146 papers published on gender-bias analysis in the NLP systems. In this paper, the authors recommend implementing proper data vetting and understanding the context of social norms and language use to proactively mitigate biases in NLP systems. Through their paper, authors present the solution of fixing the gender-bias problem from outside-in i.e., proactively understanding the context and data rather than fixing the algorithm after training [20]. Sun et al. also conducted literature review on the presence of gender-bias in NLP systems. However,

unlike the other papers, this paper focuses on the mitigation methods presented by the fellow authors to prevent gender-biases [189]. I also found other literature review papers that focus on gender biases in academic literature [23, 45, 19, 167].

4.3.6. Case Studies

Our data search resulted in five papers focused on case studies demonstrating gender biases in automated decision-making systems. Case studies are instrumental because they display the inner mechanism of implementing these opaque systems and illuminate essential details specific to that system. Thus, allowing readers to understand and follow the process of decision-making adopted by these automated systems.

For example, the paper by Prates et al. provides a detailed case study into Google Translate. Google Translate is a powerful machine translation tool that is within reach of many people. Due to its ease of access, cost-free use, and popularity (200 million users daily), it is imperative to understand if this machine translation tool is gender biased. The authors conducted a quantitative analysis of Google Translate using gender-neutral languages supported by the system, which included 14 languages including Malay, Estonian, Finnish, Hungarian, Bengali, Swahili, Chinese, and others. Using statistical translation tools, they show that Google Translate is gender-biased towards male defaults (tends to default to he/him/his pronouns more frequently) without any reason. It assigns she/her/hers pronouns when adjectives like *Shy* or *Desirable* are used, and it overwhelmingly assigns he/him/his pronouns when STEM profession words are used [155]. Authors Farkas & Nemeth extend this study with Hungarian labor data and found that occupation-related words tend to be more biased than adjectives [65].

Another case study looks into the Automated Deception Detection tool and tries to see if the prediction provided by the software is stereotyping non-verbal behaviors (NVB) cues given by people. Crockett et al. utilize raw video data collected from 32

participants to test if there is a statistically significant gender effect on the deception detection system. Through this, they find a gender effect in NVB cues generated by people, which means I cannot use a system trained on female data to detect deception on male participants and vice-versa [46]. The paper by Wang et al. also conducts a case study into career recommender systems and implementation of debiasing technique [200]. Finally, in their paper, Dutta et al. display the effect of debiasing techniques like feature hashing on the performance of automated classification systems. Although the debiasing technique resulted in a fairer system as measured by the Difference of Equal Odds metric, it causes a drop of 6.1% in the overall accuracy of the classification [60].

4.4. Discussion and Implication

Majority of the papers, except the literature reviews, have mainly focused on one specific ML/AI model or model-family, like language processing, image processing or recommender systems and so on. Although, the mitigation and detection technique discussed in these papers can be extended to other ML models as well, the specificity of these measures shows how far-reaching and nuanced ML/AI applications are. Furthermore, I have identified following topic of interest that could benefit from more future research.

4.4.1. What is Gender Bias

Most papers reviewed in this study provide valuable insights into the presence of gender bias and, practical bias mitigation and detection methods. While these insights are important, most papers do not address what constitutes gender bias. As the ML/AI assisted systems are increasingly proliferating society, these biased systems will have direct impact on users who are unaware of these biases. Also, because of the nature of gender bias and its presence in this society and this implicit choices it is tricky for

users themselves to identify a biased automated system. It is difficult to hold systems accountable for their unfair treatment when users are unable to understand what unfair treatments look like. As pointed out by researchers like Otterbacher et al., unchecked gender biases in automated systems in combination with users' implicit biases can create a regressive feedback loop that pose risk to the gender fairness movement overall [147].

Thus, the research should outline how gender biases manifests in automated decision-making systems in its varied applications. Papers by researchers like Melchiorre et al. point out that the prediction accuracy for minority population drops in the biased system [134]. However, technical terms like prediction accuracy might not be easy to understand and communicate in many users. Moreover, providing clear definition and identification of a gender-biased system can assist in effective policing and monitoring of ML/AI assisted systems that directly impacts users.

4.4.2. Algorithmic Accountability

Currently, there is a lack of a legal framework that oversees the design and development of the automated decision-making systems. This lack of rules has allowed companies and organizations to overlook their part in the gender-bias issue. As pointed out in the research by Srinivas et al. many off-the-shelf and government prescribed systems have gender biases even when the model designers claim that they have created a fair system. There are no clear legal repercussions for creating an unfair system or for making false claims [185].

Even when the model designers and creators want to take steps in vetting the data or implementing mitigation efforts, they do not have enough resources, access, or backing from either the companies or the government. As revealed by professionals in the field [37], due to these limitations and corporate agendas taking precedence, there is no improvement in the gender-bias issue even when there are multiple solutions

available. Hence, I need more research into the lack of government action and legal slump regarding gender-bias issues to push the issues further. I also need to explore the hurdles, legal and ethical dilemma faced by algorithm designers who have limited or no access to comprehensive and representative data to train a fair model.

4.4.3. Interdisciplinary Approach

The papers reviewed in the study show that gender-bias issues are present in all applications of automated-decision-making systems. These systems pose risks to users belonging to minority gender groups. As pointed out by several researchers in their papers, field knowledge is very important when creating an effective and just system. In order to understand the extent of damage and remedy, I need support from experts in these diverse fields. Their field knowledge and insight can guide the automated-systems designers and researchers to spot the risk factor and lend support to the remediation of these issues. For example, if a social program is utilizing the automated decision-support system then experts within the field of social work and policymaking should also be involved in the process of selecting/designing and vetting the decision-support system that will be implemented.

4.4.4. Missing User Perception

Through this study, I have discovered that the user perception of these ML/AI application is largely unexplored. Users play a vital role in bias recognition and success of mitigation methods. As pointed out by several papers in this study, there is great digital divide between the creators of ML/AI systems and the users who benefit from these systems. User likeability and trust into ML/AI assisted decision-making system is equally or more important than the functionality and efficiency of the system for the successful integration. As I have discovered in this study, there were only 8 papers that

leveraged user studies to understand how users perceive, comprehend, and utilize these systems.

In this study, user studies have demonstrated the implicit bias existing on users [202] and how these biases can exacerbate the gender bias further [147]. Additionally, user studies like the one conducted by Andrus et al. on ML/AI practitioners, also shed light on the ground-reality of model creators who are trying to create a fair system but are unable to do so due to various organizational and legal hurdles [6].

Considering the rapid pace at which these systems are being implemented into this lives, 8 user studies is very low. Thus, it is even more important to include user views and experiences with these applications into the larger discussion of gender biases in ML systems. Especially in the context of gender fairness, I need to conduct more studies to understand the experiences of non-binary population with these systems, as their representation is largely missing from both the algorithm design process and data used in model training.

Finally, there is no denying that automated systems provide immense advantages to us and push us forward into modern civilization. Automated systems lend us the capability to actualize the fourth industrial revolution. However, as contributors of technology society I need to be cognizant of the fact that these systems might have varied effect of population of different social strata. It is this social duty and responsibility to bring everyone along into the fold into the new age of innovation. The progress of automated systems depends on majority of population being able to understand and trust these systems. A lack of understanding and trust will result into delays and conflicts within the society. Hence, it is imperative that I strive to create a fair automated system that benefits all users.

4.5. Chapter Summary

ML/AI assisted systems have seamlessly integrated into our lives, quietly manipulating the items I buy, the entertainment I see and the doctors I visit. In this study, I reviewed n= 120 academic literature published on gender bias in automated decision-making systems. The different areas explored in the research that have demonstrated presence of gender bias are identified through this literature review. The chapter also detail the bias detection and mitigation methods proposed by the researchers. Finally, this chapter highlights the areas that require more focus in the future research to further push the conversation of gender bias in ML/AI assisted systems. In conclusion, the work here highlights the importance of the definition and identification of a gender-biased system. Also, researchers should promote algorithmic auditing and interdisciplinary approach to design and develop ML/AI systems. In conclusion, in this chapter, I find that there is a knowledge gap in digital literacy and there is a need to conduct more user studies in ML/AI systems to bring the users perception into a biased system. Hence, in the following chapters I will first outline the design and implementation of some of the commonly used ML/AI application to provide readers an simplified view of ML/AI systems.

4.6. Limitation

Although this chapter provides a comprehensive view of the gender biases explored in the literature, the study is limited due to technical limitations like lack of time, not using other possible keywords for search, not utilizing other database search platforms. The study also only considers papers published within between year 2010 and for the year 2021. This chapter only provides the analysis of bias mitigation measures from theory perspective so case studies and user studies can be use to analyze the practicality of these mitigation methods proposed.

4.7. Future Work

In the future direction, case studies and user studies by first creating a prototype addressing these bias concerns and next testing the prototype with the set of users can be conducted to explore the themes explored in this chapter.

5. Illustrative Case Studies

Prior two chapters in this paper has outlined the effect of algorithmic biases and theoretically demonstrated how biases gets introduced in ML/AI models. Although these chapters provide valuable information on the biases and offer up solutions, they are still lacking information on how ML models actually work. This chapter aims to bridge that gap of information and provide a deep dive into the implementation of ML models. To this end the following chapter demonstrates case studies into ML algorithms.

5.1. Introduction

Digital advancements and increased computational power have evolved the field of Machine Learning (ML) and Artificial Intelligence (AI) [55]. As ML/AI models are now capable of handling complex data at a very large scale as well as capable of delivering faster accurate results, these models are now changing the way we work and interact with the existing technologies [123]. Although easier in application, ML/AI models are not easy to understand because of many algorithm variables used within these learning units, making it difficult to estimate the functional relationship between the input features and the target variables even for developers [162].

Though a majority of the researchers consider bias as a static factor, algorithmic bias interacts with the users in an iterative manner which has a deep-rooted impact on the algorithm's performance. Zou and Schiebinger mentioned that when Google Translate converts news articles written in Spanish into English, it occasionally rephrases

certain phrases into 'he said' or 'he wrote' which were originally written to refer to women [219]. In this way, these models often introduce prejudices that have a detrimental impact on individuals or certain groups of people [77].

To further my understanding of the biases present in ML and AI models, in this study I have created ML models and tested it against data to demonstrate the impact of biases on the outcomes. My findings revealed that people identifying as women/other gender or minority racial/ethnic groups are mostly affected by automated decision-making due to the discriminatory behavior of ML/AI adopted systems [18, 3]. This is severely concerning; thus, it was critical for me to implement the illustrative case study on popular algorithms. The majority of papers reviewed in previous chapter 4 were discussing the presence of gender bias on NLP models like Word Embeddings or Language Processing. Hence, in this chapter I have conducted a case study on Word Embedding algorithm and Machine Translations.

5.1.1. Key Contributions

The main contributions of this chapters are as following:

- **Demonstrate ML model implementation:** One of the primary goal of this chapter is to show how ML models are created by exploring the design and implementation of such models from start to finish. This will help readers who are otherwise not well-versed in technical aspect of ML/AI systems to understand ML models in easy and simple manner. Furthermore, this helps users of these systems to understand how the systems that they daily interact with, works in a very base-level.

- **Highlight the manifestation of gender bias:** In each of the case studies, the trained model is used to create predictions based on input data. This exercise

helps to show the introduction of bias into these models and the effect of bias in the output. The aim of this cause and effect display is to encourage readers to critically analyze the outputs they receive from ML models.

In the following sections, I will describe the methodology for case study in Word Embeddings and Machine Translations in sections 5.2.1 and 5.2.1 respectively. I will discuss the results discovered in the studies in section 5.3. Thereafter, the chapter provides a summary in section 5.4. Finally, the limitation of this work and the future extension of this chapter is outlined in section 5.5 and section 5.6.

5.2. Method

Illustrative case studies are good ways to understand a problem while analyzing a tool or algorithm used in the real world [43]. To this end, I conducted illustrative case studies on two categories of ML algorithms: Word Embedding, Machine Translations, and Recommender Systems.

5.2.1. Word Embedding Algorithm

Word Embedding algorithms are highly popular algorithms that are used to perform Natural Language Processing (NLP) tasks. Usually, these algorithms are pre-trained and used in unsupervised deep learning models for various purposes. For this study, I will take a closer look at Word2Vec, a type of Word Embedding algorithm that has been in existence since 2013. Word Embedding algorithms are widely used to extract semantic relatedness, synonym detection, categorization and to perform analogies.

Word2Vec: A popular Word Embedding algorithm is Word2Vec[1]. This is an open-source algorithm that utilizes cosine distance to associate two word vectors. In

[1]More about Word2Vec

addition to association, this algorithm also supports word clustering and has pre-trained word and phrase vectors. According to author Church, the Word2Vec algorithm can be explained by the analogy; *man* is to *woman* as x is to *king*. Essentially the algorithm attempts to make an association between words by converting words into vectors and then plotting them in a vector space to find which other words it is closest to. The goal of the algorithm is to maximize the following equation [39]:

$$x = ARGMAX_{x' \in V} sim(x', king + woman - man)$$

In the equation above, similarity ($sim(a, b)$) is defined as the cosine factor of vector a and vector b. Mathematically, it is represented as:

$$sim(a, b) = cos(vec(a), vec(b)) = \frac{vec(a)\dot{vec}(b)}{|vec(a)||\dot{vec}(b)|}$$

There are two types, or as many researchers describe it flavors, of this algorithm: Continuous Bag of Words (CBoW) and Skip-Gram model. CBoW predicts words related to x based on the surrounding context words, whereas Skip-Gram predicts the surrounding words based on x and its repeated usage with x. In this case study, I used CBOW for the Word2Vec model.

Dataset: For the case study, I am using two different datasets. The first dataset used is a compilation of science science fiction stories by Jannes Klass in Kaggle. Additionally, I also used healthcare data that includes the text data created by healthcare workers while doing rounds with patients

Data Pre-processing: In order to obtain the correct predictions, the data should be pre-processed, which involves removing unnecessary characters or words from the dataset and converting continuous strings of words into individual tokens. In this study, I used the following methods to pre-process the raw data.

- First, all unnecessary characters were removed from the dataset. This included the removal of newline characters, white spaces, punctuation, HTML tags, etc.

- Next, I used the suggested *stopwords* in English from *nltk* library to remove words that appear frequently in a sentence and do not carry much meaning. This includes words such as "a", "the", "is", "and", "or", and so on. The *stopwords* are a collection of such words that a user can easily download instead of creating their own list.

- Thereafter, I converted the remaining meaningful words into *tokens* with the use of library *Tokenizer*. A program does not understand the meaning of a word, hence, to make the calculations easier I convert words into numerical values or vectors. These vectors are known as the *tokens* and the process is called *tokenization*. This way, the algorithm is able to plot these vectors/tokens in a vector space to calculate the cosine distance between two words, allowing for analysis of the association between words.

Algorithm Model: I used *Gensim*[2] library to create our Word2Vec model. *Gensim* is a Python library that provides a suite of NLP tools for topic modeling. It also allows a user to load pre-trained models for easy application of the Word2Vec algorithm. Once the words have been vectorized, as discussed above, I can use the data to train the model. The resulting model can then be used to evaluate the similarities between two words based on its vector space. The model was created and trained using the syntax shown below:

Listing 5.1: Gensim Algorithm

```
import gensim.models
sentences = MyCorpus()
```

[2]Gensim Library

```
model2 = gensim.models.Word2Vec(sentences=sentences,
size=100, window=10, min_count=2, workers=50)
model2.train( sentences, total_examples =
len(model.wv.vocab.keys()), epochs=10)
```

Here, min_count helps prune words that appear more than min_count value. Size represents the number of dimensions the algorithm maps words onto (the greater the number the more accurate the result). Workers refers to the number of training parallelization made for faster learning.

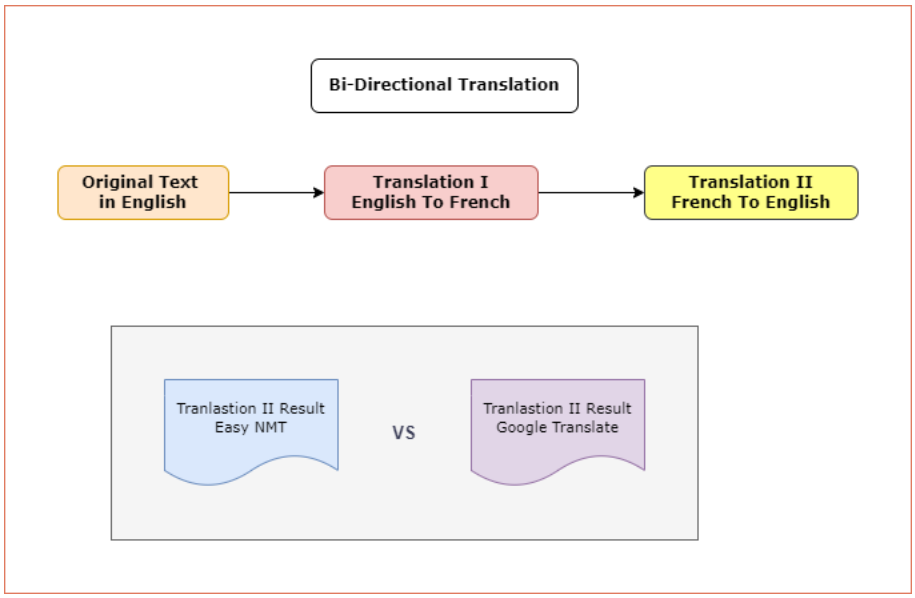

Figure 5.1.: An Overview of the Bi-directional Translation Method flow which was Implemented to Convert English Dataset to French and Back to English

5.2.2. Machine Translations

Machine Translation (MT) is the process of using automated software to translate texts, originally written in a source language, to a different language. There are three major methods proposed to conduct efficient translations[3]: Rule-based Ma-

[3]https://towardsdatascience.com/machine-translation-a-short-overview-91343ff39c9f

chine Translation (RBMT), Statistical Machine Translation (SMT), and Neural Machine Translation (NMT). Neural Machine Translation (NMT) is the newest approach to automated translation and is often referred to as an upgrade on traditional phrase-based machine translation systems. NMT employs Artificial Neural Networks (ANN) or Recurrent Neural Networks (RNN) to map the output from provided text input [209, 12]. Neural networks use multiple hidden layers to incrementally adjust the weights and derive a better outcome. GoogleNMT is especially trained on RNN, which means there is a feedback loop within the hidden layers to incrementally adjust the weights to derive correct output.

