Stefanie Urchs

# Detecting Gender Discrimination in Natural Language Processing

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht am 21.08.2025

Stefanie Urchs

# Detecting Gender Discrimination in Natural Language Processing

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Tag des Antrags auf Zulassung zum Promotionsverfahren 14.07.2025
Eingereicht am 21.08.2025

Erster Berichterstatter: Prof. Dr. Christian Heumann
Zweite Berichterstatterin: Prof. Dr. Stephanie Thiemichen
Dritte Berichterstatterin: Prof. Dr. Eirini Ntoutsi

Tag der Disputation: 08.12.2025

# Acknowledgments

*I would like to express my sincere gratitude to …*

- … *Prof. Dr. Stephanie Thiemichen, Prof. Dr. Christian Heumann, Prof. Dr. Veronika Thurner, and Dr. Matthias Aßenmacher for their supervision, open-minded mentorship, and intellectually stimulating guidance throughout this thesis. I am deeply thankful for the academic freedom I was granted, the trust they placed in my ideas, and the thoughtful support they offered whenever I sought their input.*
- … *Prof. Dr. Eirini Ntoutsi for her willingness to act as the third reviewer for my Ph.D. thesis.*
- … *Prof. Dr. Göran Kauermann and Prof. Dr. Michael Schomaker for their availability to be part of the examination panel at my Ph.D. defense.*
- … *Prof. Dr. Gabriele Fischer, Prof. Dr. Elke Wolf, Ronja Philipp, Lina Spagert, and Kushtrim Hamzaj from the Prof:inSicht project team for their truly inspiring collaboration. Working in this interdisciplinary environment was an intellectually and personally transformative experience. The many thought-provoking discussions and the open, collegial atmosphere created one of the most rewarding academic settings I have ever encountered. I am deeply grateful for everything I learned during this time and for the opportunity to be part of such a meaningful project. Without this experience, this thesis would not have been possible.*
- … *Dr. Markus Hoffmann and Dr. Jasmijn Bastings for their generous mentorship, insightful input on both my research and personal development, and the many enriching conversations we shared. I feel truly fortunate that they took the time to support and guide me along this journey.*
- … *my partner Hayato Hess, who stood by my side through every stage of this journey. Thank you for listening to my endless rambling about research, for cheering me on, and for being my anchor in difficult times. Your unwavering support meant more to me than words can express.*
- … *Prof. Dr. Isabella Graßl, with whom I had the privilege of sharing the PhD journey. Our regular calls kept me going, and her insightful ideas and thoughtful feedback were invaluable to the development of this thesis. Beyond that, she has become one of my closest friends throughout this time, and I am deeply grateful for her presence in both my academic and personal life.*
- … *my friends Jonas Albrecht, Dr. Richard Rotermund, Melanie Hörl, Verena Ertl, Anna-Katharina Nöcker, and Christina Forst, who accompanied me throughout this entire journey. Thank you for your constant encouragement, for listening patiently, and for offering the kind of personal support that kept me grounded. Without you, life (and this thesis) would have been far less fun.*
- … *my family for their genuine interest in my research and for their continuous encouragement throughout this journey. Your support has meant a great deal to me.*
- … *my cats Lilly and Lolly, who ensured I took regular breaks: whether for a cuddle, a play session, or simply to be reminded that there's more to life than writing. Their companionship was an integral (and often purring) part of this thesis.*

# Summary

Technology is an essential part of human life. It helps us be more productive and shapes the way we live. Modern technology is composed of algorithms, but modern algorithms not only benefit society, they can also harm it. By misrepresenting certain genders or reinforcing stereotypes, algorithms can contribute to discrimination. The field of natural language processing (NLP) is particularly prone to such issues. This doctoral thesis analyses gender discrimination in NLP and proposes a way to make it, specifically large language models (LLM), less gender discriminatory by improving their training data.

The first part of this thesis defines what algorithmic gender fairness means. This definition is then applied to analyse the results of information retrieval methods and the results of search algorithms. This analysis reveals that the representation of different genders in algorithmic output remains insufficient, underscoring the need to improve current approaches.

Building on the foundation of fairness in information retrieval and search, the focus then shifts to LLMs, a rapidly evolving technology that increasingly shapes everyday life. The model GPT-3 (Brown et al., 2020), as implemented in the system ChatGPT (OpenAI, 2022), is analysed with regard to how it responds to prompts in English and German from a female, male, or neutral perspective. The analysis of the prompt results shows that attempts to reduce gender discrimination after training can introduce new problems, for example, an over-representation of female personas in response to neutral prompts or an exaggerated emphasis on diversity in gendered prompts. These findings suggest that "downstream" mitigation, after model training, is not the right approach. Instead, mitigation should be done "upstream", before training, by improving the quality of the training data itself.

This is addressed in the third part of the thesis. The first publication of this part introduces a modular, language-agnostic pipeline designed to detect discrimination in English newspaper texts. This pipeline combines linguistic discourse analysis with computational techniques. Using information extraction methods, it identifies the actors mentioned in a text, how they are referred to (nomination), and how they are described (predication). Therefore, it is possible to analyse quantitative metrics for each gender in the text and, additionally, qualitative metrics like the sentiment towards actors of each gender. The pipeline is scaled up in a second publication to process an entire corpus of German newspaper articles. This work also publishes the most significant German newspaper corpus to date, spanning four decades and comprising 1.8 million texts. A third publication further extends the pipeline and utilises it to generate a gender-balanced corpus, drawing on the German newspaper corpus from the second publication.

# Zusammenfassung

Technologie ist ein zentraler Bestandteil des modernen Lebens. Sie steigert unsere Produktivität und beeinflusst maßgeblich unsere Lebensweise. Gleichzeitig bergen algorithmische Systeme nicht nur Potenziale, sondern auch Risiken für die Gesellschaft. So können sie etwa Geschlechter-representationen verzerren oder bestehende Stereotype verstärken und dadurch Diskriminierung begünstigen. Besonders im Bereich der Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) treten solche Problematiken verstärkt auf. Die vorliegende Dissertation befasst sich mit unterschiedlichen Teilbereichen des NLP und untersucht, wie geschlechtsspezifische Diskriminierung in diesen entstehen kann. Darüber hinaus wird ein Ansatz vorgestellt, um einen konkreten Bereich, Large Language Models (LLM), durch eine Analyse der Trainingsdaten weniger diskriminierend zu gestalten.

Im ersten Teil wird das Konzept der algorithmischen Gender-Gerechtigkeit definiert. Diese Definition dient anschließend als Grundlage für die Analyse von Ergebnissen von Information-Retrieval-Systemen und Suchmaschinenergebnissen. Die Analyse zeigt, dass die algorithmische Repräsentation verschiedener Geschlechter nach wie vor unzureichend ist, was auf die Notwenigkeit der Verbesserung von gängigen Vorgehen hinweist.

Auf Grundlage dieser Erkenntnisse richtet sich der Fokus im zweiten Teil der Arbeit auf LLMs, eine Technologie, die rasant in allen Lebensbereichen adaptiert wird. Konkret wird das LLM GPT-3 (Brown et al., 2020), in seiner Implementierung im System ChatGPT (OpenAI, 2022) betrachtet. Es wird analysiert, wie sich das System bei Prompts verhält, die aus weiblicher, männlicher oder neutraler Perspektive formuliert sind, sowohl auf Deutsch als auch auf Englisch. Die Analyse zeigt, dass der Versuch, geschlechtsspezifische Diskriminierung nachträglich aus dem System zu entfernen, problematisch sein kann. Die Antworten des Systems neigten zu Überkorrekturen: Bei neutralen Prompts wurden vermehrt weibliche Personen generiert und Prompts, die eine weibliche oder männliche Sichtweise einnahmen, führten zur Überbetonung von Diversität. Diese Beobachtungen machen deutlich, dass eine Korrektur nach dem Modelltraining (“downstream”) nicht ausreicht. Stattdessen sollte bereits vor dem Training (“upstream”) angesetzt werden, durch eine gezielte Aufbereitung und Verbesserung der Trainingsdaten.

Um das Problem der geschlechtsspezifischen Diskriminierung bereits vor dem Modelltraining anzugehen, befasst sich der dritte Teil der Dissertation mit sogenannten „upstream“-Mitigation-Ansätzen. In einer ersten Publikation wird eine modulare, sprach-agnostische Pipeline zur Erkennung von Diskriminierung in Zeitungstexten entwickelt. Ziel ist es, diskriminierende Muster frühzeitig in Trainingsdaten aufzudecken. Die Pipeline kombiniert Methoden der linguistischen Diskursanalyse mit informatischen Verfahren und nutzt Ansätze aus der Information Extraction, um die zentralen Akteur:innen eines Textes zu identifizieren, ihre Benennung (Nomination) sowie ihre Darstellung im Text (Prädikation) zu erfassen. Dadurch lassen sich sowohl quantitative Metriken im Bezug auf die Geschlechterverteilung im Text, als auch qualitative Aspekte wie das Sentiment gegenüber im Text genannten Geschlechtern analysieren. In einer anschließenden Publikation wird die Pipeline auf einen Korpus deutschsprachiger Zeitungstexte angewendet. Im Zuge dieser Publikation wird der bislang größte, öffentlich zugängliche, deutschsprachige Zeitungskorpus veröffentlicht. Dieser umspannt vier Jahrzehnte und besteht aus rund 1,8 Millionen Texten. Eine abschließende Publikation erweitert wiederum die Pipline der vorhergehenden Publikation und nutzt diese, um auf Grundlage des deutschprachigen Zeitungskorpuses einen Gender-ausgeglichenen Korpus zu generieren.

# Contents

# List of Figures

# Part I.

# Introduction and Background

# 1. Introduction

This thesis was written during a time of profound change, both in the realm of technology and in global politics. The emergence and widespread adoption of generative artificial intelligence (genAI), such as ChatGPT, has made artificial intelligence (AI) accessible to the broader public. For the first time, individuals without a background in computer science are directly interacting with sophisticated and powerful AI models. As this technology becomes increasingly ubiquitous, it promises to transform not only mundane tasks but also creative fields such as programming, writing, and art. At the same time, however, the political climate appears to be moving in the opposite direction. While technological progress points to the future, political discourse in many parts of the world seems to be retreating into the past. The resurgence of the political right has brought with it a renewed opposition to gender equality and the empowerment of minorities.

Against this backdrop, this thesis aims to contribute to a fairer future by addressing discrimination in training data for large language models. By identifying and mitigating discrimination in these corpora, we can work toward the development of more equitable AI systems.

## 1.1. Outline

This thesis draws on insights from social sciences, linguistics, and computer science. It combines theoretical concepts of discrimination with applied computational methods to investigate how discrimination manifests in artificial intelligence, particularly in large language models. From social sciences and linguistics, it adopts conceptual frameworks for understanding discrimination as a structural and discursive phenomenon. From computer science, it applies and develops computational approaches to analyse large textual datasets. The research was conducted within the context of the interdisciplinary Prof:inSicht project, which examines the visibility of female professors at universities of applied sciences and provided both inspiration and practical grounding for this work.

Part I lays the conceptual groundwork. It begins with ideas from the social sciences, gender (Section 2.1) and the notions of discrimination (Section 2.2). These are followed by an introduction to discourse analysis (Section 2.3), which transitions into computational approaches with a discussion of computational discrimination analysis (Section 2.4). The thesis then introduces the computer science concepts of information retrieval (Section 2.5) and large language models (Section 2.6). The research trajectory is outlined in Section 3 to provide context for the contributions.

Parts II to IV present the main contributions of this thesis. Part II introduces the concept of algorithmic gender fairness and explores its application in the context of information retrieval and search technologies. Building on this foundation, the Part III turns to large language models, examining the extent and nature of gender discrimination in the system ChatGPT. The final part (Part IV) moves from analysis to action, presenting a mitigation strategy in the form of a language-agnostic and flexible pipeline for detecting discrimination in text and gender-balancing a corpus, demonstrated through its application to a large-scale news corpus.

The thesis concludes with Part V, which summarises the main findings, outlines directions for future research and reflects on the work.

## 1.2. Motivation and Scope

This thesis is motivated by the need to make gendered discrimination in language technologies visible and addressable. It builds on insights from the interdisciplinary project Prof:inSicht, which examined the digital visibility of female professors in German universities of applied sciences and the systemic disadvantages they face. Within this context, I explored how algorithmic systems influence visibility and developed the concept of algorithmic gender fairness, which laid the groundwork for the present focus on training data in generative language models. Large language models (LLMs) increasingly shape how information is accessed and identities are represented. However, the textual corpora on which they are trained are not neutral. They often encode historical and structural inequalities, particularly in relation to gender. In Germany, for example, legal and cultural restrictions limited women's autonomy well into the 20th century, contributing to their underrepresentation in professional contexts and overrepresentation in caregiving roles (Die Bundesregierung, 2024). These historical patterns persist today: women still perform significantly more unpaid care work than men, with a current gender care gap of 44.3% (Bundesministerium für Familie, Senioren, Frauen und Jugend, 2024). Because LLMs are trained on large corpora drawn from different periods and domains, these inequalities remain embedded in the data, amplifying gendered stereotypes even when models appear neutral on the surface. The societal impact of LLMs extends beyond technical applications. These systems influence communication, knowledge production, and professional authority by shaping the public opinion (Lippmann, 1992). Yet access to them, and the ability to use them effectively, is unequally distributed. At the same time, the models themselves shape how language is used, creating a feedback loop that risks reinforcing social inequalities in who is seen as articulate or authoritative. Addressing discrimination in language technologies therefore requires more than post-hoc adjustments to model outputs. This thesis shifts the focus to training data as a site of intervention. It combines methods from linguistics, the social sciences, and computational analysis to uncover representational asymmetries at scale. The result is a modular, language-agnostic pipeline for corpus analysis and balancing. It provides interpretable reports on gender representation and tools to systematically adjust dataset composition.

The scope of this thesis is limited to gender-based discrimination in textual data. Other identity dimensions (e.g., race or class), data modalities (e.g., image or audio), and genres (e.g., social media or fiction) are not addressed. Likewise, no new model-level fairness interventions are proposed. Instead, the focus lies on developing a practical, interdisciplinary approach to identifying and mitigating gender discrimination in text corpora, with the aim of improving the data foundations on which language technologies are built.

# 2. Methodological and General Background

Given the interdisciplinary nature of this thesis, this chapter provides essential background from social sciences, linguistics, and computer science. It introduces key concepts related to gender, discrimination, and discourse analysis, followed by methodological foundations in information extraction and large language models.

The background sections are designed to provide the reader with the necessary context to understand the contributions of this thesis and how they relate to one another. The theoretical foundations are not presented exhaustively and are by no means complete.

## 2.1. Gender

The term **gender** refers to at least three different but interrelated concepts from different disciplines: **linguistic gender** from linguistics, **sex** from biology, and **social gender** from the social sciences. Each of these plays a distinct role in how people are represented and perceived, both in everyday life and in algorithmic systems.

### 2.1.1. Linguistic Gender

**Linguistic or grammatical gender** refers to the categorisation of nouns and pronouns into gendered classes, typically masculine, feminine, and occasionally neutral. These categories are only loosely related to biological sex and often follow inconsistent or non-intuitive patterns (Kramer, 2020). For example, the German word *Mädchen* (girl) is grammatically neutral, illustrating that grammatical gender does not necessarily align with the gender of the referent. Janhunen (2000) defines grammatical gender as follows: "*[...] grammatical gender in the narrow sense, which involves a more or less explicit correlation between nominal classes and biological gender (sex)*". Grammatical gender systems can influence cognition by reinforcing cultural expectations and stereotypes (Konishi, 1993; Phillips and Boroditsky, 2013). Experimental studies show that speakers often describe objects using adjectives that reflect the grammatical gender of the noun. In one study, German speakers (for whom "bridge" is feminine) described bridges as "beautiful" or "elegant," while Spanish speakers (for whom the noun is masculine) used terms like "strong" or "sturdy" (Boroditsky et al., 2003). Such findings suggest that linguistic gender categories can shape perception and subtly reinforce gendered associations even in non-human contexts.

### 2.1.2. Biological Sex

**Sex**, on the other hand, is traditionally understood as a biological categorisation, regarded as *"binary, immutable and physiological"* (Keyes, 2018). However, current research in genetics, endocrinology, and developmental biology increasingly challenges this rigid binary view. Sex is not determined by a single factor, such as chromosomes, but by the interplay of multiple biological components: chromosomal patterns, gonadal structures, hormone levels, and secondary sexual characteristics. These components do not always align in a binary fashion (Ainsworth, 2015). Intersex individuals, those with differences in sex development (DSDs), may have combinations of male and female traits, such as XY chromosomes and a uterus, or ambiguous genitalia (Carpenter, 2021). Some estimates suggest that up to 1 in 100 people may exhibit some form of DSD (Ainsworth, 2015). In addition, studies have revealed that even within one body, genetic mosaicism or chimaerism can lead to different cells having different sex chromosomal compositions (Ainsworth, 2015). These findings underscore that **biological sex** exists on a spectrum rather than as a dichotomy. The existence of intersex and transgender individuals, whose lived experience or physiological traits do not fit conventional definitions, highlights the limitations of defining sex as fixed, binary, or solely biologically determined.

