Università degli Studi di Milano-Bicocca

Dipartimento di Informatica, Sistemistica e Comunicazione

Corso di Laurea Magistrale in Data Science

Towards Fair Natural Language Processing: A Framework For Unbiased Word Embedding

Relatore: Prof. Antonio Candelieri

Co-relatore: Prof. Elisabetta Fersini

Tesi di Laurea Magistrale di:

Lorenzo Pastore

Matricola 847212

Anno Accademico 2021-2022

Summary

The rapid growth of machine learning (ML) in the last decade has inevitably ignited a debate on fairness in ML. Even though the topic has attracted much attention, there is still no universally agreed definition of the concept of fairness. Furthermore, despite the numerous attempts to address bias in Artificial Intelligence (AI) to achieve fairness, there is no framework of the different techniques to be used to achieve this end. The issue of fairness in ML is close to the field of study of explainable AI but is intertwined in a much more complex way with the nature of the data.

In this thesis, we will explain the concept of bias in AI and review the progress that the community of practice has made so far to identify and address the problem of bias. Then, we will present a structured framework of the different techniques and approaches found in the literature that have been successfully applied to make all these approaches more transparent and more accessible to the said community. Finally we will share the results of the experiment conducted to test selected metrics and debiasing techniques on pre-trained embeddings in order to highlight the relevance of textual data to promote fairness in ML.

The first chapter introduces the problem of fairness and discusses its implications. Then, it explores the fundamental concepts of discrimination, bias, and fairness, employing a desktop review of the most relevant literature on the subject to provide a more consistent and holistic view of the problem. The second chapter revises the state of the art of bias mitigation techniques and identifies a framework of methods and good practices to follow. Then, the third chapter focused on mitigating bias in textual data and presents an experiment on bias measurement and mitigation on pre-trained embeddings. In the fourth and last chapter, we draw our conclusions and indicate areas where further research might be needed.

Contents

List of Figures

List of Tables

Chapter 1

Fairness in Machine Learning

Machine learning algorithms substantially affect everyday life in areas such as education, employment, advertising, and policing. While machine learning (ML) algorithms may seem objective, the tendency to favour bias is embedded in ML essence. The widespread use of ML in a variety of sensitive fields fosters the idea that ML-based decisions are based only on facts and are not affected by human cognitive biases, discriminatory tendencies, or emotions. As matter of fact, these systems learn from data which are, directly or indirectly, shaped by human biases. There is overwhelming evidence indicating that algorithms can inherit or even perpetuate human biases in their decision-making when their underlying data contains biased human decisions [1]. In some areas, especially those having social implications, such as criminal justice, social welfare policy, hiring, and personal finance, it is important to ensure that automated decisions respect fairness principles given that datasets contain sensitive attributes (i.e., race, gender, age, disability status) and/or characteristics closely intertwined with such attributes. This means that ignoring fairness considerations may have socially unacceptable consequences. Particularly worrisome in the context of automated sequential decision making is the potential of "perpetuating injustice, i.e., when maximizing utility maintains, reinforces, or even introduces unfair dependence between sensitive features, decisions, and outcomes" [2].

As a result, the impact of bias in machine learning is growing at such a pace that makes it difficult to even realize the actual changes. The daily impact of AI systems is enormous and is not limited to companies and/or national systems but can be literally felt in our

hands every time we gather information on a browser.

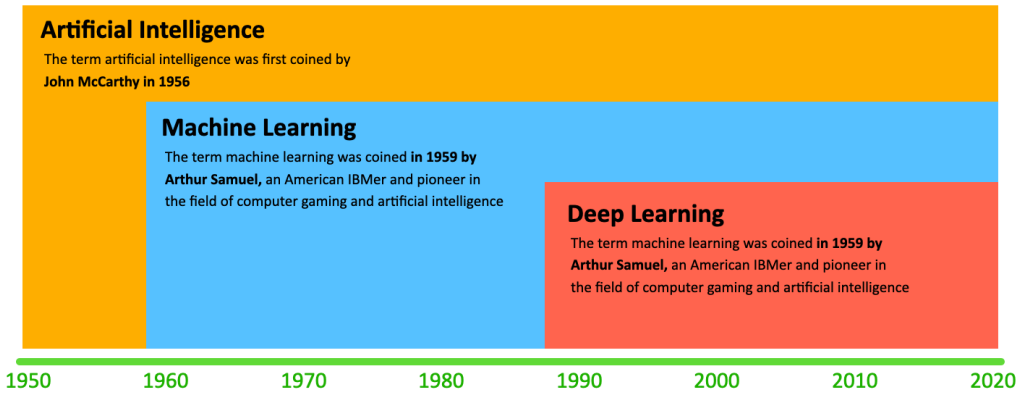

Figure 1.1: Timeline view of evolution of discipline [3]

Since ML progresses rapidly and learned predictors and risk scores increasingly support or even replace human judgment, its societal impact has come under scrutiny. In this context, there is the necessity and opportunity to identify harmful discrimination and design algorithms that avoid it. Over the last few years, researchers have introduced a rich set of definitions formalizing different fairness definitions that can be used for evaluating and designing ML systems. However, due to the complexity of the concept, researchers have found it challenging to agree on a single measure or even definition of fairness.

1.1 Evidence of the problem

In the past years, alongside the growth of artificial intelligence (AI), a few signals of the problematic of fairness have surfaced, slowly raising the scientific community and the public opinion's concern.

One of the first pieces of evidence pointing in this direction can be found in a 2004 study that found out that voice-dictation performed significantly better on men's voice [4]. Notwithstanding, for almost ten years, the rapid growth of technology obscured the topic of fairness. In 2013 an article by Latanya Sweeney pointed out the problem explicitly, speaking for the first time of "Discrimination in Online Ad Delivery" and converging the attention on the importance of this topic [5]. A few years later, another analysis developed

on COMPAS, a software used by the US state of New York to assess potential recidivism risk in criminals, revealed that the algorithm was biased against black people [6]. These, and many recent studies, show disparities in gender and race, as well as the way in which they affect society deeply and unconsciously [5], [7]. The emerge of this type of evidence has led to a call for fairness-aware ML. However, according to Tolan Songul, fairness as a complex value-driven concept is hard to formalize for ADM systems. Statistical formalization of fairness leads to a long list of criteria that can be useful in one context but can be flawed (or even harmful) in other contexts. Furthermore, the presence of bias, and tradeoffs in these criteria, makes it impossible to unify them in one general framework [8].

Although the formalization of fairness could be complex, we believe that the importance of the pursuit of fairness in ML is much more relevant and pressing. In this work, we try to go around the complexity of the problem and the widespread of misbelief that machine learning is capable of giving the best result without human intervention. This belief could be hazardous, especially among the data science community, which has access to most of the critical points of bias spreading.

1.2 Fundamental concept: Discrimination, Bias, and Fairness

While the definition of fairness seems simple on a lexical level, in the ML context and more generally in the algorithmic world, it has become very complex to find an agreed working definition. The different definitions of fairness are often incompatible with each other and with the realities of ML optimization. For example, defining *fairness* as "equality of outcomes" may simply refer to a system producing the same result for everyone, while fairness defined as "equality of treatment" might need to consider differences between individuals explicitly [9].

Although the statistical formalization of fairness is a very complex task, in the context of this thesis, it is not of our interest to identify a unified definition of fairness. Instead, we will simply refer to fairness as the ensemble of actions taken to ensure that the results are free of algorithmic bias and focus our attention on identifying and mitigating various

types of algorithmic bias and the right tools to do so.

While fairness in ML is a relatively young field of study, algorithmic bias has a much more ancient and developed history. Algorithmic bias can be described as "the systematic and repeatable errors in a computer system that create unfair outcomes, such as privileging one arbitrary group of users" [10]. In light of this definition is not hard to understand that, although fairness' definition may change depending on the context, it is strictly related to bias. However, a fundamental difference exists between these concepts: fairness is a normative concept while bias is a technical concept; in fact, the literature on fair ML algorithms mainly derives its fairness concepts from a legal context.

This section introduces the fundamental concepts of discrimination and bias that should help the reader develop a deeper understanding of fairness. Following that, some of the most relevant definitions of fairness are presented.

1.2.1 Types of discrimination

The fight against bias and discrimination has a long history in philosophy and psychology, and recently in ML. In this body of work, defining the concept of fairness has taken center stage. Philosophy and psychology had tried to define the concept of fairness long before computer science started exploring it [11]. However, no matter what definition we adopt, fairness can be incredibly difficult to achieve in practice. In this section, we will identify different kinds of discrimination to better understand the complexity of fairness.

- **Direct Discrimination:** Occurs "when individuals receive less favorable treatment explicitly based on protected attributes" [12]. An example would be rejecting a qualified female applicant to a university because of her gender [12]. Typically, some traits identified by law make discrimination illegal; usually, these traits are considered "protected", or "sensitive" attributes in computer science literature.

- **Indirect Discrimination:** The treatment is based on apparently neutral non-protected attributes but results in creating unjustified distinctions among individuals from the protected group. For example, the residential zip code of a person can be used in decision-making around issues such as loan applications. Albeit apparently neutral, this can lead to racial discrimination. Even though a zip code appears to

be a non-sensitive attribute, it may correlate with race because of the predominant racial group living in residential areas. [12].

- **Systemic Discrimination:** According to the Council of Europe [13] "systemic discrimination involves the procedures, routines and organisational culture of any organisation that, often without intent, contribute to less favourable outcomes for minority groups than for the majority of the population, from the organisation's policies, programmes, employment, and services."

- **Statistical Discrimination:** Occurs "when decision-makers use average group statistics to assess an individual belonging to that group" [11]. It usually happens when the decision-makers use an individual's prominent, recognizable characteristics as a proxy for either hidden or "more difficult to determine" characteristics that may be relevant to the outcome [11], [14].

- **Explainable Discrimination:** In some cases, differences in treatment and outcomes between groups can be justified and explained. It is not considered illegal discrimination in these situations and thus called "explainable". In [15], the authors introduce a methodology to quantify the explainable and illegal discrimination in data, highlighting that methods that do not take the explainable part of the discrimination into account may result in non-desirable outcomes. As a result, these methods often introduce reverse discrimination, which is equally harmful and undesirable. Faysal Kamiran and Indre Zliobait further explain how to quantify and measure discrimination in data or classifier's decisions which directly consider illegal and explainable discrimination [15].

- **Unexplainable Discrimination:** In contrast to explainable discrimination, unexplainable discrimination occurs when the discrimination toward a group is unjustified and therefore considered illegal. In their work, Faysal Kamiran and Indre Zliobait in [15] also present local techniques for removing only the illegal or unexplainable discrimination, allowing only for explainable differences in decisions. These are preprocessing techniques that change the training data to ensure that it does not contains any unexplainable discrimination.

The definition of different types of discrimination is beneficial when approaching unknown data because exploring and understanding data on a deeper level is essential to find the most suitable approach to mitigate the bias. Nonetheless some definitions of bias are needed to quantify fairness according to the selected definition of discrimination and fairness.

1.2.2 Types of bias

Bias can be identified in multiple shapes and forms, which can lead to unfairness in different downstream learning tasks [11]. In [16], Harini Suresh and John V. Gutta talk about sources of bias in ML with their categorizations and descriptions, to motivate future solutions to each of the sources of bias introduced in the paper. In [17], the authors prepare a complete list of different types of bias with their corresponding definitions at different stages, from data origins to its collection and processing. In [11], the authors select some of the most common and important sources of bias introduced in these two papers and add work from other existing research papers trying to seize and explain the different facets of bias. Based on the above, the following list provides an overview of the different types of biases we were able to identify from the cited papers:

- **Historical Bias:** This is *"the already existing bias and socio-technical issues in the world and can seep into the data generation process even given a perfect sampling and feature"* [11]. As an illustrative example, Mehrabi et al. explain that a 2018 image search for "women Chief Executive Officers (CEOs)" ultimately resulted in fewer female CEO images found since only 5% of Fortune 500 CEOs were women, which caused the results to be biased towards male CEOs [16]. Even if the search results reflect the reality, it is sometimes essential to consider whether or not the algorithm should reflect this reality or would cause harms to subgroups of people. In [16], the authors suggest that this type of bias should involve the evaluation of eventual representational harm possibly inflicted to certain groups.

- **Representation Bias:** Happens as a *"result of the way in which we define and sample from a population during data collection"* [11]. The lack of geographical diversity in datasets like ImageNet, which creates bias towards Western countries, is an

example of this type of bias [11].

- **Measurement Bias:** Occurs as a *"result of the way -we choose, utilize, and measure- a particular feature"* [11]. For example, in the recidivism risk prediction tool COMPAS, arrests and friend/family arrests records were used as proxy variables to assess the level of "riskiness" or "crime". Although minority communities are controlled and policed more frequently and have higher arrest rates, this does not mean that these groups are more dangerous especially when considering the differential treatment that these groups are subject to [11].

- **Evaluation Bias:** *"Happens during model evaluation"*, for instance, due to the use of inappropriate and disproportionate benchmarks for the assessment of applications like Adience and IJB-A benchmarks. For example, these benchmarks were used in the evaluation of facial recognition systems that were biased toward skin color and gender [16], [18].

- **Population Bias:** Occurs *"when statistics, demographics, representatives and user characteristics in the user population represented in the dataset/platform are different from the original target population"* [11], [17]. This type of bias can be found on different social platforms in user demographics, for example with women more likely to use Pinterest, Facebook, Instagram, while men more active in platforms like Reddit or Twitter. [11].

- **Simpson's Paradox:** *Is the manifestation of aggregation bias which can arise when the analysis involves heterogenous data.* "According to Simpson's paradox, a trend, association, or characteristic observed in underlying subgroups may be pretty different from association or characteristic observed when these subgroups are aggregated." [11]. The paradox can bias the analysis of heterogeneous data composed of subgroups or individuals with different behaviors [19]. Simpson's paradox has been noted in a variety of domains, including biology [20], psychology [21], astronomy [22], and computational social science [23].

- **Sampling Bias:** Arises as a *"result of non-random sampling of subgroups"*. Due to sampling bias, the trends estimated for one population may not generalize to data

collected from a new population.