In this study, I analyzed a lightweight implementation of NMT; EasyNMT. EasyNMT provides an implementation of the NMT mechanism with access to multiple pre-trained models. For this study, I used Opus-MT which is a pre-trained model trained on a collection of open-sourced OPUS (the Open Parallel Corpus) data.

Dataset: In this case study, I used the plot summary of a movie titled "Knives Out", obtained from a Wikipedia article, in order to demonstrate an application of EasyNMT translation.

Procedure: I conducted a bi-directional translation between French and English languages. Using EasyNMT, text that was originally written in English was translated to French. Thereafter, the French text was translated back to English using EasyNMT and GoogleNMT [4], as demonstrated in Figure 5.1. Finally, the final result from GoogleNMT and EasyNMT were compared against the original texts to analyze the differences.

[4]GoogleNMT

5.3. Results

5.3.1. *Word2Vec*

Using the *most_similar* function I can find the top 5 words associated with a specific word. The word association in Word2Vec completely relies on the text provided and the training model. Hence, if there is lack of representation of any gender in the text, then the output is going to reflect that. Furthermore, if one gender is represented more than another in the training model, this will be reflected in the output as well. As mentioned earlier, I am using CBoW version of Word2Vec, which means a word 'x' will be predicted to be related to word 'y' based on surrounding context of the word. For example, if both 'x' and 'y' are often used together with similar words then, they are predicted to be related. Table 5.1, words like "effort", "friend" and "lead" are closely associated with masculine words "male," whereas "believe" is associated with "girl".

Due to the already biased nature of science fiction stories, I also applied the same *most_similar* function to a model that is trained with healthcare data. The output of this data also displayed some similar gender biases. For Healthcare data, "male" is associated with "athletes" and the masculine pronoun "him" with positive words such as "illustrates". In contrast, "female" is associated with "preponderance" and the pronoun "her" with words like "over expression". I also tried to see if there were any results for non-binary population, but there were not any significant words associated with the pronoun "their". Despite the fact that there were a few positive associations between gendered words in the results, these word associations still have many biased outcomes. The social and technical biases that overlook women and other minority genders in the training datasets and through human error create long-lasting biases and discrimination that must be analyzed further and resolved.

Training Corpus	Searched Word	Related Words
Scifi story	male	shot, effort, ball, friend, lead
	female	russias, rodders, activators, dunlap, neri
	believe	day, why, hand, girl, that
	business	good, now, came, know, martin
	woman	you, kirk, back, lieutenant, it
	man	kirk, you, it, lieutenant, he
Health Care Data	male	female, males, females, partner, athletes
	female	male, males, females, partner, preponderance
	heart	cardiac, biventicular, chf, sedimentation, wilson
	gender	sex, race, parity, pericardiotomy, stratified
	her	neu, overexpression, pgp, progesterone, somatic
	him	orogenital, scle, illustrates, nonperfusion, urticarial
	their	the, they, those, pud, autoerythrocyte

Table 5.1.: Word Associations using Word2Vec

5.3.2. *EasyNMT*

Two paragraphs of the plot summary of movie was translated. The first paragraph and the last paragraph, just to add randomness to the translation. In a quick glance, both models translated the text correctly as demonstrated in figure 5.2 and figure 5.3. However, a closer inspection showed some flaws.

There were some peculiar and problematic differences in the language translations from both GoogleNMT and EasyNMT. The original texts had many relationship-defining words like “daughter-in-law”, “son-in-law”, “daughter”, “son”, and so on. The original text also included some gendered pronouns such as “his”, “him”, and “her”.

I found that words associated with characters identified as women were altered more than the characters identified as men. For example, the word “mansion” is referred twice in this text sample. The first time it is mentioned in relation to character Harlan: “his 85^{th} birthday party at his Massachusetts mansion”, here the translated version for this part has no changes. But later on, when the same word is mentioned in relation to character Marta: “Marta watches from the balcony of her mansion” the word *mansion* is unexpectedly altered to *private hotel*. Another instance of this is the

aforementioned term “his daughter-in-law” to “his *step-daughter*”. In the translation, GoogleNMT converted the phrase “his daughter-in-law” to “his *step-daughter*” , a similar term *son-in-law* also exists in the text sample, but that word is not altered at all. EasyNMT on the other hand converted “Harlan’s housekeeper, Fran,” to “Harlan’s *maid*, Fran,”. Table 5.2 lists all of the translation or mistranslations resulted from this experiment.

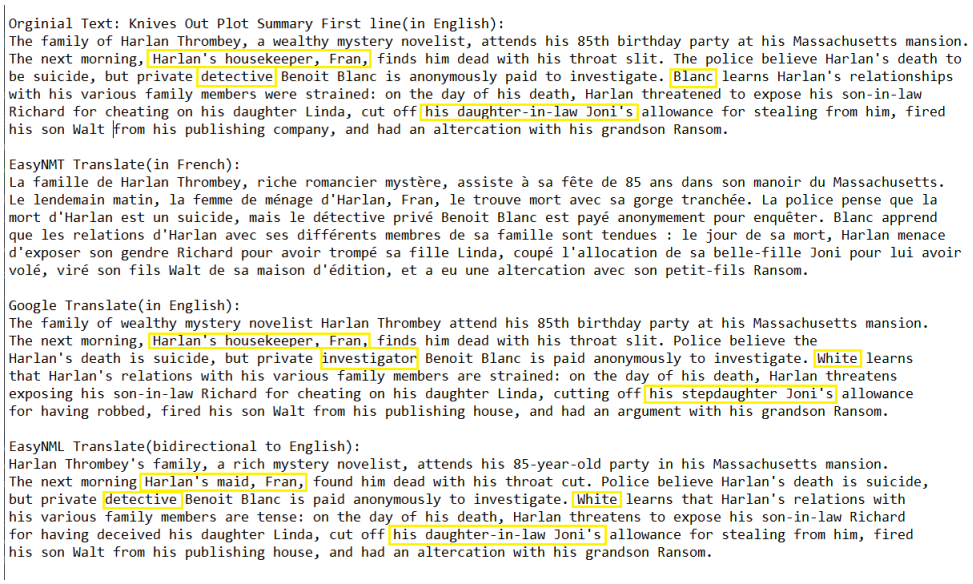

Figure 5.2.: Translation for the first couple of lines of the plot summary

Original Phrase	**GoogleNMT Translation**	**EasyNMT Translation**
Harlan’s housekeeper, Fran	Harlan’s maid, Fran	Harlan’s housekeeper, Fran
his daughter-in-law	his stepdaughter	his daughter-in-law
his son-in-law Richard	his son-in-law Richard	his son-in-law Richard
Marta tricks	Marta weaves	Marta weaves
will implicate him	will implicate her	will involve him
on her shoe	on his shoe	on her shoe
her mansion	her mansion	her private hotel

Table 5.2.: Translations Comparison: Original Phrase, GoogleNMT and EasyNMT

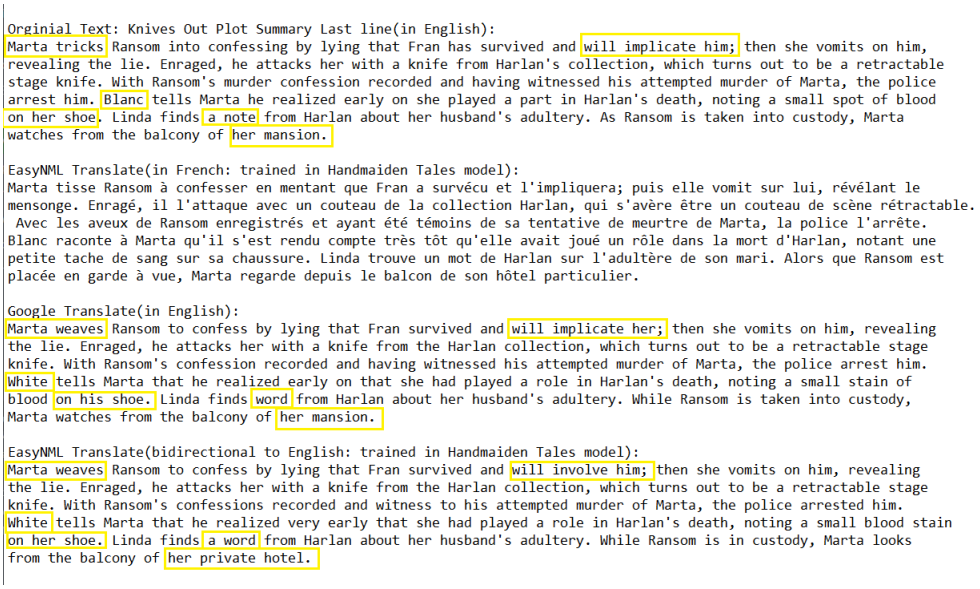

Figure 5.3.: Translation for the last couple of lines of the plot summary

5.4. Chapter Summary

In this chapter, case studies are performed in Word Embeddings and Machine Translation algorithm. These specific algorithms were selected because these algorithms of NLP family were discussed in majority of the literature in previous chapters. The case studies shed some light into how these ML models are designed and implemented, and also show how gender biases manifests into these application.

However case study do not provide an insight into the user perception of gender biases, hence the next leg of the study will discuss the gender bias in music recommender systems and present a user study conducted on a music recommender model.

Recommender system was one of the ML system discussed in many of the paper reviewed in the previous. There are some good mitigation methods proposed for such systems and these applications are immediately available to users. Music recommender systems are a sub-class of recommender systems that is available for every smartphone

users. The chapter following this will therefore discuss the recommender system in detail and demonstrate the implementation of music recommender system in detail.

5.5. Limitation

The case provide a good overview of the ML model design and implementation but this chapter only explores two algorithms and it does not dive deeper into the different types of the algorithms. The data used in these case studies are not reflective of real-life use case, this could be improved.

5.6. Future Work

The next chapter will address the limitation of these case studies by providing an in-depth view of a commonly used algorithm. It will explore different types of the algorithm and also implement bias mitigation measures introduced in previous chapters.

6. Recommender Systems: Model Creation

The previous chapter showed the design and implementation of two of the most discussed ML algorithms: Word Embeddings and Machine Translations via case studies. In this chapter, I present a detailed discussion of music Recommender System (mRS). The literature review in Chapter 4, demonstrated that recommender systems are one of the most discussed ML algorithms in literature and research also provides some good bias mitigation methods for these models. Hence, in this chapter I detail the design and implementation of mRS model. This chapter also presents implementation of three gender bias mitigation methods.

6.1. Introduction

Information overload is a genuine issue created by the rapid digitization of the world around us. While in pre-internet times people would have handful of choices prior to making a decision, now the choices are endless. Thus, resulting in emotional exhaustion and difficulty in decision-making in consumers. According to Pignatiello, when there are many options to choose from, consumers rather not make a decision to purchase [152]. Recommender systems were created as a solution to this situation. Recommender system provides a succinct and curated list of items that users can choose from, making it easier for consumers to make their decision.

The brief and curated list of options created by a recommender system can be attained from different ways. Generally, a Recommender System (RS) can be divided

into three distinct entities: users, items, and user-item matching algorithm [86]. The recommendation tasks can rely on these components of recommender system to generate accurate rating or curated suggestions [26]. Broadly the RS algorithms can be divided into three main categories as demonstrated in the paper by Tim Jones [101]. First category of algorithms focus on the method of item profiling, in such method similar items are suggested to user based on their previous purchase history. This requires the algorithm to learn about distinct items and evaluate how one item is similar or dissimilar to other items. This method is generally referred as content-based filtering. Second category of algorithms focus on the method of user profiling which means looking to a user's history and what they have preferred in the past. Based on this, highly rated items from one user's history can be suggested to another similar user. The matching algorithm in this case learns the behavior of users to understand similarity between two users. This method is generally referred as collaborative filtering. Finally, the third category of algorithms utilize some combination of the algorithms from the first and the second categories. These algorithms create a curated list or suggestions based on methods that consider both user history and item profiling, such systems are known as hybrid recommender systems.

6.1.1. Music Recommender Systems (mRS)

Music is considered one of the main sources of entertainment. Although music streaming is not a new concept, in recent times the consumption of music streaming services have grown drastically. In the year 2021, the number of premium members on music streaming platforms was 400 million, this is a huge jump from 76.8 subscribers in the year 2015 [15].

Other than providing users access to a wide selection of good quality music, music streaming platforms also provide artist and music recommendation services which

helps listeners discover new music [168]. Leading music streaming platforms employ variety of methods to accomplish the recommendation task. As discussed above music recommender tasks can be done using collaborative filtering, content-based recommendation, and employing a hybrid approach. In this study, the main goal is to understand any biases involved in user profiling thus the study uses collaborative filtering method to accomplish the music recommendation task.

6.1.2. Key Contributions

The main contributions of this chapter are as following:

- **Demystify recommender systems:** The chapter demonstrates variety of recommender systems based on their design and implementation. This gives a brief overview of how these systems are implemented in a base-level.
- **Demonstrate ML model training:** The chapter also lightly touches on the model training techniques and model validations techniques that are used in implementation of almost all ML models. This demonstration aims to provide a friendly introduction of ML model training to readers, so it eventually encourages them to discover more ML models and learn about them.
- **Present gender-bias mitigation:** The chapter also presents three different gender bias mitigation techniques that have been discussed in the research literature. These introductions showcase how bias-mitigation methods work and how they can be implemented in the ML models to combat unwanted gender bias.

In the following sections, I will describe the methodology for design of music recommender system in section 6.2. I will then demonstrate the implementation of gender bias mitigation methods in section 6.5. The summary of the chapter is provided

in section 6.6. Finally, I will conclude the chapter by outlining the limitation of this work and the future extension of this chapter is outlined in section 6.7 and section 6.8.

6.2. Methodology

6.2.1. Collaborative Filtering

Collaborative filtering relies on gathering information or preferences of many users (collaborative approach) to recommend new items to a user. For example, if user-A has given high ratings to musicA, musicB and musicC and user-B has given similar rating to musicA and musicB, the collaborative approach recommends musicC to userB because they have similar tastes regarding musicA and musicB. In this way, the recommender system relies on gathering user history and then profiling users to find similar users. In this way, in collaborative filtering user profiles are used to find similar users so the items liked by one user can be recommended to another similar user.

Collaborative filtering can be achieved by three main ways:

- Model-based Filtering: In model-based filtering different machine-learning based techniques like Bayesian networks, clustering models, singular value decomposition, Markov decision process etc. are used to generate prediction of user's rating on an unrated item.

- Memory-based Filtering: In memory-based filtering the past user-ratings for items are used to compute similarity between users or items. As suggested by the name, this method is memory heavy as it needs to learn similarity between users or items by learning user's past decisions. Calculating nearest neighbors is one of the common approaches used in this method to suggest Top-N recommendations. The memory-based filtering can suffer from a cold-start issue, which is a phenomenon in which the algorithm is unable to recommend item to a brand-new user who has no history of decisions.

In this study, memory-based collaborative filtering is utilized. For this, I will be using K-Nearest Neighbors approach and find similarity between users to create similar user profiles as shown in code snippet 6.2.1. Similar users will receive recommendations of highly rated items by other users.

Listing 6.1: Creating KNNMeans with Surprise

```
from surprise import Dataset, Reader
from surprise.model_selection import train_test_split
from surprise import KNNWithMeans
from surprise.similarities import pearson_baseline
from surprise.prediction_algorithms.knns
    import KNNBasic, KNNBaseline
myReader = Reader(line_format='user item rating',
                  sep=',',
                  rating_scale=(2,2485))
filepath = music_files+ 'model1TrainingData.csv'
data = Dataset.load_from_file(filepath,reader=myReader)
trainsetfull = data.build_full_trainset()
max_k =20
mini_k = 5
my_sim_options = {'name':'pearson',
                  'user_based':True,
                  'min_support': 3}
algo = KNNWithMeans(k=max_k,
                    min_k=mini_k,
                    sim_options=my_sim_options)
algo.fit(trainsetfull)
```

6.3. Training Data

Training data is crucial in any machine learning algorithm. It is even more important for a memory-based collaborative filtering method because the algorithm relies heavily on the past user-item rating history. Hence, in this study I are using the data presented by authors Melchiorre et al. [134]. I reached out to the authors directly to get the music listening data from LastFM site. I received several files from the authors, but for this experiment I have focused on three different types of data files. First, I had listener ratings data for all users, this file contained individual users play counts for specific track. Here the play count or how many times a track was listened by a user served as a rating of a user for that particular track (item). The file mainly contained user id, track id and play count columns. Secondly, I used the user demography file which contained demography information on the user like, username, age, gender identification, country, and timestamp. Finally, I also used the track information file which contained information on the tracks i.e., track id, track name and track artist. Unfortunately, this file didn't contain information on the track genre which might have been a good addition to this study.

For the purposes in this study, I filtered listener ratings for users who are from US and Canada regions hence I ended up with a total of $1,033,076$ rows which included $149,795$ rows for CA listeners and $88,3281$ rows for US listeners. The figure 6.1 shows the gender distribution of the listeners in the data. Each row here consists of user id, track id and play count (ratings). The play count for users ranged from 2 to 2458, which means some tracks are listened minimally twice and as many as 2458 times by users.

All the files used for the training data can be found in the Appendix section at page 151 towards the end of this paper.

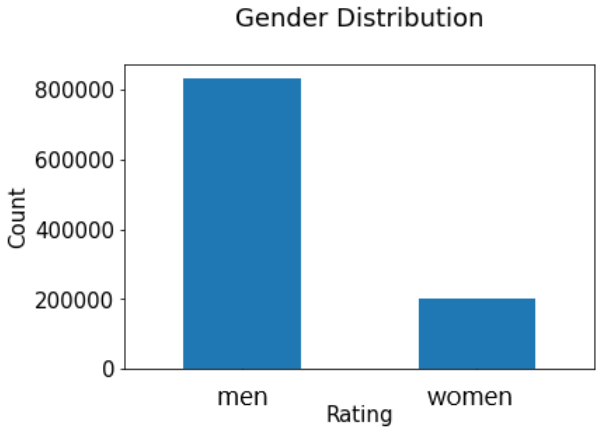

Figure 6.1.: Gender Distribution of Listeners in the Training Data

6.4. Music RS-Model

As mentioned earlier in this paper, the memory-based collaborative filtering method with KNearest Neighbors (KNN) approach is used to create the model. In this approach, the goal is to find similar users by locating top-N nearest neighbors of the users using distance methods. There are several versions of KNN approaches and a variety of distance methods that can be used with different flavors of KNN. The code set up can be viewed in the code snippet 6.4. Thankfully, there are libraries that facilitates quick and easy implementation of such complex ML models. For this study too, I utilized the **Surprise**[1] which has a variety of ready-to-use KNN models.