### 2.1.3. Social Gender

This thesis adopts the concept of **social gender**, which goes beyond biological and grammatical categories to describe gender as a socially constructed identity. Social gender is shaped through behaviour, expression, and interaction, and may change over time in alignment with an individual's self-perception. A key framework for this understanding is *doing gender*, introduced by West and Zimmerman (1987). From this perspective, gender is not a fixed attribute but an ongoing accomplishment produced through everyday activities. Individuals engage in socially interpreted gendered behaviours, which are evaluated by others according to prevailing norms. This process of accountability encourages conformity to culturally expected gender performances (West and Zimmerman, 1987). Bourdieu's theory of social practice provides a complementary structural view (Scherr, 2016). His concept of *habitus* explains how repeated exposure to norms shapes dispositions, rendering certain behaviours 'natural'. Through symbolic power, dominant groups define which forms of appearance, language, or conduct are legitimate. As a result, gendered practices, such as speaking style or body language, are not merely individual choices, but shaped by social position and internalised norms. These practices, once embodied, help reproduce structural inequalities (Scherr, 2016).

Together, these frameworks highlight gender as a dynamic, socially enacted process embedded in structures of power. Language plays a central role in this process: it not only reflects gendered norms but actively reproduces them. Yet much of natural language processing (NLP) research simplifies this complexity, conflating social gender with sex or grammatical gender and often adopting a binary, static view (Devinney et al., 2022). This excludes trans, intersex, and non-binary individuals and overlooks the interactional and structural dynamics of gender. In contrast, the present work draws on sociological and linguistic theory to treat gender as fluid, interactional, and co-constructed. It emphasises how gendered practices are shaped by context, embodied through discourse, and sustained by structural forces. This perspective provides a more inclusive and critical framework for analysing gender representation and understanding the roots of discrimination in text.

Despite this theoretical stance, the empirical analysis is constrained by data availability. The `taz2024full` corpus contains too few references beyond the binary gender spectrum, and the experimental data used in the fairness analysis in Chapter 4 reflects similar limitations. These are methodological constraints rather than theoretical positions. Where possible, the analysis remains extensible and the pipeline has been designed to accommodate more inclusive gender representations as data availability improves.

## 2.2. Discrimination

Understanding how discrimination can manifest in language technologies requires a clear distinction between the related but different concepts of **bias**, **fairness**, and **discrimination**. These terms are often used together in public and technical discussions, but in this thesis, they are treated as distinct concepts that refer to different stages and effects in the development and deployment of language models.

### 2.2.1. Definition of Bias

In machine learning, the term **bias** typically refers to imbalances or distortions in data or model behaviour. These biases may reflect historical inequalities, sampling errors, or social stereotypes (Mehrabi et al., 2021). Bias is often introduced through training data (Roselli et al., 2019), but it can also emerge from model architecture (Roselli et al., 2019), optimisation procedures (Roselli et al., 2019), or user interactions (Wolf et al., 2017). In the context of NLP, bias becomes visible in word associations (Bolukbasi et al., 2016), naming conventions (Pawar et al., 2025), or in which voices and topics are more or less represented in large-scale corpora (Naous et al., 2024). Mateo and Williams (2020) define bias as follows: "*Biases are preconceived notions based on beliefs, attitudes, and/or stereotypes about people pertaining to certain social categories that can be implicit or explicit.*". They further note that discrimination is the manifestation of such biases through behaviour and actions. **Bias** itself is not always harmful, but it becomes problematic when it leads to systematic disadvantages for certain individuals or groups. For example, models may associate women more frequently with emotions and appearance, while men are linked to professions or leadership. Such associations reflect existing social patterns but risk reinforcing them when reproduced by automated systems.

In this thesis, the focus lies on detecting *discrimination*, understood as the realisation of bias in textual behaviour, rather than on identifying bias in isolation. Even if the training data of large language models contains various biases, the methods developed here detect their discriminatory manifestations.

### 2.2.2. Definition of Fairness

**Fairness** refers to the effort to identify, understand, and mitigate differences in treatment that are considered normatively problematic. However, fairness is not a fixed or universal concept. It depends on social, cultural, and technical context, and many definitions have been proposed in the machine learning community. In classification tasks, fairness is often defined as ensuring that two otherwise similar individuals are treated similarly by the model, regardless of their group membership (Dwork et al., 2012)[1]. The term **fairness** is now widely used in the context of algorithmic decision-making and has become central to the field of fairness-aware machine learning (Caton and Haas, 2024; Verma and Rubin, 2018). However, few contributions explicitly define fairness as a philosophical concept, with some notable exceptions (Bothmann et al., 2024; Loi and Heitz, 2022; Kong, 2022). **Fairness** is typically described using related terms such as equality, justice, or the absence of bias or discrimination. A crucial element across these definitions is that fairness concerns the treatment of individuals (Aristotle, 2009; Dator, 2017).

The structure of the concept can be traced back to Aristotle's idea that fairness means treating equals equally and unequals unequally. As Bothmann et al. (2024) point out, this requires a normative definition of task-specific equality. Two individuals may be equal in one context (e.g., buying a croissant in a bakery) but unequal in another (e.g., paying taxes). The question of how to treat unequals is itself a normative decision. Protected attributes like gender or race play a central role in this discussion because they may justifiably alter what is considered equal treatment in a specific context. For example, a society may normatively decide that the gender pay gap is unjust and that this injustice should not influence other decisions. In such a case, income data could be corrected to remove the effects of gender-based pay discrimination when deciding on loan eligibility. Bothmann et al. (2024) describe this as a decision made in a "*fictitious, normatively desired (FiND) world*". They argue that it is this world, rather than the real one, that should serve as the basis for fair decision-making. Similarly, (Wachter et al., 2021) describe this type of approach as bias-transforming, aiming at substantive equality by actively correcting for existing real-world inequalities. However, it is highly questionable what a corrected or ideal version of the real world, a so-called perfect world, should look like. Notions of what is fair or desirable differ significantly between individuals, depending on factors such as upbringing, culture, place and time of birth, and socioeconomic background (Gross, 2008). Due to this diversity, assuming that a single normative ideal can represent everyone's idea of **fairness** is unrealistic. As a result, applying a bias-transforming approach that relies on a predefined notion of a better world is problematic in practice. It requires value judgments that cannot be universally agreed upon, especially in global or large-scale systems.

Bias-preserving approaches differentiate from bias-transforming approaches by aiming at formal equality. These methods try to reflect the real world as accurately as possible, without introducing new distortions (Wachter et al., 2021). Many **fairness** metrics in machine learning, such as equalised odds or predictive parity, fall into this category. They work with observed (and often biased) real-world labels but attempt to balance error rates across protected groups. Sometimes these two approaches are framed as equity (bias-transforming) versus equality (bias-preserving).

### 2.2.3. Definition of Discrimination

**Discrimination**, in contrast to bias, denotes the manifestation of unequal treatment, where individuals or groups are systematically disadvantaged due to sensitive attributes such as gender, race, or age (Mehrabi et al., 2021). In this thesis, I focus on discrimination as it appears in text, particularly how people are represented and talked about in language.

[1] A deeper discussion of statistical fairness metrics can be found in Chapter 2.4.1

To define **discrimination** in language, I draw on the definition proposed by Reisigl (2017). He describes discrimination as a situation in which someone is treated unequally based on a specific characteristic, such as gender or sexual orientation. This unequal treatment must occur through an act or process and be directed at a person or group. According to the framework by Reisigl (2017), five elements must be present:

1. **Offender**: the person or institution carrying out the act
2. **Victim or beneficiary**: the person or group who is disadvantaged or favoured
3. **Discriminatory act or process**, which I further specify following the functional model by Graumann and Wintermantel (2007). This component includes:
   - **Separate**: marking someone as different or assigning them to a specific category
   - **Distance**: emphasising social distance, for example through "us" and "them" constructions
   - **Accentuate**: exaggerating group differences
   - **Devalue**: using negative, mocking, or degrading language
   - **Typologise**: applying fixed categories or stereotypical labels
4. **Comparison group**: another group that receives different treatment
5. **Distinguishing feature**: the attribute on which the unequal treatment is based (in this case: gender)

I apply this framework to the analysis of written text. In my case, the **offender** is the author of the text, and the **victim** is a person or group mentioned in it. I detect discriminatory acts by analysing how people are talked about. This includes how they are named (nomination) and what is said about them (predication). The **comparison group** is formed by other gender groups mentioned in the same corpus. The **distinguishing feature** is gender, which I classify based on pronouns and other linguistic markers. This understanding is further shaped by the functional model of discriminatory language acts proposed by Graumann and Wintermantel (2007). According to this model, **discrimination** in language can be subtle and indirect. It may appear in the form of repeated patterns, choices of words, or associations that reflect societal hierarchies. These patterns do not need to be intentional. Language reproduces what is socially normal, and this can include inequalities.

## 2.3. Discourse Analysis

Discourse analysis offers a set of conceptual and analytical tools to examine how meaning is constructed, communicated, and negotiated through language. It provides insight into the reproduction of social structures and ideologies, particularly in the context of power relations Bendel Larcher (2015). In this thesis, I draw on two main traditions for the development of my discrimination detection pipeline: linguistic discourse analysis (LingDA), rooted in the humanities, critical social theory, and computational Discourse Analysis (CompDA), which stems from NLP and formal semantics. These approaches complement each other by enabling both close and large-scale readings of text corpora.

### 2.3.1. Linguistic Discourse Analysis

LingDA analyses how language contributes to the construction of social reality. It builds on work from pragmatics, sociolinguistics, and critical theory, and understands texts not as neutral carriers of meaning, but as elements that actively shape public discourse. In this context, discourse is defined as a collection of socially situated texts that negotiate knowledge about a given topic. These texts do not simply reflect discourse, they help constitute it (Bendel Larcher, 2015).

A central approach within LingDA is critical discourse analysis (CDA), which focuses on how language contributes to the reproduction of power relations, ideologies, and social inequality (Fairclough, 2012). Unlike other linguistic frameworks, CDA does not start with a linguistic phenomenon but with a social issue, such as sexism, racism, or classism. It examines how these issues are maintained or challenged through language. CDA examines all forms of communication about a given topic, including speech, written texts, images and all sorts of multimodal data.

Discourse can be analysed at various levels, including individual texts, large-scale corpora, and the broader societal discourse (Bendel Larcher, 2015). Central to CDA is the analysis of how social actors are constructed within discourse. According to Bendel Larcher (2015), six dimensions are particularly relevant:

1. **Perspective**: Who is speaking, and from what position?
2. **Nomination and Predication**: How are actors named and described?
3. **Topic Structure**: What is talked about and in what order?
4. **Modality**: How certain or uncertain are the statements?
5. **Evaluation**: What value judgements are made?
6. **Argumentation**: How are claims supported?

In this thesis, I focus on the nomination and predication of social actors, which are particularly relevant for analysing discrimination.

**Nomination** refers to how actors are named in discourse. Naming practices are neither neutral nor random; they position actors within social hierarchies and imply relationships of familiarity, authority, or otherness (Knobloch, 1996). Several forms of nomination can be distinguished:

- **Proper Names**: The use of full names, first names, or surnames can signal levels of respect or intimacy. In a German context, addressing unfamiliar adults by their first name can be perceived as impolite or even disrespectful (Bendel Larcher, 2015).
- **Generic Terms**: Instead of naming individuals, texts may refer to groups using general descriptors (Bendel Larcher, 2015). Some of these can be problematic or discriminatory. Reisigl (2017) identify several categories of concern:
  - **Negatively connoted general descriptions** (e.g., “wench”)
  - **Ethnonyms** used as identity reductions (e.g., “Jew”, “Muslim”)
  - **Metaphorical slurs** (e.g., “pussy”, “whore”)

  - **Animalistic metaphors** (e.g., “snake”, “parasite”)
  - **Synecdochic naming** using stereotypical proper names (e.g., “Ivan” for Russians)
  - **Relational identification** (e.g., referring to Simone de Beauvoir merely as “Sartre’s partner”)

- **Pronouns**: Pronouns like “we” and “they” can create in-groups and out-groups. Misgendering, i.e., using incorrect pronouns, constitutes a form of symbolic violence. In a German context, the “generic masculine” is particularly problematic: it does not directly address women and non-binary individuals but only implies their inclusion, leaving them unrepresented in the language (Bendel Larcher, 2015).
- **Deagentification**: The agent of an action is omitted. For example, “mistakes were made” obscures responsibility. This strategy is often used to erase agency in discourses of violence, failure, or discrimination (Bendel Larcher, 2015).

**Predication** refers to the attribution of characteristics, roles, or actions to an actor. These linguistic choices shape how the audience perceives that actor (Kamlah and Lorenzen, 1996; Reisigl, 2017). Predication can be realised through:

- **Attributes**: Adjectives or descriptive phrases (e.g., “a clever girl”, “an angry woman”)
- **Prepositional Attributes**: Phrases linked by prepositions (e.g., “the manager from Berlin”)
- **Collocations**: Recurrent word combinations that carry stereotypical connotations (e.g., “working mom”, “bossy woman”)
- **Relative Clauses**: Additional information that may convey bias (e.g., “the scientist who cried during the interview”)

Predication also includes agency patterns: Is the actor portrayed as acting or being acted upon? Passive constructions, for instance, can downplay victimhood or responsibility.

This dual lens of **nomination** and **predication** enables the detection of subtle discursive patterns that reinforce social hierarchies or stereotypes. It proofed powerful in identifying how language naturalises and legitimises discrimination.

### 2.3.2. Computational Discourse Analysis

CompDA approaches discourse from a formal and often large-scale perspective. It investigates how meaning unfolds across sentences and documents by examining structural and semantic relations between units of text (Dascalu, 2014). Unlike LingDA, which focuses on meaning-making within social contexts, CompDA prioritises measurable properties such as coherence and cohesion.

**Cohesion** refers to the surface-level connectedness of a text. It is achieved through lexical and grammatical devices that link sentences and clauses. These include coreference chains (e.g., “Maria... she...”), lexical overlap, semantic similarity, and discourse connectives (e.g., “because”, “therefore”) (Dascalu, 2014). Cohesion is often divided into:

- **Referential Cohesion**: The recurrence or semantic relatedness of terms across a text.
- **Causal Cohesion**: The use of explicit markers to indicate causal or logical relationships.

**Coherence** , by contrast, refers to the underlying semantic unity of a text. It is not merely about surface structure but about whether the reader can infer a consistent mental model across the discourse (De Beaugrande and Dressler, 1981). Dascalu (2014) distinguish two levels:

- **Informational Coherence**: Logical progression of ideas, use of lexical chains, and maintenance of topical focus.
- **Intentional Coherence**: Changes in the mental states or goals of discourse participants, often modelled via dialogue structures or narrative theory.

While cohesion can be measured with tools from NLP (e.g., coreference resolution, word embeddings, or discourse parsers), coherence remains a more elusive and interpretive property (Dascalu, 2014).

In this thesis, computational methods are used to complement the nomination and predication analysis. Building on established concepts of cohesion and coherence, I apply these methods to trace how actors and their associated characteristics are referred to across texts, enabling an examination of their consistency, prominence, and changes in portrayal. In doing so, I extend CompDA by integrating analytical categories from LingDA, specifically nomination and predication, into computational workflows. This hybrid approach allows for both qualitative and quantitative insights into how discrimination is embedded and reproduced through language.

## 2.4. Computational Discrimination Analysis

Discrimination in language is not limited to overt slurs or individual speech acts; it also emerges from subtle, large-scale patterns in how groups are represented and described. When such patterns are embedded in algorithmic systems, they can shape how people are categorised and evaluated, often reinforcing social inequalities. Computational discrimination analysis investigates how algorithmic processes and statistical models contribute to, or mitigate, these dynamics, with particular attention to gender. In contrast to interpretive approaches in linguistics or social sciences, computational methods rely on large datasets and formal metrics (Mehrabi et al., 2021).

### 2.4.1. Statistical Fairness in Classification Tasks

In supervised classification, discrimination is often understood as systematic disparities in model outcomes between groups defined by protected attributes such as gender. Within computational research, these disparities are commonly operationalised using *fairness metrics*, which offer formal criteria for detecting and mitigating unequal treatment. Although such metrics are normative approximations rather than comprehensive definitions of non-discrimination (Barr et al., 2025), they have become the standard reference point for evaluating fairness in algorithmic systems. Even though the methods developed in this thesis detect discrimination in textual data rather than measuring fairness in classification outputs, statistical fairness metrics provide a useful conceptual backdrop for situating this work in the broader discourse on algorithmic fairness. Three main families of metrics dominate this area: individual fairness, which focuses on treating similar individuals similarly; group fairness, which requires parity across demographic groups; and subgroup fairness, which extends these guarantees to more granular intersections (Mehrabi et al., 2021; Verma and Rubin, 2018).