- **Behavioral Bias:** Results *"from different user behaviors across platforms, contexts, or different datasets"* [17]. In [24] for example, the authors demonstrate how different emoji representations among platforms can lead to people displaying different reactions and behaviors, which in turn can cause communication errors.
- **Content Production Bias:** Arises *"from structural, lexical, semantic, and syntactic differences in the contents generated by users"* [17]. An example of this type of bias is discussed in [25] in relation to differences in the use of language across different gender and age groups. Differences in the use of language can also be seen across and within countries and populations.
- **Linking Bias:** Occurs *"when network attributes obtained from user connections, activities, or interactions differ and misrepresent the users' actual behavior"* [17].
- **Temporal Bias:** Results *"from differences in populations and behaviors over time"*.
- **Popularity Bias:** Is the outcome of *"increased exposure that popular items tend to get regardless of manipulations that popularity metrics could be subject to for example, by way of fake reviews or social bots"* [11]. For instance, this type of bias can be seen in search engines [26], [27] or recommendation systems where popular objects are presented more to the public, not necessarily as a sign of better quality but owing to other biases [11].
- **Algorithmic Bias:** Happens *"when the bias is not present in the input data and is added purely by the algorithm Bias"* [10].
- **User Interaction Bias:** *"This type of bias can be observed not only on the web but can also be triggered from the user interface and through the user him/herself who may impose his/her self-selected biased behavior and interaction"* [10], [11]. It can intersect with and be amplified by other types and sub types of biases, such as presentation and ranking bias.
  - **Presentation Bias:** *"Results from how information is presented"*. This determines the likelihood of certain content to be clicked or not depending on how it

is presented and whether it is visible to the user or not [11].

- **Ranking Bias:** *"Is triggered by how top-ranked results are considered the most relevant and important and likely to attract more clicks than others"*. This type of bias is clearly observable in search engines and crowdsourcing applications [10], [11].

- **Social Bias:** Occurs *"when our judgement is influenced by other people's behaviour or content"* [10]. A typical example would be someone deciding to upgrade an item based on other people's high ratings whilst her/his initial scoring would have been much lower. [11]

- **Self-Selection Bias:** *"Self-selection bias is a subtype of the selection or sampling bias in which subjects of the research select themselves"* [11]. This type of bias occurs, for example, when survey takers decide that they can appropriately participate in a study themselves.

- **Omitted Variable Bias:** Takes place *"when one or more important variables are left out of the model"* [11]. For example, if someone designs a model to accurately predict the annual percentage rate at which customers will stop subscribing to a service, which, however, was not conceived to take into account the appearance of a new strong competitor on the market offering the same service at a much lower price. The users would be cancelling their subscriptions without receiving any warning from the designed model. In this example, the competitor's appearance would be considered an omitted variable [11].

- **Cause-Effect Bias:** Can happen as a *"result of the fallacy that correlation implies causation"*. An example of this could be the case of a data analyst in a company who aims to analyze the success of a new loyalty program. It appears that customers who signed up for the loyalty program are more likely to spend money in the e-commerce store than those who did not. It would be problematic if the analyst concluded that the loyalty program is successful as it might well be the case that it attracted only the committed/loyal customers who might have planned to spend more money anyway to begin with. This type of bias can have severe consequences due to its nature and

role in sensitive decision-making policies [11].

- **Observer Bias:** Occurs *"when researchers' expectations are subconsciously projected onto the research"* [28]. This type of bias can occur when researchers (unintentionally) influence participants (during interviews and surveys) or when they cherry-pick participants or statistics that will favour their research.

- **Funding Bias:** Arises *"when biased results are used purposively to support or satisfy the funding agency or financial supporter of the research study"* [28]. For example, this can surface when employees of a company report biased results in their data and statistics to satisfy funding agencies or other parties.

The importance of identifying different types of bias is linked the need to find the most suitable solution to the case problem. Another relevant element to consider in ML is the feedback loop phenomenon, which occurs when the trained ML model decisions lead to outcomes, which affect future data to be collected for subsequent training rounds or models [29]. This phenomenon affects not only the data and the algorithm, but also algorithms and user interaction. In this context, we should try to group and place different types of bias inside this cycle between data, algorithms, and user interaction, keeping in mind that these phenomena are intertwined, and we should consider how they affect each other in this cycle (1.2) to try to address them correctly.

1.2.3 Definitions of Fairness

The issue of the formal definition of fairness has been studied by many different authors [30]–[32], along with the evolution of the concept of fairness in ML and the general public's perception of these fairness definitions. While these are interesting and widely discussed topics, they are not the main focus of this thesis as previously highlighted. [33], [34] Instead, in this thesis, we are going to provide some of the most widely used definitions of fairness in ML and AI more generally, inspired from "Fairness Definitions Explained" by Sahil Verma and Julia Rubin [35], so to provide an essential but comprehensive overview of the multiple definitions of fairness.

- **Demographic or Statistical Parity [36]:** A predictor $\hat{Y}$ satisfies demographic

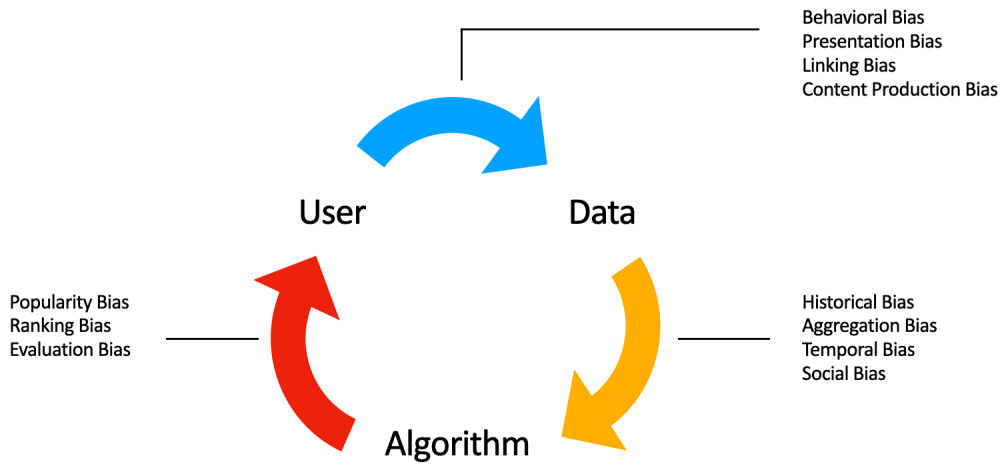

Figure 1.2: Feedback loop cycle and bias schema
Author's elaboration

parity if $P(\hat{Y} | A = 0) = P(\hat{Y}|A = 1)$. The definition is satisfied if subjects in both protected and unprotected groups have an equal probability of being assigned to the positive predicted class.

- **Conditional Statistical Parity [37]:** The definition is satisfied if subjects in both protected and unprotected groups have an equivalent probability of being assigned to the positive predicted class, given a set of legitimate factors L.

- **Equalized Odds [38]:** A predictor $\hat{Y}$ satisfies equalized odds with respect to protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on Y.

- **Equal Opportunity [37], [38]:** A binary predictor $\hat{Y}$ satisfies equal opportunity with respect to A and Y if $Pr\{\hat{Y} =1|A=0,Y =1\}=Pr\{\hat{Y} =1|A=1,Y =1\}$. The probability of assigning a person in a positive class to a positive outcome must be the same for both protected and unprotected (female and male) group members; hence both protected and unprotected groups should have equal true positive rates. Equal opportunity is a weaker, yet still interesting, notion of non-discrimination and can thus allow for better utility.

- **Treatment Equality [35]:** This definition considers the ratio of errors that the

classifier makes instead of its accuracy. "A classifier satisfies this definition if both protected and unprotected groups have an equal ratio of false negatives and false positives." [35].

- **Test Fairness [35], [39]:** "A classifier satisfies this definition if, for any predicted probability score S, subjects in protected and unprotected groups have an equal probability of belonging to the positive class truly."
- **Fairness through awareness [35], [36]:** According to this definition "Fairness is captured by the principle that similar individuals should have similar classification. The similarity of individuals is defined via a distance metric; for fairness to hold, the distance between the distributions of outputs for individuals should be at most the distance between the individuals." It is important to note that the distance metric is of fundamental importance when applying this definition and should be chosen carefully.
- **Fairness through unawareness [40]:** "An algorithm is fair so long as any protected attributes A are not explicitly used in the decision-making process." [40].
- **Counterfactual Fairness [11], [35], [40]:** A causal graph is counter-factually fair if the predicted outcome d in the graph does not depend on a descendant of the protected attribute G. "The counterfactual fairness definition is based on the assumption that a decision is fair towards an individual if it is so in both the actual world and a counterfactual world where the individual belonged to a different demographic group." [11].
- **Fairness in Relational Domains [41]:** This proposes a notion of fairness that captures the relational structure in a domain by taking attributes of individuals into consideration alongside the social, organizational, and other connections between individuals.

All these definitions can be categorized in *individual fairness*, which gives similar predictions to similar individuals, or *group fairness*, which treats different groups equally. The most prominent definitions, together with the papers that introduce them and their relative category, are shown in Tab:1.1.

Name	Reference	Category
Demographic parity	[36]	Group
Conditional statistical parity	[37]	Group
Equalized odds	[38]	Group
Equal opportunity	[38]	Group
Treatment equality	[35], [42]	Group
Test fairness	[35], [39]	Group
Fairness through awareness	[35], [36]	Individual
Fairness through unawareness	[40]	Individual
Counterfactual Fairness	[35], [40]	Individual
Fairness in Relational Domains	[41]	-

Table 1.1: Considered fairness definitions with references

The definition and categorization of the concepts of discrimination, bias, and fairness are critical to address the problem of biased data. Some equally crucial elements must also not be overlooked.

According to [43], it is possible to satisfy some of the fairness constraints at once only in highly constrained cases. Therefore, it is fundamental to consider the context and application in which fairness definitions will be used and use them accordingly [44]. Another aspect to consider is these definitions' impact on individuals or groups over time, in fact some authors show that current fairness definitions are not always helpful, do not promote improvement for sensitive groups, and can be harmful when analyzed over time [45].

Chapter 2

Bias Mitigation

This chapter focuses on bias mitigation techniques. The goal is to gather information about the conditions under which the techniques are used and to provide an organization that can be useful to the reader. It is important to emphasize that mitigating the effect of bias in AI is not an easy task; at the same time, if we consider the importance of the impact that bias can have, the difficulty of the task becomes marginal.

Data carries with it information on the socio-cultural background of those who generated it and is often affected by collective behaviours based on stereotypes. This phenomenon has always characterized the evolution of our societies and not always in a negative way. However, the recent growth of technologies has enormously fuelled realities in such ways and to the extent that the feedback loop phenomenon and the growing amount of information processed make discrimination self-feeding.

To better understand the bias mitigation task, it is essential to provide a clear overview of different cases that can be encountered and the possible approaches to mitigate each of them. In this section, following the lead of [11] we will enumerate some of the numerous attempts to address bias in different domains and sub-domains of AI.

In addition, we will also categorize these cases by type of data (visual, textual, tabular) and by the intervention phase (pre, in, and post). We believe that such categorization could practically help data scientists to find a suitable solution to a problem or get closer to identifying it.

2.1 Good practice for fair AI

Forms of discrimination act subtly in our society and hide within data in the same way. For this reason, the solution must necessarily start from re-education vis-a-vis the problem that we are facing, which also involves awareness of it in the first place. Following that, openness to explore uncharted territories and willingness to stay with the resulting discomfort will be necessary. As stated in [43], on developing new technical solutions to reach fairness, "Much of this work will require technical researchers to learn new skills or partner with social scientists, but no less a transformation is required. We must also become more comfortable with difficult or unresolvable tensions such as that between the usefulness and dangers of abstraction".

Qualitative tools help shed light on the nuances of fairness and trigger important discussions and reflections. Benefits include among others the analysis of the societal role of the AI system, the assessment of potential fairness-related harms and trade-offs and the identification of the causes of bias, alongside mitigation strategies. Furthermore, qualitative tools and resulting reflections can also help track, monitor and address fairness-related harms that might come into play [46], [47]. In this section, we present some qualitative tools for fair AI.

2.1.1 Fair ML traps

Our approach to re-education starts from the definition of some good practices to follow. For example, in [44], the authors have identified five traps to avoid when designing a new fair-ML solution. In practice, these guidelines would help assess if a technical solution:

- is appropriate to the situation. This would imply that the social contexts and its politics, including power hierarchies, have been analyzed and taken into account. "**Solutionism trap** is the failure to recognize the possibility that the best solution to a problem may not involve technology";
- affects the predictability of a social context to the extent that the technology does not effectively solve the problem it is meant to address and the situation remains unchanged. "**Ripple Effect trap** is the failure to understand how the insertion of

technology into an existing social system changes the behaviors and embedded values of the pre-existing system";

- can appropriately manage solid understandings of social requirements such as fairness, including the need for procedurality, contextuality, and contestability. "**Formalism trap** is the failure to account for the whole meaning of social concepts, such as fairness, which can be procedural, contextual, and contestable, and cannot be solved through mathematical formalisms";

- has fairly and convincingly modelled the social and technical requirements of the context in which it will be deployed. "**Portability trap** is the failure to understand how re-purposing algorithmic solutions designed for one social context may be misleading, inaccurate, or otherwise do harm when applied to a different context";

- has been conceived to include all the data and social actors relevant to the contextual question of fairness. "**Framing trap** is the failure to model the entire system over which a social criterion, such as fairness, will be enforced".

2.1.2 Fairness Analytic tool

The tool developed by Mulligan et al. aims to trigger reflections about fairness during the early stages of a project. Teams are probed to think about concepts of fairness in a multi-disciplinary way and how the significance of fairness could change for a particular AI system. The Fairness Analytic tool allows researchers to have interdisciplinary collaborations on the theme of fairness by putting the disciplinary specific definitions into conversation and thus facilitating the creation of a common research agenda [46], [48].

2.1.3 Co-designed AI fairness checklist

Microsoft and academic researchers engaged 48 individuals from 12 technology companies to co-design an AI fairness checklist [47]. The checklist, to be customized, includes items to consider at different stages of the development and deployment cycle (i.e., envision, define, prototype, build, launch, and evolve) of an AI system. The complete checklist can be found here [49].

2.2 Literature review

The literature on designing fair algorithms is extensive and interdisciplinary. Based on a desktop review of relevant literature, this section identifies and classifies the most notable solutions suggested for mitigating bias in different context. The goal of this chapter is to provide an extensive and comprehensive review for students and researchers who wants to approach different ML task from a fairness point-of-view.

From a procedural viewpoint, methods for imposing fairness can roughly be divided into three families. Methods in the first family consist of preprocessing operations or extracting representations from the data to remove undesired biases, which can then be used as input to a standard ML model. Methods in the second family involve enforcing a model to produce fair outputs through imposing fairness constraints into the learning mechanism. Finally, methods in the third family comprise post-processing the outputs of a model to make them fair.

Generally, methods that target bias fall into three categories which we will present together with a list of pros and cons for each method, extracted and summarized by Ziyuan Zhong [50]:

- **Pre-processing:** This method tries to transform the data so that the underlying discrimination is removed [51]. The basic idea is to enforce the learning of a new representation so that the information associated to the sensitive attribute is removed while retaining the information of X as much as possible [52].

Pros	Cons
Preprocessed data can be used for any downstream task.	Can only be used for optimizing Statistical Parity or Individual Fairness because, does not have the information of label Y.
No need to modify the classifier.	Inferior to the other two methods in terms of performance on accuracy and fairness measure.