[1] surprise library

Listing 6.2: Model Comparision Setup

```
from surprise import Dataset, Reader
from surprise.similarities import \
    cosine, msd, pearson, pearson_baseline
from surprise.prediction_algorithms.knns import \
    KNNBasic, KNNWithMeans, KNNWithZScore, KNNBaseline
from surprise.model_selection import \
    train_test_split, GridSearchCV, cross_validate

from surprise import accuracy
from surprise.model_selection import KFold

sim_msd = {'name':'MSD',
           'user_based':True,
           'min_support': 3}

sim_cos = {'name':'cosine',
           'user_based':True,
           'min_support': 3}

sim_pearson = {'name':'pearson',
               'user_based':True,
               'min_support': 3}

sim_pearson_baseline = {'name': 'pearson_baseline',
                        'user_based':True,
                        'min_support': 3,
                        'shrinkage': 100}

sim_options = [sim_msd, sim_cos,
    sim_pearson, sim_pearson_baseline]
list_of_ks = [10,20,40]
```

To get the best performing model for the study I experimented with KNN varieties available from the the surprise library: **KNNBasic**, **KNNWithMeans**, **KNNWithZScore** and **KNNBaseline** as demonstrated in code snippet 6.4. I also used different distance methods: **cosine**, **msd**, **pearson** and **pearson_baseline**. For each type of KNN, I conducted hyper-parameter tuning using 3-folds Cross-Validation methods using Root Mean-Squared Error (RMSE) for every distance method mentioned above. By comparing the outputs of all runs, I found KNNWithMeans with pearson_baseline distance method as the best performing model. The top five rows, based on test RMSE is shown in table 6.1. The full table with outputs from all combinations of models and distance functions is added to the Appendix.

Listing 6.3: Model Comparision

```
# KNNBasic
for curr_sim_option in sim_options[0:3]:
    for curr_k in list_of_ks:
        print("Currently calculating sim_option = " + \
            str(curr_sim_option['name']) + \
            " and k = " + str(curr_k) + ' ...' )
        algo = KNNBasic(k = curr_k,
                sim_options = curr_sim_option)
        results = cross_validate(
                    algo,
                    data,
                    measures=['RMSE'],
                    cv=3,
                    return_train_measures=True);

        with open(knn_scores, 'a') as f:
            writer = csv.writer(f)
            writer.writerow(
```

```
                ['KNNBasic', curr_sim_option['name'],
                str(curr_k),
                str(np.mean(results['train_rmse'])),
                str(np.mean(results['test_rmse']))])

# KNNWithMeans
for curr_sim_option in sim_options[0:4]:
    for curr_k in list_of_ks:
        print('Currently calculating sim_option = ' +\
                str(curr_sim_option['name']) + \
            ' and k = ' + str(curr_k) + ' ...' )
        algo = KNNWithMeans(k = curr_k,
                sim_options = curr_sim_option)
        results = cross_validate(algo, data, measures=['RMSE'],
        cv=3, return_train_measures=True);

        with open(knn_scores, 'a') as f:
            writer = csv.writer(f)
            writer.writerow(
                ['KNNWithMeans2',
                curr_sim_option['name'],
                str(curr_k),
                str(np.mean(results['train_rmse'])),
                str(np.mean(results['test_rmse']))])

# KNNWithZScore
for curr_sim_option in sim_options[0:4]:
    for curr_k in list_of_ks:
        print('Currently calculating sim_option = ' + \
            str(curr_sim_option['name']) + \
            ' and k = ' + str(curr_k) + ' ...' )
        algo = KNNWithZScore(k = curr_k,
                sim_options = curr_sim_option)
```

```
results = cross_validate(
                algo,
                data,
                measures=['RMSE'],
                cv=3,
                return_train_measures=True);

with open(knn_scores, 'a') as f:
    writer = csv.writer(f)
    writer.writerow(
        ['KNNWithZScore',
        curr_sim_option['name'],
        str(curr_k),
        str(np.mean(results['train_rmse'])),
        str(np.mean(results['test_rmse']))])
```

Model Type	Distance Option	K Value	Train RMSE	Test RMSE
KNNWithMeans	pearson_baseline	20	1.372586	13.681448
KNNWithMeans	pearson_baseline	40	1.351994	13.730725
KNNWithMeans	pearson_baseline	10	1.225266	13.899492
KNNWithZScore	pearson_baseline	20	1.219249	14.116749
KNNWithMeans	MSD	10	0.527981	14.146706

Table 6.1.: Train RMSE values for different KNN models and distance options

The RMSE for this model was 13.68 which means the model was able to predict play counts(ratings) for user-item with 13 accuracy. Considering the play counts in this dataset ranged from $2 - 2485$, I deemed RMSE of 13 to be an acceptable accuracy. I derived the methods that are described above based on an article written by Mate Pocs[2].

[2]article written by Mate Pocs

6.5. Bias Mitigation Methods

Based on the prior research into gender bias in machine learning models, I identified several bias mitigation methods. However, for mRS I had consistently seen following methods suggested as viable bias mitigation methods:

- Data Rebalancing: Data re-balancing refers to creating a balanced dataset that can be used to train the models. Here, for gender bias mitigation the balancing is done for all genders represented so that the model treat all genders equally. This can be achieved by either upsampling the minority population or downsampling majority population.

- Counterfactual Intervention: Counterfactual intervention refers to providing counterfactual data for the model to train on. In case of gender bias mitigation, this means countering the gender proportion in the dataset. For example, re-labeling the gender of data created by population that identifies as men to women or vice versa, this way the model may be biased against the minority population which has been intentionally misgendered.

- Posterior Regularization: Posterior Regularization refers to methods that impose constraints on the posterior distribution [99] by way of increasing weights if the prediction produced by the algorithm is not accurate. There are different ways to achieve posterior regularization. In this study, I am leveraging the Stochastic Gradient Descent (SGD), but other popular methods include Alternating Latent Squares are also available.

Thus in this study, I am employing data re-sampling, counterfactual intervention and posterior regularization using SGD as the three bias mitigation methods to create mRS models. Since data re-sampling and counterfactual intervention mostly involves altering the underlying data these tasks were accomplished using the KNN models I

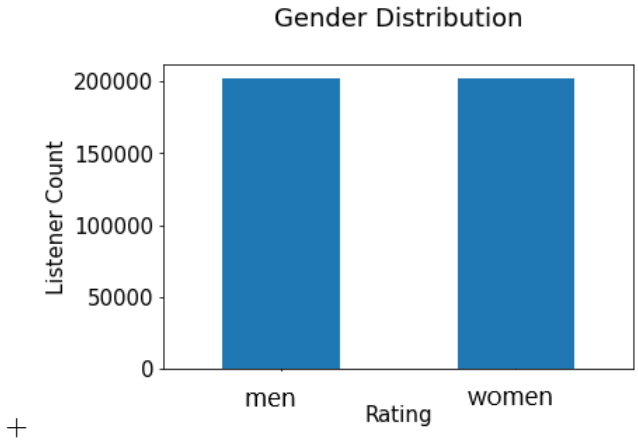

Figure 6.2.: Gender Distribution After Down Sampling Training Data

had evaluated earlier. However, for the SGD posterior regularization I ran a hyper-parameter tuning to get the optimum values of learning rate and regularization value. A brief view of the cross-validation is presented in code snippet 6.5. I used KNNBaseline model to tune the hyperparameter and based on the output's optimum values for the model. All the outputs for the tuning are shown in table 6.2.

Listing 6.4: SGD Cross-validation

```
# Choosing with best model with cross validation.
sgd_bsl_options = [
    {'method':'sgd', 'reg': 0.05, 'learning_rate': 0.006},
    {'method':'sgd', 'reg': 0.06, 'learning_rate': 0.006},
    {'method':'sgd', 'reg': 0.1, 'learning_rate': 0.006},
    {'method':'sgd', 'reg': 0.07, 'learning_rate': 0.01},
    {'method':'sgd', 'reg': 0.05, 'learning_rate': 0.01},
    {'method':'sgd', 'reg': 0.1, 'learning_rate': 0.01}]
for curr_bsl_option in sgd_bsl_options:
    for curr_k in list_of_ks:
        print('Currently calculating k = ' \
            + str(curr_k) + ' ...')
```

```
algo = KNNBaseline(k = curr_k,
                   sim_options = sim_pearson_baseline,
                   bsl_options = curr_bsl_option)
results = cross_validate(
                   algo,
                   data,
                   measures=['RMSE'],
                   cv=3,
                   return_train_measures=True);

with open(knn_baseline_sgd_score, 'a') as f:
    writer = csv.writer(f)
    writer.writerow(
        [curr_bsl_option['reg'],
        curr_bsl_option['learning_rate'],
        str(curr_k),
        str(np.mean(results['train_rmse'])),
        str(np.mean(results['test_rmse']))])
```

In the end, I created four total mRS model to create recommendations for the users. I will be able to conduct a control versus experimental study with these models. The four models are:

- Model-1: This is the control model that doesn't have any debiasing method implemented. This model is created with KNNWithMeans algorithm using pearson_baseline similarity function and k-value of 20. The LFM dataset was used as-is for training.
- Model-2: This is the first debiased model which uses balanced dataset to train the model. The downsampling is achieved by running the code shown in the the

Regularization Value	Learning Rate	K Value	Train RMSE	Test RMSE
0.10	0.010	10	1.192063	13.367878
0.05	0.010	10	1.113168	13.399422
0.07	0.010	10	1.196827	13.407844
0.05	0.010	40	1.171015	13.428545
0.05	0.010	20	1.113662	13.450819
0.05	0.006	40	1.166568	13.463796
0.10	0.010	20	1.083498	13.483932
0.07	0.010	20	1.203795	13.489803
0.06	0.006	20	1.197359	13.519747
0.05	0.006	10	1.214216	13.530841
0.07	0.010	40	1.047058	13.541944
0.10	0.010	40	1.220588	13.554132
0.10	0.006	10	1.278343	13.560243
0.05	0.006	20	1.198668	13.573240
0.06	0.006	40	1.193005	13.585327
0.06	0.006	10	1.254257	13.624414
0.10	0.006	20	1.177910	13.628031
0.10	0.006	40	1.306508	13.646247

Table 6.2.: Train RMSE values for Hyper-parameter Tuning for KNNBaseline using SGD

snippet 6.5. The fig 6.2 shows the distribution of gender after the downsampling of the training data, this gives equal representation of listeners who identify as men and women. This model is also created with KNNWithMeans algorithm using pearson_baseline similarity function and k-value of 20.

- Model-3: This is the second debiased model where I have implemented counterfactual intervention. This model is created with KNNWithMeans algorithm using pearson_baseline similarity function and k-value of 20.

- Model-4: The final debiased model is created with SGD regularization. The model uses KNNWithBaseline with pearson_baseline distance function. The regularization is 0.10, the learning rate is 0.010 and the k-value is 20.

Listing 6.5: Data Downsampling

```
from sklearn.utils import resample

# Separate majority and minority classes
df_majority = combine_track_rating[combine_track_rating.gender=='m']
df_minority = combine_track_rating[combine_track_rating.gender=='f']

# Downsample majority class
df_majority_downsampled =
    resample(df_majority,
    replace=False,
# sample without replacement
    n_samples=201897, # to match minority class
    random_state=123) # reproducible results

# Combine minority class with downsampled majority class
df_downsampled = pd.concat([df_majority_downsampled, df_minority])
```

6.6. Chapter Summary

This chapter provides a detailed summary of recommender systems. It also provides an implementation of music Recommender Systems (mRS) using user-based collaborative filtering. The chapter demonstrates the selection of best model using cross-validation techniques like Root Mean Squared Error (RMSE). Additionally, three different methods of gender bias mitigation are discussed and applied to the resulting music recommender model. The trained models now can be used in creating curated track lists to users in the user-study. The chapter following this will discuss the details of implementation of the models created here and the user-study conducted with these models.

6.7. Limitation

This chapter provides a detailed view of music recommender system design and development but it doesn't dive deeper into the different types of the recommender systems. The chapter implements merely three types of bias mitigation methods. The data used to train the model doesn't contain genre information and up-to-date playlists. The chapter is lacking user perspective of such music recommender systems and the evaluation of the mitigation methods.

6.8. Future Work

The next chapter will address the limitation of this chapter by conducting a user study that provides user perception of music recommender systems and analysing the gender bias mitigation methods.

7. Music Recommender Systems: User Study

As discussed in the prior chapter, music streaming platforms are gaining subscriptions rapidly across variety of user groups. Through such platforms, music Recommender System (mRS) algorithms engage with millions of users everyday. Thus, it is important for us to examine the impact of these algorithms on users decision making process, especially when there are little to no oversight committees that keep track of the ethical and legal aspect of such algorithms. For this same reason, research into mRS systems has been gaining popularity in the research community as well. There are many papers that has successfully demonstrated the presence of algorithmic bias in mRS in the recent times [57, 69]. In this chapter too, I intend to analyze the effect of gender bias present in music created by the mRS on users.

Gender bias in mRS can be analyzed from two different perspective: gender bias against the music artists [70] and gender bias in the music suggestion based on listeners perceived gender. Although in this chapter, I am focusing on the later both types of gender bias pose equal risk towards gender stereotyping, throttling opportunities for minority gender and ultimately amplifying the implicit biases present in the people.

Popularity bias is gender bias from the perspective of artists, and it is also one of the most researched topic in the realm of algorithmic biases in mRS [118, 111]. This type of bias results when already popular artists are favored more by the algorithm over the new and upcoming artists. Since, the popular artists have already gained exposure within the listeners, their music is heard more and thus suggested more. Thus, making

it difficult for the new and upcoming artist to gain momentum. Reaching the listeners is even more difficult when the artist represents a minority group (women and non-binary).

Hence, in this chapter I conduct a user study to explore the gender bias present in mRS, listeners perception of gender bias and their experience with music recommendation. With the help of control versus experiment model testing, I will also be evaluating the effectiveness of gender bias mitigation methods from users' end.

7.1. Experiment Setup

In this study, I analyze the effectiveness of gender bias mitigation methods with the aid of users' experience and perception of gender bias in music Recommender Systems (mRS). To achieve this, I have employed model testing of different mRS, control versus experimental models. I asked users to interact with four different mRS models and I ascertain their experience of such models and mRS in general by performing a semi-structured interview with the users. For the study, I reached out to all users who have varying level of exposure to music recommending systems in their everyday life. In this study, I recruited a total of 20 participants through recruitment efforts within the university via campus-wide emails, word-of-mouth, and social media postings. This study is approved by the ethical review board of the university and details of the study was filed under IRB# 1921637.

7.2. User Study Design

I have discussed in depth about the mRS developed for this study in the previous chapter. To summarize, for this study I have created four mRS models which will aid in conducting control versus experimental model testing with help of users. Model-1 is the control model which is trained in the LFM dataset with to alteration in the algorithm. Model-2 is the model trained in a balanced LFM dataset, the balancing was achieved

by upsampling minority user rating data (user data for women). Model-3 is the model trained on data created by using counterfactual intervention and finally, Model-4 is trained by using Posterior Regularization to prevent bias against minority gender (in this case, women). Here, it is imperative to mention that the KNN model training only uses the ratings of individual item by a user which means the gender of the user is not directly used during the training of the model.

Once the models were created and vetted, I created a User-Interface (UI) that takes user's input. Figure 7.1 shows the UI that participants used to enter their data. The study was conducted in an on-campus location of the university. The participants interacted with the UI with the investigator present. Although, in the beginning I considered performing a think-a-loud protocol to get insights into user's experience, however this protocol didn't work for this study as well as I expected. The users would often get distracted by the questions and would be unable to focus on their task at hand hence, users were allowed to finish providing their information first, and then I interviewed them afterwards. Once user was done entering the information the UI created user information data, I used this data to train the different models while I interviewed the users about their experience of using mRS applications. All the models created recommendation of top ten tracks and created one file for the user to review. Once the file was available, the recommendation from four different models were displayed to the user and I interviewed them further on these music recommendations created. The interview involved open-ended questions as well as yes/no questions. The study concluded when the final questions were asked, and each iteration of the study took around 30-40 minutes.

7.2.1. Recruitment Process

The recruitment process began by posting about the study in internal social media sites, then department-wide emails were sent out to students and faculties in computer science and engineering departments. I also implemented word-of-mouth advertisement of the study. I sent an individual follow-up recruitment email to any candidate that showed interest in the study which directed them to select a timeslot to conduct the in-lab study. In this email, the participants were again informed about the study, their eligibility requirement, and the procedure. In this way, in total 20 participants were recruited for the study.

7.2.2. Study Questionnaire

The participants were asked questions in two ways. Firstly, participants are asked to enter their information and their favorite music tracks through the UI. Secondly, a semi-structure interview was conducted with participants to understand their experience with music Recommender System (mRS) and their perception of such systems.