**Individual Fairness.** Individual fairness assesses whether similar individuals receive similar treatment. Unlike group-based approaches, it operates on a per-individual basis and requires either predefined similarity metrics or causal reasoning about identity.

Let $X$ denote the set of individuals, $A$ a protected attribute (e.g., gender) with values $a, a' \in A$, and $f : X \to \Delta(Y)$ a (possibly probabilistic) classifier mapping individuals to outcome distributions. A key requirement is that $f$ treats similar individuals $x, x' \in X$ similarly, where similarity is defined by a domain-specific distance function $d : X \times X \to \mathbb{R}_{\geq 0}$. Prediction similarity is measured using a distance $D(P, Q)$ between outcome distributions $P$ and $Q$, such as total variation distance $D_{\mathrm{TV}}(P, Q) = \frac{1}{2}\sum_{y \in Y} |P(y) - Q(y)|$ or relative infinity distance $D_\infty(P, Q) = \sup_{y \in Y} \log\left(\max\left\{\frac{P(y)}{Q(y)}, \frac{Q(y)}{P(y)}\right\}\right)$ (Dwork et al., 2012).

*Fairness Through Awareness* (Dwork et al., 2012) requires that similar individuals receive similar treatment. Formally, for a classifier $f$ over input space $X$ and output space $Y$, the following condition must hold:

$$D(f(x), f(x')) \leq d(x, x') \quad \forall x, x' \in X$$

where $d(x, x')$ measures the similarity between individuals and $D(f(x), f(x'))$ the difference in their predicted outcomes. As an example, let's consider two loan applicants $x$ and $x'$ with nearly identical financial profiles e.g. incomes of €40,000 and €41,000, and a similar SCHUFA score. A fair model should assign them similar loan approval probabilities. If $f(x) = 0.75$ and $f(x') = 0.60$, the output difference $|0.75 - 0.60| = 0.15$ must not exceed $d(x, x')$, the feature-based dissimilarity. If it does, the model violates the fairness constraint by treating similar individuals too differently, potentially due to proxy signals for protected attributes.

*Fairness Through Unawareness* (Grgic-Hlaca et al., 2016) simplifies fairness by excluding protected attributes from the input:

$$f(x) = f(x') \quad \text{if } x_{\setminus A} = x'_{\setminus A}$$

That is, two individuals who differ only in $A$ (e.g. gender, race) must receive identical outcomes. For example, let's consider two loan applicants with the same income, credit history, and employment status, but different genders. Under fairness through unawareness, the model must return the same decision for both, as it does not see gender. However, this approach is vulnerable to *proxy discrimination*: if features correlated with gender, such as name or occupation, remain in the input, the model may still treat individuals unequally despite not explicitly accessing the protected attribute (Grgic-Hlaca et al., 2016).

*Counterfactual Fairness* (Kusner et al., 2017) uses causal models to ensure that outcomes are unaffected by changes to protected attributes $A$ and for all $y$ and all attainable values $a'$ of $A$:

$$P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a)$$

Here, $\hat{Y}_{A\leftarrow a}(U)$ denotes the prediction that would be made if $A$ were set to $a$ through a counterfactual intervention, keeping all latent background variables $U$ constant. This definition expresses individual-level fairness: a decision is fair if it would have remained the same had the individual belonged to a different demographic group. For example, in a loan approval setting, counterfactual fairness requires that an applicant's gender does not influence the decision, even indirectly. If a woman with a certain income, credit score, and employment history is denied a loan, the model must ensure that a counterfactual version of her with the same qualifications but male gender would also be denied. If this is not the case, the decision is deemed unfair, even if the model does not explicitly use gender as a feature, because gender may have affected intermediate variables (e.g. income history) in a biased way. While this approach offers strong guarantees, it relies on a well-specified causal model to separate permissible and impermissible influences.

While individual fairness provides fine-grained control and avoids arbitrary group boundaries, its implementation depends on assumptions about similarity, causality, or feature independence that may be hard to justify in practice.

**Group Fairness.** Group fairness metrics assess whether individuals from different demographic groups, defined by a protected attribute $A$ (e.g., gender or race, with values $a, a'$), receive equitable treatment by a classifier. Let $\hat{Y} \in \{0, 1\}$ be the predicted label and $Y \in \{0, 1\}$ the true label. These metrics compare aggregate statistics across groups and reflect different fairness notions.

*Demographic Parity* (Mehrabi et al., 2021; Verma and Rubin, 2018) requires that:

$$P(\hat{Y} = 1 \mid A = a) = P(\hat{Y} = 1 \mid A = a')$$

ensuring equal rates of favourable predictions across groups, regardless of $Y$. For example, the proportion of approved loans should be the same across all genders. If 70% of men get a loan, then 70% of women or non-binary individuals should get one too. This measure does not consider the qualification of loan applicants, only their protected attribute, in this case, their gender.

*Equalised Odds* (Mehrabi et al., 2021; Barr et al., 2025) strengthens this by conditioning on the true label:

$$P(\hat{Y} = 1 \mid Y = y, A = a) = P(\hat{Y} = 1 \mid Y = y, A = a') \quad \forall y \in \{0, 1\}$$

requiring parity in both true positive and false positive rates. This means that accuracy should be the same across all groups. Not only should all qualified applicants for a loan have the same chance of getting one, independent of their gender, but also all unqualified applicants should have the same chance of not getting a loan, independent of their gender.

*Equal Opportunity* (Mehrabi et al., 2021) is a relaxation that focuses only on equal true positive rates:

$$P(\hat{Y} = 1 \mid Y = 1, A = a) = P(\hat{Y} = 1 \mid Y = 1, A = a')$$

In the loan example, this would mean that if 90% of all qualified men get a loan, then 90% of all women and non-binary individuals should get one too. This is similar to the demographic parity but also considers the true label $Y$.

*Predictive Parity* (Mehrabi et al., 2021) ensures equal precision across groups:

$$P(Y = 1 \mid \hat{Y} = 1, A = a) = P(Y = 1 \mid \hat{Y} = 1, A = a')$$

If 80% of approved men repay their loans, the same should be true for approved women and non-binary individuals.

*Conditional Statistical Parity* (Mehrabi et al., 2021) allows outcome differences only when justified by task-relevant features $L$ and a fixed combination of values for those features $\ell$:

$$P(\hat{Y} = 1 \mid A = a, L = \ell) = P(\hat{Y} = 1 \mid A = a', L = \ell)$$

Among applicants with the same income and credit score, all genders should have the same approval rate.

*Treatment Equality* (Mehrabi et al., 2021) balances the ratio of false negatives (FN) to false positives (FP), in other words, the burden of errors:

$$\frac{FN_a}{FP_a} = \frac{FN_{a'}}{FP_{a'}}$$

If men are wrongly denied loans (false negatives) just as often as they are wrongly approved (false positives), the same should hold for women and non-binary individuals.

*False Positive Rate Balance* (Verma and Rubin, 2018) avoids favouring one group among unqualified applicants:

$$P(\hat{Y} = 1 \mid Y = 0, A = a) = P(\hat{Y} = 1 \mid Y = 0, A = a')$$

If 20% of unqualified men are incorrectly approved, the same should hold for women and non-binary individuals.

*False Negative Rate Balance* (Verma and Rubin, 2018) ensures that qualified individuals are not systematically rejected more often in one group:

$$P(\hat{Y} = 0 \mid Y = 1, A = a) = P(\hat{Y} = 0 \mid Y = 1, A = a')$$

If 10% of qualified women are wrongly denied, the same rate should apply to men and non-binary individuals.

*Overall Accuracy Equality* (Verma and Rubin, 2018) requires that the classifier is equally accurate across all groups:

$$P(\hat{Y} = Y \mid A = a) = P(\hat{Y} = Y \mid A = a')$$

If predictions are correct 85% of the time for men, they should be 85% correct for women and non-binary persons too.

Group fairness metrics are intuitive and widely used but often mutually incompatible (Verma and Rubin, 2018), sensitive to unequal base rates (Barr et al., 2025), and limited in capturing within-group variation and structural inequality (Mehrabi et al., 2021).

**Subgroup Fairness.** Subgroup fairness strengthens classical group fairness by requiring that fairness constraints hold not just across a few coarse demographic groups, but across a rich collection of subgroups defined by protected attributes. This addresses *fairness gerrymandering*, where a classifier appears fair at the group level but discriminates against smaller, intersectional populations (Kearns et al., 2018). Suppose a classifier approves loans only for Black men and White women. It approves 50% of applicants by gender and 50% by race, which satisfies demographic parity for each attribute individually. However, it never approves loans for White men or Black women, revealing unfairness at the intersection of race and gender.

Let $D : \mathcal{X} \to \{0, 1\}$ be a binary classifier, and let $P$ denote the distribution over inputs $x \in \mathcal{X}$ and labels $y \in \{0, 1\}$. Subgroups are defined via a collection $\mathcal{G}$ of indicator functions $g : \mathcal{X} \to \{0, 1\}$, where $g(x) = 1$ iff $x$ belongs to subgroup $g$.

*Statistical Parity (SP) Subgroup Fairness (Kearns et al., 2018)*
Fix any classifier $D$, distribution $P$, collection of group indicators $\mathcal{G}$, and parameter $\gamma \in [0, 1]$. For each $g \in \mathcal{G}$, define:

$$\alpha_{\mathrm{SP}}(g, P) = \Pr_{x \sim P}[g(x) = 1], \quad \beta_{\mathrm{SP}}(g, D, P) = \left|\mathrm{SP}(D) - \mathrm{SP}(D, g)\right|,$$

where

$$\mathrm{SP}(D) = \Pr_{x \sim P}[D(x) = 1], \quad \mathrm{SP}(D, g) = \Pr_{x \sim P}[D(x) = 1 \mid g(x) = 1].$$

We say that $D$ satisfies $\gamma$-SP subgroup fairness with respect to $P$ and $\mathcal{G}$ if

$$\forall g \in \mathcal{G}, \quad \alpha_{\mathrm{SP}}(g, P) \cdot \beta_{\mathrm{SP}}(g, D, P) \leq \gamma.$$

This definition requires that the statistical parity difference between each subgroup $g$ and the full population is bounded, weighted by the subgroup's prevalence in the data. This ensures that the fraction of positive predictions in every subgroup is close to the overall average, unless the subgroup is very small.

If the general loan approval rate is 60%, then subgroups like “black women under 30” or “middle-aged Muslim men” should not deviate too far from this rate, unless their population size is so small that a large deviation is statistically insignificant.

*False Positive (FP) Subgroup Fairness (Kearns et al., 2018)*
Fix any classifier $D$, distribution $P$, collection of group indicators $\mathcal{G}$, and parameter $\gamma \in [0, 1]$. For each $g \in \mathcal{G}$, define:

$$\alpha_{\mathrm{FP}}(g, P) = \Pr_{x \sim P}[g(x) = 1,\, y = 0], \quad \beta_{\mathrm{FP}}(g, D, P) = \left|\mathrm{FP}(D) - \mathrm{FP}(D, g)\right|,$$

where

$$\mathrm{FP}(D) = \Pr_{x \sim P}[D(x) = 1 \mid y = 0], \quad \mathrm{FP}(D, g) = \Pr_{x \sim P}[D(x) = 1 \mid g(x) = 1,\, y = 0].$$

We say that $D$ satisfies $\gamma$-FP subgroup fairness with respect to $P$ and $\mathcal{G}$ if

$$\forall g \in \mathcal{G}, \quad \alpha_{\mathrm{FP}}(g, P) \cdot \beta_{\mathrm{FP}}(g, D, P) \leq \gamma.$$

This version of subgroup fairness targets error rates, ensuring that false positive disparities are bounded proportionally to how often each subgroup occurs among the negatively labelled data.

This focuses on unjustified positive predictions (i.e., false positives) and ensures that their rate is balanced across subgroups.

If the overall false positive rate is 10%, then "middle-aged Muslim men" or "single mothers with low credit scores" should not have a significantly higher false positive rate, unless the group is tiny and statistical fluctuations are expected.

Both definitions share the same structure: they bound the product of subgroup prevalence ($\alpha$) and fairness deviation ($\beta$) by a small threshold $\gamma$. This design ensures that violations in small subgroups are only penalised if the deviation is large, and vice versa. While subgroup fairness offers strong guarantees, enforcing it exactly over large or infinite $\mathcal{G}$ is computationally hard, motivating approximate or relaxable alternatives (Kearns et al., 2018).

Formal fairness definitions express normative goals in mathematical terms and enable integration into model development via constraints or regularisation. However, their application is often limited by the need for discrete group labels, stable similarity notions, and sufficient subgroup data. Requirements that are difficult to satisfy in the presence of intersectional or fluid identities such as non-binary gender. Fairness metrics may thus oversimplify the social complexities they aim to model. Moreover, different definitions can be mutually incompatible (Chouldechova, 2017). Achieving fairness therefore entails trade-offs between competing metrics and underlying normative assumptions. Crucially, statistical definitions tend to treat groups as symmetric and ahistorical, neglecting structural inequalities and cumulative disadvantage. Without contextual awareness, even seemingly neutral metrics like statistical parity may reinforce existing inequities. For example, enforcing equal outcomes across groups with unequal access to education or healthcare can create the illusion of fairness while ignoring deeper systemic disparities (Barr et al., 2025).

### 2.4.2. Discrimination in Text Classification

Discrimination in computational systems often becomes visible in text classification tasks, where models learn to associate linguistic patterns with labels in ways that may reinforce social stereotypes. In such tasks, a classifier $f$ maps an input text $x$ to a predicted label $\hat{y} = f(x)$, with $y$ denoting the true label (e.g., sentiment, profession, or toxicity). Discrimination can occur when predictions are skewed against texts that include features correlated with a protected attribute $A$ (such as gender or ethnicity), where $a$ and $a'$ represent different group values (e.g., male, female and non-binary). Let $\phi(x)$ denote the linguistic features extracted from $x$ (such as tokens or part-of-speech tags). Discrimination is present when the probability of receiving a label $y$ given the features $\phi(x)$ differs across groups:

$$P(\hat{Y} = y \mid \phi(x), A = a) \neq P(\hat{Y} = y \mid \phi(x), A = a')$$

This implies that even with identical linguistic input, the model's prediction varies based on group membership. Such disparities violate fairness expectations, especially when the protected attribute $A$ should not influence the outcome beyond what is explained by task-relevant features (Mehrabi et al., 2020). In practical terms, such discrimination is often observed in applications like named entity recognition, sentiment analysis, and hate speech detection. For instance, Mehrabi et al. (2020) demonstrate that male names are more often classified into professional categories, while female names tend to be associated with family or location categories. This reflects the implicit associations encoded in training data. In many NLP models, stereotypes present in the input data are not only learned but amplified. This means the model's predictions exhibit even greater disparities between groups than those found in the original data (Blodgett et al., 2020). Text classification models are highly sensitive to context. Das and Paik (2021) show that gender attribution in named entities can shift drastically with minor lexical changes. Formally, if $x$ and $x'$ are minimal edits of each other differing only in neutral context, but

$$f(x) \neq f(x') \quad \text{and} \quad A(x) = A(x')$$

then $f$ is context-sensitive in a potentially problematic way.

**Mitigation Strategies.** Several mitigation strategies have been proposed. Entropy-based Attention Regularisation introduces a penalty term to the loss function:

$$L_{\text{total}} = L_{\text{task}} + \lambda \cdot H(\alpha)$$

where $H(\alpha) = -\sum_i \alpha_i \log \alpha_i$ is the entropy of the attention distribution $\alpha$ over tokens in $x$, and $\lambda$ is a regularisation weight (Attanasio et al., 2022).

A low entropy attention distribution indicates that the model focuses heavily on a few tokens, which may include biased cues (e.g. gendered terms like "nurse" or "CEO"). In contrast, high entropy encourages the model to distribute its attention more evenly across the input, reducing over-reliance on potentially biased signals. When $H(\alpha)$ is small, the penalty is minimal and the model behaves as usual. A larger $H(\alpha)$ increases the penalty, encouraging broader attention. The regularisation strength $\lambda$ determines how much this influences learning: a high $\lambda$ strongly promotes distributed attention, while a low $\lambda$ allows the task loss to dominate, making the regularisation less effective. In essence, this approach discourages the model from "latching onto" biased shortcuts, nudging it towards more balanced reasoning (Attanasio et al., 2022).