- **In-processing:** The idea is to add a constraint or a regularization term to the existing optimization objective [52]. A large body of work in the research literature falls in this category because these methods can be used to optimize any fairness definition.

In-processing techniques aim to change state-of-the-art learning algorithms to remove discrimination during the model training process [51].

Pros	Cons
Good performance on accuracy and fairness measures.	Method in this category is task-specific.
Higher flexibility in trade-off between accuracy and fairness measures (depends on specific the algorithm).	Need to modify classifier, which may not be possible in many scenarios.

- **Post-processing:** This method is normally performed after training by accessing a holdout set that was not involved during the training of the model [51]. It can optimize most fairness definitions (except counterfactual fairness). When the algorithm treats the learned model as a black box without the ability to modify the training data or learning algorithm, then post-processing can be used to reassign the original labels assigned by the black-box model based on a function during the post-processing phase [50], [52].

Pros	Cons
Can be applied after any classifiers.	Requires test-time access to the protected attribute.
Relatively good performance, especially fairness measures.	Lacks the flexibility of picking any accuracy–fairness tradeoff.
No need to modify the classifier.	

While facilitating model design, not imposing constraints on the full shapes of relevant distributions can be restrictive and problematic. Also, most often, the goal of these methods is to create a fair model from scratch on a specific task. However, in many real-world applications using the same model or part of it over different tasks might be desirable. It is necessary to consider the learning problem in a multitask/lifelong learning framework to ensure that fairness properties generalize to multiple tasks [53].

Since in this context it can be beneficial to take a cross-domain view, this table (2.1) has been created with the purpose to give the reader a broader view of the many possible approaches to achieve fairness, especially in classification. Although the best approach could vary based on the context, data type, and goal, we think that the given organization of these articles might be insightful for the user to have a broader and more conscious approach to the problem. It is important to notice that the categorization in pre, in, and post-processing can be helpful, but it is not exhaustive of all the possible approaches.

Phase	**Data type**	**Task**	**Technique**	**Refs**
Pre	Tabular	Classification	Adversarial training	[54]
Pre	Tabular	Classification	Independency constraint	[55]
Pre	Tabular	Classification	Fair decision tree	[56]
Pre	Tabular	Classification	k-NN variation	[57]
Pre	Tabular	Classification	Bayesian networks	[58]
Pre	Tabular	Classification	Causal graph/network	[12], [59]
Pre	Tabular	Classification	Random Repair	[60]
Pre	Tabular	Classification	Combinatorial and Geometrical repair	[61], [62]
Pre	Tabular	Classification	Shifted decision boundary	[63], [64]
Pre	Tabular	Classification	Direct/Indirect Rule Protection	[65]
Pre	Tabular	Classification	Privacy and discrimination protected patterns	[66]
Pre	Tabular	Classification	Univariate transformations	[67]
Pre	Tabular	Classification	Non-discriminatory constraints	[68]
Pre	Tabular	Clustering	Fairlet decompositions	[69]

Pre	Tabular	Classification	Adversarial fair representations	[70]–[73]
Pre	Tabular	Classification	Local conditional discrimination	[74]
Pre	Tabular	Classification	Communities from Lowly-connected Attributed Nodes	[75]
Pre	Visual	Classification	Adversarial fair representations	[76], [77]
Pre	Visual	Classification	VAE and Adversarial Censoring	[78]
Pre	Visual	Classification	Variational fair autoencoder	[79]
Pre	Textual	Word embeddings	Differential Bias for GloVe	[80]
In	Tabular	Regr and Class	Lagrangian multipliers methods	[81]–[87]
In	Tabular	Classification	Adversary weights methods	[88]
In	Tabular	Classification	Penalties to objective methods	[89]–[97]
In	Multi	Reinforcement learning	Penalties to objective methods	[98]
Post	Tabular	Classification	Obscuring features	[99]
Post	Tabular	Classification	Updating classifiers loss-adversively	[100]
Post	Tabular	Classification	Group-dependent threshold	[101]
Post	Tabular	Classification	Derived predictor	[38]

Post	Tabular	Classification	Selecting near-optimal metric multifair predictions	[102]
Post	Tabular	Classification	Counterfactual fairness procedure	[40]
Post	Tabular	Classification	Active feature acquisition	[103]
Post	Tabular	Classification	Active feature acquisition	[103]
Post	Tabular	Classification	Fair forest	[104]
Post	Tabular	Classification	LogLinear model coefficients	[105]
Post	Textual	Word embedding	Neutralize and equalize	[106]
Post	Visual	Classification	Multiaccuracy boost algorithm for auditing	[107]
Multi	Tabular	Classification	Multi-Max Mistreatment Pareto optimal solution	[108]

Table 2.1: Categorization of fairness mitigation approaches with references

2.2.1 Available tools from the web

Various AI fairness tools exist that can help data scientists analyze and mitigate discrimination and bias in ML models. It is important to underscore that these should be used in combination with qualitative tools, as highlighted previously. Technical solutions are critical but not sufficient to understand the societal context of biases and therefore to mitigate them. Furthermore, they risk perpetuating the "misleading notion that ML systems can achieve "fairness" or be "unbiased". Technical tools include:

- **IBM's AI Fairness 360 Toolkit:** a Python toolkit focusing on technical solutions through fairness metrics and algorithms to help users examine, report, and mitigate discrimination and bias in ML models.
- **Google's What-If Tool:** a tool to explore a models' performance on a dataset, including examining several preset definitions of fairness constraints (i.e., equality of opportunity). This tool is interesting as it allows users to explore different definitions of fairness.
- **Microsoft's fairlean.py:** a Python package that implements a variety of algorithms that seek to mitigate "unfairness" in supervised machine learning.
- **Facebook** is developing a "Fairness Flow" internal tool to identify bias in ML models.

Whether the focus is on data or the broader AI system lifecycle, these tools tend to use a technical lens and center on technical solutions. A purely technical solution tool would not have captured the COMPAS algorithm's discrimination nuances.

Chapter 3

Natural Language Processing

As ML tools become increasingly popular, it becomes vital to recognize their role in shaping societal biases and stereotypes. Unintended bias in ML can manifest as systemic differences in performance for different demographic groups, potentially compounding existing challenges to fairness in society at large. Natural Language Processing (NLP) models have proven successful in modeling various applications; however, they can amplify bias found in text corpora and social norms.

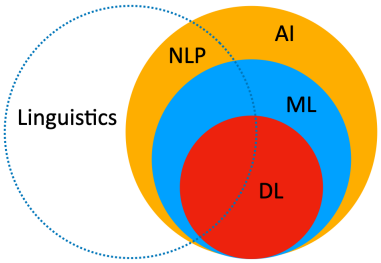

Figure 3.1: Relationship between AI, ML, DL and NLP
Author's elaboration

In this chapter, we will focus expressively on gender bias in NLP and introduce some methods to mitigate this type of bias.

3.1 Introduction to gender bias in NLP

Gender bias can be defined as the preference or prejudice toward one gender over the other. While deeply rooted in social and cultural norms, these biases are a reflection of human and societal biases and stereotypes that algorithms learn from training data.

ML systems are increasingly impacting millions of people every day, and the examples shared in this thesis are just the tip of the iceberg; there are countless back-end systems below the surface often applying off the shelf ML as a service system that can propagate bias in ways that do not have a customer front end.

Corpora of human language are regularly fed into ML systems to give them critical insights and information about the world. NLP plays an important role in many powerful applications, including speech recognition, text translation, and autocomplete, and defines many prominent automated decision systems making crucial recommendations about our lives and our future world [109].

While NLP systems can carry gender bias in many of their components, leading to biased or even completely wrong predictions. The extent of the impact of gender bias in NLP is still to be fully appreciated, including how the trickle down effects of gender bias find their way into downstream applications, impacting real people, particularly women, as suggested in [110]. The spread of gender bias in NLP algorithms poses the threat of reinforcing damaging stereotypes, with real-life consequences. For example, automatic resume filtering systems could give preference to male when the only discriminating factor is the applicant's gender [11], [55]. This is often the consequence of overlooking the implications of and considering bias as a purely technical issue, missing out on its socio-political and power dimensions which fundamentally shape social and gender inequalities. An example of NLP model-derived bias are browser auto-completion systems, which tend to change output depending on user information and localization, potentially contributing to reinforcing misbeliefs and stereotypes. Furthermore, since nowadays children start using these systems at an early age, this phenomenon is getting potentially even more harmful.

In Table 3.1 we use use examples found in the literature to categorize [110] representation bias in NLP tasks into the following four groups: (D)enigration, (S)tereotyping, (R)ecognition, (U)nder-representation.

Task	Example of gender representation bias	D	S	R	U
Machine Translation	Translating "He is a nurse. She is a doctor." to Hungarian and back to English results in "She is a nurse. He is a doctor." (Douglas, 2017)		✓	✓	
Caption Generation	An image captioning model incorrectly predicts the agent to be male because there is a computer nearby (Burns et al., 2018)		✓	✓	
Speech Recognition	Automatic speech detection works better with male voices than female voices (Tatman, 2017)			✓	✓
Sentiment Analysis	Sentiment Analysis Systems rank sentences containing female noun phrases to be indicative of anger more often than sentences containing male noun phrases (Park et al., 2018)		✓		
Language Model	"He is doctor" has a higher conditional likelihood than "She is doctor" (Lu et al., 2018)		✓	✓	✓
Word Embedding	Analogies such as "man : woman :: computer programmer : homemaker" are automatically generated by models trained on biased word embeddings (Bolukbasi et al., 2016)	✓	✓	✓	✓

Table 3.1: Categorization in four groups of gender representation bias in NLP tasks [110]

3.2 Bias in NLP

A recent body of work on gender bias in NLP has focused the attention on quantifying bias through various approaches, spanning from psychological tests and performance differences for various tasks, to the geometry of vector spaces [110]. This subsection provides an overview of some of these evaluation methods (metrics) and discusses different definitions used for bias detection in NLP methods.

As pointed out in the previous chapter, quantifying the amount of bias is essential to define which type of bias we are trying to identify and quantify. Alongside the different types of bias described in the first chapter, here we list some forms of biases specific to NLP application. In 2019 Hitti et al. defined "gender bias in a text as the use of words or syntactic constructs that connote or imply an inclination or prejudice against one gender", highlighting that gender bias can evidence itself structurally, contextually, or in both forms. **Structural bias** occurs when the construction of sentences shows patterns closely tied to the presence of gender bias. On the other hand, **contextual bias** can happen in the tone, words, or context of a sentence. Unlike structural bias, this type of bias is not evident in grammatical structure but requires contextual background information and human perception [111]. Therefore, gender bias can be ascertained using both linguistic and extra-linguistic cues and can manifest itself in subtle or explicit ways, with differing degrees of intensity, which in turn makes this type of research challenging [112].

Furthermore, gender bias can easily propagate to models and downstream tasks, causing harm to the end-users [106]. As seen above, these forms of harm can emerge as representational and allocation harms and gender gaps. In the context of NLP, **allocation harm** is when models perform better on data related with the "majority" gender; this often happens in machine translation and co-reference resolution [113]. **Representation harm** is noted when associations between gender and certain concepts are captured in word embeddings and model parameters [106], [110]. Moreover, the gender gap influences gender bias in the text. As women are underrepresented in most areas of society, there follows that available texts mainly discuss and quote men, leading, for example, to biased corpora that researchers train their models on. As a result, the work of women researchers and academics tend to be quoted much less than that of male researchers [114].

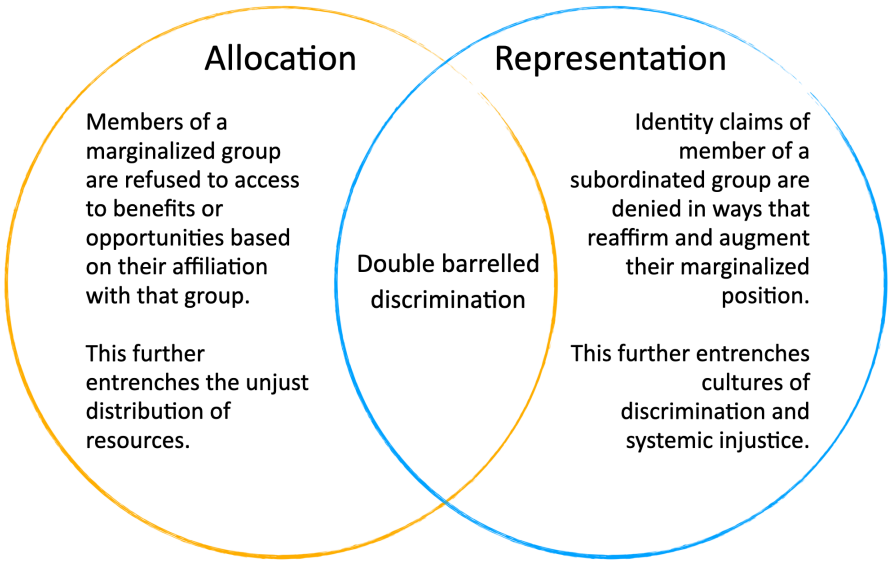

Figure 3.2: Allocation vs Representation harm recap
Author's elaboration

3.2.1 Measuring bias in word embeddings

In recent years numerous publications have approached the quantification of bias in word embedding. Complex and large structures as word embedding can suffer inherently of historical bias. Due to the prevalence of NLP systems and their increasing application areas, researchers have focused their attention on developing measures to uncover gender biases encoded in these methods. Therefore there are many studies reviewing the different measures according to different tasks [11], [110], [112]. In this section, for the experimental purpose of this thesis, we will focus only on the bias measurement for word embedding that we will include in our experiment.

Projection-based Measures

In [106], while using state-of-the-art word embedding in word analogy tests, the authors noticed that "man" would be associated with "computer programmer" and "woman" would be mapped to "homemaker." The authors then proposed debiasing word embeddings using a method that respects the embeddings for gender-specific words but debiases embeddings for gender-neutral words. The procedure is based on two steps:

- Identify gender subspace: find the direction of the embedding that captures the bias.

$$DirectBias_c = \frac{1}{|N|} = \sum_{w \in N} |cos(\overrightarrow{w}, g)|^c$$

- Hard or soft debiasing:
  - Hard debiasing: neutralize and equalize.
  - Soft bias correction: tries to move as little as possible to retain similarity to the original embedding while reducing the gender bias.

Bolukbasi et al. (2016) definition of gender bias reads as "the correlation between the magnitude of the projection onto the gender subspace of a word embedding representing a gender-neutral word and that word's bias rating, as rated by crowd workers." [106]. Their approach consists in building a linear support vector machine to identify the gender subspace, as a set of gender-specific and gender-neutral words based on a training set of hand picked gender-specific words. Then they identify a "gender direction" by aggregating ten gender pairs (i.e., she-he, woman-man, etc.) and use PCA to find a single eigenvector that shows significantly greater variance than the rest [112].