Through UI, participants entered their name, age, gender identification and top tracks that they listen to. The interview questionnaire on the other hand consisted of many questions that addressed different aspects of the study. The interview questions can be broadly divided into four different groups:

- Questions regarding participant's music listening habits: Here, the goal was to gauge a participant's music listening habits and also to create a rapport with the participant. These included questions on frequency of music listening using an online application, their favorite music streaming application, their preferred genre of music, and why they preferred one application over the others. Questions are as follows:

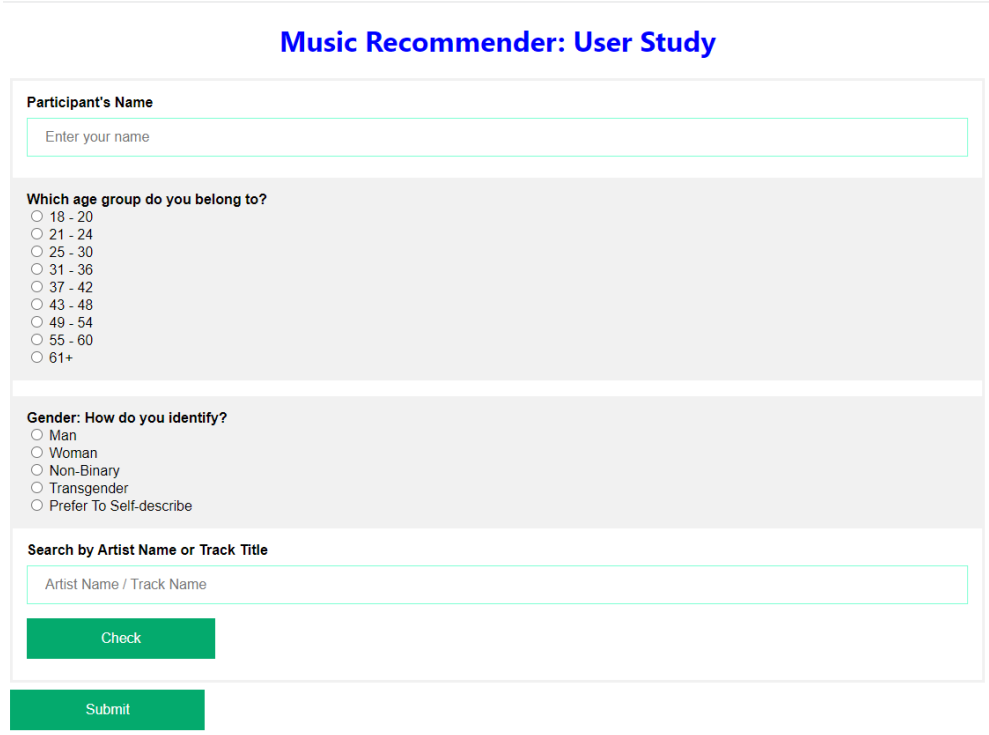

Figure 7.1.: User Study: Landing Page

1. How often do you listen to music via online applications like iTunes, Spotify, Bandcamp, etc.?
2. Which is your favorite music application and why?
3. What is your favorite genre of music and why?
4. Additional follow-up questions on music listening habits based on participant's answers to above questions.

- Questions regarding participant's experience with music recommendations: These questions were to examine the participant's familiarity with mRS. These included questions on their current experience with music recommendations provided by the streaming platforms, their satisfaction with the recommendation they are receiving, the frequency they use such recommendations and so on. Questions are as follows:
  1. How do you find new music in the [favorite music application]?
  2. How often do you find new music through music recommendations from [favorite music application]?
  3. What are your thoughts on the kinds of music recommended in these apps?
  4. What are your thoughts on how do these apps know your music tastes?
  5. Any additional follow-up questions based on participant's answers to above questions.
- Questions regarding the recommendation they have received during study: Once the models were trained and gave music recommendation, the recommendations from different models were displayed to the participant and I questioned them on these recommendations. There were four different version of recommendation, and I asked participants about their preferred recommendation and their familiarity or interest with tracks presented in each recommendation. Finally I asked

participants to rank the recommendation on the scale of $1-4$, where rank 1 represented most favorite/interesting list of music recommended and rank 4 indicated the least favorite. Here, I also showed them the most listened tracks from listeners who identified as men and listeners who identified as women from the training data and asked participants to choose which one they most related to. Questions are as follows:

1. Do you see any new songs/music recommended here in [Model 1/2/3/4]?
2. Do you recognize any artist or tracks in this recommendation from [Model 1/2/3/4]?
3. Do you think you will listen to these songs if they were recommended to you in your [favorite music application]?
4. If you had to rank these recommendations as 1: I really liked these tracks to 4:I don't like any of these tracks, how would you rank these recommendations?
5. Any additional follow-up questions based on participant's answers to above questions.

- Questions regarding bias perception in mRS: I also randomly asked participants some question on the effect of bias in music recommendation systems. Questions are as follows:

  1. Do you think this recommendation will be different if you provide any different/additional information about yourself?

7.3. Study Results

7.3.1. Demography

In this study I recruited 20 participants who took the study within a closed environment. Among the participants 8 (40%) belonged to age group $18-20$, 7 (35%) belonged to age range $25-30$, 4 (20%) belonged to age range $21-24$ and 1 (5%) belonged to age range $37-42$. The distribution of participants in different age range is also shown in figure 7.2.

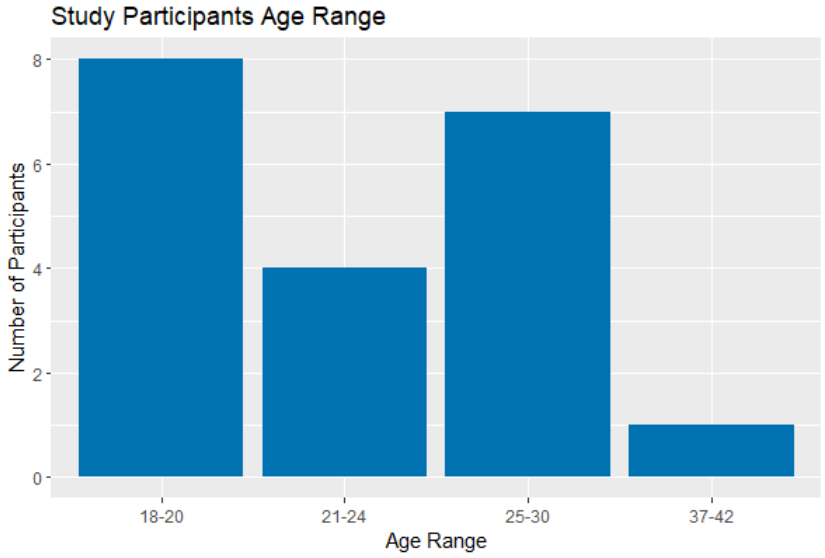

Figure 7.2.: Age Distribution of Study Participants

Majority of the participants, 13 (65%) in this study self-identified as men. However, there were 2 (10%) participants who self-identified as non-binary and transgender. Rest of the participants, 5 (25%) self-identified as women. The gender identification distribution of participants can be viewed in fig 7.3.

I asked participants about their music listening habits in online platform. Majority of participants, 11 (55%) reported listening to music daily, 8 (40%) reported to listening to music most days and only one participant reported to listening to music us-

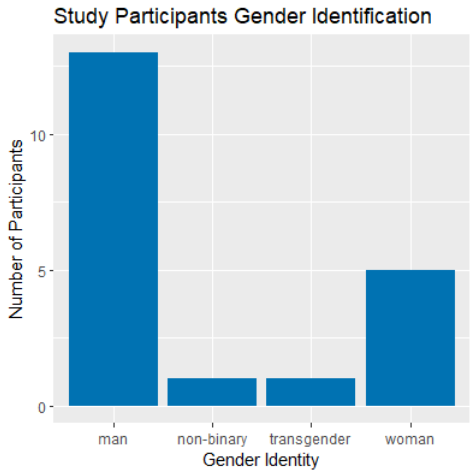

Figure 7.3.: Gender Distribution of Study Participants

ing applications once in every month. Almost all participants reported using *Spotify* as their go-to application for music listening, except two participants who reported using *iTunes* and *YouTube*.

7.3.2. Music Recommendation Ranking

Next, I asked participants to rate the recommendations received from different models. I used numerical ranking from $1-4$, in which 1 is the recommendation they like most and 4 is the recommendation they like the least. Most participants ranked Model-1 (control) model and Model-2 in favorable lights, compared to Model-3 and Model-4.

Model-3 and Model-4 were mostly rated 4 (least favorite) by 11 participants and 12 participants respectively. Only one participant rated Model-3 as their most favorite(1) and only two participants rated Model-4 as their most favorite recommendation. The Fig 7.6 and Fig 7.7 shows the ranking distribution by gender identification for Model-3 and Model-4 respectively. On the other hand, 7 participants ranked Model-1 and 9 participants ranked Model-2 as their most favorite recommendation. Although

Figure 7.4.: Model-1 Ranking with Gender Distribution

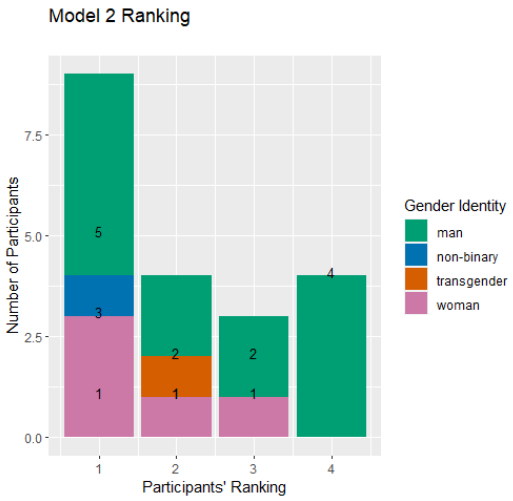

Figure 7.5.: Model-2 Ranking with Gender Distribution

in a initial glance, it looks like both Model-1, the control model and Model-2, the mitigation method implemented model did equally good with the participants, the picture changes when these rankings are broken down by gender identification. Model-1 seems to have done well for merely 2 participants who identified as woman and transgender, other than that most of its 1 ranking comes from participants who identified as men, as shown in figure 7.4. In contrast, Model-2 was ranked 1 favorably by participants who identified as women, men and non-binary, clearly demonstrated in figure 7.5. Even more interestingly, only participants who identified as men ranked Model-2 as least favorite (4). This evaluation shows that Model-2 is the more fairer recommendation for a diverse group of participants.

7.3.3. Bias Perception

I asked some questions to the participants in this study with the aim to understand their perception of gender bias in music recommender systems. These were open-ended questions and I also asked participants follow-up questions if they gave more

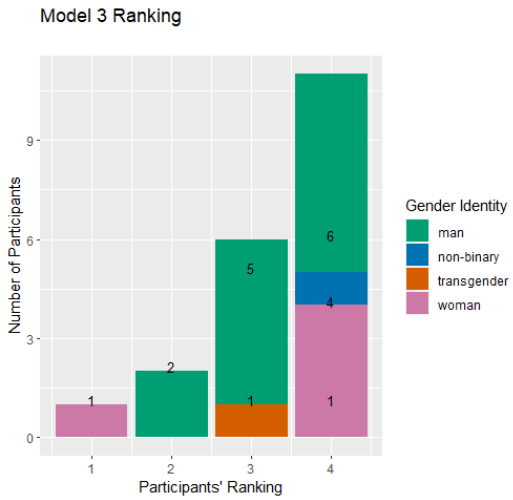

Figure 7.6.: Model-3 Ranking with Gender Distribution

Figure 7.7.: Model-4 Ranking with Gender Distribution

than a yes/no response. I also asked a scenario-based questions to participants thinking about the effect of a user's identity in the recommendations presented to them.

Q. What are your thoughts on how these apps know your music tastes? Majority of participants (12) responded that the Recommender Systems learn about user's tastes based on user's listening history. Eight participants also said that they thought recommender systems recommended songs to users based on similar listeners and similar or related artists that user has previously listened to. Only very few mentioned uses of genre, connected social media app activity and popular songs to create recommendations for listeners. One participant who had prior experience of working on a music recommender system project pointed out that music qualifiers for example the mood of the music like happy, melancholic, nostalgic etc., related to a specific track can be used to profile the user and recommend similar tracks.

Q. Scenario: Suppose there is a person out there with very similar or exactly similar listening history as you, but they belong to a different demography than you. For example, they are of different race, gender identity, age

group and location. Do you think they will get similar music recommended to them?

Participants had a variety of responses to this question. Almost everyone began with stating that the recommendation should be similar because the listening history was similar. However, as they expanded on their initial answer, some users provided contradicting opinions. Here I am going to divide the responses into two parts: responses from participants who identified as men and responses participants who did not identify as men.

I asked this question to six participants who identified as men, some of these participants believed that the recommendations should be similar regardless of the difference in user's identity, but others had a little different view. Participant 16 pointed out that music recommender system do not have users profile picture so they wouldn't be able to profile a user based on their demography. He further explained,

> *"I'm not sure. I don't think it should affect it, but I could see that maybe it would, but it, I don't think it should. Okay. Yeah, I think, I think that if I change like my age country and like gender, I think it would've changed. I feel like, I mean obviously I, I don't think it really should matter that much. Right. But I think that, I think age plays a role in what music people listen to..."*

Participant 11 also responded similarly and then mentioned,

> *"I think so. Probably, um, a lot of the things you mentioned, our region, because I would assume if someone's in Nashville, they'd probably wanna listen to more country music or something. Mm-hmm..... Um, but at the same time, I, I would, I would guess that if YouTube knows I'm a white middle-aged male, they may throw more white middle-aged music at me or something. It could be my, be my guess."*

However, rest of the participants who identified as men (4) responded that they either didn't know if apps can track that kind of information or that differing demography effects music recommendation. In contrast, the answers provided by participants who identified as women/non-binary had a diverse response to this question. I asked this question to six participants who identified as women/non-binary. While two participants responded that the recommendations should be similar even if the user's demography is different and did not elaborate further, two other participants acknowledged that there might be slight difference but nothing too drastic. When asked to consider recommendation received by a listener in Nashville versus a listener in Denver, participant 13 said,

> *"Um, I think in that case I would say, um, probably like there might be like a slight difference but not like a huge difference, you know? Yeah. Okay. I just also think like Spotify is always collecting data on like literally everything you do. So I feel like, you know, like they, they can definitely like probably try out different music than I do. So like there might be like a slight difference where it's like, oh, you tried this out, but like they didn't like it where I wouldn't like it either. Um, but like they would see like Spotify would take that data and then like make a recommendation on it. So I think like a slight difference but not like a huge difference."*

Similarly, participant 14 also initially began to answer that demography and user's identity shouldn't make any difference but the answer gave an insight into their view on the role gender identity plays in recommendations. She said,

> *"Maybe not mm-hmm... I think there will be a good amount of overlap... But, um, I don't think like the entire, like, it won't be like a, like an exact match. – Uh, yeah, I don't know if they actually collect that information. Effect of demography: Yeah, Some effect for sure, because I think the songs*

that I listen to mm-hmm.– many girls actually listen to or, you know, are crazy about those particular artists. Okay. Okay. So I think like compared to the male or you know, men mm-hmm. – So that is one thing. Another thing is the age. So I think usually, you know, the people like who are 18 to 30 years old listen to a, um, a lot of rock music or, you know, pop and all of that."

Most interestingly the participant who identified as non-binary mentioned that race and ethnicity might affect music recommendation, they also touched on the difference in music preferences for users who belong to LGBTQIA+ demography. They responded,

"Yeah. Um, but I feel like race and like ethnicity definitely plays a role too... – just cuz like a lot of my like Hispanic friends mm-hmm. Who are also like Mexican, like – Um, they tend to listen to more Spanish artists than Right. Yeah. So like, definitely plays a part for sure. Okay. And then there's also like music that very specific, like they're very specific to like gay communities as well. – So like, there's like a lot of like, um, like Lofi and like bedroom pop that's like really popular in like, in the LGBTQ community. Definitely like that."

7.3.4. Music Taste Relatedness

At the end of the study many participants were also asked to look at two different list of tracks. These were most listened tracks for listeners who identified as men as shown in figure 7.1 and most listened tracks of listeners who identified as women shown in figure 7.2.

Participants were asked to review the tracks and elect which list of tracks they were more likely to listen to or most related to. It came as a surprise that majority of participants who were asked this question selected track list 2, which was the most

Track Artists	Genre	Track Name
The Black Dahlia Murder	Melodic Death Metal	Verminous
Armin van Buuren	Trance	A State Of Trance
Above & Beyond	EDM	Group Therapy
Megadeth	Thrash Metal	Tornado Of Souls - Remastered 2004
Weather Factory	NA	Deep Thunderstorm
Decidic FX	NA	Breeze Rain
Weather Factory	NA	Calm Rain & Thunder
Uppermost	Electronic Music	Visions
Talamasca	Psychedelic Trance	Super Hero - Original Mix

Table 7.1.: Top Tracks For Listeners (Men)

Track Artists	Genre	Track Name
Kendrick Lamar	Rap	These Walls
Father John Misty	Folk	Real Love Baby
Loona	Korean Pop	Eclipse
Wet	Pop	Deadwater
Kitty	Synthpop	Drink Tickets
YESEO	Korean Pop	Last Touch
Mini Mansions	Indie + Psychedellic	Works Every Time
Allah-Las	Rock	Place In The Sun
NCT U	Korean Pop	BOSS
SF9	Korean Dance Pop	[foreign language track]

Table 7.2.: Top Tracks For Listeners (Women)

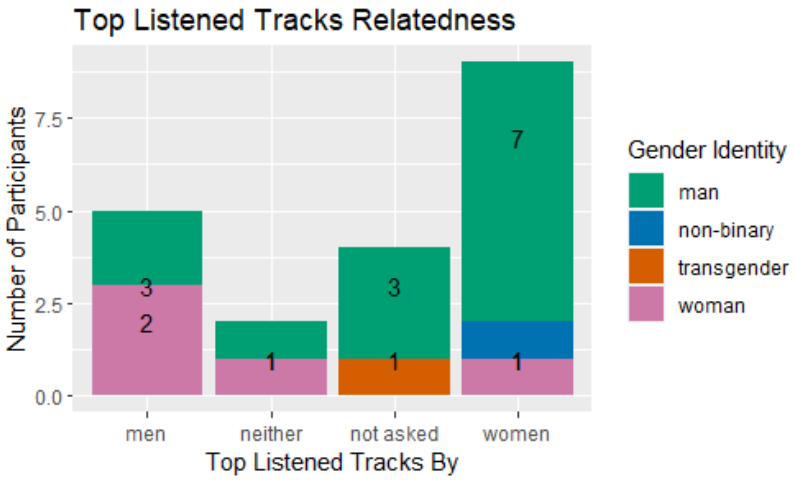

Figure 7.8.: Participants Selection of Track List

listened tracks of listeners who identified as women. It is even more intriguing that some of the tracks in this 2 list were in foreign language. Figure 7.8 provides visual representation of this rating.

7.4. Discussion & Implication

The user study has provided some interesting insights into the user perception of gender bias in music recommender systems. Here, I will discuss the main takeaways from the results of the user study and elaborate on the implications.