**Auditing and Benchmarking.** Benchmark datasets such as `Bias-in-Bios` (De-Arteaga et al., 2019) and `CrowS-Pairs` (Nangia et al., 2020) expose differential model behaviour across demographic groups by measuring prediction disparities on semantically controlled sentence pairs. Evaluation also extends to domain-specific settings, e.g., Zhang et al. (2020) on clinical notes, and Breitfeller et al. (2019) on microaggressions in social media, revealing genre- and platform-specific manifestations of discrimination. Tools such as the Automatic Misuse Detector (Cai et al., 2022) further aid in identifying systematic misclassifications that disproportionately affect marginalised groups.

### 2.4.3. Discrimination in Word Embeddings

Word embeddings map each word $w$ in a vocabulary $V$ to a vector $\vec{w} \in \mathbb{R}^d$ (with $d$ components), positioning semantically similar words close together in a continuous space. While effective for capturing linguistic regularities, these representations also encode and perpetuate social biases present in training corpora (Pennington et al., 2014; Mikolov et al., 2013).

**Word Embedding Association Test (WEAT).** To quantify stereotypical associations in word embeddings, Caliskan et al. (2017) introduce the *Word Embedding Association Test (WEAT).* It compares the relative association between two sets of target words, $X$ and $Y$ (for example, male and female names), and two sets of attribute words, $A$ and $B$ (such as career- and family-related terms). Each word $w$ is represented by a vector $\vec{w}$ in the embedding space, and similarity is measured using cosine similarity $\cos(\vec{u}, \vec{v})$. The association of a target word $w$ with the attribute sets is defined as

$$s(w, A, B) = \mathrm{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \mathrm{mean}_{b \in B} \cos(\vec{w}, \vec{b}),$$

which reflects how much more $w$ is associated with $A$ than with $B$ (Caliskan et al., 2017).

**Hard Debiasing.** To mitigate gender bias in word embeddings, Bolukbasi et al. (2016) propose a post-processing algorithm called *hard debiasing*, which consists of two main steps.

*Step 1: Identify gender subspace.*
Given a vocabulary $W$, a collection of defining word sets $D_1, D_2, \ldots, D_n \subseteq W$, and an integer parameter $k \geq 1$, the first step is to compute the *bias subspace* $B$ that captures the directions along which gender is expressed.

*Step 2: Hard de-biasing (neutralize and equalize).*
In the second step, the embeddings are updated to remove and symmetrise gender information. This step requires two inputs: a set $N \subseteq W$ of gender-neutral words, and a family of equality sets $\mathcal{E} = \{E_1, E_2, \ldots, E_m\}$ where each $E_i \subseteq W$ contains words that should be treated equally with respect to gender.

For each word $w \in N$, let $\vec{w}$ denote its original embedding and let $\vec{w}_B$ be its projection onto the bias subspace $B$. The neutralised embedding is computed as:

$$\vec{w} := (\vec{w} - \vec{w}_B)/\|\vec{w} - \vec{w}_B\|.$$

For each equality set $E \in \mathcal{E}$, define the mean vector

$$\mu := \sum_{w \in E} \vec{w}/|E|, \quad \nu := \mu - \mu_B,$$

where $\mu_B$ is the projection of $\mu$ onto $B$. Then for each $w \in E$, the updated embedding is:

$$\vec{w} := \nu + \sqrt{1 - \|\nu\|^2} \cdot \frac{\vec{w}_B - \mu_B}{\|\vec{w}_B - \mu_B\|}.$$

This operation ensures that all words in $E$ lie at an equal distance from gendered concepts, while being symmetric with respect to the gender direction. The final output consists of the subspace $B$ and the updated embeddings $\{\vec{w} \in \mathbb{R}^d \mid w \in W\}$.

**Limits of Directional Debiasing.** While methods like hard debiasing explicitly remove components along a predefined gender direction, subsequent work questions whether this approach is sufficient. Papakyriakopoulos et al. (2020) demonstrate that bias in word embeddings is not confined to a single direction but is instead distributed across multiple dimensions of the embedding space. As a result, even after projecting embeddings onto the subspace orthogonal to the identified gender direction, residual information about protected attributes remains. Through classification experiments, the authors show that it is still possible to predict gender from supposedly debiased embeddings with accuracy significantly above chance. This suggests that discriminatory signals persist in a more diffuse and less interpretable form. To better characterise this phenomenon, Papakyriakopoulos et al. (2020) introduce the concept of *bias anisotropy*, which quantifies how directionally concentrated gender bias is, and propose the *isotropic projection loss* to measure how much of this bias remains after debiasing. These findings highlight the limitations of direction-based methods and motivate the need for more comprehensive, geometry-aware approaches to fairness in embedding spaces (Papakyriakopoulos et al., 2020).

In sum, while linear debiasing reduces surface-level associations, it does not remove deeper structural patterns. Word embeddings reflect not only linguistic regularities but also cultural and social hierarchies. Eliminating bias without disrupting semantic coherence remains a major challenge in NLP.

### 2.4.4. Gender Discrimination in Language Models

Large language models (LLMs) such as BERT (Devlin et al., 2019) and the GPT family (Radford et al., 2018, 2019b; Brown et al., 2020) are trained on massive corpora that reflect historical inequalities. This makes them prone to reproducing gendered stereotypes, even when explicit gender information is not present. Gender discrimination manifests in skewed completions of neutral prompts, unequal sentiment distributions, and biased occupational or behavioural inferences.

**Discrimination Detection.** Bias in language models is typically detected by comparing model behaviour across systematically varied inputs. One common approach is *perturbation-based testing*, where semantically equivalent prompts differing only in gendered terms are compared. Ma et al. (2020); Li et al. (2024) show that such metamorphic testing can reveal discrepancies in tone, sentiment, or factual content solely due to gendered wording. To evaluate these effects at scale, several *benchmark frameworks* have been introduced. *GenderCARE* (Tang et al., 2024) and *SOFA* (Manerba et al., 2024) assess model bias along lexical, syntactic, and semantic dimensions, targeting stereotypical associations such as linking women to caregiving or men to authority. *Cross-linguistic evaluations* extend this analysis to other languages. *TWBias* (Hsieh et al., 2024) tests models in Traditional Chinese by comparing perplexity between gendered variants, using statistical measures to identify subtle group disparities. Finally, bias can have *downstream consequences* in applied settings. *JobFair* (Wang et al., 2024) models discrimination in job recommendation pipelines, demonstrating how minor differences in model scores for otherwise similar candidates can compound across stages, leading to unequal hiring outcomes.

**Mitigation Strategies.** Mitigating gender bias in LLMs involves interventions at various stages of the modelling pipeline, including prompt design, representation learning, decoding, and model training.

*Input-level interventions* include prompt augmentation and counterfactual generation, which expose models to balanced representations during training or evaluation. One formal method is causal front-door adjustment (Zhang et al., 2025), which introduces intermediate variables (e.g., chain-of-thought prompts) to mediate the effect of biased inputs $X$ on outputs $A$. The adjusted output distribution is estimated as:

$$P(A \mid do(X)) = \sum_r P(r \mid X) \cdot \sum_x P(x) \cdot P(A \mid r, x)$$

This approach avoids direct manipulation of sensitive attributes by intervening through mediators $r$ where $do(X)$ denotes an intervention in the causal sense, meaning that $X$ is set to a specific value independently of its natural causes. Alternatively, counterfactual prompts $(x, x')$ that differ only in protected attributes (e.g., gender) are used to assess or reduce a model's reliance on such features while preserving semantic content (Zhang et al., 2025).

*Representation-level debiasing* leverages contrastive learning to distinguish biased from neutral language. Park et al. (2024) train a sentence encoder to bring stereotype-free (positive) and anchor embeddings closer together, while pushing away stereotypical (negative) examples. They use a margin-based triplet loss:

$$\mathcal{L}_{\mathrm{CL}} = \frac{1}{|x|} \sum_{i=1}^{|x|} \max\{0, \rho - s(a_i, a_i^+) + s(a_i, a_i^-)\}$$

where $|x|$ denotes the batch size, $s$ denotes cosine similarity, $a_i$ is the anchor, $a_i^+$ the positive (fair), and $a_i^-$ the negative (biased) example and $\rho$ a predefined margin. This encourages the model to separate biased and unbiased meanings in the embedding space by enforcing that the similarity between anchor and fair examples $s(a_i, a_i^+)$ exceeds that of anchor and biased examples $s(a_i, a_i^-)$ by at least a margin $\rho$. If this condition is not met, the loss increases, pushing the embeddings apart until the margin is satisfied(Park et al., 2024).

*Decoding-time strategies* like *FairFlow* (Cheng and Amiri, 2024) rerank candidate outputs based on both likelihood and fairness. The method penalises completions that include biased continuations (e.g., gendered stereotypes) and defers decisions by assigning a reserved "undecided" label when no fair candidate meets the confidence criteria. This allows the model to avoid committing to potentially discriminatory outputs under uncertainty (Cheng and Amiri, 2024).

*Training-level interventions* such as *FairDistillation* (Delobelle and Berendt, 2023) reduce bias by distilling fairness-aware knowledge from a large teacher model into a smaller student model. During training, the teacher's masked language model (MLM) predictions are modified to satisfy fairness constraints, such as assigning equal probabilities to gendered terms, before being passed to the student. The student then minimises a composite loss function:

$$\mathcal{L} = \alpha_{\mathrm{ce}}\mathcal{L}_{\mathrm{ce}} + \alpha_{\mathrm{mlm}}\mathcal{L}_{\mathrm{mlm}} + \alpha_{\mathrm{cos}}\mathcal{L}_{\mathrm{cos}}$$

balancing cross-entropy loss $\mathcal{L}_{\mathrm{ce}}$, MLM loss $\mathcal{L}_{\mathrm{mlm}}$, and cosine similarity loss $\mathcal{L}_{\mathrm{cos}}$. The weighting parameters $\alpha_{\mathrm{ce}}, \alpha_{\mathrm{mlm}}, \alpha_{\mathrm{cos}}$ allow tuning between accuracy, fairness, and representational alignment (Delobelle and Berendt, 2023).

Despite these advances, mitigation remains challenging due to the diffuse and context-sensitive nature of learned biases (Tang et al., 2024; Ma et al., 2020).

### 2.4.5. Detecting Discrimination in Text

In contrast to model-focused approaches, some studies analyse the texts themselves for patterns of discrimination. This strand of research draws on traditions in linguistics and discourse analysis, examining how power and inequality are reflected in language use. A common focus lies in detecting forms of discriminatory language such as hate speech (Fortuna and Nunes, 2018; Paz et al., 2020), microaggressions (Breitfeller et al., 2019), and ambivalent sexism (Jha and Mamidi, 2017). These tasks typically require annotated corpora with context-sensitive labels, as discriminatory language is often indirect, euphemistic, or context-dependent. Methodologically, supervised classifiers trained on manually curated examples are widely used, but recent work also incorporates weak supervision, crowd-sourced judgments, or linguistic pattern mining. To detect microaggressions, for instance, Breitfeller et al. (2019) construct a dataset from social media and apply syntactic filtering to isolate indirect expressions of bias. Similarly, ambivalent sexism detection (Jha and Mamidi, 2017) leverages linguistic features such as sentiment polarity, modality, and target group references to differentiate between hostile and benevolent forms of sexism. In hate speech detection, approaches range from keyword-based heuristics to deep learning models that incorporate discourse-level features and external knowledge sources (Fortuna and Nunes, 2018).

Beyond individual utterances, researchers have also explored structural patterns in longer texts and media corpora. Wagner et al. (2021) examine gender disparities in Wikipedia biographies by analysing content length, topical focus, and linguistic framing, showing that biographies of women tend to be shorter and more focused on personal life. Madaan et al. (2018) apply sentiment analysis and character-role mapping to identify recurring gender stereotypes in Bollywood films. In narrative domains, Fast et al. (2016) analyse fan fiction for differences in verb usage, point of view, and descriptions across male and female characters. Work on recommendation letters (TRIX and PSENKA, 2003) combines linguistic profiling with qualitative coding to identify gendered patterns in attribution, agency, and praise.

These studies demonstrate that computational discrimination analysis benefits from combining quantitative text analysis with concepts from sociolinguistics and critical theory. Purely metric-based approaches risk overlooking the discursive and structural dimensions of discrimination that manifest through framing, omission, or evaluative language.

## 2.5. Information Extraction

Information Extraction (IE) refers to computational techniques that automatically identify and structure predefined types of information from unstructured text (Grishman, 2015; Sarawagi, 2008). Unlike general language understanding, IE targets specific signals: such as who is mentioned, how often, and in what context and serves as the foundation of the pipeline developed in this thesis.

### 2.5.1. Named Entity Recognition

At the core of IE lies **named entity recognition** (NER), the task of detecting and classifying spans in text that refer to real-world entities such as persons, locations, or organisations. Given a token sequence $X = (x_1, \ldots, x_n)$, NER assigns each token a label $y_i \in C$, where $C$ is a set of entity types (e.g. $C = \{\texttt{PER}, \texttt{LOC}, \texttt{ORG}\}$). NER does not resolve coreference, meaning expressions like “Angela Merkel” and “Chancellor Merkel” are treated independently (Jurafsky and Martin, 2025).

**Methods.** NER methods have evolved significantly over time. *Rule-based systems* rely on manually defined patterns, dictionaries, and orthographic features, offering strong performance in narrow domains but limited generalisability (Alharbi and Tiun, 2015). *Statistical approaches*, such as Hidden Markov Models (HMM), Support Vector Machines (SVM), and Conditional Random Fields (CRF), treat NER as a sequence labelling task, estimating the conditional probability

$P(Y \mid X)$, where $X = (x_1, \ldots, x_n)$ is the input token sequence and $Y = (y_1, \ldots, y_n)$ the corresponding sequence of entity labels. This formulation models the likelihood of label sequences given observed tokens. These methods depend on hand-crafted features like part-of-speech tags and character patterns, which limit adaptability (Guo et al., 2020). *Deep learning methods*, learn contextual representations end-to-end using architectures such as BiLSTM-CRF (Wu et al., 2019) or Transformer-based models like BERT (Devlin et al., 2019). Although deep learning achieves state-of-the-art results, it typically requires large annotated datasets and may struggle in domain transfer. Transfer learning techniques partially address this by fine-tuning pre-trained models on target data (Warto et al., 2024; Guo et al., 2020).

**Challenges.** NER remains difficult in domain-specific and low-resource settings. In languages like Chinese word boundaries are hard to detect (Guo et al., 2020). Biomedical texts pose particular challenges due to specialised vocabulary and limited labelled data (Alharbi and Tiun, 2015). Robust NER requires models that generalise across domains, languages, and annotation schemes.

### 2.5.2. Syntactic Processing

While Named Entity Recognition (NER) identifies entity mentions, **syntactic processing** determines how these entities function within sentence structure. It involves analysing grammatical relations such as subject (`nsubj`), object (`dobj`), and modifier roles (`amod`), enabling a fine-grained understanding of who does what to whom. For example, the distinction between “Marie Curie criticised the report” and “the report criticised Marie Curie” reflects a syntactic difference essential for inferring agency and polarity. These structures are derived through syntactic parsing and serve as a foundation for downstream tasks in IE (Jurafsky and Martin, 2025).

**Methods**. Syntactic analysis typically begins with tokenisation, segmenting input text ($T = (t_1, \ldots, t_n)$) into linguistic units. This is followed by part-of-speech (POS) tagging , and then dependency parsing, which constructs a directed graph ($G = (V, E)$), where $V$ is the set of tokens and $E$ the set of labelled syntactic dependencies (edges) between them, such as `nsubj`$(x, y)$ indicating that token $(x)$ is the syntactic subject of token $(y)$ (Jurafsky and Martin, 2025; Radishevskii et al., 2018). Some systems also use constituency parsing to produce hierarchical phrase structures (Wan and Xia, 2017). Parsing strategies vary in depth: shallow parsing identifies chunks like noun or verb phrases, while deep parsing builds full parse trees (Nallapati, 2004). Modern approaches integrate syntax with machine learning, for example using composite kernels in SVMs or shortest dependency paths in neural networks for relation extraction (Nallapati, 2004; Campos et al., 2013). Semantic role labelling can be layered on top of syntactic analysis to capture predicate–argument structures such as agents (`ARG0`) and patients (`ARG1`) (Chen et al., 2011).

**Challenges.** Syntactic parsing remains computationally and linguistically challenging. Long or ambiguous sentences are costly to process and often require robust grammars. Traditional parsers based on context-free grammars scale poorly, although techniques like generalised LR parsing (left-to-right, rightmost derivation) improve efficiency by handling ambiguities in parallelised settings (Radishevskii et al., 2018). Language-specific challenges include tokenisation in Chinese or complex morphology in Tibetan (Wan and Xia, 2017). Parsing errors in early stages (e.g. tokenisation, POS tagging) can propagate and degrade downstream accuracy. Moreover, syntactic parsers trained on general corpora often struggle with domain adaptation, underperforming in specialised contexts such as biomedical or legal texts. Finally, despite recent advances, scalability remains a concern for real-time or large-scale applications (Radishevskii et al., 2018).