Since the above definitions are clear-cut and geometrically grounded, they have often been employed to quantify gender bias in word embeddings. However, Gonen and Goldbert (2019) [115] highlight that this method falls short of capturing the full picture of gender bias in vector spaces, as word embeddings representing words with similar biases still cluster them together after removing the projections of word embeddings representing gender-neutral words onto the gender subspace. Therefore, they introduce the notion of cluster bias of a word (W), which can be calculated as the percentage of male and/or female stereotypical words among the k nearest neighbors of W's embedding where the male or female stereotypical words are obtained through human annotation.

Word Embedding Association Test

The definition of WEAT comes from the paper "Semantics derived automatically from language corpora necessarily contain human biases" by Caliskan et al. (2017) [116]. The idea is to use the Implicit Association Test (IAT) core concept in order to quantify gender bias in word embeddings through the difference in strength of association of concepts. In

psychology the Implicit Association Test (IAT) is used to assess the presence of subconscious gender bias in humans. This can be defined as "the difference in time and accuracy that humans take to categorize words related to two concepts they find similar versus two concepts they find different" [116], [117]. In detail, the WEAT compares sets of identified concepts (i.e., male and female words), denoted as X and Y (each of equal size N), with a set of attributes, denoted as A and B, in order to measure bias over social attributes and roles (i.e., career/family words). The resulting test statistics is defined as a permutation test over X and Y :

$$S(X, Y, A, B) = [mean_{x \in X} sim(x, A, B) - mean_{y \in Y} sim(y, A, B)]$$

where sim is the cosine similarity. The resulting effect size is then the measure of association:

$$d = \frac{S(X, Y, A, B)}{std_{t \in X \cup Y} s(t, A, B)}$$

The null hypothesis suggests that there is no difference between X and Y in terms of their relative similarity to A and B. In Caliskan et al. [2017], the null hypothesis is tested through a permutation test, i.e., the probability that there is no difference between X and Y (in relation to A and B) and, therefore, that the word category is not biased. However, we note that results obtained with WEAT should be treated with caution since Ethayarajh et al. [2019] prove that WEAT systematically overestimates bias [118].

Relative Negative Sentiment Bias

Relative Negative Sentiment Bias (RNSB) measure was introduced by Chris Sweeney and Maryam Najafian (2019) [119]. The metric measures fairness in word embeddings via the relative negative sentiment associated with terms from various protected groups. The idea is to use the embedding model that we are trying to evaluate to initialize vectors for an unbiased positive/negative word sentiment dataset. Using this dataset, a logistic classification algorithm is trained to predict the probability of any word being a negative sentiment word. After training, a selected set of neutral identity terms from a protected group (i.e., national origin) is taken to predict the probability of negative sentiment for each word in the set. Neutral identity terms that are unfairly entangled with negative sentiment in the word embeddings will be classified like their neighboring sentiment words

from the sentiment dataset. Finally, we leverage this set of negative sentiment probabilities to summarize unintended demographic bias using RNSB.

For gold standard labeled positive/negative sentiment words, (x_i, y_i), in the training set, S, where x_i is a word vector from a possibly biased word embedding model, we find the minimizer, $f^*(x_i) = (w^T x_i)$, for the logistic loss, l, and learned weights, w.

$$min_{w \in R^d} \sum_{i=0}^{n} l(y_i, w^T x_i) + \lambda \|w\|^2, \lambda > 0$$

Then for a set, $K = k_1, ..., k_t$, of t demographic identity word vectors from a particular protected group (i.e., national origin, religion, etc.), we define a set, P, containing the predicted negative sentiment probability via minimizer, f, normalized to be one probability mass.

$$P = \left\{ \frac{f^*(k_1)}{\sum_{i=1}^{t} f^*(k_i)}, ..., \frac{f^*(k_t)}{\sum_{i=1}^{t} f^*(k_i)} \right\}$$

The metric $RNSB(P)$ is defined as the KL divergence of P from U, where U is the uniform distribution from the t elements.

$$RNSB(P) = D_{KL}(P \| U)$$

The RNSB metric captures the distance, via KL divergence, between the current distribution of negative sentiment and the fair uniform distribution. So the more fair a word embedding model with respect to sentiment bias, the lower the RNSB metric [119].

Embedding Coherence Test

It quantifies the amount of explicit bias $B_E = T_1, T_2, A$ (Dev and Phillips 2019). Unlike WEAT, it compares vectors of target sets T_1 and T_2 (averaged over the constituent terms) with vectors from a single attribute set A. ECT first computes the mean vectors for the target sets T_1 and T_2:

$$\mu_1 = \frac{1}{|T_1|} \sum_{t_1 \in T_1} t_1$$

and

$$\mu_2 = \frac{1}{|T_2|} \sum_{t_2 \in T_2} t_2$$

Next, for both μ_1 and μ_2 it computes the (cosine) similarities with vectors of all $a \in A$. Finally, the two resultant vectors of similarity scores, $s1$ (for $T1$) and $s2$ (for $T2$), are used to obtain the final ECT score. It is the Spearman's rank correlation between the rank orders of $s1$ and $s2$ – the higher the correlation, the lower the bias.

3.2.2 Debiasing method overview

While removing all the bias from language or NLP algorithms is an impossible task, researchers have taken significant steps towards developing fair systems in recent years. Since the impact of bias on many different applications is not arguable, given the potential risk of using ML algorithms that amplify gender stereotypes, the main challenge in debiasing tasks is to strike a balance between maintaining model performance on downstream tasks while reducing the encoded gender bias [120].

Table 3.2 summarizes the identified lines of gender bias mitigation methods together with the respective publications, a valuable classification of gender bias mitigation methods by Karolina Stanczak and Isabelle Augenstein [112] following the work of Sun et al. (2019) [110].

Data Manipulation			
Data Augmentation	**Gender Tagging**	**Balanced Fine-Tuning**	**Adding Context**
Madaan et al. [2018]; Park et al. [2018] Hall Maudslay et al. [2019]; Zhao et al. [2018a] Emami et al. [2019]; Zmigrod et al. [2019] Bartl et al. [2020]; Zhao et al. [2019] de Vassimon Manela et al. [2021]; Sen et al. [2021]	Moryossef et al. [2019]; Vanmassenhove et al. [2018] Habash et al. [2019]; Stafanovičs et al. [2020] Saunders et al. [2020]	Park et al. [2018]; Saunders and Byrne [2020] Costa-jussà and de Jorge [2020]	Basta et al. [2020]
Methodological Adjustment			
Projection-Based Debiasing	**Adversarial Learning**	**Constraining Output**	**Other**
Bolukbasi et al. [2016]; Schmidt [2015] Bordia and Bowman [2019]; Park et al. [2018] Ethayarajh et al. [2019]; Sahlgren and Olsson [2019] Karve et al. [2019]; Sedoc and Ungar [2019] Liang et al. [2020]; Prost et al. [2019] Dev et al. [2020]; Kaneko and Bollegala [2021a]	Li et al. [2018]; Zhang et al. [2018]	Ma et al. [2020]; Zhao et al. [2017]	Qian et al. [2019]; Zhao et al. [2018b] Jin et al. [2021]; Kaneko and Bollegala [2019]

Table 3.2: Classification of gender bias mitigation methods with reference [112]

Projection-Based Debiasing

To the best of our knowledge, the first paper to suggest the first word embedding debiasing algorithm was published in 2015 by Schmidt [121]. His approach consisted in removing similarity to the gender subspace by developing "a genderless framework" using cosine similarity and orthogonal vectors. The genderless framework, however, may be inconsistent because the semantic definition of a given word is likely to intertwined with its gender component. At the same time, it could also be argued that the gender component

of a word should have a role in shaping its semantic definition. Future research should be conducted in collaboration with social scientists to deepen understanding of this topic, as suggested by [109], [110].

Instead of completely removing gender information, Bolukbasi et al. [106] call for an approach that shifts word embeddings to be equally male and female in terms of their vector direction and propose to modify the embedding space by removing the gender component only from gender-neutral words. For instance, a debiased embedding for grandmother and grandfather will be equally close to babysitting without neglecting the analogy to gender. More specifically, Bolukbasi et al. propose two debiasing methods, hard and soft-debiasing [112]. Both approaches start from the identification of a list of gender-neutral words and a direction of the embedding that catches the bias. **Hard-debiasing** (or 'Neutralise and Equalise method') ensures that gender-neutral words are zero in the gender subspace and equalizes sets of words outside the subspace. In so doing, it enforces the property that any neutral word is equidistant to all words in each equality set (a set of words which differ only in the gender component). "For instance, taking (grandmother, grandfather) and (guy, gal) as two equality sets, after equalization, "babysit" would result to be equidistant to grandmother and grandfather and gal and guy, but closer to the grandparents and further away from the gal and guy." [106]. This approach is appropriate for applications where one does not wishes to display any bias in any such pair with respect to neutral words. The disadvantage of equalizing sets of words outside the subspace is that it removes certain specific distinctions that may be of value in specific applications. For instance, Bolukbasi et al. highlight that one may wish a language model to assign a higher probability to the phrase such as 'grandfather a regulation' since it is an idiom, unlike 'grandmother a regulation'.

The **soft-debiasing** algorithm reduces differences between sets whilst maintaining as much similarity as possible to the original embedding , with a parameter that controls for this trade-off. More specifically, soft-debiasing applies a linear transformation that seeks to preserve pairwise inner products between all the word vectors while minimizing the projection of the gender-neutral words onto the gender subspace [112].

Both hard and soft-debiasing approaches have been applied in research to word embeddings and language models. Bordia and Bowman [123] validate the soft-debiasing approach

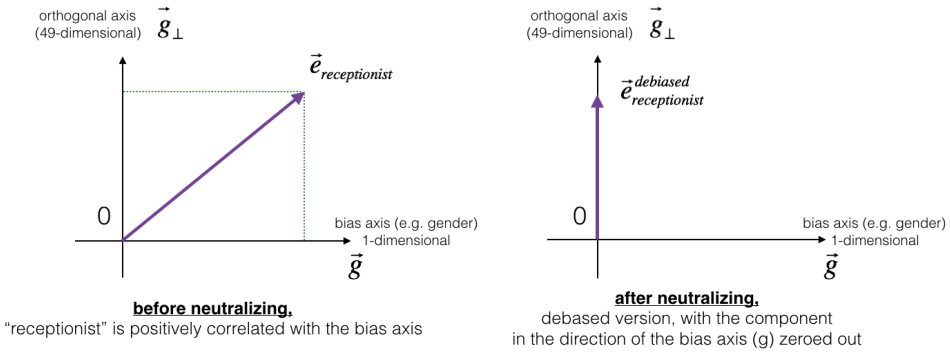

Figure 3.3: The word vector for "receptionist" before and after neutralization [122]

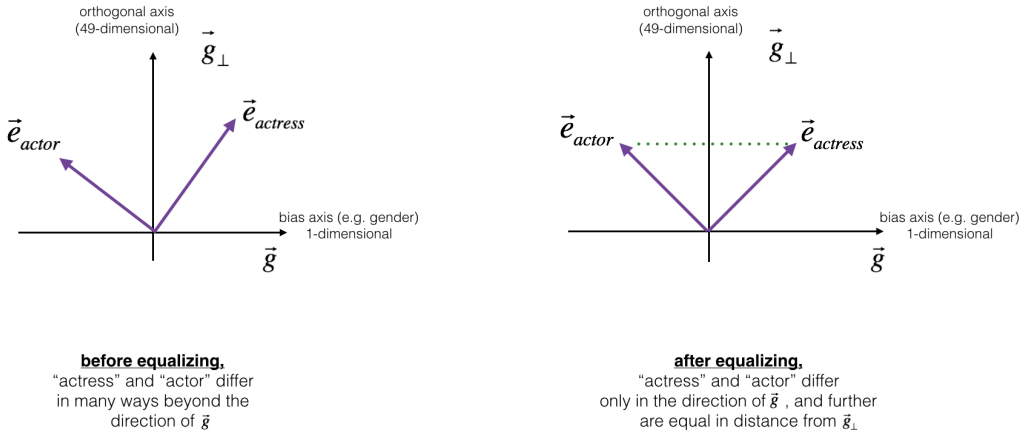

Figure 3.4: The word vectors "actor" and "actress" before and after equalization [122]

to mitigate bias in LSTM based word-level language models. Park et al. [124] compare hard-debiasing methods to other methods in abusive language detection. Sahlgren and Olsson [125] apply hard-debiasing to Swedish word embeddings and demonstrate that this method does not produce the desired outcome when tested on selected downstream tasks. Interestingly, Ethayarajh et al. [118] prove that post hoc word embeddings debiasing, using subspace projection, is equivalent to training on an unbiased corpus, under certain conditions. Similarly, Bolukbasi et al. [106] strive to identify the bias subspace in word embeddings by way of a set of target words and a debiasing conceptor, that is, a mathematical

representation of subspaces that can be operated on and composed by logic-based manipulations. It is important to highlight, however, that these methods are strongly dependent on pre-defined lists of gender-neutral words. Moreover, Zhao et al. [126] prove that an error in identifying gender-neutral words affects the performance of the downstream model. Bordia and Bowman in [123] and Zhao et al. in [126] notice the existence of a trade-off between perplexity and gender bias since male and female words are predicted with an equal probability in an unbiased model. This, however, can be undesirable in domains such as social science and medicine.

A new method - GN-GloVe – suggested by Zhao et al. [126] consists in isolating gender information to train the word embeddings, while maintaining gender-neutral information in other dimensions. This is achieved by "(1) minimizing the negative difference (maximizing the difference) between the gender dimension in male and female definitional word embeddings and (2) maximizing the difference between the gender direction and the other neutral dimensions in the word embeddings." [11], [126]. As a result, gender dimensions can be used or neglected making room for greater flexibility.

While Gonen and Goldberg [115] assert that debiasing is fundamentally superficial since a lot of the allegedly removed bias can still be recovered due to the geometry of the word representation of the gender neutralized words. Prost et al. [127] show that soft-debiasing can even increase the bias of a downstream classifier since it removes noise from gender-neutral words and ultimately provides a less noisy communication channel for gender clues. Recently, Dev et al. [128] employed an orthogonal projection to gender direction to debias contextualized embeddings and test it on a NLP task with pairs of gender-biased hypothesis. Notheless, this method can only be applied to the non-contextualized layers of the model. Kaneko and Bollegala [129] address this limitation in a fine-tuning setting, applying an orthogonal projection only in the hidden layers, and outperforming Dev et al. Furthermore, this method is independent of model architectures or their pre-training method. However, it requires a list of attribute words (i.e., she, man, her) and target words (i.e., occupations) in order to extract relevant sentences from the corpus, making the method highly reliant on this list [112].