7.4.1. Secondary Gender Identifiers

As mentioned earlier in this chapter, the gender identity of the user is not used in the model training. The methods utilized in this experiment to create the four models were some form of K-Nearest Neighbors (KNN) algorithm. In this type of training the data used in the actual training is just a normalized ratings of items by users. In this way, gender identity of the users was not directly used in the training of the model. Still, the models created were able to learn the embedded gender bias within

the dataset, possibly due to lack of equal representation of the both genders (men and women) in the dataset and the presence of secondary gender identifiers within the dataset. The data that directly identifies the gender of a person is known as primary gender identifiers, any other attributes that impart gender inference of a person are secondary gender identifiers [74]. Secondary identifiers of genders include one's name (feminine vs. masculine sounding name), physical attributes like hair (long vs. short or plain vs. colorful), gait, voice (hoarse vs. shrill) and so on. In this study, the algorithm was able to learn gender typing of participants even when gender identification was not provided at the time of model training. This shows that secondary gender identifiers plays a very important role in contributing unintentional gender bias into algorithms.

7.4.2. Diversity in Training Data

Out of three mitigation methods that I implemented in this project, the balancing of data achieved by down-sampling worked the best with the users. By down-sampling the majority gender data (users who identified as men), the balanced training data created representative sampling thus it was able to create a recommendation that was equally liked by the diverse participants in this study. The other mitigation methods: counterfactual intervention and posterior regularization with SGD did not work as well with the users. This shows the importance of using a balanced and representative dataset to train the ML models. The dataset used in the Model-2 has representative data samples from users who identified as men and users who identified as women, it didn't have data from other gender identity. Still, it was one of the better ranked solutions. Hence, even if it is difficult to collect representative data from minority populations ML model designers should utilize the re-sampling and re-labeling techniques to create a balanced dataset for training.

7.4.3. Importance of Users Perspective

This responses received by participants in this study highlights the different viewpoints of diverse users. Only two out of 6 participants who identified as men talked about the potential effect of demography in music recommender systems at all. However, three participants out of 5 who identified as women mentioned the possibility of some slight differences. The only non-binary participant also gave an interesting view on genres that are popular with LGBTQIA+ communities. This shows that users who belong to the minority population, in this case participants who do not identify as men, are more aware of different tastes with changing demography. Only three participants in this study talked about stereotypes and tracks that are popular with a section of population like a man participant mentioned music for "middle-aged white men", a woman participant mentioned different artists that "many girls actually listen to" and a non-binary participant discussed tracks popular with "LGBTQ community". There differing viewpoints show that ML model creators should take into account that every user is aware of interests and preferences of the gender identity they belong to. Although they may be generalizing a bit, it is still important to conduct these types of user study to get a holistic view of different needs of users based on their identity. Mainly ML model creators should move away from designing models with a default gender in mind (generally a cisgender man).

7.4.4. Users Gender Bias

The answers provided by the participants in this study show that gender bias is implicit and usually embedded in our thinking process. In their answers participants unknowingly stereotyped the music based on a listener's gender identity. In the last question, in which participants were asked to select which list of tracks they were most likely to listen to or most related to, majority of men participants selected the second

list without knowing that it was most listened tracks for women. Majority of women participants selected the first list, which was the most listened tracks for men. This shows that if genre of the music, which is highly associated with gender identity, is not present users prefer music from differing genres.

7.4.5. Social and Cultural Component Inclusion

We identified that only three out of 30 papers conducted any user study to understand the algorithmic knowledge among the users. Their findings showed that in general people show distrust towards automated decision-making systems. However, end-users are inclined to know about the decision-making process as well as how their information is being used by these systems. To learn about algorithmic knowledge gaps, Cotter and Reisdorf mentioned that users learn about algorithmic knowledge through social media and their regular interaction with algorithmic platforms [40]. Besides, people with high social and economic status as well as people whose jobs require computer programming tend to exhibit greater algorithmic knowledge.

The technological solution cannot fully address machine learning biases. New design strategies are required so that technical solutions can be implemented which can jointly work with possible policy changes [9]. Furthermore, by linking both social science and computer science, it is possible to address the machine learning biases [83]. We suggest that developers should take into account the impact of cultural and human factors while designing machine learning models. Moreover, software developers and ML researchers also play a key role. We identified that researchers have warned if developers/the users of the system are unaware of the decision process of the system, then the possibility of potential biases can go unnoticed. Hence, both parties should be aware of the usage and scope of the system [31].

7.4.6. Interdisciplinary Research Approach

As ML/AI models are now applied in diverse sectors, these problems are very much context-dependent which makes it very challenging for the developers to propose any general solution for achieving 100% fair results. Therefore, we suggest incorporating people from the corresponding background while developing automated systems. For example, while building systems for offering services in the healthcare sector, healthcare professionals as well as researchers from medical fields should be included in the algorithm development process. This will help developers to identify which sets of data are required to make a final prediction as well as to understand how humans make the decision. This will also prevent the collection of unnecessary private data from the end-users.

7.4.7. Policy Reforms to Regulate Model Design and Deployment

Thus, there is a need for legislation to regulate data collection, data management, and ML Biases monitoring at current times. Bills proposed in the U.S. Congress provide a glimmer of hope for tighter regulation of ML systems and seek to assign accountability to companies that have access to user data.

- Algorithmic Accountability Act 2019 : Seeks to require periodic assessments of ML systems to prevent biases, discrimination and mishandling of personal data.
- Commercial Facial Recognition Privacy Act of 2019 : Seeks to prohibit collection, process, storage and usage of facial recognition data without documentation and consent.
- No Biometric Barriers Act 2019 : Prohibits federally assisted rental units to collect biometric data on any tenants.

Such policies and acts should be passed and enforced to ensure privacy and ownership of data by the users. This will also compel companies to take accountability, regulate their data management and usage, and mitigate biases in automated systems.

7.4.8. Digital Literacy and User Awareness

In the wake of Cambridge Analytica scandal, many educators and policymakers have reflected on the need for critical digital literacy amongst users. Prior to this, digital literacy was explained as the ability to understand and interact with technology. However, it is also important to evaluate the content presented and understand the basic mechanism of the technology [153]. Although some efforts have been made in the WEIRD (White Educated Industrialized Rich Democratic) nations to advance digital literacy, the rest of the world still remains in the dark about such technologies and their problems. A user's ability to understand the technology they are using daily has been an under-researched area, especially in the countries that make-up most of the users for digital companies [198]. If AI and ML systems are going to be used in all major sectors of societies like government, transportation, retail, entertainment, etc., then the need for digital literacy of all users becomes even more dire. The existence of ML systems that exacerbate biases within our society makes it crucial to foster and promote user awareness about how data is being collected, used, shared, and stored, as well as how this relates to automated systems [126].

7.5. Chapter Summary

This chapter provides a detailed summary of recommender systems. It also provides an implementation of music recommender systems using user-based collaborative

filtering. The chapter demonstrates the selection of best model using cross-validation techniques like Root Mean Squared Error (RMSE).

Additionally, three different methods of gender bias mitigation are discussed and applied to the resulting music recommender model. The trained models now can be used in creating curated track lists to users in the user-study. The chapter following this will discuss the details of implementation of the models created here and the user-study conducted with these models.

8. Limitations

This work comprises of systematic literature review into algorithmic gender bias in ML/AI systems. It presents the overview of ML/AI implementation with the aid of case studies and also tries to highlight the user-end of biases with user study on music recommender system. However, this work is not without its limitations. During the compilation of this thesis I have recognized following limitations:

8.1. Literature Reviews Limitations

Although I have done my due diligence to collect relevant literature papers for the systematic literature reviews discussed in Chapter 3 and Chapter 4, it is very possible that I might have missed some papers which addressed and analyzed algorithmic and gender biases. Papers published in different language and works-in-progress paper were also not considered for the systematic review. This work can be further expanded to include these papers to present a diverse and complete view of the research in algorithmic and gender biases in ML/AI application.

8.2. Case Studies Limitations

I have presented case studies into two most widely used ML algorithms in this paper. However, this can be expanded to include other different types of ML implementations like facial analysis algorithm, search optimization algorithm, healthcare related automated systems and so on. The dataset used in the case study can also be updated

to include a more complex set of data than what is used currently. Additionally, more case studies can be conducted into different ML algorithms which are easy to implement with the help of libraries.

8.3. User Study Limitations

In this paper the user study was conducted on 20 participants, due to the small number of participants it was difficult to draw statistically significant conclusion of the study. More participants could be recruited to get a better sense of the user perception of gender bias. The training data used for the user study was gathered from LFM2b data which did not include gender information, however the study could be expanded to use more data from popular music recommendation application like Spotify or iTunes.

In summary, there are limitations to this work that could be addressed in future to gain more insight into gender biases in recommender systems.

9. Future Work

Gender bias in ML/AI applications is a critical issue as it affects more than half of the population. In this paper, music recommender systems is used a vehicle to introduce the concept of gender bias and to understand the impact of the bias on user decision making. More work is required to further explore the extent of gender bias ML/AI applications that have immediate effect on users. Hence, in the future extension of this work, I aim to explore the impact of gender bias in ML/AI implementations like job recommender system or online dating/relationship applications.

Additionally, the research into gender bias can immensely benefit from more user studies into various ML/AI assisted applications. Thus, in the upcoming extensions of this work I intend to conduct more user studies and recruit users from diverse backgrounds and gender identity.

10. Conclusion

This paper addresses the issue of gender bias in the ML/AI application with the help of literature reviews, case studies and a user study. Two systematic literature review was conducted on algorithmic biases and gender biases in ML models which gives a detailed view of different biases present in ML/AI systems, the mitigation methods proposed to resolve these biases and non-technical solutions proposed to tackle the issue. Three case studies in different commonly used ML algorithms demonstrate how ML models are designed and implemented. Finally, a user study on music recommender system with 20 participants present the user perception of gender biases and the effectiveness of bias mitigation methods in improving the recommendation. Overall this work aims to give an overview on gender biases and provide a user-end perspective of importance of equitable ML/AI systems that benefits all stakeholders.

In conclusion, this paper demonstrates that gender bias in automated decision making systems is a crucial issue that needs immediate attention from ML/AI model creators, policymakers and researchers. Although the users might not be able to readily discern the effect of gender bias in everyday applications such as music recommendation, presence of such biases pose a threat to disadvantage minority populations and undo the progress made in gender equality movement.

Bibliography

[1] Aniya Aggarwal et al. “Black Box Fairness Testing of Machine Learning Models”. In: *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ESEC/FSE 2019. Tallinn, Estonia: Association for Computing Machinery, 2019, pp. 625–635. ISBN: 9781450355728. DOI: 10.1145/3338906.3338937. URL: https://doi-org.du.idm.oclc.org/10.1145/3338906.3338937.

[2] Muhammad Aurangzeb Ahmad et al. “Fairness in machine learning for healthcare”. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 3529–3530.

[3] Doris Allhutter et al. “Algorithmic profiling of job seekers in Austria: how austerity politics are made effective”. In: *Frontiers in big Data* 3 (2020), p. 5.

[4] Jafar Alzubi, Anand Nayyar, and Akshi Kumar. “Machine learning from theory to algorithms: an overview”. In: *Journal of physics: conference series*. Vol. 1142. 1. IOP Publishing. 2018, p. 012012.

[5] Saleema Amershi et al. “Guidelines for Human-AI Interaction”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–13. ISBN: 9781450359702. DOI: 10.1145/3290605.3300233. URL: https://doi.org/10.1145/3290605.3300233.

[6] McKane Andrus et al. “What We Can’t Measure, We Can’t Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 249–260. ISBN: 9781450383097. DOI: 10.1145/3442188.3445888. URL: https://doi.org/10.1145/3442188.3445888.

[7] Marija Antic and Ivana Radacic. “The evolving understanding of gender in international law and gender ideology pushback 25years since the Beijing conference on women”. In: *Women’s Studies International Forum* 83 (2020), p. 102421. ISSN: 0277-5395. DOI: https://doi.org/10.1016/j.wsif.2020.102421. URL: https://www.sciencedirect.com/science/article/pii/S0277539520308001.

[8] Maria De-Arteaga et al. “Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* ’19. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 120–128. ISBN: 9781450361255. DOI: 10.1145/3287560.3287572. URL: https://doi.org/10.1145/3287560.3287572.

[9] Abolfazl Asudeh, HV Jagadish, and Julia Stoyanovich. “Towards Responsible Data-driven Decision Making in Score-Based Systems.” In: *IEEE Data Eng. Bull.* 42.3 (2019), pp. 76–87.

[10] Marzieh Babaeianjelodar et al. “Quantifying Gender Bias in Different Corpora”. In: *Companion Proceedings of the Web Conference 2020*. WWW ’20. Taipei, Taiwan: Association for Computing Machinery, 2020, 752
bibrangedash 759. ISBN: 9781450370240. DOI: 10.1145/3366424.3383559. URL: https://doi.org/10.1145/3366424.3383559.

[11] Ricardo Baeza-Yates. “Bias in Search and Recommender Systems”. In: *Fourteenth ACM Conference on Recommender Systems*. New York, NY, USA: Association

for Computing Machinery, 2020, p. 2. ISBN: 9781450375832. URL: https://doi-org.du.idm.oclc.org/10.1145/3383313.3418435.

[12] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (2014).

[13] Michiel A Bakker et al. "Fair Enough: Improving Fairness in Budget-Constrained Decision Making Using Confidence Thresholds." In: *SafeAI AAAI*. 2020.

[14] Guha Balakrishnan et al. "Towards Causal Benchmarking of Biasin Face Analysis Algorithms". In: *Deep Learning-Based Face Analytics*. Ed. by Nalini K. Ratha, Vishal M. Patel, and Rama Chellappa. Cham: Springer International Publishing, 2021, pp. 327–359. ISBN: 978-3-030-74697-1. DOI: 10.1007/978-3-030-74697-1_15. URL: https://doi.org/10.1007/978-3-030-74697-1_15.

[15] Ian Baracskay et al. "The Diversity of Music Recommender Systems". In: *27th International Conference on Intelligent User Interfaces*. IUI '22 Companion. Helsinki, Finland: Association for Computing Machinery, 2022, pp. 97–100. ISBN: 9781450391450. DOI: 10.1145/3490100.3516474. URL: https://doi-org.du.idm.oclc.org/10.1145/3490100.3516474.

[16] Giorgio Barnabò et al. "Algorithms for Fair Team Formation in Online Labour Marketplaces". In: *Companion Proceedings of The 2019 World Wide Web Conference*. WWW '19. San Francisco, USA: Association for Computing Machinery, 2019, pp. 484–490. ISBN: 9781450366755. DOI: 10.1145/3308560.3317587. URL: https://doi.org/10.1145/3308560.3317587.

[17] Emily M. Bender and Batya Friedman. "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science". In: *Transactions of the Association for Computational Linguistics* 6 (Dec. 2018), pp. 587–604. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00041. eprint: https://

direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00041/1567666/tacl_a_00041.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00041.

[18] Sebastian Benthall and Bruce D Haynes. "Racial categories in machine learning". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019, pp. 289–298.

[19] Sarah Bird et al. "Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned". In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. WSDM '19. Melbourne VIC, Australia: Association for Computing Machinery, 2019, pp. 834–835. ISBN: 9781450359405. DOI: 10.1145/3289600.3291383. URL: https://doi-org.du.idm.oclc.org/10.1145/3289600.3291383.

[20] Su Lin Blodgett et al. "Language (technology) is power: A critical survey of" bias" in nlp". In: *arXiv preprint arXiv:2005.14050* (2020).

[21] Keith Bonawitz et al. "Towards federated learning at scale: System design". In: *arXiv preprint arXiv:1902.01046* (2019).

[22] Brandon M Booth et al. "Bias and fairness in multimodal machine learning: A case study of automated video interviews". In: *Proceedings of the 2021 International Conference on Multimodal Interaction*. 2021, pp. 268–277.

[23] Ludovico Boratto et al. "Report on the International Workshop on Algorithmic Bias in Search and Recommendation (Bias 2020)". In: *SIGIR Forum* 54.1 (Feb. 2021), pp. 1–5. ISSN: 0163-5840. DOI: 10.1145/3451964.3451973. URL: https://doi.org/10.1145/3451964.3451973.

[24] Martim Brandao. "Age and gender bias in pedestrian detection algorithms". In: (2019). DOI: 10.48550/ARXIV.1906.10490. URL: https://arxiv.org/abs/1906.10490.

[25] Anna Brown et al. “Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–12.

[26] Jiajun Bu et al. “Music recommendation by unified hypergraph: combining social media information and music content”. In: *Proceedings of the 18th ACM international conference on Multimedia*. Firenze, Italy: ACM, 2010, pp. 391–400.

[27] Joy Buolamwini and Timnit Gebru. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. New York, NY, USA: PMLR, 23–24 Feb 2018, pp. 77–91. URL: https://proceedings.mlr.press/v81/buolamwini18a.html.

[28] Keith T Butler et al. “Machine learning for molecular and materials science”. In: *Nature* 559.7715 (2018), pp. 547–555.

[29] Yang Trista Cao and Hal Daumé III. “Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle”. In: *Computational Linguistics* 47.3 (2021), pp. 615–661.

[30] Giuseppe Carleo et al. “Machine learning and the physical sciences”. In: *Reviews of Modern Physics* 91.4 (2019), p. 045002.

[31] Noel Carroll, Ita Richardson, and Raja Manzar Abbas. “"The Algorithm Will See You Now": Exploring the Implications of Algorithmic Decision-making in Connected Health.” In: *HEALTHINF*. 2020, pp. 758–765.

[32] Joymallya Chakraborty et al. “Fairway: A way to build fair ml software”. In: *Proceedings of the 28th ACM Joint Meeting on European Software Engineering*

Conference and Symposium on the Foundations of Software Engineering. 2020, pp. 654–665.

[33] Irene Y Chen, Peter Szolovits, and Marzyeh Ghassemi. "Can AI help reduce disparities in general medical and mental health care?" In: *AMA journal of ethics* 21.2 (2019), pp. 167–179.

[34] Irene Y Chen et al. "Ethical Machine Learning in Healthcare". In: *Annual Review of Biomedical Data Science* 4 (2020).

[35] Yan Chen et al. "Gender Bias and Under-Representation in Natural Language Processing Across Human Languages". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 24–34. ISBN: 9781450384735. URL: https://doi-org.du.idm.oclc.org/10.1145/3461702.3462530.

[36] Won Ik Cho et al. "Towards Cross-Lingual Generalization of Translation Gender Bias". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 449–457. ISBN: 9781450383097. DOI: 10.1145/3442188.3445907. URL: https://doi.org/10.1145/3442188.3445907.

[37] Alexandra Chouldechova and Aaron Roth. "A snapshot of the frontiers of fairness in machine learning". In: *Communications of the ACM* 63.5 (2020), 82 bibrangedash 89.