### 2.5.3. Coreference Resolution

**Coreference resolution** is the task of identifying expressions that refer to the same real-world entity within a text. Given a document $D$ with a sequence of mentions $M = \{m_1, m_2, \dots, m_n\}$, the goal is to partition $M$ into equivalence classes $C = \{c_1, c_2, \dots, c_k\}$ such that all mentions in $c_i$ refer to the same entity. For example, "Dr. Ruth Harriet Bleier," "Dr. Bleier," "R. H. Bleier," and "she" may all appear in the same document and refer to the same person. Without resolution, linguistic signals such as sentiment or frequency can be distributed across partial identities, reducing coherence and interpretability in downstream tasks (Kozlova et al., 2025; Lu and Ng, 2018). Coreference resolution thus enables document-level information extraction, supporting more complete and accurate linking of entities across sentences and paragraphs (Nam et al., 2020; Kilicoglu and Demner-Fushman, 2016).

**Methods.** Early rule-based systems rely on syntactic and semantic constraints such as gender and number agreement or apposition structures (Park et al., 2016). While interpretable, these systems struggle with generalisation. More robust are mention-pair models, which predict whether a pair of mentions $(m_i, m_j)$ is coreferent by encoding each mention's context using deep neural architectures (Park et al., 2016). Clustering approaches extend this by learning a similarity function $s(m_i, m_j)$ and optimising mention groupings holistically (Kozlova et al., 2025). Alternatively, graph-based frameworks represent mentions as nodes $M$ and candidate coreference links as edges $E$ in a graph $G = (M, E)$, enabling joint inference with related IE tasks (Zheng and Tuan, 2023). In neural formulations, each mention $m_i$ is embedded as a vector $\vec{m}_i$, and a scoring function $s(\vec{m}_i, \vec{m}_j)$ estimates the likelihood of coreference. The objective is typically to maximise the sum of coreference scores:

$$\max \sum_{(i,j)\in E} s(\vec{m}_i, \vec{m}_j) \cdot y_{ij}$$

where $y_{ij} = 1$ if $m_i$ and $m_j$ are coreferent, and 0 otherwise (Kozlova et al., 2025; Lu and Ng, 2018).

**Challenges.** Coreference resolution remains a difficult task due to multiple sources of ambiguity. Pronouns like "he" or "it" often have multiple plausible antecedents[2], and different types of coreference, such as anaphora[3], cataphora[4], and apposition[5], require distinct resolution strategies. Models trained on one domain often underperform in others, especially where annotated data is scarce. Moreover, coreference resolution is sensitive to upstream NLP errors: inaccurate tokenisation, POS tagging, or parsing can propagate and undermine resolution accuracy. These challenges are particularly pronounced in low-resource settings or in texts with complex discourse structures (Lu and Ng, 2018; Kozlova et al., 2025).

[2] An antecedent is the earlier expression in a text that a later expression (often a pronoun or another referring phrase) refers back to.

[3] Anaphora is a linguistic phenomenon where a word or phrase (often a pronoun) refers back to something mentioned earlier in the text: its antecedent.

[4] Cataphora is the opposite of anaphora: it's when a word or phrase refers to something that appears later in the text.

[5] Apposition is when two noun phrases appear next to each other and refer to the same entity, with the second phrase providing more information about the first.

### 2.5.4. Semantic Processing

Semantic processing refers to computational techniques for deriving structured representations of meaning from natural language text (Sarawagi, 2008; Jurafsky and Martin, 2025). Its goal is to identify entities, the roles they play, and the relationships between them, often represented as predicate–argument structures that answer the question "who did what to whom, when, where, and how". Formally, given a corpus $D = \{d_1, d_2, \ldots, d_n\}$ and a set of extracted linguistic units (e.g. tokens, phrases, entities), semantic processing maps each sentence $s \in d$ to a structured tuple

$$t = (p_r, a_1, a_2, \ldots, a_m)$$

where $p_r$ is a predicate (often a verb or relational noun) and each $a_j$ is an argument filling a semantic role $r_j$ from a predefined inventory such as PropBank (`ARG0` for agent, `ARG1` for patient, etc.). Aggregating such tuples across a corpus yields a semantic graph $G = (V, E)$, where $V$ contains entities and concepts and $E$ encodes labelled semantic relations (Sarawagi, 2008; Jurafsky and Martin, 2025).

**Methods.** A central approach to semantic processing is **semantic role labelling** (SRL), which identifies predicates, locates their arguments, and classifies each according to its semantic role (Jurafsky and Martin, 2025). Early SRL systems used hand-crafted lexical, syntactic, and semantic features with statistical classifiers. Modern approaches employ contextualised embeddings from Transformer-based models, treating SRL as a sequence labelling or span classification task. Outputs from SRL can be enriched with relation extraction heuristics to capture evaluative or affective modifiers (Pandian et al., 2008; Assal et al., 2011), or integrated with ontology-based information extraction pipelines (Wang et al., 2008) that constrain relations to domain-specific schemas. More advanced pipelines construct semantic graphs directly, with nodes representing entities or concepts and edges encoding extracted relations (Wimalasuriya and Dou, 2010; Dörpinghaus and Stefan, 2019; Zhao et al., 2023).

**Challenges.** Despite their utility, semantic processing systems face several limitations. Ambiguity in role assignment arises frequently, especially in complex or elliptical constructions. Domain-specific variation in predicates and argument structures reduces the transferability of models trained on general corpora. Rule-based or shallow SRL systems often underperform in detecting nuanced or implicit relations. Lastly, scalability remains a concern: reliable semantic extraction across large corpora demands efficient yet interpretable techniques (Wang et al., 2008; Venugopal et al., 2023).

### 2.5.5. Cross-Document Coreference Resolution

**Cross-document coreference resolution** (CCR) is the task of identifying mentions of the same real-world entity across multiple documents. While within-document coreference focuses on resolving references locally, CCR consolidates scattered mentions into coherent entity profiles that span a corpus $D = \{d_1, d_2, \ldots, d_n\}$. Each document $d_i$ contains a set of mentions $M_i = \{m_{i1}, \ldots, m_{ik}\}$. The union $\bigcup_i M_i$ denotes the set of all mentions from all documents in the corpus, obtained by combining every $M_i$ into a single set. The goal is to cluster these mentions into equivalence classes $C = \{c_1, \ldots, c_l\}$ such that all $m \in c_j$ refer to the same underlying entity (Grishman, 2015; Huang et al., 2009a). CCR is fundamental to large-scale NLP tasks such as knowledge graph construction, media monitoring, event tracking, and cross-document summarisation (Beheshti et al., 2017).

**Methods.** A common CCR pipeline begins by constructing entity profiles for each cluster candidate. These profiles aggregate lexical, syntactic, semantic, and metadata features from all associated mentions. Each profile $e_j$ is represented as a tuple:

$$e_j = (\texttt{name}, \texttt{aliases}, \texttt{features}, \texttt{context})$$

Similarity between two profiles $e_i$ and $e_j$ is computed using a kernel function $K(e_i, e_j)$, which feeds into clustering algorithms (e.g. hierarchical or density-based clustering) to form global equivalence classes (Huang et al., 2009a,b). Hierarchical models have proven especially effective, enabling joint reasoning across multiple levels of granularity and scaling efficiently via distributed inference (Singh et al., 2011). Relational clustering combines unary features (e.g. gender or string similarity) with binary relations (e.g. document co-occurrence) to improve resolution accuracy. Summarisation-based techniques enhance disambiguation by condensing relevant context (Gao et al., 2010), and recent advances leverage large language models for query diversification and improved evidence selection across documents (Wang et al., 2025).

**Challenges.** Despite significant progress, CCR remains a difficult problem. Ambiguity and polysemy frequently lead to errors, especially when common names (e.g. "Michael Müller") refer to different individuals (Upadhyay et al., 2016). Sparse or underspecified mentions lack sufficient contextual clues, making disambiguation hard. Evaluation is also challenging, as it is often unclear whether an error stems from within-document or cross-document misresolution (Beheshti et al., 2017). Moreover, scalability is a persistent concern: real-world applications may involve millions of mentions across large corpora, requiring highly efficient inference algorithms (Singh et al., 2011).

## 2.6. Large Language Models

Large language models (LLM) have become central to the field of natural language processing and underpin many recent advances in text generation, understanding, and interaction. Their development is deeply intertwined with architectural innovations, growing computational scale, and evolving training objectives.

### 2.6.1. Transformer Architecture

The foundation for today's large language models (LLMs) was laid by Vaswani et al. (2017), who introduced the **transformer** architecture and its core innovation: self-attention. This mechanism enables each token to weigh the relevance of all other tokens in the input sequence, capturing long-range dependencies and allowing for parallel processing. The model follows an encoder–decoder structure (Figure 2.1). Both encoder and decoder consist of $N$ identical layers. Each encoder layer contains multi-head self-attention and a position-wise feed-forward network. Decoder layers add a third component: cross-attention over the encoder outputs. All sub-layers are followed by residual connections and layer normalisation (Vaswani et al., 2017).

Attention operates over queries $Q$, keys $K$, and values $V$, each of shape $\mathbb{R}^{n \times d_k}$, where $n$ is the sequence length and $d_k$ is the dimensionality of keys and queries. It is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

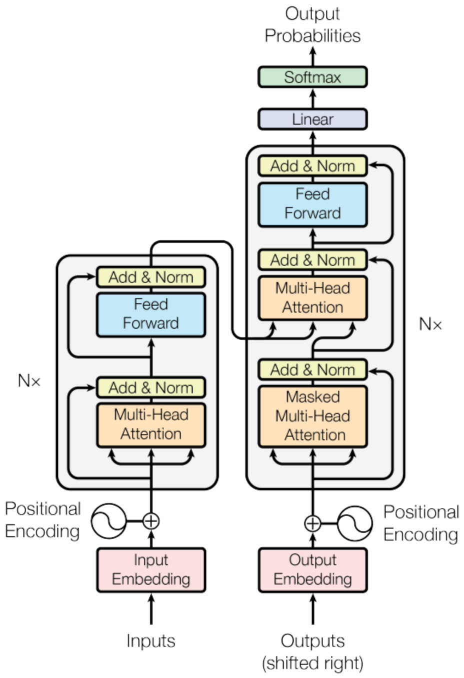


Figure 2.1.: The transformer architecture as proposed by Vaswani et al. (2017).

This mechanism assigns weights based on similarity between $Q$ and $K$ to extract relevant information from $V$ (Vaswani et al., 2017).

Instead of performing a single attention function, the model uses $h$ parallel attention heads to capture information from different subspaces:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

Here, $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_k}$ are learned projection matrices for each head, and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$ projects the concatenated outputs back to the model dimensionality $d_{\text{model}}$ (typically 512) (Vaswani et al., 2017).

Each layer also contains a fully connected feed-forward network applied to each position independently:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

Here, $W_1 \in \mathbb{R}^{d_{\text{model}} \times d_{\text{ff}}}$ and $W_2 \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{model}}}$ are the feed-forward weight matrices, $b_1 \in \mathbb{R}^{d_{\text{ff}}}$ and $b_2 \in \mathbb{R}^{d_{\text{model}}}$ are the bias terms, and $d_{\text{ff}}$ is the inner-layer dimensionality (typically 2048) (Vaswani et al., 2017).

### 2.6.2. BERT: Contextual Embeddings and Bidirectional Attention

Devlin et al. (2019) introduced BERT (Bidirectional Encoder Representations from Transformers), a deep bidirectional model based on the transformer encoder. Unlike traditional unidirectional models, BERT uses self-attention to condition each token's representation on both its left and right context across all layers. $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$ consist of $L = 12$ or 24 Transformer encoder layers, each with hidden size $H = 768$ or 1024 and $A = 12$ or 16 attention heads. Input sequences are tokenised using WordPiece and enriched with embeddings $E$ that sum token, segment, and positional information. Special tokens include $[CLS]$ (prepended for classification), $[SEP]$ (to separate sentences), and $[MASK]$ (for pretraining) (Devlin et al., 2019).

BERT is pretrained using two self-supervised tasks:

*Masked Language Modelling (MLM)* randomly masks 15% of input tokens $x = (x_1, \ldots, x_n)$. Of these, 80% are replaced with $[MASK]$, 10% with a random token, and 10% remain unchanged. The model predicts the original token using the corresponding hidden vector $T_i \in \mathbb{R}^H$:

$$\hat{y}_i = \text{softmax}(WT_i + b)$$

where $W$ and $b$ are task-specific weights and biases (Devlin et al., 2019).

*Next Sentence Prediction (NSP)* classifies whether segment B follows segment A. The model uses the final hidden state $C \in \mathbb{R}^H$ of the $[CLS]$ token to compute:

$$\hat{y}_{\text{NSP}} = \text{softmax}(W_{\text{nsp}}C + b_{\text{nsp}})$$

(Devlin et al., 2019) Each token is represented by an embedding $E = E_{\text{token}} + E_{\text{segment}} + E_{\text{position}}$, where segment embeddings distinguish sentence A from B, and positional encodings indicate token order. BERT supports end-to-end fine-tuning for various tasks. For classification, the $[CLS]$ representation $C$ is used. For token-level tasks (e.g., NER or QA), the relevant $T_i$ vectors serve as input to task-specific heads. The architecture requires minimal modification, as its self-attention layers naturally accommodate both single and paired inputs (Devlin et al., 2019).

### 2.6.3. T5 and the Text-to-Text Paradigm

To unify NLP tasks under a single framework, Raffel et al. (2020) introduced T5 (Text-to-Text Transfer Transformer), which reformulates all tasks, including classification, summarisation, translation, and question answering, as text-to-text problems. Both input $x$ and output $y$ are treated as sequences of text, enabling a single sequence-to-sequence model to perform diverse tasks without task-specific architecture changes (Raffel et al., 2020).

T5 uses a standard Transformer encoder–decoder setup with $L$ layers, hidden size $H$, and $A$ attention heads per layer. Each input is tokenised via SentencePiece (32k vocabulary), embedded as $E = E_{\text{token}} + E_{\text{position}}$, and prefixed with a task-specific prompt (e.g., “summarise:” or “translate English to German:”). The decoder autoregressively generates each output token $y_i$ conditioned on the previous tokens $y_{<i}$ and the input $x$, producing hidden representations $T_i \in \mathbb{R}^H$ (Raffel et al., 2020). The model is trained via maximum likelihood estimation:

$$\mathcal{L}_{\text{T5}} = -\sum_{i=1}^{|y|} \log p(y_i \mid y_{<i}, x)$$

where $p(y_i \mid y_{<i}, x)$ denotes the conditional generation probability (Raffel et al., 2020).

T5 variants (e.g., $\text{T5}_{\text{BASE}}$, $\text{T5}_{\text{LARGE}}$, $\text{T5}_{\text{11B}}$) share the same encoder–decoder design but vary in depth ($L$) and width ($H$). The decoder generates output token-by-token while attending to both its prior outputs and the encoder’s contextualised states (Raffel et al., 2020).

T5 enables multitask learning by simply changing the task prefix in the input. For instance: “sst2 sentence: this movie was great” → “positive” (classification), “squad question: What is the capital of France? context: Paris is the capital of France.” → “Paris” (question answering), “summarize: This paper explores...” → “This paper proposes...” (summarisation) (Raffel et al., 2020).

This flexible formulation laid the foundation for prompt-based learning and influenced many subsequent LLM designs.

### 2.6.4. The GPT Family: Scaling and Generalisation

The GPT (Generative Pretrained Transformer) family adopts a unidirectional, decoder-only transformer architecture for autoregressive language modelling. Given an input sequence $x = (x_1, \ldots, x_t)$, the model predicts the next token $x_{t+1}$ by computing hidden states $T_t \in \mathbb{R}^H$ and projecting them via learned weights and biases $(W, b)$:

$$\hat{x}_{t+1} = \text{softmax}(WT_t + b), \quad \mathcal{L}_{\text{GPT}} = -\sum_{t=1}^{n} \log p(x_t \mid x_{<t})$$

This left-to-right training objective, scaled across $L$ Transformer layers with $A$ attention heads and hidden size $H$, enables flexible text generation and in-context learning Radford et al. (2018).

GPT-2 Radford et al. (2019a) showed that large models trained on diverse corpora can generalise in zero-shot settings. GPT-3 scaled this to 175 billion parameters, demonstrating few-shot learning via prompts without weight updates (Brown et al., 2020). This shifted the focus from fine-tuning to prompt design, enabling models to adapt to new tasks at inference time. GPT-4 further improved multilingual, logical, and programming abilities. Although its architecture remains undisclosed, evaluations suggest signs of general-purpose intelligence and strong performance across a wide range of tasks (Bubeck et al., 2023).