Moving the discussion to a qualitative point of view, although the word embedding model is essential in many NLP systems and mitigating bias in embeddings helps reducing

bias's spread to downstream tasks, it is debatable if debiasing word embeddings has to be considered a step towards mitigating bias. For example, Caliskan et al. argue that "debiasing word embeddings blind an AI agent's perception rather than teaching it to perform appropriate actions" [116].

It is also important to recognize that entirely removing gender bias from the embedding space is difficult. While existing methods successfully mitigate bias with respect to projection onto the gender subspace in some degrees, Gonen and Goldberg (2019) show that gender bias based on more subtle metrics such as cluster bias still exists.

3.3 Experiment

This chapter presents our framework for understanding and evaluating unintentional gender bias in word embeddings. We first illustrate the flow of our framework in Figure 3.5. Then, we briefly describe the pre-trained word embedding and the dataset used for their evaluation. Finally the experimental results are presented and commented in sub-section 3.3.3.

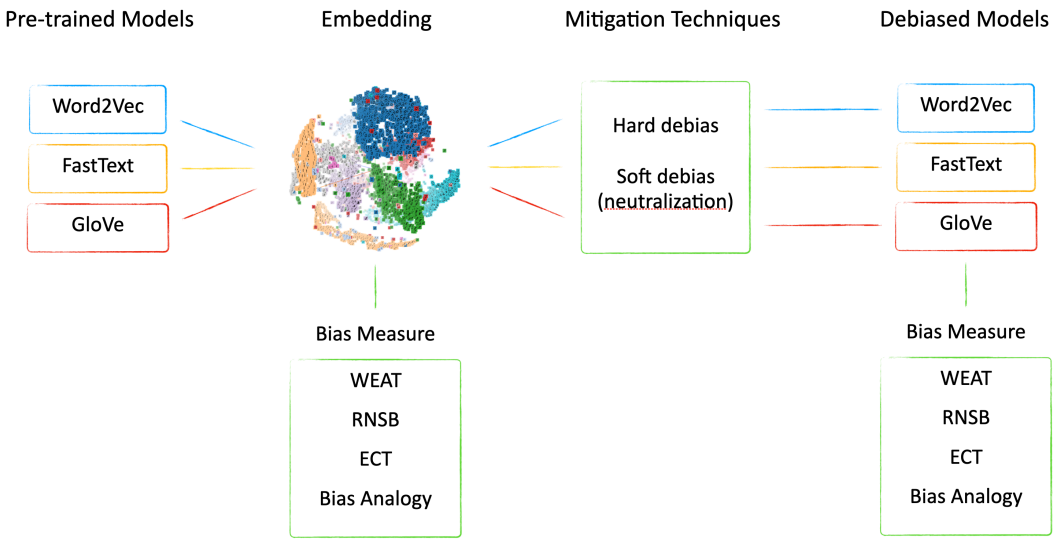

Figure 3.5: The general framework for our analysis
Author's elaboration

The experiment consists in evaluating three different pre-trained word embedding using

the measures described previously, applying the debiasing methods proposed by Bolukbasi [106], and finally evaluating the debiased embeddings. The experiment was entirely delevoped in python 3.7 on a 8 core 16 GB machine.

3.3.1 Pre-trained word embeddings

In this section we describe the considered pre-trained word embeddings. In particular we select three of the most well known pre-trained word embeddings based on their widespread use within academia and on our computational ability.

- **Word2Vec:** 300-dimensional embeddings for ca. 3M words learned from Google News corpus [130]
- **FastText:** 300-dimensional embeddings for ca. 1M words learned from Wikipedia 2017, UMBC web base corpus, and statmt.org news [131]
- **Glove:** 300-dimensional embeddings for ca. 2.2M words learned from the Common Crawl [132]

These three models belong to two different families. Both families learn geometrical encodings (vectors) of words from their co-occurrence information. However, they differ because word2vec and fastText are "predictive" models, whereas GloVe is a "count-based" model [133].

Predictive models learn their vectors in order to improve their predictive ability of Loss(target_word|context_words; vectors), i.e., the loss of predicting the target words from the context words given the vector representations. In word2vec, this is cast as a feed-forward neural network and optimized as such using SGD, etc.

In *Word2Vec* representation of words can be done using two main methods:

- **Continuous Bag of Words (CBOW)**: This method completes an incomplete sentence by "predicting the words that can be fitted into the middle of the sentence based on the surrounding context of the words. The prediction context depends on the few words before and after the predicted word." This name is due to the fact that the order of words in the context is irrelevant [133].

- **Skip-Gram**: This method helps to predict the context words or surrounding words based on a current word in the same sentence. "The Skip-gram model takes each word of the large corpus as the input, and the hidden or embedding layer using the embedding weights predicts the context words" [133].

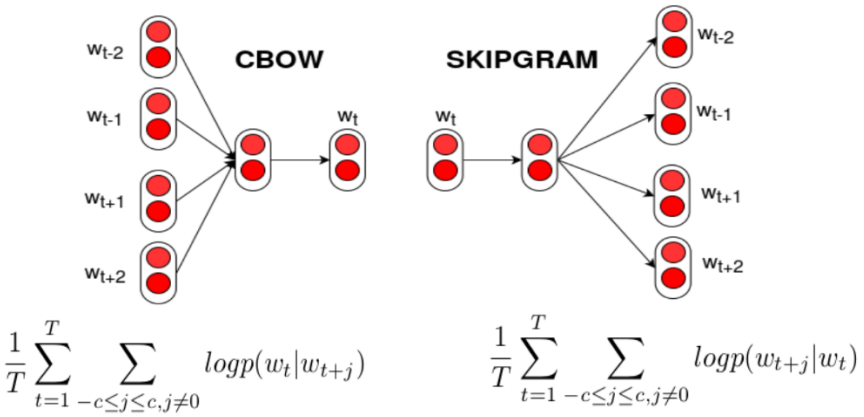

Figure 3.6: The basic architecture of CBOW and Skip-Gram models [133]

Count-based models learn their vectors by essentially making dimensionality reduction on the co-occurrence counts matrix. They first construct a large matrix of co-occurrence, in example, for each "word" (rows), they count how frequent this word is in some "context" (columns) in a large corpus. The number of "contexts" is large since it is essentially combinatorial in size. So, they factorize this matrix to yield a lower-dimensional matrix, where each row yields a vector representation for each word. In general, this is done by minimizing a "reconstruction loss" which tries to find the lower-dimensional representations which can explain most of the variance in the high-dimensional data. In the specific case of GloVe, the counts matrix is pre-processed by normalizing the counts and log-smoothing them. This turns out to be a positive thing in terms of the quality of the learned representations [133].

However, as pointed out, when we control for all the training hyper-parameters, the embeddings generated using the two methods tend to perform very similarly in downstream NLP tasks. The additional benefit of GloVe over word2vec is that it is easier to parallelize the implementation, which means it is faster to train over more data, which, with these models, is always a good thing.

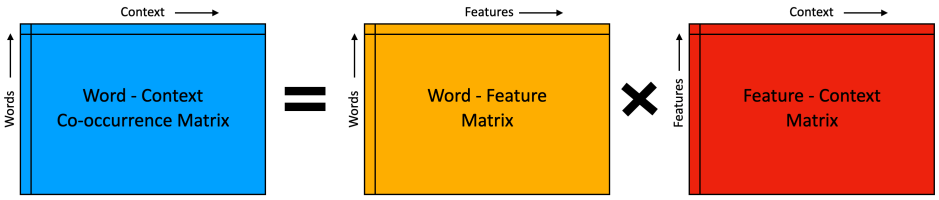

Figure 3.7: Conceptual model for the GloVe model's implementation [134]

3.3.2 Dataset

In order to measure gender bias impact in NLP methods and downstream applications, multiple datasets have been developed. Karolina Stanczak and Isabelle Augenstein in [112] list the well-established datasets shown in 3.3 together with the tasks they can probe and biases they provide a testbed for.

Dataset	Size	Data	Gender	Task	Bias
EEC [Kiritchenko and Mohammad 2018]	8 640 sent.	sent. templates	b	SA	stereotyping
WinoBias [Zhao et al. 2018a]	3 160 sent.	sent. templates	nb	cor. res.	occ. bias
WinoGender [Rudinger et al. 2018]	720 sent.	sent. templates	b	cor. res.	occ. bias
WinoMT [Stanovsky et al. 2019]	3 888 sent.	sent. templates	b	MT	occ. bias
Occupations Test [Escudé Font and Costa-jussà 2019]	2 000 sent.	sent. templates	b	MT	occ. bias
GAP [Webster et al. 2018]	8 908 ex.	Wikipedia	b	cor. res.	stereotyping
KNOWREF	8 724 sent.	Wikipedia & other	b	cor. res.	stereotyping
BiosBias [De-Arteaga et al. 2019]	397 340 bios	CommonCrawl	b	classification	occ. bias
GeBioCorpus	2 000 sent.	Wikipedia	b	MT	occ. bias
StereoSet [Nadeem et al. 2021]	2 022 sent.	human-generated	b	probing LMs	stereotyping
CrowS-Pairs	1508 ex.	human-generated	b	probing LMs	stereotyping

Table 3.3: List of common probing datasets for gender bias in language [112]

Template-based datasets

Several studies accounting for gender bias in NLP have been conducted on benchmark datasets that consist of template sentences of simple structures, for example "He/She is a/an [occupation/adjective]", where the [person/adjective] is populated with occupations or positive/negative descriptors. Similarly, the **EEC dataset** [135] includes sentence templates such as [Person] feels [emotional state word]. Another multilingual dataset has been proposed by Bianchi et al. [136] who created a template-based dataset consisting of a subject and a predicate in 6 languages (English, Italian, French, Portuguese, Romanian, and Spanish).

Another strain of work has utilized the structure of Winograd Schemas [137]: **WinoBias** [126], **WinoGender** [138], and **WinoMT** [139]. Since Winograd Schema Challenge is a co-reference resolution task with human-generated sentence templates that require reasoning with commonsense knowledge, it has been used to analyze if the reasoning of the co-reference system is dependent on the gender of a pronoun in a sentence, and to measure stereotypical and non-stereotypical gender associations for different occupations. WinoBias contains two types of sentences that require linking gendered pronouns to either male or female stereotypical occupations. While none of the examples can be disambiguated by the gender of the pronoun, this cue could potentially mislead the model. The sentences in WinoBias have been formulated so that there is no objective way to choose between different gender pronouns in the absence of stereotypes. In parallel, Rudinger et al. develop a WinoGender dataset. Each sentence contains three variables in the WinoBias dataset: occupation, person, and pronoun. Winogender includes two similar sentence templates for each occupation: one pronoun is coreferent with occupation, and one coreferent with a person. It is worth noting that WinoGender sentences, differently from WinoBias, also include gender-neutral pronouns. Finally, sentences in WinoGender cannot be solved from syntax alone, unlike in the WinoBias dataset, which might better isolate the effect of gender bias. Based on WinoGender and WinoBias, Stanovsky et al. [139] curate WinoMT, a probing dataset for machine translation, with sentences that have stereotypical and non-stereotypical gender-role assignments. WinoMT has become widely used as a challenge dataset for gender bias detection in MT systems, with Saunders et al. [140] developing a version of the WinoMT dataset with binary templates filled with singular they pronouns.

SemBias analogy dataset

The SemBias Dataset was created by Zhao et al. [126] to identify the correct analogy of "he-she" from four s of words. Each case in the dataset contain four word pairs:

- a gender-definition word pair (Definition; i.e., "waiter - waitress")
- a gender-stereotype word pair (Stereotype; i.e., "doctor - nurse")
- two other pairs of words that have similar meanings (None; i.e., "dog - cat", "cup - lid")

The dataset contains 20 gender-stereotype word pairs and 22 gender-definition word pairs and uses their Cartesian product to generate 440 instances. Among the 22 gender-definitional word pairs, two word pairs are not used as the seed words during the training. Following Zhao et al. to test the model's generalization ability, we generate a subset of data (SemBias (subset)) of 40 instances linked with these two pairs.

The relational similarity between (he,she) and (a,b) in SemBias is computed using the cosine similarity between the "he-she" gender directional vector and "a-b" using the word embeddings under evaluation. For the four word-pairs in each instance in SemBias, we select the word-pair with the highest cosine similarity with "he-she" as the predicted answer.

3.3.3 Experimental results

To evaluate the pre-trained word embeddings, we use the metrics described in 3.2.1, comparing the results from the original and debiased embeddings.

Below are WEAT, RNSB, and ECT results, according to seven sets of different target words and multiple males and females attribute words. For every metric, we computed the values obtained with each model for the (**o**)riginal embedding, the (**s**)oft debiased, and the (**h**)ard debiased ones.

The first measure we evaluate is the Word Embedding Association Test by Caliskan et al. [116], described in 3.2.1. where, for each target group, we computed the association with the set of male and female attribute words (pronouns).

	Word2Vec			**FastText**			**GloVe**		
	o	**s**	**h**	**o**	**s**	**h**	**o**	**s**	**h**
Career-Family	0.35	−0.12	**0.03**	0.38	0.03	**0.03**	0.41	−0.10	**0.01**
Math-Arts	0.71	−0.20	**-0.09**	0.66	0.19	**0.01**	0.38	**-0.01**	−0.03
Science-Arts	0.90	−0.01	**0.00**	0.89	0.29	**0.09**	1.06	−0.07	**-0.06**
Intel.-Appearance	1.18	**-0.12**	−0.21	0.94	0.16	**-0.14**	0.96	**0.04**	−0.09
Intel.-Sensitive	0.91	0.21	**-0.07**	0.45	0.12	**-0.06**	0.69	**0.03**	−0.07
Pos-Neg words	−0.40	−0.30	**-0.18**	−0.32	−0.27	**-0.13**	−0.42	−0.23	**-0.05**
Man-Woman roles	1.83	0.97	**0.74**	1.81	1.06	**0.78**	1.78	0.87	**0.82**

Table 3.4: WEAT values for target words group with respect to male and female terms
Author's elaboration - *simplest solution is preferred in the case of equal values

In table 3.4 we highlight in **bold** the best results obtained for each model. At first glance, it seems that the considered debiasing operations have affected the WEAT value for all the embeddings. Compared to the original version, all three embeddings show a clear improvement in both soft and hard debiased embeddings. Nevertheless, Word2Vec and FastText have a noticeable tendency to the hard debiased embedding, while Glove has very similar values for the soft and hard embeddings. Since as stated by Ethayarajh et al. [118] WEAT systematically overestimates bias we now evaluate further metrics.