[38] Evangelia Christodoulou et al. "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models". In: *Journal of clinical epidemiology* 110 (2019), pp. 12–22.

[39] Kenneth Ward Church. "Word2Vec". In: *Natural Language Engineering* 23.1 (2017), pp. 155–162. DOI: 10.1017/S1351324916000334.

[40] Kelly Cotter and Bianca C Reisdorf. "Algorithmic Knowledge Gaps: A New Dimension of (Digital) Inequality". In: *International Journal of Communication* 14 (2020), pp. 745–765.

[41] Rachel Courtland. "The bias detectives". In: *Nature* 558.7710 (2018), pp. 357–360.

[42] Bo Cowgill et al. "Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics". In: *Proceedings of the 21st ACM Conference on Economics and Computation*. EC '20. Virtual Event, Hungary: Association for Computing Machinery, 2020, pp. 679–681. ISBN: 9781450379755. DOI: 10.1145/3391403.3399545. URL: https://doi.org/10.1145/3391403.3399545.

[43] Melissa H Cragin and Kalpana Shankar. "Scientific data collections and distributed collective practice". In: *Computer Supported Cooperative Work (CSCW)* 15.2 (2006), pp. 185–204.

[44] Henriette Cramer et al. "Assessing and Addressing Algorithmic Bias in Practice". In: *Interactions* 25.6 (Oct. 2018), pp. 58–63. ISSN: 1072-5520. DOI: 10.1145/3278156. URL: https://doi-org.du.idm.oclc.org/10.1145/3278156.

[45] Henriette Cramer et al. "Translation, Tracks &; Data: An Algorithmic Bias Effort in Practice". In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI EA '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–8. ISBN: 9781450359719. DOI: 10.1145/3290607.3299057. URL: https://doi-org.du.idm.oclc.org/10.1145/3290607.3299057.

[46] Keeley Crockett, James O'Shea, and Wasiq Khan. "Automated Deception Detection of Males and Females From Non-Verbal Facial Micro-Gestures". In: *2020 International Joint Conference on Neural Networks (IJCNN)*. Glasgow, UK, 2020, pp. 1–7. DOI: 10.1109/IJCNN48605.2020.9207684.

[47] Catherine D'Ignazio et al. "Toward Equitable Participatory Design: Data Feminism for CSCW amidst Multiple Pandemics". In: *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 437–445. ISBN: 9781450380591. URL: https://doi-org.du.idm.oclc.org/10.1145/3406865.3418588.

[48] Abhijit Das, Antitza Dantcheva, and Francois Bremond. "Mitigating Bias in Gender, Age and Ethnicity Classification: A Multi-task Convolution Neural Network Approach". In: *ECCV Workshops (1)*. Munich, Germany, 2018, pp. 573–585. URL: https://doi.org/10.1007/978-3-030-11009-3_35.

[49] Sanchari Das et al. "Evaluating User Perception of Multi-Factor Authentication: A Systematic Review". In: *Proceedings of the Thriteenth International Symposium on Human Aspects of Information Security & Assurance (HAISA 2019)*. Nicosia, Cyprus, 2019.

[50] R. Dass et al. "It's Not Just Black and White: Classifying Defendant Mugshots Based on the Multidimensionality of Race and Ethnicity". In: *2020 17th Conference on Computer and Robot Vision (CRV)*. Los Alamitos, CA, USA: IEEE Computer Society, May 2020, pp. 238–245. DOI: 10.1109/CRV50864.2020.00039. URL: https://doi.ieeecomputersociety.org/10.1109/CRV50864.2020.00039.

[51] Chris DeBrusk. "The risk of machine-learning bias (and how to prevent it)". In: *MIT Sloan Management Review* (2018).

[52] Yashar Deldjoo et al. "A flexible framework for evaluating user and item fairness in recommender systems". In: *User Modeling and User-Adapted Interaction* (2020), pp. 1–47.

[53] Audrey Desjardins, Ron Wakkary, and William Odom. "Investigating genres and perspectives in HCI research on the home". In: *Proceedings of the 33rd Annual*

ACM Conference on Human Factors in Computing Systems. 2015, pp. 3073–3082.

[54] Prithviraj Dhar et al. "Towards gender-neutral face descriptors for mitigating bias in face recognition". In: *arXiv preprint arXiv:2006.07845* (2020). DOI: 10.48550/ARXIV.2006.07845. URL: https://arxiv.org/abs/2006.07845.

[55] Dennis M Dimiduk, Elizabeth A Holm, and Stephen R Niezgoda. "Perspectives on the impact of machine learning, deep learning, and artificial intelligence on materials, processes, and structures engineering". In: *Integrating Materials and Manufacturing Innovation* 7.3 (2018), pp. 157–172.

[56] Emily Dinan et al. "Multi-dimensional gender bias classification". In: *arXiv preprint arXiv:2005.00614* (2020).

[57] Karlijn Dinnissen and Christine Bauer. "Fairness in music recommender systems: A stakeholder-centered mini review". In: *Frontiers in big Data* 5 (2022).

[58] N Donnelly and L Stapleton. "Digital Enterprise Technologies: Do Enterprise Control and Automation Technologies Reinforce Gender Biases and Marginalisation?" In: *IFAC-PapersOnLine* 54.13 (2021), pp. 551–556.

[59] Julia Dressel and Hany Farid. "The accuracy, fairness, and limits of predicting recidivism". In: *Science advances* 4.1 (2018), eaao5580.

[60] Ritik Dutta, Varun Gohil, and Atishay Jain. "Effect of Feature Hashing on Fair Classification". In: *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*. CoDS COMAD 2020. Hyderabad, India: Association for Computing Machinery, 2020, pp. 365–366. ISBN: 9781450377386. DOI: 10.1145/3371158.3371230. URL: https://doi-org.du.idm.oclc.org/10.1145/3371158.3371230.

[61] Sanghamitra Dutta et al. "An Information-Theoretic Quantification of Discrimination with Exempt Features." In: *AAAI*. 2020, pp. 3825–3833.

[62] Tim Dwyer and Jonathon Hutchinson. “Through the Looking Glass: The Role of Portals in South Korea’s Online News Media Ecology.” In: *Journal of Contemporary Eastern Asia* 18.2 (2019).

[63] Avriel Epps-Darling, Henriette Cramer, and Romain Takeo Bouyer. “Artist gender representation in music streaming.” In: *ISMIR*. Montreal, Canada: digital, 2020, pp. 248–254.

[64] Motahhare Eslami et al. “"Be Careful; Things Can Be Worse than They Appear": Understanding Biased Algorithms and Users’ Behavior Around Them in Rating Platforms"”. In: *ICWSM*. 2017, pp. 62–71.

[65] Anna Farkas and Renáta Németh. “How to measure gender bias in machine translation: Real-world oriented machine translators, multiple reference points”. In: *Social Sciences & Humanities Open* 5.1 (2022), p. 100239.

[66] Golnoosh Farnad, Behrouz Babaki, and Michel Gendreau. “A Unifying Framework for Fairness-Aware Influence Maximization”. In: *Companion Proceedings of the Web Conference 2020*. WWW ’20. Taipei, Taiwan: Association for Computing Machinery, 2020, pp. 714–722. ISBN: 9781450370240. DOI: 10.1145/3366424.3383555. URL: https://doi.org/10.1145/3366424.3383555.

[67] Tal Feldman and Ashley Peake. “End-To-End Bias Mitigation: Removing Gender Bias in Deep Learning”. In: (2021). DOI: 10.48550/ARXIV.2104.02532. URL: https://arxiv.org/abs/2104.02532.

[68] José Luis Fernández-Alemán et al. “Security and privacy in electronic health records: A systematic literature review”. In: *Journal of biomedical informatics* 46.3 (2013), pp. 541–562.

[69] Andres Ferraro, Xavier Serra, and Christine Bauer. “What is fair? Exploring the artists’ perspective on the fairness of music streaming platforms”. In: *IFIP*

Conference on Human-Computer Interaction. Springer. Switzerland: Springer, Cham, 2021, pp. 562–584.

[70] Andres Ferraro et al. *Artist and style exposure bias in collaborative filtering based music recommendations*. 2019. DOI: 10.48550/ARXIV.1911.04827. URL: https://arxiv.org/abs/1911.04827.

[71] Will Fleisher. "What's Fair about Individual Fairness?" In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '21. Virtual Event, USA: Association for Computing Machinery, 2021, pp. 480–490. ISBN: 9781450384735. DOI: 10.1145/3461702.3462621. URL: https://doi.org/10.1145/3461702.3462621.

[72] Finn Folkerts et al. "Analyzing Sentiments of German Job References". In: *2019 IEEE International Conference on Humanized Computing and Communication (HCC)*. IEEE Computer Society. Laguna Hills, CA, USA, 2019, pp. 1–6.

[73] Joel Escudé Font and Marta R Costa-Jussa. "Equalizing gender biases in neural machine translation with word embeddings techniques". In: *arXiv preprint arXiv:1901.03116* (2019).

[74] Eduard FoschVillaronga et al. "Gendering algorithms in social media". In: *ACM SIGKDD Explorations Newsletter* 23.1 (2021), 24
bibrangedash 31.

[75] Pete Fussey, Bethan Davies, and Martin Innes. "'Assisted'facial recognition and the reinvention of suspicion and discretion in digital policing". In: *The British Journal of Criminology* 61.2 (2021), pp. 325–344.

[76] Nufar Gaspar. *Intelligence reduces costs and accelerates time to market paper*. June 2018. URL: https://www.intel.com/content/dam/www/public/us/en/

documents/white-papers/artificial-intelligence-reduces-costs-and-accelerates-time-to-market-paper.pdf.

[77] Gabriel Geiger. *Court Rules Deliveroo Used 'Discriminatory' Algorithm*. 2021. URL: https://www.vice.com/en/article/7k9e4e/court-rules-deliveroo-used-discriminatory-algorithm.

[78] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. "Fairness-aware ranking in search & recommendation systems with application to linkedin talent search". In: *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*. 2019, pp. 2221–2231.

[79] Milena A Gianfrancesco et al. "Potential biases in machine learning algorithms using electronic health record data". In: *JAMA internal medicine* 178.11 (2018), pp. 1544–1547.

[80] Thomas Krendl Gilbert and Yonatan Mintz. "Epistemic Therapy for Bias in Automated Decision-Making". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '19. Honolulu, HI, USA: Association for Computing Machinery, 2019, pp. 61–67. ISBN: 9781450363242. DOI: 10.1145/3306618.3314294. URL: https://doi.org/10.1145/3306618.3314294.

[81] Bruce Glymour and Jonathan Herington. "Measuring the Biases That Matter: The Ethical and Casual Foundations for Measures of Fairness in Algorithms". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT '19. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 269–278. ISBN: 9781450361255. DOI: 10.1145/3287560.3287573. URL: https://doi.org/10.1145/3287560.3287573.

[82] Bryce W Goodman. "A step towards accountable algorithms? algorithmic discrimination and the european union general data protection". In: *29th Conference*

on Neural Information Processing Systems (NIPS 2016), Barcelona. NIPS Foundation. 2016.

[83] Bryce W Goodman. "Economic models of (algorithmic) discrimination". In: *29th Conference on Neural Information Processing Systems*. Vol. 6. 2016.

[84] Nina Grgic-Hlaca et al. "Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning." In: *AAAI*. Vol. 18. 2018, pp. 51–60.

[85] Jonathan Grudin. "AI and HCI: Two fields divided by a common focus". In: *Ai Magazine* 30.4 (2009), pp. 48–48.

[86] Alexander AS Gunawan, Derwin Suhartono, et al. "Music recommender system based on genre using convolutional recurrent neural networks". In: *Procedia Computer Science* 157 (2019), pp. 99–109.

[87] Wei Guo and Aylin Caliskan. "Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 122–133. ISBN: 9781450384735. URL: https://doi-org.du.idm.oclc.org/10.1145/3461702.3462536.

[88] Miren Gutierrez. "New Feminist Studies in Audiovisual Industries| Algorithmic Gender Bias and Audiovisual Data: A Research Agenda". In: *International Journal of Communication* 15.0 (2021). ISSN: 1932-8036.

[89] Philipp Hacker and Emil Wiedemann. "A continuous framework for fairness". In: *arXiv preprint arXiv:1712.07924* (2017).

[90] Hoda Heidari and Andreas Krause. "Preventing Disparate Treatment in Sequential Decision Making." In: *IJCAI*. 2018, pp. 2248–2254.

[91] Mar Hicks. “Hacking the Cis-tem”. In: *IEEE Annals of the History of Computing* 41.1 (2019), pp. 20–33. DOI: 10.1109/MAHC.2019.2897667.

[92] Yasmeen Hitti et al. “Proposed Taxonomy for Gender Bias in Text; A Filtering Methodology for the Gender Generalization Subtype”. In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 8–17. DOI: 10.18653/v1/W19-3802. URL: https://aclanthology.org/W19-3802.

[93] Yasmeen Hitti et al. “Proposed taxonomy for gender bias in text; a filtering methodology for the gender generalization subtype”. In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. 2019, pp. 8–17.

[94] Junyuan Hong et al. “Federated Adversarial Debiasing for Fair and Transferable Representations”. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. KDD ’21. Virtual Event, Singapore: Association for Computing Machinery, 2021, pp. 617–627. ISBN: 9781450383325. DOI: 10.1145/3447548.3467281. URL: https:%5C/%5C/doi.org%5C/10.1145%5C/3447548.3467281.

[95] Ayanna Howard and Jason Borenstein. “The ugly truth about ourselves and our robot creations: the problem of bias and social inequity”. In: *Science and engineering ethics* 24.5 (2018), pp. 1521–1536.

[96] Debra Howcroft and Jill Rubery. “’Bias in, Bias out’: gender equality and the future of work debate”. In: *Labour & Industry: a journal of the social and economic relations of work* 29.2 (2019), pp. 213–227.

[97] Alejandro Jaimes et al. “Guest Editors’ Introduction: Human-Centered Computing–Toward a Human Revolution”. In: *Computer* 40.5 (2007), pp. 30–34.

[98] Lauren Kirchner Jeff Larson Surya Mattu and Julia Angwin. *Machine Bias*. 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[99] Shengyu Jia et al. "Mitigating gender bias amplification in distribution by posterior regularization". In: *arXiv preprint arXiv:2005.06251* (2020).

[100] Elizabeth E Joh. "Artificial intelligence and policing: First questions". In: *Seattle UL Rev.* 41 (2017), p. 1139.

[101] M Tim Jones. "Recommender systems, Part 1: Introduction to approaches and algorithms". In: *IBM DeveloperWorks* 12 (2013).

[102] Marzieh Karimi-Haghighi and Carlos Castillo. "Enhancing a Recidivism Prediction Tool with Machine Learning: Effectiveness and Algorithmic Fairness". In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. ICAIL '21. São Paulo, Brazil: Association for Computing Machinery, 2021, pp. 210–214. ISBN: 9781450385268. DOI: 10.1145/3462757.3466150. URL: https://doi.org/10.1145/3462757.3466150.

[103] Kimmo Kärkkäinen and Jungseock Joo. "Fairface: Face attribute dataset for balanced race, gender, and age". In: *arXiv preprint arXiv:1908.04913* (2019). DOI: 10.48550/ARXIV.1908.04913.

[104] Michael Katell et al. "Toward Situated Interventions for Algorithmic Equity: Lessons from the Field". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT '20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 45–55. ISBN: 9781450369367. DOI: 10.1145/3351095.3372874. URL: https://doi.org/10.1145/3351095.3372874.

[105] T Kehrenberg, Z Chen, and N Quadrianto. "Tuning Fairness by Balancing Target Labels. Front". In: *Artif. Intell* 3 (2020), p. 33.

[106] Ashraf Khalil et al. “Investigating Bias in Facial Analysis Systems: A Systematic Review”. In: *IEEE Access* 8 (2020), pp. 130751–130761. DOI: 10.1109/ACCESS.2020.3006051.

[107] Svetlana Kiritchenko and Saif M Mohammad. “Examining gender and race bias in two hundred sentiment analysis systems”. In: *arXiv preprint arXiv:1805.04508* (2018).

[108] Rob Kling and Susan Leigh Star. “Human centered systems in the perspective of organizational and social informatics”. In: *ACM SIGCAS Computers and Society* 28.1 (1998), pp. 22–29.

[109] Ansgar Koene, Liz Dowthwaite, and Suchana Seth. “IEEE P7003 Standard for Algorithmic Bias Considerations: Work in Progress Paper”. In: *Proceedings of the International Workshop on Software Fairness*. FairWare ’18. Gothenburg, Sweden: Association for Computing Machinery, 2018, pp. 38–41. ISBN: 9781450357463. DOI: 10.1145/3194770.3194773. URL: https://doi.org/10.1145/3194770.3194773.

[110] Jakub Konecny et al. *Federated Learning: Strategies for Improving Communication Efficiency*. 2017. arXiv: 1610.05492 [cs.LG].

[111] Dominik Kowald, Markus Schedl, and Elisabeth Lex. “The unfairness of popularity bias in music recommendation: A reproducibility study”. In: *European conference on information retrieval*. Springer. Switzerland: eprint arXiv:1912.04696, 2020, pp. 35–42.

[112] Anoop Krishnan, Ali Almadan, and Ajita Rattani. “Understanding Fairness of Gender Classification Algorithms Across Gender-Race Groups”. In: *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2020, pp. 1028–1035. DOI: 10.1109/ICMLA51294.2020.00167.

[113] S. Kruger and B. Hermann. "Can an Online Service Predict Gender? On the State-of-the-Art in Gender Identification from Texts". In: *2019 IEEE/ACM 2nd International Workshop on Gender Equality in Software Engineering (GE)*. Los Alamitos, CA, USA: IEEE Computer Society, May 2019, pp. 13–16. DOI: 10.1109/GE.2019.00012. URL: https://doi.ieeecomputersociety.org/10.1109/GE.2019.00012.

[114] Anja Lambrecht and Catherine Tucker. "Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads". In: *Management science* 65.7 (2019), pp. 2966–2981.

[115] Susan Leavy. "Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning". In: *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*. GE '18. Gothenburg, Sweden: Association for Computing Machinery, 2018, pp. 14–16. ISBN: 9781450357388. DOI: 10.1145/3195570.3195580. URL: https://doi.org/10.1145/3195570.3195580.

[116] Katrin Leberfinger. "Gender Bias in Content-Based Music Recommendation Systems-submitted by Katrin Leberfinger". In: *Institute of Computational Perception* 0 (2022), p. 68.