To better match human expectations, recent models use reinforcement learning from human feedback (RLHF) (Christiano et al., 2017). A reward model trained on human preferences guides fine-tuning, with Kullback–Leibler (KL) penalties ensuring the updated policy remains close to the original distribution (Christiano et al., 2017). RLHF has improved summarisation quality (Stiennon et al., 2020) and is core to assistant models like ChatGPT.

The launch of ChatGPT integrated a GPT model with RLHF and a user interface for dialogue-based interaction (OpenAI, 2022). This made LLMs accessible to the public, driving adoption in writing, education, and programming. It also raised new concerns around ethics, misinformation, and AI governance.

### 2.6.5. Alternative Closed-Source LLMs

While the GPT family dominates public discourse, several other proprietary LLMs have significantly advanced the state of the art across various modalities and benchmarks. **Anthropic’s Claude 3** model family (Anthropic, 2024) includes Claude 3 Haiku, Sonnet, and Opus, and focuses on safe, aligned language generation using constitutional AI. These models excel at reasoning, coding, and multimodal tasks, including interpreting complex visual input. **Google DeepMind’s Gemini series** (Team et al., 2025) offers highly capable multimodal models that combine text, image, audio, and video understanding. Gemini models have demonstrated state-of-the-art performance across a wide array of academic and industrial benchmarks. Mistral AI has released proprietary models like **Mistral Large** and **Mistral Small** (Mistral AI, 2024b), which power their conversational interface **Le Chat** (Mistral AI, 2024a). Le Chat provides multilingual capabilities and system-level moderation, targeting both general-purpose and instruction-tuned use cases. The underlying models are also described in detail in Mistral's documentation (Mistral AI, 2024b). **Amazon's Titan models** (Amazon Web Services, 2024) are pretrained on vast datasets and support text generation, classification, information extraction, and question answering. Delivered through Amazon Bedrock, they integrate with AWS infrastructure and are designed for scalable, production-level applications.

These closed-source models, while often highly performant, raise similar concerns to GPT-style systems regarding transparency, reproducibility, and dataset opacity.

### 2.6.6. Open-Source Alternatives to Proprietary Models

As proprietary language models have become increasingly powerful but opaque, open-source alternatives have gained traction for promoting reproducibility, transparency, and equitable access to advanced AI systems. Prominent examples include **LLaMA** (Touvron et al., 2023), **PaLM** (Chowdhery et al., 2023), and **Zephyr** (Tunstall et al., 2023), which vary in scale and architecture but are united in their commitment to community-driven research and responsible deployment. A particularly notable recent development is **DeepSeek LLM** (DeepSeek-AI, 2024), an open model family trained from scratch on 2 trillion English and Chinese tokens. Available in 7B and 67B parameter sizes, it is released under the permissive MIT License. DeepSeek outperforms LLaMA-2 70B on several benchmarks, especially in reasoning, mathematics, and coding tasks. Its instruction-tuned variant, *DeepSeek Chat* (DeepSeek-AI, 2024), leverages supervised fine-tuning and Direct Preference Optimisation (Rafailov et al., 2023) to improve dialogue alignment, achieving performance on par with or above GPT-3.5 in open-ended tasks (DeepSeek-AI, 2024).

Overall, open-source models have closed the performance gap with earlier proprietary systems and now compete on standard evaluation tasks. Nonetheless, transparency in pretraining data and fine-tuning methods remains limited for many models, including those released under open licences.

# 3. Research Trajectory

This chapter outlines the conceptual development of the thesis and explains how the three main parts and five publications are connected. The aim is to provide a coherent meta-narrative that illustrates the progression of the research, its underlying motivations, and how the components interact and build upon one another.

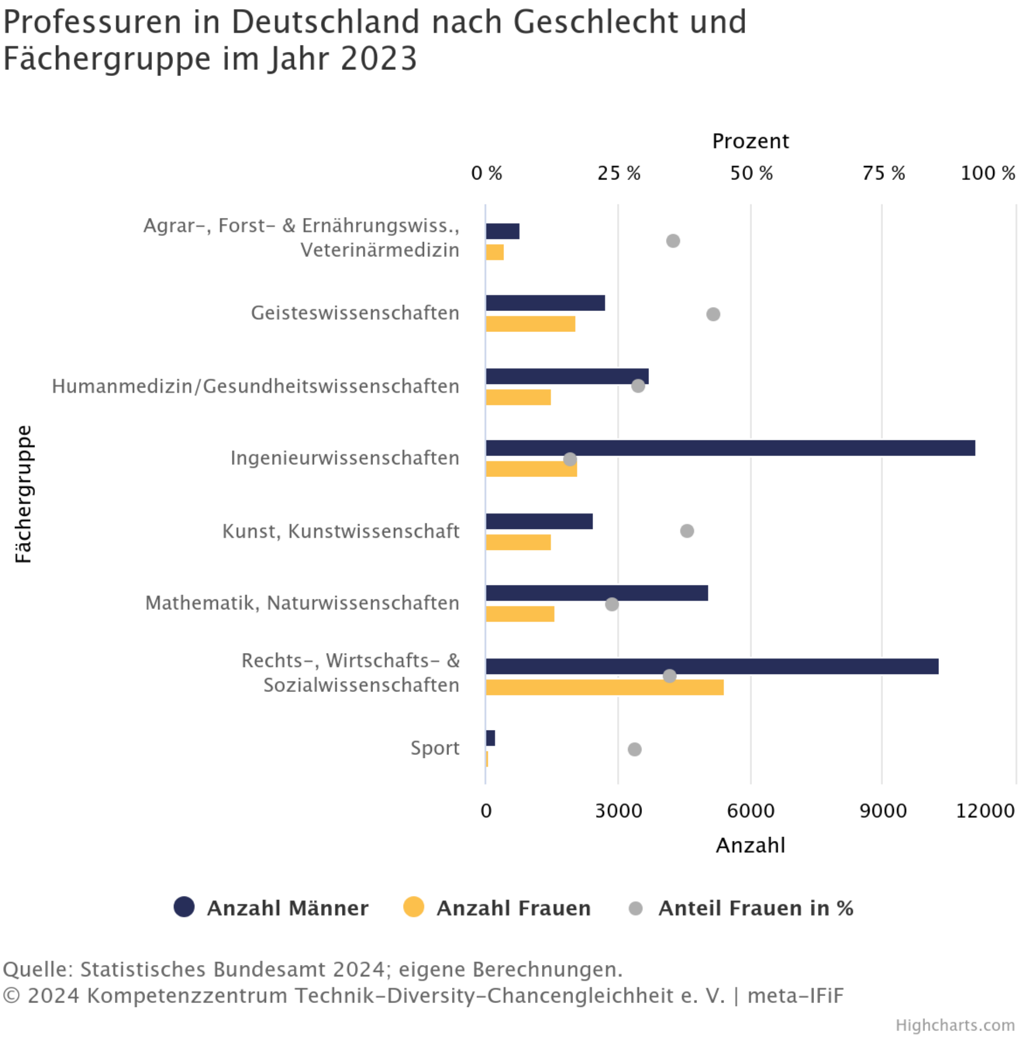
## 3.1. From Visibility to Discrimination Detection

The starting point of this thesis was the interdisciplinary project **Prof:inSicht**, which investigated the visibility of female professors at German universities of applied sciences (UAS). These institutions, specific to the German and Austrian academic system, prioritise applied teaching and collaboration with industry and are often perceived as less prestigious than traditional research universities. The project explored how lower institutional prestige and gendered disparities in recognition interact to shape the public and professional visibility of female professors. The project's empirical focus lay on two contrasting disciplines: computer science and the social sciences. These fields differ markedly in gender composition, publishing cultures, and visibility practices. According to national statistics (see Figure 3.1), the social sciences (classified under *Rechts-, Wirtschafts- & Sozialwissenschaften)* have the highest number of female professors, whereas computer science (classified under *Ingenieurwissenschaften*) has the highest number of male professors. Visibility practices also vary: computer scientists often publish in conference proceedings and use platforms such as LinkedIn, whereas social scientists tend to publish in journals and follow different norms of media engagement (Spagert and Wolf, 2025). These disciplinary contrasts made the fields ideal case studies for exploring how visibility is shaped by intersecting dynamics of gender and academic culture.

The project adopted a interdisciplinary analytical framework, integrating perspectives from social sciences, economics, and computer science. Visibility was conceptualised not merely as representation, but as a process embedded in academic hierarchies and disciplinary norms. A shared conceptual foundation was developed using the terms **dimensions**, **economies**, and **practices**, inspired by Bourdieu's field theory (Scherr, 2016) and the interactional approach of *doing gender* (West and Zimmerman, 1987). This led to the concept of *doing visibility* (Fischer et al., 2024), which conceptualises visibility as a co-constructed, ongoing process: individuals enact visibility through specific practices that must be recognised and interpreted by others.

In my first work-package I focused on visibility in digital infrastructures. From a computer science perspective, I examined how female professors are discovered and represented in systems such as Google and academic databases. This included a comparative analysis of publication accessibility across the ACM Digital Library (computer science), Beltz (social sciences), and SpringerLink (interdisciplinary), as well as the representation of experts in search engine results. These findings contributed to a broader conceptual discussion on **algorithmic gender fairness** (cf. Chapter

4, providing a theoretical foundation for the rest of the thesis. My second work package focused on generative language technologies. Here, I analysed how women are represented in the outputs of large language models (LLM), using ChatGPT (OpenAI, 2023) as a case study. While overt discrimination was rare, I found that post-hoc fairness interventions could inadvertently reinforce gender stereotypes. These insights revealed the limitations of downstream debiasing (cf. Chapter 5). Furthermore, I developed a language-agnostic pipeline for detecting and mitigating gender discrimination in corpora to help LLM developers create fairer LLMs. The pipeline is detailed in the sections below.



Quelle: Statistisches Bundesamt 2024; eigene Berechnungen.
© 2024 Kompetenzzentrum Technik-Diversity-Chancengleichheit e. V. | meta-IFiF

Figure 3.1.: Number of male and female professors in different fields at German universities and UAS in 2023.

## 3.2. Identifying Discrimination Without Judging It

Building on the findings in ChatGPT, I turned my attention to the training data of LLMs, where gender discrimination is often embedded long before model outputs are generated. Analysing such data required a shift in methodological approach. Most existing techniques for detecting discrimination in text, such as hate speech detection (Paz et al., 2020), ambivalent sexism analysis (Jha and Mamidi, 2017), microaggression detection (Breitfeller et al., 2019), condescending language detection (Wang and Potts, 2019), stereotype identification (Joseph et al., 2017), or analyses of the portrayal of women (Wagner et al., 2021), tend to be highly specialised or language-specific, limiting their applicability to large-scale or multilingual datasets.

To develop a more general and language-agnostic method, I turned to work at the intersection of social sciences and linguistics, where discrimination in language has been systematically studied for nearly a century (Myrdal et al., 1944; Razran, 1950; Allport et al., 1954). A central resource in this context was the overview of Reisigl (2017) of linguistic discrimination research, which defines discrimination as a social act that disadvantages or privileges someone based on a distinguishing feature such as gender, race, or sexual orientation. From this definition, five key components of discrimination can be identified:

1. The **offender**, who carries out the act,
2. The **victim** or beneficiary (in the case of positive discrimination),
3. The **disadvantaging or favouring act or process**,
4. A **comparison group** that is treated differently,
5. The **distinguishing feature** (e.g. gender) that grounds the unequal treatment.

Building on this framework, and supported by the functional approach to discriminatory speech acts proposed by Graumann and Wintermantel (2007), I conceptualise discrimination in written text as a manifestation of social discrimination. In this setting, the author of a text functions as the offender, and actors mentioned in the text may become victims. The comparison group is implicit in the analysis, and the distinguishing feature is the actor's gender. The disadvantaging act is reflected in quantitative and qualitative asymmetries in representation.

To scope the work, I focus exclusively on gender discrimination. While other forms of discrimination, especially in intersectional combinations, are highly relevant, they are not addressed in this thesis. The goal is not to label texts as discriminatory or non-discriminatory, but to generate structured discrimination reports that summarise gender representation. These reports offer a transparent, interpretable basis for human judgment.

A methodological fit for this approach can be found in linguistic discourse analysis, which investigates how actors are referred to and described in texts. Two key mechanisms in this pipeline are:

- **Nomination**: how actors are named,
- **Predication**: how actors are described.

Because these mechanisms align with tasks commonly addressed in information extraction (IE), I can draw on established IE methods to automatically extract nomination and predication. This enables the development of a computational pipeline that extends existing approaches in computational discourse analysis in a novel direction.

More complex tasks such as analysing argumentation strategies, stance detection, or assessing whether discriminatory content is reinforced or mitigated are explicitly out of scope, as their inclusion would have exceeded the scope of this dissertation. These aspects may be explored in future work.

The pipeline was first implemented and tested on individual English texts, including both news articles and generative model outputs (cf. Chapter 6).

The resulting pipeline consists of four main steps (see Figure 3.2):

1. **Actor extraction**: identifying individuals mentioned in a text.
2. **Gender approximation**: estimating the likely gender of each actor using only internal textual information (pronouns).
3. **Predication extraction**: collecting the linguistic context surrounding each actor.
4. **Discrimination analysis**: analysing nomination and predication for potential markers of unequal representation and compiling the results into a discrimination report.

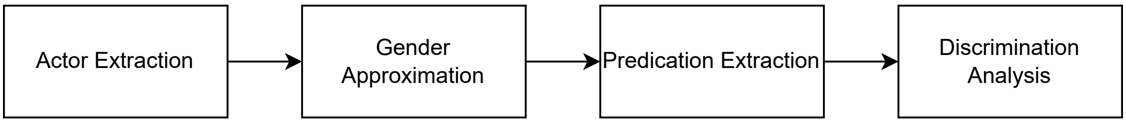


Figure 3.2.: Flexible language-agnostic pipeline developed in this thesis.

A central strength of the pipeline is that it operates without the need for external knowledge bases, pretrained gender classifiers, or labelled datasets. All information is derived directly from the text, making the method robust, flexible, and applicable to large corpora in diverse domains and languages.

To capture patterns of unequal treatment, the discrimination analysis component relies on a set of linguistically motivated markers. These markers reflect different aspects of representational asymmetries:

- **Count of actors:** Number of actors identified as women, men, non-binary, or undefined, both overall and per actor. This captures the visibility and presence of different genders.
- **Count of mentions:** Number of times actors are mentioned, aggregated by gender. This metric distinguishes between texts where few actors are frequently mentioned and those where many actors appear only once. For example, one woman mentioned ten times is not equivalent to ten men mentioned once each, even if both result in ten gendered references. The mention count thus captures the narrative prominence of individuals, not just their presence.
- **Sentiment:** Sentiment associated with each actor and aggregated per gender, providing insights into evaluative framing.

- **Gender-coded language:** Frequency of feminine- and masculine-coded terms in predications, capturing subtle stereotypical framings.
- **Abusive language:** Instances of abusive terms directed at actors of different genders (if present).

Each of these markers aligns with one or more components of the theoretical discrimination definition, particularly the disadvantaging act, the comparison group, and the distinguishing feature. They are designed to surface structural asymmetries and support critical interpretation without making normative claims.

A related initiative is the *Gender Gap Tracker* by Asr et al. (2021), which analyses gender disparities in quotation practices within Canadian news media. Similar to the approach described in Chapter 6, the system first extracts `PERSON` entities from the text and then clusters them into actor representations. For gender identification, however, the authors deliberately avoid using pronouns, as not all clusters contain sufficient pronominal cues. Instead, they rely on an external service that infers gender based on names. While this practice is widely considered problematic within the NLP community due to its cultural biases and lack of inclusivity, the authors explicitly acknowledge these concerns. They justify their decision as a pragmatic compromise, noting the absence of better alternatives for their specific use case (Asr et al., 2021). In contrast, the approach developed in this thesis relies exclusively on internal textual cues to approximate gender, avoiding external resources and thereby enabling a more robust, self-contained, and language-agnostic solution.

## 3.3. From Single Texts to Large Corpora

While the proof of concept demonstrated that the pipeline could successfully analyse discrimination at the level of individual texts, studying training data for LLMs required significant scaling. In the next phase of this thesis, I extended the pipeline to process full corpora and adapted it for application to German-language newspaper texts (cf. Chapter 7).

Since no full-text German newspaper corpus was publicly available, I contacted multiple publishers directly. Only *taz* (Die Tageszeitung) granted permission for both the (free of charge) use and publication of their data. This made it possible to compile a large-scale corpus of over 1.8 million articles published between 1980 and 2024, which now represents the largest publicly available German newspaper corpus (at the time this thesis is published).