	Word2Vec			**FastText**			**GloVe**		
	o	**s**	**h**	**o**	**s**	**h**	**o**	**s**	**h**
Career-Family	.0059	**.0057**	.0065	.0026	**.0022**	.0031	.0075	.0047	**.0036**
Math-Arts	.0008	**.0006**	.0007	.0008	.0006	**.0005**	.0012	.0011	**.0010**
Science-Arts	.0005	.0006	**.0003**	.0005	.0005	**.0004**	.0006	.0006	**.0004**
Intel.-Appearance	.0069	**.0035**	.0037	.0062	**.0035**	.0042	.0100	.0059	**.0048**
Intel.-Sensitive	.0022	.0019	**.0016**	.0021	**.0014**	.0020	.0024	**.0016**	.0018
Pos-Neg words	.0204	.0165	**.0134**	.0499	.0454	**.0404**	.0339	.0324	**.0293**
Man-Woman roles	.0076	**.0011**	.0012	.0029	.0006	**.0003**	.0051	.0008	**.0005**

Table 3.5: RNSB values for target words group with respect to male and female terms
Author's elaboration - *simplest solution is preferred in the case of equal values

The Relative Negative Sentiment Bias (RNSB) metric can be interpreted as the distance between the current distribution of negative sentiment and the fair, uniform distribution. Therefore, the fairer a word embedding model is with respect to sentiment bias, the lower the RNSB metric. The results in table 3.5, although RNSB is not directly comparable with WEAT, seem to be consistent with the ones shown in table 3.4. All models seem to be improving in the debiased embedding. However, it is necessary to make some consideration with respect to the WEAT results: 1) the relative improvement from the original to the hard debiased version is much more moderate in RNSB than in WEAT. 2) In contrast with WEAT values, GloVe's best embeddings in terms on RNSB is the hard debiased one, while Word2Vec and FastText's best model seems to swing between soft and hard debiasing.

	Word2Vec			**FastText**			**GloVe**		
	o	**s**	**h**	**o**	**s**	**h**	**o**	**s**	**h**
Career	.714	**1.00**	1.00	**.952**	.929	.952	.976	.976	**1.00**
Family	.762	.833	**1.00**	.952	.976	**.976**	.905	.976	**1.00**
Science	.571	.857	**1.00**	.976	.976	**1.00**	.976	**1.00**	1.00
Arts	.810	.952	**.976**	.833	.929	**1.00**	.929	.952	**.952**
Appearance	.363	.879	**.904**	.507	.833	**.858**	.448	.952	**.965**
Intelligence	.744	.976	**.998**	.841	.943	**.991**	.916	.990	**.999**
Pleasant	.733	.978	**.983**	.943	966	**.989**	.938	.978	**.997**
Unpleasant	.800	.962	**.984**	.872	.912	**.976**	.900	.976	**.985**
Positive words	.771	.972	**.994**	.925	.982	**.997**	.936	.992	**.999**
Negative words	.791	.964	**.993**	.939	.981	**.997**	.954	.992	**.999**
Man roles	.972	.986	**.993**	.979	.972	**1.00**	.958	.958	**.993**
Woman roles	.747	**.956**	.879	.780	.885	**.901**	.511	**.923**	.736

Table 3.6: ECT values for target words group with respect to male and female terms
Author's elaboration - *simplest solution is preferred in the case of equal values

The Embedding Coherence Test quantifies the amount of explicit bias and returns the Spearman's rank correlation between the vectors of similarity scores for the attribute words set with the gender target sets; hence the higher is the correlation, the lower the bias. The results in table 3.6 seem to confirm the previously obtained values from WEAT and RNSB since all models have improved from the original embedding version. In particular, they find their best embedding in the hard debiased version. Nevertheless, we notice that ECT's values are extremely high in the soft or even the original embedding for some attribute words.

The three figures below show the evolution of the gender direction for different occupations for each pre-trained model for each embedding version. Although, according to these plots, there is an explicit improvement for all models from the original to the hard debiased embedding, in the hard debiased embedding it is still possible to identify some ambiguous words in the "she" direction, such as "maid", "waitress" and "housewife". The presence of these ambiguous words does not constitute a form of bias, but the absence of male equivalent terms is a warning sign for the presence of bias.

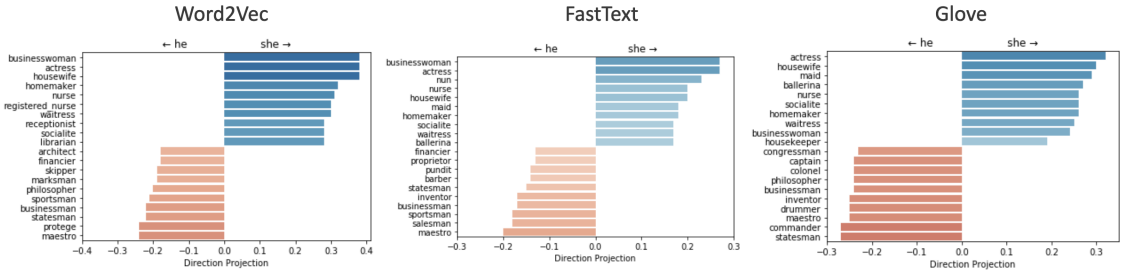

Figure 3.8: Gender direction for occupations in Soft embeddings
Generated using Matplotlib

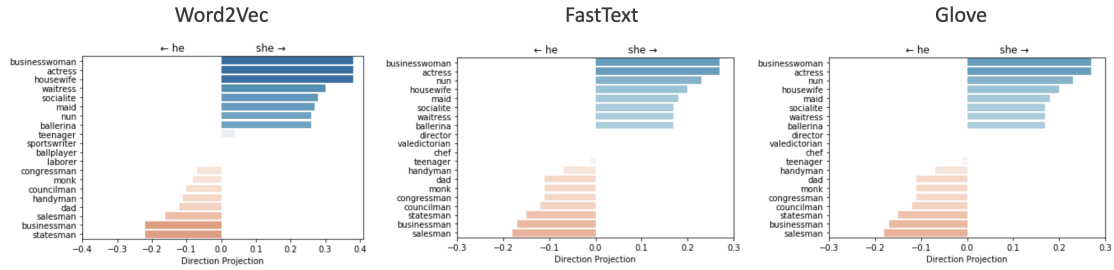

Figure 3.9: Gender direction for occupations in Soft embeddings
Generated using Matplotlib

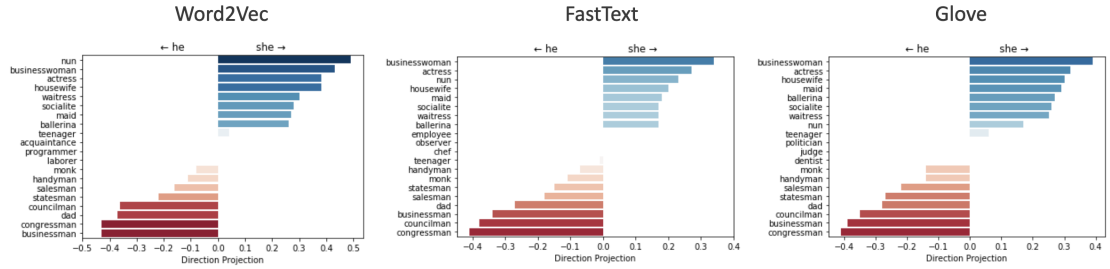

Figure 3.10: Gender direction for occupations in Hard embeddings
Generated using Matplotlib

Although the analysis carried out to this point seems to confirm that the embeddings have been successfully debiased, the qualitative evaluation of the results has brought out some concerns regarding the actual presence of bias. Therefore, we decide to evaluate the embeddings using a fourth metric to address this issue. The selected metric uses the SemBias Dataset, explained in 3.3.2, to quantify the correctness of the analogy of "he-she". To do so, for each set of word pairs (Definition, Stereotype, None) the percentage of times that each class of pair is on the top based on a word embedding model is computed.

Hence, if the word embedding has been correctly debiased, we would expect higher values in Definition and lower values in Stereotype and None.

Metrics	Word2Vec	FastText	GloVe
definition	**.827**/.823/.795	**.911**/.777/.820	**.834**/.770/.809
stereotype	.134/**.102**/.116	.065/**.047**/.061	.115/**.077**/.079
none	**.039**/.075/.089	**.023**/.175/.118	**.050**/.152/.111
sub-definition	.600/**.700**/.500	**.825**/.500/.700	**.675**/.525/.500
sub-stereotype	.300/**.200**/.275	.125/.125/**.100**	.275/**.125**/.225
sub-none	.100/**.100**/.225	**.050**/.375/.200	**.050**/.350/.275

Table 3.7: SemBias analogy values for pre-trained models
Author's elaboration - *simplest solution is preferred in the case of equal values

The doubt that emerged from the qualitative analysis seems to be fueled by the results shown in table 3.7. As a matter of fact, the debiased models are showing lower values in **definition** than the original embedding, suggesting the presence of bias. In particular, looking at the upper side of table 3.7, for each pre-trained model, the only improvement from the original embedding appears to be in the **Stereotype** values of the soft embedding. While the bottom sub-metrics spotlight a bad generalization ability for all the embeddings.

The experiment allowed us to identify and address, albeit only partially, some of the biases present in the word embeddings. Furthermore, it helped to highlight analogies and dissimilarities among metrics. This points to the importance of using different type of measures to tackle different type of biases in ML.

Chapter 4

Conclusions

The first goal of this thesis was to provide an extensive overview of fairness concepts, metrics and debiasing techniques, while its second objective was to test some selected metrics and debiasing techniques on three pre-trained embeddings in order to highlight the relevance of textual data to promote fairness in ML.

The initial exploration on the fundamental concepts and definitions in the ML fairness research area, followed by the literature review, has provided a picture of an area of growing interest, populated by tools, which is however also bungled and confusing. Most of the papers we analyzed lacked indeed a clear formulation of the concepts of fairness, bias and discrimination as they had taken into account. Furthermore, the analysis of the results did not always offer a crystal clear understanding of which bias' type was addressed and according to which fairness definition, potentially causing presentation bias and enhancing other forms of bias.

The analysis of the result of the experiment shows that the different word embeddings are consistent between the embeddings but inconsistent across the different metrics. Although WEAT, RNSB and ECT values are coherent, the visual evaluation and the SemBias analogy values seem to reflect the presence of bias in the supposed debiased models. In this context it is clear how, without an appropriate evaluation, the results of this thesis could have easily been affected by a form of observer bias, in which "researchers' expectations are subconsciously projected onto the research" [28]. As unfortunately, many authors may tend to do.

As we reached the end of our analysis, the results obtained raised the question "whether and to what extent the bias metrics can estimate the actual impact of untreated bias in downstream application" defining in practice the starting point of the follow-up task.

A second framework was then designed to estimate the fairness quality of the embeddings. The main idea was to test two applications for each version of the embedding: a TensorFlow model on the "next word" prediction to evaluate the presence of bias from a qualitative perspective; and sentiment analysis to be assessed with the EEC dataset, described in 3.3.2.

Significant future improvements that could be made to this work include integrating Elmo and Bert as contextualized word embeddings, which we could not conduct due to the lack of computational ability and the testing of different bias mitigation techniques.

We ascertained the trickiness of some bias measures and mitigation techniques and tried to gather as much helpful knowledge on fairness as possible, aiming to provide an overview of the right tools that are needed to approach the task with the appropriate amount of caution. Finally, we developed an experiment in the NLP context to test some of the available metrics and techniques and raise awareness on the relevance of textual data in the pursuit of fairness in ML.

The decision to treat this specific type of data was motivated by our will to point out its relevance in the pursuit of fairness. These NLP applications find their way into our everyday life in different forms, from auto-completion in browsers to sentiment analysis for recommendation systems, among others. As far as we could assess, these systems tend to be affected by popularity and representation bias, which highly contribute to the propagation and rooting of misbeliefs and stereotypes. However, as emerged from the review summarized in chapter 2.2, the research and practitioner community's attention seems to focus mainly on tabular data and classification tasks. In our view, this is highly problematic and highlights that more research is needed on the topic, and especially one that promotes intersectional analyses and multi-disciplinarity. Furthermore, it also signals that much more awareness still needs to be raised at all levels, including in the general public, of the implications of bias in ML.

Bibliography

[1] S. Barocas and A. D. Selbst, "Big Data's Disparate Impact", *California Law Review*, vol. 104, no. 3, pp. 671–732, 2016, ISSN: 00081221. [Online]. Available: http://www.jstor.org/stable/24758720.

[2] R. Nabi, D. Malinsky, and I. Shpitser, *Learning Optimal Fair Policies*, 2019. arXiv: 1809.02244 [cs.LG].

[3] D. Banerjee, *Natural language processing (NLP)*, 2020. [Online]. Available: https://datascience.foundation/sciencewhitepaper/natural-language-processing-nlp-simplified-a-step-by-step-guide.

[4] J. Rodger and P. Pendharkar, "A field study of the impact of gender and user's technical experience on the performance of voice-activated medical tracking application", *Int. J. Hum.-Comput. Stud.*, vol. 60, pp. 529–544, May 2004. DOI: 10.1016/j.ijhcs.2003.09.005.

[5] L. Sweeney, "Discrimination in Online Ad Delivery", *CoRR*, vol. abs/1301.6822, 2013. arXiv: 1301.6822. [Online]. Available: http://arxiv.org/abs/1301.6822.

[6] Julia Angwin, Jeff Larson, Surya Mattu, Lauren Kirchner. (2016). "Machine Bias, There's software used across the country to predict future criminals. And it's biased against blacks", [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[7] M. Mansoury, H. Abdollahpouri, J. Smith, A. Dehpanah, M. Pechenizkiy, and B. Mobasher, "Investigating Potential Factors Associated with Gender Discrimination in Collaborative Recommender Systems", in *FLAIRS Conference*, 2020.

[8] S. Tolan. (Jan. 2019). "Fair and Unbiased Algorithmic Decision Making: Current State and Future Challenges".

[9] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "On the (im)possibility of fairness", *CoRR*, vol. abs/1609.07236, 2016. arXiv: 1609.07236. [Online]. Available: http://arxiv.org/abs/1609.07236.

[10] R. Baeza-Yates, "Bias on the Web", *Commun. ACM*, vol. 61, no. 6, 54–61, May 2018, ISSN: 0001-0782. DOI: 10.1145/3209581. [Online]. Available: https://doi.org/10.1145/3209581.

[11] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning", *CoRR*, vol. abs/1908.09635, 2019. arXiv: 1908.09635. [Online]. Available: http://arxiv.org/abs/1908.09635.

[12] L. Zhang, Y. Wu, and X. Wu, *A causal framework for discovering and removing direct and indirect discrimination*, 2016. arXiv: 1611.07509 [cs.LG].

[13] N. Crowley, *Identifying and Preventing Systemic Discrimination at the Local Level*, 2018.

[14] E. S. Phelps, "The Statistical Theory of Racism and Sexism", *The American Economic Review*, vol. 62, no. 4, pp. 659–661, 1972, ISSN: 00028282. [Online]. Available: http://www.jstor.org/stable/1806107.