[117] Nicol Turner Lee. "Detecting racial bias in algorithms and machine learning". In: *Journal of Information, Communication and Ethics in Society* (2018).

[118] Oleg Lesota et al. "Analyzing Item Popularity Bias of Music Recommender Systems: Are Different Genders Equally Affected?" In: *Proceedings of the 15th ACM Conference on Recommender Systems*. RecSys '21. Amsterdam, Netherlands: Association for Computing Machinery, 2021, pp. 601–606. ISBN: 9781450384582. DOI: 10.1145/3460231.3478843. URL: https://doi.org/10.1145/3460231.3478843.

[119] Hector J. Levesque, Ernest Davis, and Leora Morgenstern. "The Winograd Schema Challenge". In: *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*. KR'12. Rome, Italy: AAAI Press, 2012, pp. 552–561. ISBN: 9781577355601.

[120] Jinyang Li, Yuval Moskovitch, and H. V. Jagadish. "DENOUNCER: Detection of Unfairness in Classifiers". In: *Proc. VLDB Endow.* 14.12 (July 2021), pp. 2719–2722. ISSN: 21508097. DOI: 10.14778/3476311.3476328. URL: https://doi%5C-org.du.idm.oclc.org/10.14778/3476311.3476328.

[121] Li Li et al. "A review of applications in federated learning". In: *Computers & Industrial Engineering* 149 (2020), p. 106854. ISSN: 0360-8352. DOI: https://doi.org/10.1016/j.cie.2020.106854. URL: https://www.sciencedirect.com/science/article/pii/S0360835220305532.

[122] Tian Li et al. "Federated learning: Challenges, methods, and future directions". In: *IEEE Signal Processing Magazine* 37.3 (2020), pp. 50–60.

[123] Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. "Does mitigating ML's impact disparity require treatment disparity?" In: *Advances in Neural Information Processing Systems*. 2018, pp. 8125–8135.

[124] Zachary C Lipton, Alexandra Chouldechova, and Julian McAuley. "Does mitigating ml's disparate impact require disparate treatment". In: *arXiv preprint arXiv:1711.07076* (2017).

[125] Haochen Liu et al. "Mitigating gender bias for neural dialogue generation with adversarial learning". In: *arXiv preprint arXiv:2009.13028* (2020).

[126] Duri Long and Brian Magerko. "What is AI Literacy? Competencies and Design Considerations". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing

Machinery, 2020, pp. 1–16. ISBN: 9781450367080. DOI: 10.1145/3313831.3376727. URL: https://doi.org/10.1145/3313831.3376727.

[127] Sarah Lopez et al. "Investigating Implicit Gender Bias and Embodiment of White Males in Virtual Reality with Full Body Visuomotor Synchrony". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–12. ISBN: 9781450359702. DOI: 10.1145/3290605.3300787. URL: https://doi.org/10.1145/3290605.3300787.

[128] Kaiji Lu et al. "Gender Bias in Neural Natural Language Processing". In: Cham: Springer International Publishing, 2020, pp. 189–202. ISBN: 978-3-030-62077-6. DOI: 10.48550/ARXIV.1807.11714. URL: https://arxiv.org/abs/1807.11714.

[129] *Machine learning (ML) market size, share &; covid-19 impact analysis, by component (solution, and services), by Enterprise Size (smes, and large enterprises), by deployment (cloud and on-premise), by end-user (healthcare, retail, it and Telecommunication, BFSI, Automotive and transportation, advertising and media, manufacturing, and others), and Regional Forecast, 2022-2029*. Mar. 2022. URL: https://www.fortunebusinessinsights.com/machine-learning-market-102226.

[130] Cristina Manresa-Yee and Silvia Ramis. "Assessing Gender Bias in Predictive Algorithms Using EXplainable AI". In: *Proceedings of the XXI International Conference on Human Computer Interaction*. Interacción '21. Málaga, Spain: Association for Computing Machinery, 2021. ISBN: 9781450375979. DOI: 10.1145/3471391.3471420. URL: https://doi.org/10.1145/3471391.3471420.

[131] Rowan Hall Maudslay et al. "It's all in the name: Mitigating gender bias with name-based counterfactual data substitution". In: *arXiv preprint arXiv:1909.00871* (2019). DOI: 10.48550/ARXIV.1909.00871.

[132] Melissa D McCradden et al. “Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning”. In: *Journal of the American Medical Informatics Association* 27.12 (2020), pp. 2024–2027.

[133] Ninareh Mehrabi et al. “A survey on bias and fairness in machine learning”. In: *ACM Computing Surveys (CSUR)* 54.6 (2021), 1
bibrangedash 35.

[134] Alessandro B Melchiorre et al. “Investigating gender fairness of recommendation algorithms in the music domain”. In: *Information Processing & Management* 58.5 (2021), p. 102666.

[135] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. “Recommender systems and their ethical challenges”. In: *AI & SOCIETY* (2020), pp. 1–11.

[136] Arul Mishra, Himanshu Mishra, and Shelly Rathee. “Examining the presence of gender bias in customer reviews using word embedding”. In: *arXiv preprint arXiv:1902.00496* (2019).

[137] David A Molina, Leonardo Causa, and Juan Tapia. “Reduction of Bias for Gender and Ethnicity from Face Images using Automated Skin Tone Classification”. In: *2020 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE. Darmstadt, Germany, 2020, pp. 1–5.

[138] Aythami Morales et al. “Sensitivenets: Learning agnostic representations with application to face images”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.6 (2020), pp. 2158–2164.

[139] Akhila Narla et al. “Automated Classification of Skin Lesions: From Pixels to Practice”. In: *Journal of Investigative Dermatology* 138.10 (2018), pp. 2108–2110.

ISSN: 0022-202X. DOI: https://doi.org/10.1016/j.jid.2018.06.175. URL: https://www.sciencedirect.com/science/article/pii/S0022202X18322930.

[140] Arpita Nayak and Kaustubh Dutta. "Impacts of machine learning and artificial intelligence on mankind". In: *2017 International Conference on Intelligent Computing and Control (I2C2)*. IEEE. 2017, pp. 1–3.

[141] Gregory S. Nelson. "Bias in Artificial Intelligence". In: *North Carolina Medical Journal* 80.4 (2019), pp. 220–222. ISSN: 0029-2559. DOI: 10.18043/ncm.80.4.220. eprint: https://www.ncmedicaljournal.com/content/80/4/220.full.pdf. URL: https://www.ncmedicaljournal.com/content/80/4/220.

[142] Jakob Nielsen and Rolf Molich. "Heuristic evaluation of user interfaces". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1990, pp. 249–256.

[143] Mutale Nkonde. "Automated anti-blackness: facial recognition in Brooklyn, New York". In: *Harvard Journal of African American Public Policy* 20 (2019), pp. 30–36.

[144] Alejandro Noriega et al. "Algorithmic fairness and efficiency in targeting social welfare programs at scale". In: *Bloomberg Data for Good Exchange Conference*. 2018.

[145] Alamir Novin and Eric Meyers. "Making Sense of Conflicting Science Information: Exploring Bias in the Search Engine Result Page". In: *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. CHIIR '17. Oslo, Norway: Association for Computing Machinery, 2017, pp. 175–184. ISBN: 9781450346771. DOI: 10.1145/3020165.3020185. URL: https://doi.org/10.1145/3020165.3020185.

[146] Vikas N O'Reilly-Shah et al. "Bias and ethical considerations in machine learning and the automation of perioperative risk assessment". In: *British Journal of Anaesthesia* 125.6 (2020), pp. 843–846.

[147] Jahna Otterbacher et al. "Investigating User Perception of Gender Bias in Image Search: The Role of Sexism". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR '18. Ann Arbor, MI, USA: Association for Computing Machinery, 2018, pp. 933–936. ISBN: 9781450356572. DOI: 10.1145/3209978.3210094. URL: https://doi.org/10.1145/3209978.3210094.

[148] Nicolas Papernot et al. "SoK: Security and privacy in machine learning". In: *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE. 2018, pp. 399–414.

[149] Annunziata Paviglianiti and Eros Pasero. "VITAL-ECG: a de-bias algorithm embedded in a gender-immune device". In: *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT*. IEEE. Roma, Italy, 2020, pp. 314–318.

[150] Alejandro Pena et al. "Bias in Multimodal AI: Testbed for Fair Automatic Recruitment". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, WA, USA, 2020, pp. 129–137. DOI: 10.1109/CVPRW50498.2020.00022.

[151] Andi Peng et al. ""What you see is what you get? The impact of representation criteria on human bias in hiring"". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 7. 1. 2019, pp. 125–134.

[152] Grant A Pignatiello, Richard J Martin, and Ronald L Hickman Jr. "Decision fatigue: A conceptual analysis". In: *Journal of health psychology* 25.1 (2020), pp. 123–135.

[153] Gianfranco Polizzi. "Digital literacy and the national curriculum for England: Learning from how the experts engage with and evaluate online content". In: *Computers & Education* 152 (2020), p. 103859.

[154] Mirjam Pot, Nathalie Kieusseyan, and Barbara Prainsack. "Not all biases are bad: equitable and inequitable biases in machine learning and radiology". In: *Insights into imaging* 12.1 (2021), pp. 1–10.

[155] Marcelo OR Prates, Pedro H Avelar, and Lus C Lamb. "Assessing gender bias in machine translation: a case study with google translate". In: *Neural Computing and Applications* 32.10 (2020), pp. 6363–6381.

[156] Flavien Prost, Nithum Thain, and Tolga Bolukbasi. "Debiasing embeddings for reduced gender bias in text classification". In: *arXiv preprint arXiv:1908.02810* (2019).

[157] Manish Raghavan et al. "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT '20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 469–481. ISBN: 9781450369367. DOI: 10.1145/3351095.3372828. URL: https://doi.org/10.1145/3351095.3372828.

[158] Daniella Raz et al. "Face Mis-ID: An Interactive Pedagogical Tool Demonstrating Disparate Accuracy Rates in Facial Recognition". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 895–904. ISBN: 9781450384735. URL: https://doi-org.du.idm.oclc.org/10.1145/3461702.3462627.

[159] Navid Rekabsaz and Markus Schedl. "Do Neural Ranking Models Intensify Gender Bias?" In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '20. Virtual Event, China: Association for Computing Machinery, 2020, pp. 2065–2068. ISBN: 9781450380164.

DOI: 10.1145\/3397271.3401280. URL: https:%5C/%5C/doi.org%5C/10.1145%5C/3397271.3401280.

[160] Ludovic Righetti, Raj Madhavan, and Raja Chatila. "Unintended consequences of biased robotic and artificial intelligence systems [ethical, legal, and societal issues]". In: *IEEE Robotics & Automation Magazine* 26.3 (2019), pp. 11–13.

[161] Drew Roselli, Jeanna Matthews, and Nisha Talagala. "Managing bias in AI". In: *Companion Proceedings of The 2019 World Wide Web Conference*. 2019, pp. 539–544.

[162] Cynthia Rudin and Joanna Radin. "Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition". In: *Harvard Data Science Review* 1.2 (2019).

[163] Rachel Rudinger et al. "Gender bias in coreference resolution". In: *arXiv preprint arXiv:1804.09301* (2018).

[164] Keisuke Sakaguchi et al. "WinoGrande: An Adversarial Winograd Schema Challenge at Scale". In: *Commun. ACM* 64.9 (Aug. 2021), pp. 99–106. ISSN: 0001-0782. DOI: 10.1145/3474381. URL: https://doi-org.du.idm.oclc.org/10.1145/3474381.

[165] Brenda Salenave Santana, Vinicius Woloszyn, and Leandro Krug Wives. "Is there Gender bias and stereotype in Portuguese Word Embeddings?" In: *arXiv preprint arXiv:1810.04528* (2018). DOI: 10.48550/ARXIV.1810.04528.

[166] Daniel Sarraf et al. "Use of artificial intelligence for gender bias analysis in letters of recommendation for general surgery residency candidates". In: *The American Journal of Surgery* 222.6 (2021), pp. 1051–1059.

[167] Beatrice Savoldi et al. "Gender bias in machine translation". In: *Transactions of the Association for Computational Linguistics* 9 (2021), pp. 845–874.

[168] Markus Schedl et al. “Music recommender systems”. In: *Recommender System Handbook* 0.0 (2015), pp. 453–492.

[169] Ron Schmelzer. *The Achilles’ Heel Of AI*. 2019. URL: https://www.forbes.com/sites/cognitiveworld/2019/03/07/the-achilles-heel-of-ai/?sh=700407187be7.

[170] Jakob Schoeffer and Niklas Kuehl. “Appropriate Fairness Perceptions? On the Effectiveness of Explanations in Enabling People to Assess the Fairness of Automated Decision Systems”. In: *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. CSCW ’21. Virtual Event, USA: Association for Computing Machinery, 2021, pp. 153–157. ISBN: 9781450384797. DOI: 10.1145/3462204.3481742. URL: https://doi-org.du.idm.oclc.org/10.1145/3462204.3481742.

[171] Klaus Schwab. *The Fourth Industrial Revolution: What It Means and how to respond*. Jan. 2016. URL: https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/.

[172] Carsten Schwemmer et al. “Diagnosing gender bias in image recognition systems”. In: *Socius* 6 (2020), p. 2378023120967171.

[173] Ignacio Serna et al. “InsideBias: Measuring bias in deep networks and application to face gender biometrics”. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. Milan, Italy, 2021, pp. 3720–3727.

[174] William Seymour. “Detecting bias: does an algorithm have to be transparent in order to Be Fair?” In: *Jo Bates Paul D. Clough Robert Jäschke* (2018), p. 2.

[175] Rubina Shaheen and Mir Kasi. “Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies, a Case of USA”. In: *European Journal of Technology* 5.1 (Jan. 2021), pp. 1–15. DOI: 10.47672/ejt.641. URL: https://ajpojournals.org/journals/index.php/EJT/article/view/641.

[176] Dougal Shakespeare et al. “Exploring artist gender bias in music recommendation”. In: *arXiv preprint arXiv:2009.01715* (2020). DOI: 10.48550/ARXIV.2009.01715.

[177] Sam Shead. *Researchers: Are we on the cusp of an ’AI winter’?* 2020. URL: https://www.bbc.com/news/technology-51064369.

[178] Esther Shein. “The Dangers of Automating Social Programs”. In: *Commun. ACM* 61.10 (Sept. 2018), pp. 17–19. ISSN: 0001-0782. DOI: 10.1145/3264627. URL: https://doi.org/10.1145/3264627.

[179] Nisha Shekhawat, Aakanksha Chauhan, and Sakthi Balan Muthiah. “Algorithmic Privacy and Gender Bias Issues in Google Ad Settings”. In: *Proceedings of the 10th ACM Conference on Web Science*. WebSci ’19. Boston, Massachusetts, USA: Association for Computing Machinery, 2019, pp. 281–285. ISBN: 9781450362023. DOI: 10.1145/3292522.3326033. URL: https://doi.org/10.1145/3292522.3326033.

[180] Keng Siau and Yin Yang. “Impact of artificial intelligence, robotics, and machine learning on sales and marketing”. In: *Twelve Annual Midwest Association for Information Systems Conference (MWAIS 2017)*. 2017, pp. 18–19.

[181] Vivek K Singh and Connor Hofenbitzer. “Fairness across network positions in cyberbullying detection algorithms”. In: *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE. Vancouver, BC, Canada, 2019, pp. 557–559.

[182] Vivek K Singh et al. “Female librarians and male computer programmers? Gender bias in occupational images on digital media platforms”. In: *Journal of the Association for Information Science and Technology* 71.11 (2020), pp. 1281–1294.

[183] Andrew Slavin Ross et al. “Ensembles of Locally Independent Prediction Models”. In: *arXiv* (2019), arXiv–1911.

[184] Philip Smith and Karl Ricanek. "Mitigating Algorithmic Bias: Evolving an Augmentation Policy that is Non-Biasing". In: *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. Snowmass, CO, USA, 2020, pp. 90–97. DOI: 10.1109/WACVW50321.2020.9096905.

[185] Nisha Srinivas et al. "Exploring Automatic Face Recognition on Match Performance and Gender Bias for Children". In: *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. Waikoloa, HI, USA, 2019, pp. 107–115. DOI: 10.1109/WACVW.2019.00023.

[186] Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. "Evaluating gender bias in machine translation". In: *arXiv preprint arXiv:1906.00591* (2019).

[187] Elizabeth Stowell et al. "Designing and Evaluating MHealth Interventions for Vulnerable Populations: A Systematic Review". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–17. ISBN: 9781450356206. DOI: 10.1145/3173574.3173589. URL: https://doi.org/10.1145/3173574.3173589.

[188] Mangala Subramaniam. "Whose interests? Gender issues and wood-fired cooking stoves". In: *American Behavioral Scientist* 43.4 (2000), pp. 707–728.

[189] Tony Sun et al. "Mitigating gender bias in natural language processing: Literature review". In: *arXiv preprint arXiv:1906.08976* (2019).

[190] Wenlong Sun, Olfa Nasraoui, and Patrick Shafto. "Evolution and impact of bias in human and machine learning algorithm interaction". In: *Plos one* 15.8 (2020), e0235502.

[191] Wenlong Sun, Olfa Nasraoui, and Patrick Shafto. "Iterated Algorithmic Bias in the Interactive Machine Learning Process of Information Filtering." In: *KDIR*. 2018, pp. 108–116.

[192] Pål Sundsøy et al. “Big data-driven marketing: how machine learning outperforms marketers’ gut-feeling”. In: *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*. Springer. 2014, pp. 367–374.

[193] Shiliang Tang et al. “Gender bias in the job market: A longitudinal analysis”. In: *Proceedings of the ACM on Human-Computer Interaction* 1.CSCW (2017), pp. 1–19.

[194] Adi L Tarca et al. “Machine learning and its applications to biology”. In: *PLoS computational biology* 3.6 (2007), e116.

[195] Eva Thelisson, Kirtan Padh, and L Elisa Celis. “Regulatory mechanisms and algorithms towards trust in AI/ML”. In: *Proceedings of the IJCAI 2017 workshop on explainable artificial intelligence (XAI), Melbourne, Australia*. 2017.

[196] Mike Thelwall. “Gender bias in sentiment analysis”. In: *Online Information Review* (2018).

[197] Florian Tramer et al. “Fairtest: Discovering unwarranted associations in data-driven applications”. In: *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE. Paris, France, 2017, 401
bibrangedash 416.