I made several methodological and technical adjustments to adapt the pipeline for this new context. These changes allowed the analysis to scale from single texts to thousands at once, while also accounting for linguistic features specific to German. The following modifications were introduced:

- **Gender Assumption:** As in the English-language proof of concept, gender is not assumed based on names or external databases. Instead, I determine the gender of actors based on the primary pronoun used to refer to them. For the German adaptation, this approach was adjusted to reflect German pronoun usage. To account for non-binary identities, I scanned the corpus for German neo-pronouns[1] but identified only five instances. Consequently, the analysis focuses on the dominant binary pronouns *sie* (she) and *er* (he). For reporting purposes, actors primarily referred to with *sie* are categorised as women, and those referred to with *er* as men.

- **Generic Masculine and Gender-Neutral Language Markers:** I introduce two binary markers to detect linguistic practices that are particularly relevant in German: one indicating the use of the *generic masculine* and one for the presence of explicitly *gender-neutral language*.
- **Pairwise Mutual Information (PMI):** To better understand how actors are described, I compute PMI scores for adjectives appearing in their predications. PMI measures the association between words based on their co-occurrence probabilities (Jurafsky and Martin, 2025). For each actor, I calculate PMI scores for all adjectives (excluding stop words) that appear in their predication context. The top 10 adjectives with the highest PMI values are identified as the most characteristic descriptions. PMI is defined as:

$$\mathsf{PMI}(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

- **Aggregated Report:** To support interpretation at scale, I developed a human-readable aggregated discrimination report. This format facilitates year-by-year or category-based comparisons and provides a practical tool for corpus-wide analysis.
- **Multiprocessing:** I implemented multiprocessing to enable parallel execution of independent processing steps. This was necessary to handle the size of the corpus efficiently, as the initial sequential implementation led to unreasonable long runtimes. By distributing tasks across multiple CPU cores, the pipeline can now process large volumes of text substantially faster, making it feasible to run analyses on the full dataset.

In addition to adapting the system for the German language and larger scale, I also refined the architecture. The abusive word module was removed, as slurs are rarely found in formal journalistic writing. With the addition of PMI scoring and language-specific gender markers, the pipeline was now equipped to detect long-term trends in actor representation and gendered language use.

These enhancements preserve the core principle of the pipeline: offering a structured, interpretable representation of gender portrayal in text, without making normative judgments about whether a given article or corpus is discriminatory. By operating independently of external knowledge sources, the pipeline remains transparent and applicable to a wide range of domains and languages.

## 3.4. Closing the Loop: From Detection to Intervention

Having scaled the pipeline for large-scale corpus analysis, the final phase of this thesis shifted the focus from diagnosis to intervention. The aim was to explore whether representational asymmetries in language use, particularly regarding who is named, quoted, or attributed agency, can be reduced

[1] I used the list of neo-pronouns published at `https://gleichstellung.tu-dortmund.de/projekte/klargestellt/neo-pronomen`.

directly at the level of training data. Rather than relying on downstream debiasing techniques, this stage introduced an upstream filtering and balancing framework that curates corpora based on transparent, user-configurable criteria (cf. Chapter 8).

To support this, I developed a multi-stage extension of the pipeline that allows for targeted article exclusion based on actor-level discrimination metrics, followed by global corpus-level balancing. The process unfolds in four key steps:

- **Text-Level Filtering:** Articles are flagged for exclusion based on four framing asymmetries: sentiment disparity, syntactic agency imbalance, quotation style differences, and referential asymmetries (named vs. pronominal mentions). Each metric is calculated per article using Laplace-smoothed ratios to reduce instability. Thresholds are configurable by the user during runtime, with default values set to detect pronounced asymmetries: a sentiment gap above 0.3, a subject-to-object ratio difference exceeding 0.5, an indirect-to-direct quotation ratio exceeding 0.5 and a named-to-pronominal mention ratio exceeding 0.5. Texts exceeding a user-defined number of these thresholds are excluded from the corpus.

- **Impact-Aware Corpus Balancing:** After text-level filtering, a second exclusion step is applied at the corpus level to restore referential parity between male- and female-coded actors. The user defines an acceptable **global equilibrium range** for actor and mention ratios (default: $[0.75, 1.25]$ meaning that each gender may occur up to 25% more than the other one in the corpus). Articles that contribute most to the imbalance are iteratively removed until both ratios fall within the specified equilibrium. This step ensures that residual skews from individual articles do not result in systemic underrepresentation.

- **Visualisation and Reporting:** Both filtering stages are accompanied by diagnostic histograms and time-series plots, showing the distribution of gender representation across articles. The system also generates structured exclusion logs, enabling full reproducibility and transparency. Updated gender ratio distributions illustrate the effects of each intervention step.

- **Corpus Reconstruction:** All articles marked for exclusion are removed from the original dataset. The resulting corpus is stored separately and represents a more balanced and equitable dataset for downstream use. Importantly, the core document structure remains unchanged, enabling easy substitution in training pipelines.

In addition to the introduction of filtering and balancing mechanisms I implemented, several methodological and architectural enhancements to improve the analytical depth and accessibility of the pipeline:

- **Syntactic Role Annotation:** The pipeline now detects whether an actor appears in grammatical subject or object position. This enables the analysis of discursive agency, since subjects typically perform actions while objects are acted upon (Halliday, 2004).

- **Quotation Style Detection:** Using punctuation and reporting verbs, the pipeline now distinguishes direct from indirect speech. This makes it possible to assess whether actors are quoted in their own words or paraphrased, a difference often linked to perceived authority and narrative presence (Bendel Larcher, 2015).

- **Naming vs. Pronominal Reference:** The system now tracks whether actors are referred to by name or only via pronouns. Named references often indicate higher salience and individuation, while exclusive pronoun usage may suggest de-individuation or backgrounding (Bendel Larcher, 2015).
- **Extended PMI Scoring:** The calculation of PMI was expanded beyond adjectives to also include the top ten **nouns** and **verbs** per actor. This enriches the analysis of thematic and role-specific framing tied to gendered actors.
- **Structured Yearly Reports:** The yearly analysis output was redesigned for greater clarity and accessibility. Reports now include dedicated sections for summary statistics, syntactic roles, sentiment, and PMI results, with consistent formatting across gender groups to facilitate interpretation.

The balancing extension reflects a central principle of this thesis: rather than imposing fixed thresholds for fairness, the pipeline provides users with the flexibility to define, detect, and mitigate discrimination in ways that are appropriate to their domain and goals. This enables the creation of training datasets with reduced gender asymmetries while preserving the essential role of human oversight in the decision-making process.

By avoiding rigid labels or universal rules, the approach supports nuanced, context-sensitive corpus curation. It recognises that what constitutes discriminatory imbalance may vary across applications, domains and regions. Rather than replacing human judgment, the pipeline is designed to enhance it: through transparent metrics, interpretable reports, and practical tools for critically engaging with textual data.

Ultimately, this final stage closes the loop of the research trajectory. Beginning with a definition of fairness and a case study on representation in search, the work has gradually shifted towards upstream interventions. Through an iterative methodological approach, the thesis offers a language-agnostic, actor-centred pipeline capable of both detecting and reducing representational asymmetries in large textual corpora.

# Part II.

# Algorithmic Gender Fairness

# 4. Are All Genders Equal in the Eyes of Algorithms? - Analysing Search and Retrieval Algorithms for Algorithmic Gender Fairness

Chapter 4 defines algorithmic gender fairness in the context of search and information retrieval systems, grounded in the concept of equality. The definition is tested by comparing the digital visibility of female and male professors in computer science and social work/social pedagogy at universities of applied sciences and universities in Germany. Using self-reported university profiles as a baseline, the study analyses publication database results and Google search outcomes. The findings reveal subtle but consistent gender differences in visibility and representation, highlighting the need for more transparent and fair algorithmic systems.

### Contributing article:

Urchs, S., Thurner, V., Aßenmacher, M., Bothmann, L., Heumann, C. and Thiemichen, S.(2025). Are All Genders Equal in the Eyes of Algorithms? - Analysing Search and Retrieval Algorithms for Algorithmic Gender Fairness. Accepted at the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KDIR)

### Copyright information:

### Author contributions:

The idea for the publication is based on the BMBF-funded Prof:inSicht project, which was acquired by Veronika Thurner. Stefanie Urchs was responsible for designing the publication, collecting the data, conducting the experiments, researching the literature and writing the manuscript. Ludwig Bothmann wrote the chapter on fairness. Veronika Thurner, Matthias Aßenmacher, Christian Heumann and Stephanie Thiemichen provided support in the form of content discussions, technical support and correction of the manuscript.

### Supplementary material available at:

- arXiv link: http://arxiv.org/abs/2508.05680

# Are All Genders Equal in the Eyes of Algorithms? - Analysing Search and Retrieval Algorithms for Algorithmic Gender Fairness

Stefanie Urchs[1,2][a], Veronika Thurner[1][b], Matthias Aßenmacher[2,3][c], Ludwig Bothmann[2], Christian Heumann[2][d], Stephanie Thiemichen[1][e]

[1]*Faculty for Computer Science and Mathematics, Hochschule München University of Applied Sciences, Munich, Germany*

[2]*Department of Statistics, LMU Munich, Munich, Germany*

[3]*Munich Center for Machine Learning (MCML), LMU Munich, Munich, Germany*

{*stefanie.urchs, veronika.thurner, stephanie.thiemichen*}*@hm.edu,*
{*matthias, ludwig.bothmann, christian.heumann@stat.uni-muenchen.de*}

Keywords: algorithmic fairness, academic visibility, information retrieval, search engines, gender fairness

Abstract: Algorithmic systems such as search engines and information retrieval platforms significantly influence academic visibility and the dissemination of knowledge. Despite assumptions of neutrality, these systems can reproduce or reinforce societal biases, including those related to gender. This paper introduces and applies a bias-preserving definition of algorithmic gender fairness, which assesses whether algorithmic outputs reflect real-world gender distributions without introducing or amplifying disparities. Using a heterogeneous dataset of academic profiles from German universities and universities of applied sciences, we analyse gender differences in metadata completeness, publication retrieval in academic databases, and visibility in Google search results. While we observe no overt algorithmic discrimination, our findings reveal subtle but consistent imbalances: male professors are associated with a greater number of search results and more aligned publication records, while female professors display higher variability in digital visibility. These patterns reflect the interplay between platform algorithms, institutional curation, and individual self-presentation. Our study highlights the need for fairness evaluations that account for both technical performance and representational equality in digital systems.

## 1 INTRODUCTION

Algorithms are increasingly embedded in nearly every aspect of our daily lives, shaping the information we encounter and influencing our perceptions and decisions. From social media recommendations to online shopping suggestions, algorithmic processes impact what we see, how we engage, and ultimately how we make choices. Among these, algorithms in search engines and publication databases have significant power in determining which information, content, and experts are made visible to users, directly influencing public knowledge, career opportunities, and academic visibility. For instance, studies have shown that job advertisements displayed by search engines can be targeted by gender (Datta et al., 2015; Eren et al., 2021), image search results prefer white individuals (Makhortykh et al., 2021) and text-based search results sexualise woman, especially from the global south (Urman and Makhortykh, 2022), raising significant concerns about the presence and impact of gender-based bias in these systems. Such examples underscore the urgency of examining and defining algorithmic fairness, particularly regarding gender representation, as these biases risk perpetuating and amplifying existing societal inequities.

Algorithmic gender fairness is essential because these biases are not merely technical flaws but reflections of deeper societal structures embedded in data and system design. Algorithms do not operate in isolation; they are shaped by the data they are trained on, the objectives they are optimised for, and the societal context in which they function. Addressing gender

[a] https://orcid.org/0000-0002-1118-4330
[b] https://orcid.org/0000-0002-9116-390X
[c] https://orcid.org/0000-0003-2154-5774
[d] https://orcid.org/0000-0002-4718-595X
[e] https://orcid.org/0009-0001-8146-9438

fairness requires navigating the intersection of mathematical criteria and social implications, as technical fixes alone cannot resolve biases rooted in historical and structural inequalities. Without a well-defined framework for fairness, efforts to mitigate algorithmic discrimination risk being inconsistent or even counterproductive. Therefore, a clear and robust definition of algorithmic gender fairness is crucial, not only to prevent direct and indirect discrimination but also to establish transparency, accountability, and trust in automated systems.

Building upon existing research in algorithmic fairness and algorithmic gender fairness, this work contributes to the ongoing discourse by proposing and empirically testing a definition of algorithmic gender fairness. While many studies have explored fairness in algorithms, our approach focuses on evaluating two influential types of systems: publication database retrieval algorithms and Google's search engine. These algorithms play a crucial role in shaping public visibility and access to information, making them particularly impactful subjects for analysis. By applying our fairness definition to these systems, we aim to offer insights into their performance, identify improvement areas, and contribute to developing more transparent, accountable, and inclusive algorithmic designs.

## 2 BACKGROUND

To define algorithmic gender fairness, we begin by outlining how we understand the core concepts of gender and fairness. Given the interdisciplinary nature of this work, the section is deliberately extensive. In the final part of the section, we first introduce how information retrieval and search engines work in general, providing the necessary technical background for readers unfamiliar with the field. We then review existing research on algorithmic fairness in these domains and highlight how our approach differs from previous work.

### 2.1 Gender

The term "gender" encompasses at least three distinct concepts: linguistic gender, sex, and social gender. Each concept has unique implications in various professional and private contexts, especially when considering algorithmic representation, identity, and fairness issues. Linguistic or grammatical gender is defined as "*[...] grammatical gender in the narrow sense, which involves a more or less explicit correlation between nominal classes and biological gender (sex).*" (Janhunen, 2000). In many languages, nouns and pronouns are assigned a gender, classified as feminine, masculine, or neutral, often loosely correlated with perceived biological characteristics (Kramer, 2020). This linguistic categorisation can affect the way gender roles and identities are understood culturally, as language shapes and reinforces social expectations (Konishi, 1993; Phillips and Boroditsky, 2013).

"Sex", on the other hand, is traditionally understood as a biological categorisation, regarded as "*binary, immutable and physiological*" (Keyes, 2018). However, a strict binary framework is increasingly recognised as insufficient for representing the full spectrum of human diversity. Intersex individuals, who may not fit the conventional definitions of feminine or masculine due to variations in physiological characteristics (Carpenter, 2021), and transgender individuals, whose gender identity differs from their sex assigned at birth (Beemyn and Rankin, 2011), exemplify the limitations of this binary, immutable perspective. The presence of these identities challenges the conventional definitions of sex.

In our work, we embrace the concept of social gender, which goes beyond biological and linguistic classifications to encompass a socially constructed identity shaped by behaviours, expressions, and self-presentation. Social gender is fluid, non-binary, and co-constructed through social interactions, allowing it to evolve over time in alignment with an individual's sense of self. This perspective aligns with research that views gender not as an inherent or static characteristic but as a performative act shaped by personal expression and social context (West and Zimmerman, 1987; Devinney et al., 2022).

Although we adopt this inclusive understanding of gender, our study faces limitations due to the constraints in our data. The available information only allows for analysing participants within the binary gender spectrum, and we were thus unable to identify trans or intersex individuals in the dataset. As a result, our empirical analysis focuses on binary gender categories. However, the underlying framework of our proposed definition of algorithmic gender fairness remains rooted in the concept of social gender – emphasising its non-binary, flexible, and socially co-constructed nature. We aim to contribute to a broader, more inclusive understanding of gender fairness in algorithmic systems, even as we acknowledge the current limitations of our dataset.

### 2.2 Fairness

The term "fairness" is increasingly used in the field of algorithmic decision-making and "fairness-aware machine learning" [fairML, surveys can be found, e.g., in (Caton and Haas, 2024; Verma and Rubin, 2018)]. However, few contributions concretely define the meaning of this term as a philosophical concept, with positive exceptions to be found in (Bothmann et al., 2024; Loi and Heitz, 2022; Kong, 2022). Fairness is usually described by synonyms such as equality, justice, or the absence of bias or discrimination.

A crucial component of fairness as a philosophical concept is that it concerns the treatment of individuals (Aristotle, 2009; Dictionary, 2022; Dator, 2017; Kleinberg et al., 2017). The basic structure of the concept can be traced back to Aristotle and relates fairness to equality: A decision or treatment is fair if equals are treated equally and unequals are treated unequally. As (Bothmann et al., 2024) point out, this requires the normative definition of task-specific equality, that is: Two individuals may be equal in one task (e.g., buying a croissant in a bakery), but unequal in another task (e.g., paying taxes). Deciding how to treat unequals is also a normative task.