[15] F. Kamiran and I. Zliobaite, "Explainable and Non-explainable Discrimination in Classification", English, in *Discrimination and Privacy in the Information Society*, ser. Studies in Applied Philosophy, Epistemology and Rational Ethics. International: Springer, 2013, vol. 3, pp. 155–170, ISBN: 978-3-642-30486-6. DOI: 10.1007/978-3-642-30487-3_8.

[16] H. Suresh and J. V. Guttag, "A Framework for Understanding Unintended Consequences of Machine Learning", *CoRR*, vol. abs/1901.10002, 2019. arXiv: 1901.10002. [Online]. Available: http://arxiv.org/abs/1901.10002.

[17] A. Olteanu, C. Castillo, F. D. Diaz, and E. Kıcıman, "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries", *Frontiers in Big Data*, vol. 2, 2019.

[18] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", in *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, S. A. Friedler and C. Wilson, Eds., ser. Proceedings of Machine Learning Research, vol. 81, PMLR, 2018, pp. 77–91. [Online]. Available: http://proceedings.mlr.press/v81/buolamwini18a.html.

[19] C. R. Blyth, "On Simpson's Paradox and the Sure-Thing Principle", *Journal of the American Statistical Association*, vol. 67, no. 338, pp. 364–366, 1972, ISSN: 01621459. [Online]. Available: http://www.jstor.org/stable/2284382.

[20] J. S. Chuang, O. Rivoire, and S. Leibler, "Simpson's paradox in a synthetic microbial system", *Science*, vol. 323, no. 5911, pp. 272–275, 2009.

[21] R. Kievit, W. Frankenhuis, L. Waldorp, and D. Borsboom, "Simpson's Paradox in Psychological Science: A Practical Guide", *Frontiers in psychology*, vol. 4, p. 513, Aug. 2013. DOI: 10.3389/fpsyg.2013.00513.

[22] I Minchev, G Matijevic, D. W. Hogg, G Guiglion, M Steinmetz, F Anders, C Chiappini, M Martig, A Queiroz, and C Scannapieco, "Yule-Simpson's paradox in Galactic Archaeology", *Monthly Notices of the Royal Astronomical Society*, 2019, ISSN: 1365-2966. DOI: 10.1093/mnras/stz1239. [Online]. Available: http://dx.doi.org/10.1093/mnras/stz1239.

[23] K. Lerman, "Computational Social Scientist Beware: Simpson's Paradox in Behavioral Data", *CoRR*, vol. abs/1710.08615, 2017. arXiv: 1710.08615. [Online]. Available: http://arxiv.org/abs/1710.08615.

[24] ""Blissfully Happy" or "Ready toFight": Varying Interpretations of Emoji", vol. 10, [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14757.

[25] D.-P. Nguyen, R. Gravel, R. Trieschnigg, and T. Meder, ""How old do you think I am?": A study of language and age in Twitter", English, in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, ICWSM 2013*, eemcs-eprint-23604 ; 7th International AAAI Conference on Weblogs and Social Media,

ICWSM 2013, ICWSM ; Conference date: 08-07-2013 Through 10-07-2013, AAAI Press, Jul. 2013, pp. 439–448, ISBN: 978-1-57735-610-3. [Online]. Available: http://www.icwsm.org/2013/.

[26] L. Introna and H. Nissenbaum, "Defining the Web: the politics of search engines", *Computer*, vol. 33, no. 1, pp. 54–62, 2000. DOI: 10.1109/2.816269.

[27] A. Nematzadeh, G. L. Ciampaglia, F. Menczer, and A. Flammini, "How algorithmic popularity bias hinders or promotes quality", *CoRR*, vol. abs/1707.00574, 2017. arXiv: 1707.00574. [Online]. Available: http://arxiv.org/abs/1707.00574.

[28] A. Jaokar, *23 sources of data bias for machine learning and deep learning*, 2020. [Online]. Available: https://www.datasciencecentral.com/23-types-of-bias-in-data-for-machinelearning-and-deeplearning/.

[29] A. Chouldechova and A. Roth, "The Frontiers of Fairness in Machine Learning", *CoRR*, vol. abs/1810.08810, 2018. arXiv: 1810.08810. [Online]. Available: http://arxiv.org/abs/1810.08810.

[30] R. Binns, "Fairness in Machine Learning: Lessons from Political Philosophy", in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, S. A. Friedler and C. Wilson, Eds., ser. Proceedings of Machine Learning Research, vol. 81, PMLR, 2018, pp. 149–159. [Online]. Available: https://proceedings.mlr.press/v81/binns18a.html.

[31] M. Srivastava, H. Heidari, and A. Krause, "Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning", in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, ser. KDD '19, Anchorage, AK, USA: Association for Computing Machinery, 2019, 2459–2468, ISBN: 9781450362016. DOI: 10.1145/3292500.3330664. [Online]. Available: https://doi.org/10.1145/3292500.3330664.

[32] A. Woodruff, S. E. Fox, S. Rousso-Schindler, and J. Warshaw, "A Qualitative Exploration of Perceptions of Algorithmic Fairness", in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2018, 1–14, ISBN: 9781450356206. [Online]. Available: https://doi.org/10.1145/3173574.3174230.

[33] B. Hutchinson and M. Mitchell, "50 Years of Test (Un)Fairness: Lessons for Machine Learning", in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* '19, Atlanta, GA, USA: Association for Computing Machinery, 2019, 49–58, ISBN: 9781450361255. DOI: 10.1145/3287560.3287600. [Online]. Available: https://doi.org/10.1145/3287560.3287600.

[34] A. Ignatiev, M. C. Cooper, M. Siala, E. Hebrard, and J. Marques-Silva, "Towards Formal Fairness in Machine Learning", in *Principles and Practice of Constraint Programming*, H. Simonis, Ed., Cham: Springer International Publishing, 2020, pp. 846–867, ISBN: 978-3-030-58475-7.

[35] S. Verma and J. Rubin, "Fairness Definitions Explained", in *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, 2018, pp. 1–7. DOI: https://dl.acm.org/doi/10.1145/3194770.3194776.

[36] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, *Fairness Through Awareness*, 2011. arXiv: 1104.3913 [cs.CC].

[37] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness", *CoRR*, vol. abs/1701.08230, 2017. arXiv: 1701.08230. [Online]. Available: http://arxiv.org/abs/1701.08230.

[38] M. Hardt, E. Price, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning", in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

[39] A. Chouldechova, *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments*, 2016. arXiv: 1610.07524 [stat.AP].

[40] M. Kusner, C. Russell, J. Loftus, and R Silva, "Counterfactual Fairness", Jan. 2017.

[41] G. Farnadi, B. Babaki, and L. Getoor, "Fairness in Relational Domains", in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '18, New Orleans, LA, USA: Association for Computing Machinery, 2018, 108–114,

ISBN: 9781450360128. DOI: 10.1145/3278721.3278733. [Online]. Available: https://doi.org/10.1145/3278721.3278733.

[42] RichardBerk, HodaHeidari, ShahinJabbari, MichaelKearns, and AaronRoth, "Fairness in Criminal Justice Risk Assessments: The State of the Art.".

[43] J. Kleinberg, S. Mullainathan, and M. Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, 2016. arXiv: 1609.05807 [cs.LG].

[44] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and Abstraction in Sociotechnical Systems", in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* '19, Atlanta, GA, USA: Association for Computing Machinery, 2019, 59–68, ISBN: 9781450361255. DOI: 10.1145/3287560.3287598. [Online]. Available: https://doi.org/10.1145/3287560.3287598.

[45] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt, *Delayed Impact of Fair Machine Learning*, 2018. arXiv: 1803.04383 [cs.LG].

[46] G. Smith, *What does "fairness" mean for machine learning systems?*, 2020.

[47] M. A. Madaio, L. Stark, J. Wortman Vaughan, and H. Wallach, "Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI", in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20, Honolulu, HI, USA: Association for Computing Machinery, 2020, 1–14, ISBN: 9781450367080. DOI: 10.1145/3313831.3376445. [Online]. Available: https://doi.org/10.1145/3313831.3376445.

[48] D. K. Mulligan, J. A. Kroll, N. Kohli, and R. Y. Wong, "This Thing Called Fairness", *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, 1–36, 2019, ISSN: 2573-0142. DOI: 10.1145/3359221. [Online]. Available: http://dx.doi.org/10.1145/3359221.

[49] M. et al., "AI Fairness Checklist". [Online]. Available: https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4t6dA.

[50] Z. Zhong, "A Tutorial on Fairness in Machine Learning", 2018. [Online]. Available: https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb.

[51] B. d'Alessandro, C. O'Neil, and T. LaGatta, "Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification", *Big Data*, vol. 5, no. 2, 120–134, 2017, ISSN: 2167-647X. DOI: 10.1089/big.2016.0048. [Online]. Available: http://dx.doi.org/10.1089/big.2016.0048.

[52] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*, 2018. arXiv: 1810.01943 [cs.AI].

[53] L. Oneto and S. Chiappa, "Fairness in Machine Learning", *Studies in Computational Intelligence*, 155–196, 2020, ISSN: 1860-9503. DOI: 10.1007/978-3-030-43883-8_7. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-43883-8_7.

[54] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi, *Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations*, 2017. arXiv: 1707.00075 [cs.LG].

[55] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints", English, in *Proceedings ICDM'09 Workshop on Domain Driven Data Mining (Miami FL, USA, December 6, 2009)*, United States: Institute of Electrical and Electronics Engineers, 2009, pp. 13–18, ISBN: 978-1-4244-5384-9. DOI: 10.1109/ICDMW.2009.83.

[56] H. Jeong, H. Wang, and F. P. Calmon, *Fairness without Imputation: A Decision Tree Approach for Fair Prediction with Missing Values*, 2021. arXiv: 2109.10431 [cs.LG].

[57] B. L. Thanh, S. Ruggieri, and F. Turini, "k-NN as an implementation of situation testing for discrimination discovery and prevention", in *KDD*, 2011.

[58] K. Mancuhan and C. Clifton, "Combating Discrimination Using Bayesian Networks", *Artificial Intelligence and Law*, vol. 22, no. 2, pp. 211–238, 2014. DOI: 10.1007/s10506-014-9156-4.

[59] L. Zhang, Y. Wu, and X. Wu, "Achieving Non-Discrimination in Data Release", in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '17, Halifax, NS, Canada: Association for Computing Machinery, 2017, 1335–1344, ISBN: 9781450348874. DOI: 10.1145/3097983.3098167. [Online]. Available: https://doi.org/10.1145/3097983.3098167.

[60] P. Gordaliza, E. D. Barrio, G. Fabrice, and J.-M. Loubes, "Obtaining Fairness using Optimal Transport Theory", in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., ser. Proceedings of Machine Learning Research, vol. 97, PMLR, 2019, pp. 2357–2365. [Online]. Available: https://proceedings.mlr.press/v97/gordaliza19a.html.

[61] M. Feldman, "Computational Fairness: Preventing Machine-Learned Discrimination". [Online]. Available: https://scholarship.tricolib.brynmawr.edu/handle/10066/17628.

[62] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, *Certifying and removing disparate impact*, 2015. arXiv: 1412.3756 [stat.ML].

[63] B. Fish, J. Kun, and A. D. Lelkes, "Fair boosting: a case study", Citeseer.

[64] B. Fish, J. Kun, and Ádám D. Lelkes, *A Confidence-Based Approach for Balancing Fairness and Accuracy*, 2016. arXiv: 1601.05764 [cs.LG].

[65] S. Hajian and J. Domingo-Ferrer, "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining", *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 7, pp. 1445–1459, 2013. DOI: 10.1109/TKDE.2012.72.

[66] S. Hajian, J. Domingo-Ferrer, A. Monreale, D. Pedreschi, and F. Giannotti, "Discrimination- and privacy-aware patterns", *Data Min. Knowl. Discov.*, vol. 29, no. 6, 1733–1782, 2015. DOI: 10.1007/s10618-014-0393-7. [Online]. Available: http://dx.doi.org/10.1007/s10618-014-0393-7.

[67] J. E. Johndrow and K. Lum, *An algorithm for removing sensitive information: application to race-independent recidivism prediction*, 2017. arXiv: 1703.04957 [stat.AP].

[68] F. Kamiran and T. Calders, "Data Pre-Processing Techniques for Classification without Discrimination", *Knowledge and Information Systems*, vol. 33, Oct. 2011. DOI: 10.1007/s10115-011-0463-8.

[69] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii, *Fair Clustering Through Fairlets*, 2018. arXiv: 1802.05733 [cs.LG].

[70] D. Madras, E. Creager, T. Pitassi, and R. Zemel, *Learning Adversarially Fair and Transferable Representations*, 2018. arXiv: 1802.06309 [cs.LG].

[71] D. McNamara, C. S. Ong, and R. C. Williamson, "Costs and Benefits of Fair Representation Learning", in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '19, Honolulu, HI, USA: Association for Computing Machinery, 2019, 263–270, ISBN: 9781450363242. DOI: 10.1145/3306618.3317964. [Online]. Available: https://doi.org/10.1145/3306618.3317964.

[72] J. Song, P. Kalluri, A. Grover, S. Zhao, and S. Ermon, *Learning Controllable Fair Representations*, 2020. arXiv: 1812.04218 [cs.LG].

[73] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning Fair Representations", in *Proceedings of the 30th International Conference on Machine Learning*, S. Dasgupta and D. McAllester, Eds., ser. Proceedings of Machine Learning Research, vol. 28, Atlanta, Georgia, USA: PMLR, 2013, pp. 325–333. [Online]. Available: https://proceedings.mlr.press/v28/zemel13.html.

[74] I. Zliobaite, F. Kamiran, and T. Calders, "Handling conditional discrimination", English, in *Proceedings 11th IEEE International Conference on Data Mining (ICDM'11, Vancouver BC, Canada, December 11-14, 2011)*, United States: Institute of Electrical and Electronics Engineers, 2011, pp. 992–1001, ISBN: 978-1-4577-2075-8. DOI: 10.1109/ICDM.2011.72.

[75] N. Mehrabi, F. Morstatter, N. Peng, and A. Galstyan, *Debiasing Community Detection: The Importance of Lowly-Connected Nodes*, 2019. arXiv: 1903.08136 [cs.SI].

[76] H. Edwards and A. Storkey, *Censoring Representations with an Adversary*, 2016. arXiv: 1511.05897 [cs.LG].

[77] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez, *Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations*, 2019. arXiv: 1811.08489 [cs.CV].

[78] Y. Wang, T. Koike-Akino, and D. Erdogmus, *Invariant Representations from Adversarially Censored Autoencoders*, 2018. arXiv: 1805.08097 [cs.LG].

[79] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, *The Variational Fair Autoencoder*, 2017. arXiv: 1511.00830 [stat.ML].