[198] Trung Tran et al. “How Digital Natives Learn and Thrive in the Digital Age: Evidence from an Emerging Economy”. In: *Sustainability* 12.9 (2020). ISSN: 2071-1050. DOI: 10.3390/su12093819. URL: https://www.mdpi.com/2071-1050/12/9/3819.

[199] Sriram Vasudevan and Krishnaram Kenthapadi. “LiFT: A Scalable Framework for Measuring Fairness in ML Applications”. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. CIKM ’20. Virtual Event, Ireland: Association for Computing Machinery, 2020, pp. 2773–

2780. ISBN: 9781450368599. DOI: 10.1145/3340531.3412705. URL: https://doi.org/10.1145/3340531.3412705.

[200] Clarice Wang et al. “Bias: Friend or foe? user acceptance of gender stereotypes in automated career recommendations”. In: *UMBC Student Collection* (2021).

[201] Hao Wang et al. “"Avoiding disparate impact with counterfactual distributions"”. In: *NeurIPS Workshop on Ethical, Social and Governance Issues in AI*. 2018.

[202] Ningxia Wang and Li Chen. “User Bias in Beyond-Accuracy Measurement of Recommendation Algorithms”. In: *Proceedings of the 15th ACM Conference on Recommender Systems*. RecSys ’21. Amsterdam, Netherlands: Association for Computing Machinery, 2021, pp. 133–142. ISBN: 9781450384582. DOI: 10.1145/3460231.3474244. URL: https://doi.org/10.1145/3460231.3474244.

[203] Tianlu Wang et al. “Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 5310–5319.

[204] Tianlu Wang et al. “Double-hard debias: tailoring word embeddings for gender bias mitigation”. In: *arXiv preprint arXiv:2005.00965* (2020).

[205] Zijian Wang et al. “Demographic Inference and Representative Population Estimates from Multilingual Social Media Data”. In: *The World Wide Web Conference*. WWW ’19. San Francisco, CA, USA: Association for Computing Machinery, 2019, pp. 2056–2067. ISBN: 9781450366748. DOI: 10.1145/3308558.3313684. URL: https://doi-org.du.idm.oclc.org/10.1145/3308558.3313684.

[206] Damien Patrick Williams. “Fitting the description: historical and sociotechnical elements of facial recognition and anti-black surveillance”. In: *Journal of Responsible Innovation* 7.sup1 (2020), pp. 74–83.

[207] Terry Winograd. “Shifting viewpoints: Artificial intelligence and human–computer interaction”. In: *Artificial intelligence* 170.18 (2006), pp. 1256–1258.

[208] Wenying Wu et al. “Gender classification and bias mitigation in facial images”. In: *12th acm conference on web science*. 2020, pp. 106–114.

[209] Yonghui Wu et al. “Google’s neural machine translation system: Bridging the gap between human and machine translation”. In: *arXiv preprint arXiv:1609.08144* (2016).

[210] Yongkai Wu, Lu Zhang, and Xintao Wu. “Counterfactual Fairness: Unidentification, Bound and Algorithm.” In: *IJCAI*. 2019, pp. 1438–1444.

[211] Ke Yang et al. “"A nutritional label for rankings”. In: *Proceedings of the 2018 international conference on management of data*. 2018, pp. 1773–1776.

[212] Zekun Yang and Juan Feng. “A causal inference method for reducing gender bias in word embedding relations”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. New York, USA, Apr. 2020, pp. 9434–9441. DOI: 10.1609/aaai.v34i05.6486. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6486.

[213] Muhammad Bilal Zafar et al. “Fairness Constraints: A Flexible Approach for Fair Classification.” In: *J. Mach. Learn. Res.* 20.75 (2019), pp. 1–42.

[214] Daniel Yue Zhang, Ziyi Kou, and Dong Wang. “FairFL: A Fair Federated Learning Approach to Reducing Demographic Bias in Privacy-Sensitive Classification Models”. In: *2020 IEEE International Conference on Big Data (Big Data)*. 2020, pp. 1051–1060. DOI: 10.1109/BigData50022.2020.9378043.

[215] Jieyu Zhao et al. “Learning gender-neutral word embeddings”. In: *arXiv preprint arXiv:1809.01496* (2018).

[216] Jieyu Zhao et al. “Men also like shopping: Reducing gender bias amplification using corpus-level constraints”. In: *arXiv preprint arXiv:1707.09457* (2017).

[217] Alina Zhiltsova, Simon Caton, and Catherine Mulway. “Mitigation of Unintended Biases against Non-Native English Texts in Sentiment Analysis.” In: *AICS*. 2019, pp. 317–328.

[218] Zining Zhu, Jekaterina Novikova, and Frank Rudzicz. “Deconfounding age effects with fair representation learning when assessing dementia”. In: *arXiv preprint arXiv:1807.07217* (2018).

[219] James Zou and Londa Schiebinger. “AI Can Be Sexist and Racist-It’s Time to Make It Fair”. In: *Nature* 559.7714 (2018), pp. 324–326.

A. Appendix

IRB Documents

Letter of Recruitment

Dear Prospective Participant,

We are inviting you to participate in this semi-structured study on music recommender systems. Graduate student Sunny Shrestha is conducting a study, with the guidance from Professor Sanchari Das (Ph.D.), regarding Music Recommender Systems. This is a campus wide email sent only to the current DU students, no personal information, except the email address, on students has been retained.

You are eligible to be in this study if you:

- currently reside in US,
- are 18 years of age or older,
- and be able to attend the study in person at DU campus location.

Participation in this study is voluntary. In this study you will have an opportunity to interact with music recommender systems. The systems will recommend new songs to you based on your listening history. We are interested in your thoughts and opinions on the music recommendation you get from the system. As an appreciation for your participation, two participants will be selected through raffle to get a $25 gift card.

If you choose to participate, please reply to this email stating your intention to participate or send an email at sunny.shrestha@du.edu. Any personally identifiable

information (email address) collected on you will be deidentified and stored in a DU server within a password protected folder that is only accessible to the authorized research conductors.

On the day of the study, please arrive at Ritchie School of Computer Science and Engineering, in your selected time slot. The interview-based study will take 50-60 minutes. Please answer the questions to your comfort level. At the conclusion of the study, two participants will receive the gift card.

Thank you for your consideration.

Sincerely,

Sunny Shrestha

MS in Computer Science

Inclusive Security and Privacy-focused Innovative Research in Information Technology (InSPIRIT) Lab

Email: inspirit.lab@du.edu

Follow-up Recruitment Letter

Dear Prospective Participant,

You are receiving this letter because you expressed your interest in participating in the following study:

Music Recommender Systems − A User Study

Please follow the link here and select the time slot that works for you and provide your email address. On the day of the study, please arrive at ECS 360, in your selected time slot. The interview-based study will take 50-60 minutes. Please answer the questions to your comfort level.

To re-iterate, the primary investigator of the study is Dr. Sanchari Das, Assistant Professor at Ritchie School Engineering and Computer Science, and graduate student

Sunny Shrestha will be conducting the study.

- currently reside in US,
- are 18 years of age or older,
- and be able to attend the study in person at DU campus location.

Participation in this study is voluntary. This is a semi-structured study in which you will interact with a music recommender system and provide your thoughts and opinions on the music recommendation you get from the system. As an appreciation for your participation, two participants will be selected through raffle to get a $25 gift card.

Any personally identifiable information (email address) collected on you will be de-identified and stored in a DU server within a password protected folder that is only accessible to the authorized research conductors.

Thank you for your consideration.

Sincerely,

Sunny Shrestha
MS in Computer Science
Inclusive Security and Privacy-focused Innovative Research in Information Technology (InSPIRIT) Lab
Email: inspirit.lab@du.edu

Verbal Consent Script

Introduction

We are Dr. Sanchari Das an Assistant Professor of Computer Science and Sunny Shrestha a graduate student in the Department of Ritchie School of Engineering and Computer Science at the University of Denver. Thank you for expressing your interest

in participating in this study.

Subjects Rights

Your participation in this research study is completely voluntary. You can withdraw at any time. Choosing not to be in this study or to stop being in this study will not result in any penalty to you or loss of benefit to which you are entitled. Your choice to not be in this study will not negatively affect any rights to which you are otherwise entitled, including your right to any present or future treatment your class standing, or your present or future employment at DU.

Description of the study and study procedures

Dr. Das and I are conducting a research study to understand user interaction with Music Recommender systems. The name of the study is Music Recommender Systems - A User Study. The IRB Project Number is 1921637-1. The principal investigator of the study is Dr. Sanchari Das, Assistant Professor at Ritchie School of Engineering and Computer Science. In this study, you will be presented with two music recommender algorithms you will provide your information (age, gender and listening history) to these algorithms and they will give you a list of recommended music.

If you agree to participate, you will be asked to answer a couple of questions. Once you provide your consent, I will begin an audio-recording of this session. Then you will be asked to interact with the Music Recommender algorithms. Throughout the study I will ask you some questions, regarding your experience with these algorithms and the outputs you have received. Please keep the raffle ticket safe as two people will be winning a gift card through raffle draw. Finally, if you win the raffle draw you will receive one of the two gift cards.

Risks

Your participation in this study does not involve any physical or emotional risk to you beyond that of everyday life.

Benefits

The possible benefits to you from this study is a chance to contribute to our research of trying to understand the interaction between user and music recommender systems.

Alternatives

You may choose to not participate in this research study at any point during this study.

Financial Information

Participation in this study will involve no cost to you. But you do have a chance to win one of the two $25 electronic gift cards. At the conclusion of this session, a raffle ticket will be provided to you. Each raffle ticket has a unique number, we will note this number on your ticket with the email address that you provide to us. Your email address and ticket number will be saved in a password protected folder in DU servers, no one except the Dr. Das (Principal Investigator) and me (student researcher) will have access to this data. Once the study is complete, two tickets will be selected winner via lottery drawing. We will email the email address associated with the winning ticket number the electronic gift cards using the address provided to us. All the email addresses will be deleted once winning participants receive their prize.

Confidentiality

Study records that can identify you will be kept confidential by keeping the data in the DU servers in password protected folders that is only accessible to the study investigators. All your identifying information will be de-identified and personal information and the audio recording of this study will be deleted once the study is complete. The audio recordings will also be stored in DU servers in password protected folders that is only accessible to the study investigators.

The results of the research study may be published, but your personally identifiable information will not be used.

Whom to contact with questions

If you have any questions or problems during your time on this study, you should contact Dr. Sanchari Das at inspirit.lab@du.edu. If you have any questions regarding your rights as a research subject, please contact the the University of Denver's Institutional Review Board (IRB) Office at $(303)871 - 2121$.

Study Design & Questionnaire For User Study Sessions

Study Design:

Study format: In person session within DU Campus location

Study duration: $50 - 60$ minutes Study director: Dr. Sanchari Das

Study procedure:

Participants will be recruited via announcement using mass emails, posted flyers and social media posts. The interested participants will sign up for the study by selecting a scheduled time slots using their preferred email address. The study will be conducted within DU Campus location. Once the participant has arrived at the location, the student researcher will verify the participant had signed up for that time slot. At the arrival of each participant, the student researcher will hand-out raffle tickets to each participant. One half of the ticket is given to participant and the other half will be entered in a raffle to select participants who will be getting a chance to win gift cards. Participants will be made aware of the raffle at the beginning of the study. The student researcher will greet each participant and answer any questions participant might have. The director will then read aloud the consent document to remind participant that their participation is voluntary and there will be no consequences to their views expressed during this meeting time. The student researcher will record the session from here on out, participant will be made aware of this, if there is hesitation on participant's end, handwritten notes will be taken by the

student researcher present at the meeting. The student researcher will continue with the study when the participant provides the verbal consent or the participants who do not provide this consent leave the room. Participant will also be made aware that everything shared during this study is confidential. At the study location, a desktop/computer device will be already prepared for the participant to interact with. The participant will now be allowed to interact with a music recommender application running at this lab computer. The participant will be asked to provide their music preference and listening history to the application and the application in return will give a sorted list of music recommended to the participant. There will be two or more of these applications that participant will interact with. Throughout this participant and music recommender application interaction, student researcher will be asking questions to the participant to understand their perspective of such application and their experience. At the 40 mins mark, the participant will be asked for any final views on the experience and their response will be recorded. Thereafter, the participant will be reminded that they will receive an email on the winning lottery in the email they have provided. This part of the study will complete with the participant leaves.

Study Questionnaire:

Section: Online Music Listening experiences

The questions for this section are meant to warm up the participants to understand their over experience with the online music applications.

1. How often do you listen to music via online applications like iTunes, Spotify, Bandcamp, etc.?
2. Which is your favorite music application and why?
3. What is your favorite genre of music and why?
4. Additional following questions on music recommendation based on participant's

answer to Q2 and Q3.

Section: Music Recommendations Received

The questions for this section try to understand participants familiarity with music recommender systems.

1. How do you find new music in the [favorite music application]?
2. How often do you find new music through music recommendation from these apps?
3. What are your thoughts on the kind of music recommended in these apps?
4. What are your thoughts on how do these apps know your music tastes?
5. Any additional questions based on answers given by participants in previous questions

Section:Current music recommendations

The question for this section is regarding the music application that the participant is interacting with during this study.

1. Do you see any new songs/music recommended from this app?
2. Do you think you will listen to [top 3 recommendation from app]?
3. Do you notice any difference between music recommended from app version A vs. app version B/C?
4. Is the music genre recommended consistent with your preferred genre?
5. Do you think this recommendation will be different if you provide any different information about yourself?

6. Rank the recommendations: 1, 2, 3 or not listen to at all. (Likert scale of the recommendation).

7. Top 5 tracks men and women. (Likert scale) Why did you choose to rank one track over the other? Do you think participants identity like age, gender etc. plays a role in music selection?

Section: Final Thoughts

1. Do you have any final parting thoughts that you would like this research to consider?

Music Recommender User Study- UI page

Model comparison outputs:

UI Page:

Model Type	Distance Option	K Value	Train RMSE	Test RMSE
KNNWithMeans	pearson_baseline	20	1.372586	13.681448
KNNWithMeans	pearson_baseline	40	1.351994	13.730725
KNNWithMeans	pearson_baseline	10	1.225266	13.899492
KNNWithZScore	pearson_baseline	20	1.219249	14.116749
KNNWithMeans	MSD	10	0.527981	14.146706
KNNWithZScore	pearson_baseline	40	1.243186	14.168157
KNNWithZScore	pearson_baseline	10	1.220669	14.174269
KNNWithMeans2	MSD	20	0.557066	14.197135
KNNWithMeans2	MSD	40	0.568812	14.218592
KNNWithMeans2	cosine	20	7.176648	14.299351
KNNWithMeans2	cosine	40	7.203940	14.303731
KNNWithMeans2	cosine	10	7.066077	14.307751
KNNWithZScore	cosine	20	7.096316	14.349520
KNNWithZScore	cosine	40	7.11931	14.40085
KNNWithZScore	cosine	10	6.948528	14.403419
KNNWithZScore	pearson	40	4.069086	14.462074
KNNWithMeans2	pearson	40	4.544860	14.466834
KNNWithZScore	pearson	10	4.013093	14.539822
KNNWithMeans2	pearson	20	4.422556	14.542815
KNNWithZScore	pearson	20	3.968004	14.547523
KNNWithMeans2	pearson	10	4.439397	14.566339
KNNWithZScore	MSD	10	0.671338	14.676978
KNNBasic	MSD	10	0.330020	14.722837
KNNWithZScore	MSD	40	0.696537	14.744869
KNNBasic	MSD	40	0.388819	14.761170
KNNWithZScore	MSD	20	0.715879	14.781041
KNNBasic	MSD	20	0.374570	14.807151
KNNBasic	cosine	10	7.525762	14.965820
KNNBasic	cosine	20	7.656895	14.974374
KNNBasic	cosine	40	7.675246	14.983878
KNNBasic	pearson	10	4.705137	15.263756
KNNBasic	pearson	40	4.736578	15.371474
KNNBasic	pearson	20	4.754848	15.525080

Table A.1.: Full output of Train RMSE values for Different KNN models and distance options

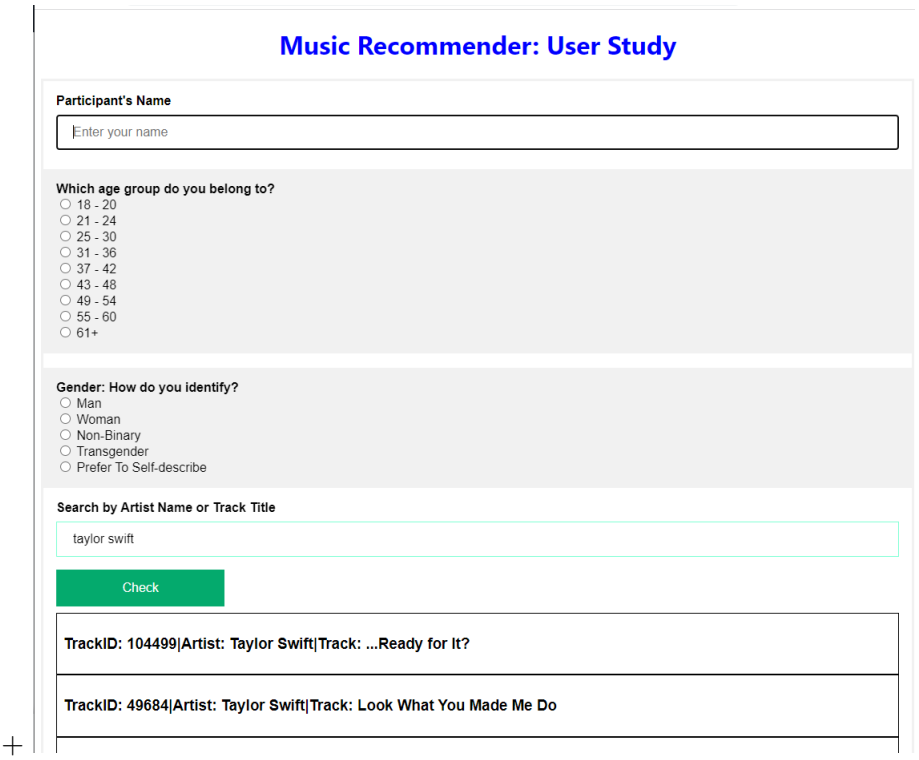

Figure A.1.: User Study: Search Page