[80] M.-E. Brunet, C. Alkalay-Houlihan, A. Anderson, and R. Zemel, *Understanding the Origins of Bias in Word Embeddings*, 2019. arXiv: 1810.03611 [cs.LG].

[81] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. M. Wallach, "A Reductions Approach to Fair Classification", *CoRR*, vol. abs/1803.02453, 2018. arXiv: 1803.02453. [Online]. Available: http://arxiv.org/abs/1803.02453.

[82] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. J. Kearns, J. Morgenstern, S. Neel, and A. Roth, "A Convex Framework for Fair Regression", *CoRR*, vol. abs/1706.02409, 2017. arXiv: 1706.02409. [Online]. Available: http://arxiv.org/abs/1706.02409.

[83] A. Cotter, M. Gupta, H. Jiang, N. Srebro, K. Sridharan, S. Wang, B. Woodworth, and S. You, *Training Well-Generalizing Classifiers for Fairness Metrics and Other Data-Dependent Constraints*, 2018. arXiv: 1807.00028 [cs.LG].

[84] G. Goh, A. Cotter, M. Gupta, and M. P. Friedlander, "Satisfying Real-world Goals with Dataset Constraints", in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper/2016/file/dc4c44f624d600aa568390f1f1104aa0-Paper.pdf.

[85] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness", in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings

of Machine Learning Research, vol. 80, PMLR, 2018, pp. 2564–2572. [Online]. Available: https://proceedings.mlr.press/v80/kearns18a.html.

[86] H. Narasimhan, "Learning with Complex Loss Functions and Constraints", in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, A. Storkey and F. Perez-Cruz, Eds., ser. Proceedings of Machine Learning Research, vol. 84, PMLR, 2018, pp. 1646–1654. [Online]. Available: https://proceedings.mlr.press/v84/narasimhan18a.html.

[87] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, "Fairness Constraints: A Flexible Approach for Fair Classification", *Journal of Machine Learning Research*, vol. 20, no. 75, pp. 1–42, 2019. [Online]. Available: http://jmlr.org/papers/v20/18-262.html.

[88] B. H. Zhang, B. Lemoine, and M. Mitchell, *Mitigating Unwanted Biases with Adversarial Learning*, 2018. arXiv: 1801.07593 [cs.LG].

[89] T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang, "Controlling Attribute Effect in Linear Regression", *2013 IEEE 13th International Conference on Data Mining*, pp. 71–80, 2013.

[90] D. Alabi, N. Immorlica, and A. Kalai, "Unleashing Linear Optimizers for Group-Fair Learning and Optimization", in *Proceedings of the 31st Conference On Learning Theory*, S. Bubeck, V. Perchet, and P. Rigollet, Eds., ser. Proceedings of Machine Learning Research, vol. 75, PMLR, 2018, pp. 2043–2066. [Online]. Available: https://proceedings.mlr.press/v75/alabi18a.html.

[91] Y. Bechavod and K. Ligett, *Penalizing Unfairness in Binary Classification*, 2018. arXiv: 1707.00044 [cs.LG].

[92] K. FUKUCHI, T. KAMISHIMA, and J. SAKUMA, "Prediction with Model-Based Neutrality", *IEICE Transactions on Information and Systems*, vol. E98.D, no. 8, pp. 1503–1516, 2015. DOI: 10.1587/transinf.2014EDP7367.

[93] H. Heidari, M. Loi, K. P. Gummadi, and A. Krause, *A Moral Framework for Understanding of Fair ML through Economic Models of Equality of Opportunity*, 2018. arXiv: 1809.03400 [cs.LG].

[94] L. Hu and Y. Chen, *Fair Classification and Social Welfare*, 2019. arXiv: 1905.00147 [cs.LG].

[95] R. Williamson and A. Menon, "Fairness risk measures", in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., ser. Proceedings of Machine Learning Research, vol. 97, PMLR, 2019, pp. 6786–6797. [Online]. Available: https://proceedings.mlr.press/v97/williamson19a.html.

[96] L. Oneto, M. Donini, A. Elders, and M. Pontil, "Taking Advantage of Multitask Learning for Fair Classification", *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

[97] M. Olfat and A. Aswani, *Spectral Algorithms for Computing Fair Support Vector Machines*, 2017. arXiv: 1710.05895 [cs.LG].

[98] S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, and A. Roth, "Fairness in Reinforcement Learning", in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., ser. Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, pp. 1617–1626. [Online]. Available: https://proceedings.mlr.press/v70/jabbari17a.html.

[99] P. Adler, C. Falk, S. A. Friedler, G. Rybeck, C. Scheidegger, B. Smith, and S. Venkatasubramanian, *Auditing Black-box Models for Indirect Influence*, 2016. arXiv: 1602.07043 [stat.ML].

[100] J. Ali, P. Lahoti, and K. P. Gummadi, "Accounting for Model Uncertainty in Algorithmic Discrimination", *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021. DOI: 10.1145/3461702.3462630. [Online]. Available: http://dx.doi.org/10.1145/3461702.3462630.

[101] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil, *Leveraging Labeled and Unlabeled Data for Consistent Fair Binary Classification*, 2020. arXiv: 1906.05082 [math.ST].

[102] M. Kim, O. Reingold, and G. Rothblum, "Fairness Through Computationally-Bounded Awareness", in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/c8dfece5cc68249206e4690fc4737a8d-Paper.pdf.

[103] A. Noriega-Campero, M. A. Bakker, B. Garcia-Bulle, and A. Pentland, *Active Fairness in Algorithmic Decision Making*, 2018. arXiv: 1810.00031 [cs.CY].

[104] E. Raff, J. Sylvester, and S. Mills, *Fair Forests: Regularized Tree Induction to Minimize Model Bias*, 2017. arXiv: 1712.08197 [stat.ML].

[105] Y. Wu and X. Wu, "Using Loglinear Model for Discrimination Discovery and Prevention", in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2016, pp. 110–119. DOI: 10.1109/DSAA.2016.18.

[106] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*, 2016. arXiv: 1607.06520 [cs.CL].

[107] M. P. Kim, A. Ghorbani, and J. Zou, *Multiaccuracy: Black-Box Post-Processing for Fairness in Classification*, 2018. arXiv: 1805.12317 [cs.LG].

[108] A. Roy, V. Iosifidis, and E. Ntoutsi, *Multi-Fair Pareto Boosting*, 2021. arXiv: 2104.13312 [cs.LG].

[109] A. Matthews, I. Grasso, C. Mahoney, Y. Chen, E. Wali, T. Middleton, M. Njie, and J. Matthews, "Gender Bias in Natural Language Processing Across Human Languages", in *Proceedings of the First Workshop on Trustworthy Natural Language Processing*, Online: Association for Computational Linguistics, Jun. 2021, pp. 45–54. DOI: 10.18653/v1/2021.trustnlp-1.6. [Online]. Available: https://aclanthology.org/2021.trustnlp-1.6.

[110] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, and W. Y. Wang, "Mitigating Gender Bias in Natural Language Processing: Literature Review", in *Proceedings of the 57th Annual Meeting of the*

Association for Computational Linguistics, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1630–1640. DOI: 10.18653/v1/P19-1159. [Online]. Available: https://aclanthology.org/P19-1159.

[111] Y. Hitti, E. Jang, I. Moreno, and C. Pelletier, "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. [Online]. Available: https://aclanthology.org/W19-3802.

[112] K. Stanczak and I. Augenstein, "A Survey on Gender Bias in Natural Language Processing", *CoRR*, vol. abs/2112.14168, 2021. arXiv: 2112.14168. [Online]. Available: https://arxiv.org/abs/2112.14168.

[113] K. Webster, M. Recasens, V. Axelrod, and J. Baldridge, "Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns", *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 605–617, Dec. 2018, ISSN: 2307-387X. DOI: 10.1162/tacl_a_00240. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00240/1567674/tacl_a_00240.pdf. [Online]. Available: https://doi.org/10.1162/tacl_a_00240.

[114] F. T. Asr, M. Mazraeh, A. Lopes, V. Gautam, J. Gonzales, P. Rao, and M. Taboada, "The Gender Gap Tracker: Using Natural Language Processing to measure gender bias in media", *PLOS ONE*, vol. 16, no. 1, pp. 1–28, Jan. 2021. DOI: 10.1371/journal.pone.0245533. [Online]. Available: https://doi.org/10.1371/journal.pone.0245533.

[115] H. Gonen and Y. Goldberg, "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them", in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 609–614. DOI: 10.18653/v1/N19-1061. [Online]. Available: https://aclanthology.org/N19-1061.

[116] A. C. Islam, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora necessarily contain human biases", *CoRR*, vol. abs/1608.07187, 2016. arXiv: 1608.07187. [Online]. Available: http://arxiv.org/abs/1608.07187.

[117] G. AG, M. DE, and S. JL, "Measuring individual differences in implicit cognition: the implicit association test".

[118] K. Ethayarajh, D. Duvenaud, and G. Hirst, "Understanding Undesirable Word Embedding Associations", *CoRR*, vol. abs/1908.06361, 2019. arXiv: 1908.06361. [Online]. Available: http://arxiv.org/abs/1908.06361.

[119] C. Sweeney and M. Najafian, "A Transparent Framework for Evaluating Unintended Demographic Bias in Word Embeddings", in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1662–1667. DOI: 10.18653/v1/P19-1162. [Online]. Available: https://aclanthology.org/P19-1162.

[120] D. de Vassimon Manela, D. Errington, T. Fisher, B. van Breugel, and P. Minervini, "Stereotype and Skew: Quantifying Gender Bias in Pre-trained and Fine-tuned Language Models", in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online: Association for Computational Linguistics, Apr. 2021, pp. 2232–2242. DOI: 10.18653/v1/2021.eacl-main.190. [Online]. Available: https://aclanthology.org/2021.eacl-main.190.

[121] B. Schmidt, "Rejecting the gender binary: a vector-space operation", *Ben's Bookworm Blog*, 2015.

[122] G. Devansh, *tackling-gender-bias-in-ordord-embeddings*, 2020. [Online]. Available: https://towardsdatascience.com/tackling-gender-bias-in-word-embeddings-c965f4076a10.

[123] S. Bordia and S. R. Bowman, "Identifying and Reducing Gender Bias in Word-Level Language Models", *CoRR*, vol. abs/1904.03035, 2019. arXiv: 1904.03035. [Online]. Available: http://arxiv.org/abs/1904.03035.

[124] J. H. Park, J. Shin, and P. Fung, "Reducing Gender Bias in Abusive Language Detection", in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 2799–2804. DOI: 10.18653/v1/D18-1302. [Online]. Available: https://aclanthology.org/D18-1302.

[125] M. Sahlgren and F. Olsson, "Gender Bias in Pretrained Swedish Embeddings", in *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, Turku, Finland: Linköping University Electronic Press, 2019, pp. 35–43. [Online]. Available: https://aclanthology.org/W19-6104.

[126] J. Zhao, Y. Zhou, Z. Li, W. Wang, and K. Chang, "Learning Gender-Neutral Word Embeddings", *CoRR*, vol. abs/1809.01496, 2018. arXiv: 1809.01496. [Online]. Available: http://arxiv.org/abs/1809.01496.

[127] F. Prost, N. Thain, and T. Bolukbasi, "Debiasing Embeddings for Reduced Gender Bias in Text Classification", in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 69–75. DOI: 10.18653/v1/W19-3810. [Online]. Available: https://aclanthology.org/W19-3810.

[128] S. Dev, T. Li, J. M. Phillips, and V. Srikumar, "On Measuring and Mitigating Biased Inferences of Word Embeddings", *CoRR*, vol. abs/1908.09369, 2019. arXiv: 1908.09369. [Online]. Available: http://arxiv.org/abs/1908.09369.

[129] M. Kaneko and D. Bollegala, "Debiasing Pre-trained Contextualised Embeddings", in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online: Association for Computational Linguistics, Apr. 2021, pp. 1256–1266. DOI: 10.18653/v1/2021.eacl-main.107. [Online]. Available: https://aclanthology.org/2021.eacl-main.107.

[130] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality", in *Advances in Neural Information Processing Systems*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., vol. 26, Curran Associates, Inc., 2013.

[Online]. Available: https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

[131] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information", *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017. DOI: 10.1162/tacl_a_00051. [Online]. Available: https://aclanthology.org/Q17-1010.

[132] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation", in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. [Online]. Available: https://aclanthology.org/D14-1162.

[133] S. Dipanjan, *Word2Vec vs GloVe – A Comparative Guide to Word Embedding Techniques*, 2018. [Online]. Available: https://analyticsindiamag.com/word2vec-vs-glove-a-comparative-guide-to-word-embedding-techniques/.

[134] Y. Verma, *Implementing deep learning methods feature engineering text data glove*, 2021. [Online]. Available: https://www.kdnuggets.com/2018/04/implementing-deep-learning-methods-feature-engineering-text-data-glove.html.

[135] S. Kiritchenko and S. Mohammad, "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems", in *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 43–53. DOI: 10.18653/v1/S18-2005. [Online]. Available: https://aclanthology.org/S18-2005.

[136] F. Bianchi, S. Terragni, D. Hovy, D. Nozza, and E. Fersini, "Cross-lingual Contextualized Topic Models with Zero-shot Learning", in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online: Association for Computational Linguistics, Apr. 2021, pp. 1676–1683. DOI: 10.18653/v1/2021.eacl-main.143. [Online]. Available: https://aclanthology.org/2021.eacl-main.143.

[137] H. Levesque, "The Winograd Schema Challenge.", Jan. 2011.

[138] R. Rudinger, J. Naradowsky, B. Leonard, and B. Van Durme, "Gender Bias in Coreference Resolution", in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 8–14. DOI: 10.18653/v1/N18-2002. [Online]. Available: https://aclanthology.org/N18-2002.

[139] G. Stanovsky, N. A. Smith, and L. Zettlemoyer, "Evaluating Gender Bias in Machine Translation", in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1679–1684. DOI: 10.18653/v1/P19-1164. [Online]. Available: https://aclanthology.org/P19-1164.

[140] D. Saunders, R. Sallis, and B. Byrne, "Neural Machine Translation Doesn't Translate Gender Coreference Right Unless You Make It", *CoRR*, vol. abs/2010.05332, 2020. arXiv: 2010.05332. [Online]. Available: https://arxiv.org/abs/2010.05332.

[141] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, S. A. Friedler and C. Wilson, Eds., ser. Proceedings of Machine Learning Research, vol. 81, New York, NY, USA: PMLR, 2018, pp. 77–91. [Online]. Available: http://proceedings.mlr.press/v81/buolamwini18a.html.

[142] K. Crawford, *The trouble with bias*, 2017. [Online]. Available: https://blog.revolutionanalytics.com/2017/12/the-trouble-with-bias-by-kate-crawford.html.