The role of protected attributes such as gender or race is that they can normatively alter the definition of task-specific equality. For example, a society may decide that the grievance of the gender pay gap is not the responsibility of an individual and that in deciding whether to grant a loan, income should therefore be fictitiously corrected for this real-world bias; (Bothmann et al., 2024) call this a fictitious, normatively desired (FiND) world, and advocate making decisions using data from this world rather than real-world data. (Wachter et al., 2021) describe such an approach as "bias-transforming", aiming at "substantive equality", because a real-world bias should be "actively eroded" to make the world fairer.

In contrast, (Wachter et al., 2021) describes approaches as "bias-preserving", aiming at "formal equality", if they try to reflect the real world as accurately as possible, i.e., without introducing new biases that may even increase the real-world biases. Many fairML metrics, such as equalised odds or predictive parity, can be categorised as bias-preserving because they measure against real-world labels but try to balance the errors thus measured across levels of the protected attribute. Sometimes the concept of bias-transforming methods is referred to as aiming for "equity", while bias-preserving approaches are referred to as aiming for "equality". In our work, we will follow a bias-preserving approach to adequately or "correctly" reflect individuals in the real world while prohibiting the introduction of gender bias by information retrieval algorithms or search engines (in addition to the already existing gender bias in the real world).

## 2.3 Information Retrieval and Search Engines

Information Retrieval (IR) focuses on finding relevant material, typically text documents, to satisfy a user's information need. An information need represents the user's underlying intention or goal when seeking information. At the same time, a query explicitly represents this need, usually entered as keywords or phrases in a search engine. These concepts are fundamental in bridging the gap between human intentions and computational processing, ensuring that search systems accurately interpret and address user needs (Schütze et al., 2008).

An Information Retrieval System (IRS) is a software system that efficiently stores, manages, and retrieves information from large datasets. An IRS relies on indexing and searching algorithms to match user queries with relevant documents. Retrieval systems can be categorised based on their retrieval models, with the two primary examples being Boolean Retrieval and Vector Space Retrieval. Boolean Retrieval allows users to formulate queries using logical operators such as AND, OR, and NOT, ensuring that documents are returned only if they satisfy the Boolean expression. On the other hand, the Vector Space Model represents documents and queries as vectors in a multi-dimensional space, using similarity measures like cosine similarity to rank results by relevance. In document retrieval, user queries are matched against different parts of documents, such as title, keywords, author name(s), and abstract. These metadata fields often provide valuable signals for relevance, enabling the system to prioritise results more effectively. An IRS typically employs inverted indexes, which map each term to a list of documents containing it, facilitating rapid query processing. Additionally, ranking algorithms ensure that results are retrieved and presented in an order reflecting their relevance to the user's query (Schütze et al., 2008).

Recent research highlights a critical issue within IRS: the presence of biases in their structure and outcomes (Fang et al., 2022). These biases can emerge from relevance judgment datasets, neural representations, and query formulation. Relevance judgment datasets, often regarded as gold-standard benchmarks, may carry stereotypical gender biases, propagating into ranking algorithms when IRS are trained on such datasets (Bigdeli et al., 2022). Additionally, neural embeddings used for query and document representations, pre-trained on large corpora, are susceptible to inheriting societal biases present in those datasets (Bolukbasi et al., 2016). Retrieval methods, especially those using neural architectures, have shown a tendency to intensify pre-existing gender biases (Francazi et al., 2024). Bias-aware ranking strategies, such as adversarial loss functions, bias-aware negative sampling, and query reformulation techniques (e.g., AdvBERT), have been proposed to reduce these biases while maintaining retrieval effectiveness. Researchers emphasise the importance of balancing retrieval performance with fairness, advocating for systematic evaluation metrics and datasets explicitly designed for measuring and mitigating gender biases in IRS (Bigdeli et al., 2022). Prior work on fairness in information retrieval has largely focused on technical interventions in ranking systems (e.g., (Singh and Joachims, 2018); (Geyik et al., 2019)) or on consumer-side fairness (Ekstrand et al., 2022), typically evaluating search and recommendation systems in general-purpose digital platforms. In contrast, few empirical studies have investigated gender fairness in academic retrieval contexts. Our work bridges this gap by conducting a fairness audit of academic visibility, applying a bias-preserving fairness perspective to both domain-specific publication databases and general-purpose search engines. In doing so, we extend the methodological orientation of studies like (Bigdeli et al., 2022) and (Fang et al., 2022) to a new sociotechnical domain. For instance, Singh and Joachims (Singh and Joachims, 2018) propose formal fairness constraints on exposure in rankings, ensuring that protected groups receive visibility proportional to their relevance. Their framework relies on probabilistic rankings to balance user utility and provider fairness in expectation.

Search engines are advanced Information Retrieval Systems tailored for web-scale datasets. They consist of three primary components: crawling, indexing, and query processing. Crawlers systematically fetch web pages indexed using data structures like inverted indexes. Query processing involves parsing the user's input and matching it with indexed documents. The PageRank algorithm, introduced by Google, revolutionised web search by considering the hyperlink structure of the web. Each webpage is assigned a numerical score based on the quantity and quality of incoming links. The algorithm models a "random surfer" who follows hyperlinks or randomly jumps to other pages. This behaviour is mathematically represented using Markov Chains, and steady-state probabilities are computed iteratively to determine the importance of each page. Search engines blend PageRank with other ranking factors, including content relevance, term proximity, and user-specific data, creating a hybrid scoring system that delivers highly accurate search results (Schütze et al., 2008).

However, search engines are not immune to biases. Biases in search engines can emerge from data sources, crawling strategies, and ranking algorithms, resulting in the reinforcement of stereotypes, underrepresentation of marginalised groups, or discriminatory exposure of content. Biases may also be amplified over time through dynamic adaptation mechanisms, where user interactions create feedback loops that reinforce pre-existing biases. Addressing these biases requires mitigation strategies such as bias-aware re-ranking algorithms, adversarial training, and query reformulation techniques (Ekstrand et al., 2022).

Additionally, fairness concerns in search engines align with consumer fairness (ensuring users receive equally relevant and satisfying results across diverse groups) and provider fairness (ensuring content creators or document providers receive equitable exposure in rankings). Evaluation methodologies play a key role in addressing these concerns, often combining relevance metrics with fairness-aware metrics to strike a balance between accuracy and equity (Ekstrand et al., 2022). In industrial applications, (Geyik et al., 2019) present a fairness-aware re-ranking framework deployed at scale in LinkedIn Talent Search. Their system enforces minimum representation thresholds through post-processing algorithms, demonstrating that fairness and utility can coexist in production systems. However, their approach is grounded in fairness-transforming principles such as demographic parity.

In practice, search engines represent a complex interplay between technical architecture, algorithmic fairness, and societal values. Continuous research and refinement are essential to ensure these systems meet efficiency and fairness criteria simultaneously (Ekstrand et al., 2022).

## 3 ALGORITHMIC GENDER FAIRNESS

To define algorithmic gender fairness, we build upon the theoretical framework presented in Section "Fairness" and the practical insights discussed in Section "Information Retrieval and Search Engines". Our approach adopts a bias-preserving perspective, aiming to reflect real-world distributions without introducing new distortions or exacerbating existing gender biases.

Bias in algorithmic systems can arise from several sources, including biased training datasets, pre-existing societal inequalities, and the interaction between users and algorithmic feedback loops (for Justice et al., 2021). Gender biases, in particular, are often perpetuated through historical inequalities encoded in data, proxies that stand in for protected attributes, and opaque decision-making processes inherent to many machine-learning systems.

At the data stage, biases can emerge from training datasets that reflect societal inequalities, including historical gender pay gaps or occupational stereotypes. These biases are often amplified when algorithms learn patterns from these datasets without critical oversight. From a bias-preserving perspective, systems should strive to reflect gender distributions accurately without further entrenching societal disparities. However, achieving this requires ongoing monitoring and transparency to detect and address unintended distortions.

At the algorithmic stage, gender biases can manifest in ranking systems, recommendation algorithms, or classification processes. Proxy variables, such as zip codes, browsing behaviour, or inferred demographic data, often serve as indirect markers for gender, leading to indirect discrimination. Mitigating these biases involves identifying such proxies and adjusting algorithmic models to ensure they do not disproportionately disadvantage individuals based on gender (for Justice et al., 2021).

From a bias-transforming perspective, algorithms may be adjusted proactively to counteract historical inequalities and actively reshape outcomes. Such approaches aim for substantive equality, where systems not only avoid perpetuating existing biases but actively correct for them by introducing calibrated adjustments to outputs (for Justice et al., 2021). Such fairness interventions are often formalised as constrained optimisation problems, where utility (e.g., accuracy or public safety) is maximised subject to fairness constraints. (Corbett-Davies et al., 2017) demonstrate that implementing common fairness definitions, such as statistical parity or predictive equality, typically requires group-specific decision thresholds, a trade-off that can reduce utility or violate principles of equal treatment.

Transparency and explainability remain central challenges in algorithmic gender fairness. The opacity of many systems, particularly those based on deep-learning architectures, makes it difficult to detect and address gender biases effectively. Without clear explanations of how decisions are reached, it becomes challenging to hold systems accountable for gender-discriminatory outcomes.

Additionally, intersectionality plays a crucial role in algorithmic gender fairness. Gender does not exist in isolation but intersects with other protected attributes such as race, age, or socio-economic status, leading to compounded forms of bias and discrimination. Addressing intersectionality requires fairness-aware metrics that account for these overlapping dimensions (for Justice et al., 2021).

Our approach focuses on bias-preserving fairness as the guiding principle, ensuring that algorithmic systems in information retrieval and search engines reflect real-world gender distributions without introducing additional biases. While our approach focuses on preserving bias patterns as they exist in real-world data, many prior works have proposed alternative fairness frameworks. Žliobaitė (Žliobaitė, 2017) offers a systematic overview of such fairness definitions in algorithmic decision-making, highlighting group fairness notions such as statistical parity, conditional parity, and predictive parity, as well as individual fairness principles based on similarity of treatment. While bias-transforming approaches, which aim to correct historical inequalities proactively, offer an appealing vision of fairness, they require defining an ideal dataset or outcome, a "perfect world", to serve as a benchmark. However, defining such an ideal world is inherently challenging, given the vast diversity of cultural, social, and political value systems across the globe. Even if we attempted to define it, measuring an ideal world would remain an insurmountable task, as no dataset could comprehensively capture such a reality.

Given these constraints, we adopt a bias-preserving approach, which evaluates whether algorithms accurately reflect and replicate the analogue reality within the digital domain without amplifying existing biases. This approach leverages measurable real-world data, allowing us to assess algorithmic outcomes in relation to observed societal distributions.

Therefore, we define algorithmic gender fairness as:

> *The ability of algorithmic systems, particularly in information retrieval and search engines, to accurately reflect real-world gender distributions and representations in their outputs without introducing, amplifying, or reinforcing existing biases.*

In this paper, we apply the above definition of algorithmic gender fairness to evaluate real-world systems that mediate academic visibility. While prior studies have primarily focused on technical fairness interventions or theoretical proposals, our contribution lies in conducting a fairness audit grounded in this definition, using empirical data from both domain-specific academic databases and a general-purpose search engine. By doing so, we extend the application of fairness frameworks to a previously underexplored domain: the digital representation of academic expertise.

## 4 EXPERIMENTS

We test our notion of algorithmic gender fairness by analysing the online visibility of professors through two distinct types of algorithmic systems: search algorithms, exemplified by Google, and information retrieval algorithms used in academic publication databases. While Google clearly ranks results through its proprietary search algorithm, the publication databases return results based on ”relevance”, a criterion that remains undefined by the platforms. Consequently, we do not compare the results directly but instead analyse each system separately to explore how algorithmic structures may influence visibility across gender lines.

### 4.1 Data

The data for this study stems from a broader research project that investigated the visibility of female professors at universities of applied sciences (UAS)[1] in Germany. The full dataset includes professors from different institutional types (universities and universities of applied sciences) and academic disciplines (computer science and social work/social pedagogy). This heterogeneity was intentional: to capture a broad spectrum of academic visibility, we aimed for maximum variation within the German academic landscape. Including both institutional types reflects structural differences in prestige, mission, and digital presence. Moreover, computer science and social sciences follow distinct publication cultures: computer science is predominantly conference-driven, while social scientists typically publish in journals.

As the main focus of the project lay on female professors at UAS, we manually collected the full population of women professors working in the departments of computer science and social work at these institutions. To provide a meaningful comparison, we additionally included random samples of male professors at UAS, as well as female and male professors from traditional universities in comparable fields. For university-level social science, we focused on social pedagogy, as the field of social work is not formally established at universities. The comparison samples were drawn from all German UAS and universities that host relevant departments in the selected disciplines. Table 1 summarises the resulting sample sizes for both the full dataset and the balanced subsample used in downstream analyses.

For the Google-based analysis, we used the full dataset. For the publication database analysis, we drew on a balanced subsample of 80 professors (40 female, 40 male), randomly selected to ensure equal representation across institutional types. We relied on a subsample of the full dataset because extracting publication lists required manual effort. Since each professor curated their own list individually and in non-standardised formats, the extraction process could not be automated.

Gender was inferred from the presentation on university profiles and treated as binary due to the limitations of available data. Public websites typically included names and profile pictures only, so gender was manually inferred based on these attributes. We acknowledge that this is not best practice, as it does not allow individuals to self-identify. However, contacting each professor individually was not feasible. The student responsible for data collection was instructed to assign a gender only when absolutely certain; otherwise, entries were to be marked as *unknown*. In practice, no such cases occurred.

Table 1: Sample size of professors of the full data set and the subsample. The full dataset contains all female professors at UAS in the departments of computer science and social work. For all other categories, a random sample of 50 professors was used. The random sample was used as a comparison group for the main focus of the project, female professors at UAS.

| | Full Dataset | Subsample |
|---|---|---|
| Female Professors UAS Computer Science | 219 | 10 |
| Female Professors UAS Social Work | 863 | 10 |
| Female Professors Uni Computer Science | 50 | 10 |
| Female Professors Uni Social Pedagogy | 50 | 10 |
| Male Professors UAS Computer Science | 50 | 10 |
| Male Professors UAS Social Work | 50 | 10 |
| Male Professors Uni Computer Science | 50 | 10 |
| Male Professors Uni Social Pedagogy | 50 | 10 |

For each professor, the following information was

[1] UAS are a distinct feature of the German higher education system. They focus on practice-oriented teaching and maintain close ties to industry. Compared to traditional universities, they generally have smaller student groups and place less emphasis on theoretical research. Within the German academic system, traditional universities often view UAS as less prestigious due to their more applied, less theory-driven orientation.

collected:

- Name and title
- Gender (inferred)
- Institutional affiliation
- Reported keywords
- Presence of a CV and/or picture on the university profile
- Publication list on the university profile

Because professors manage their own profiles and present publication lists in diverse formats, all data was manually extracted.

## 4.2 Experimental Design

To examine gendered visibility in digital environments, we analyse three interconnected layers of representation: Google search results, academic publication databases, and university profiles. Each serves a distinct role in how professors are made visible, discovered, and contextualised online.

**Google Search Results.** We began our analysis by examining broader forms of digital visibility through Google Search. For each professor in the full sample, we conducted a name-based search that included their institutional affiliation and collected the first 100 search results. These results were categorised into the following types:

- university
- social media
- research institutes
- newspapers/media
- research profiles
- publication databases/preprint servers

We analysed the number, type, and ranking position of these results to identify gendered patterns in digital visibility, with a particular focus on whether algorithmic search systems shape differential representations of female and male professors. Given the broader reach of search engines and the structured nature of Google results, this part of the analysis serves as the primary basis for evaluating algorithmic gender fairness in our study.

**Publication Databases.** To complement the Google-based visibility analysis, we also examined how academic content is retrieved in publication databases, a step that reflects common search strategies used by science journalists and other knowledge intermediaries. It is a typical workflow to begin by querying databases for topic-relevant keywords, and only after identifying promising names, turn to search engines like Google for more context. To honour this process, we conducted an additional exploratory analysis based on academic keyword searches.

For this analysis, we focused on a balanced subsample of 80 professors. From their university profiles, we compiled all self-reported keywords and queried them individually in three major academic databases: the ACM Digital Library[2] (used for publication in computer science), Springer Link[3] (used for publication in computer science and social sciences), and Beltz[4] used for publication in social sciences). For each professor in our subsample, we extracted all self-reported keywords from their university profiles and compiled them into a single list. Each keyword in the list was queried individually in the respective databases, and for each query, we collected the top 1,000 results. We then attempted to match retrieved publications to professors based on their names. We attempted to match retrieved publications to professors based on their names, using either the full first and last name or the first initial and last name. Given the limited available information, this was the most feasible matching strategy, despite the potential for false positives. However, a manual review of the matches confirmed that they appeared valid.

Because publication lists were not uniformly available for all individuals in the full dataset, this analysis was limited to a balanced subsample of 80 professors. While this sample size does not support generalisable claims, it provides initial insights into how academic content is retrieved and associated with named individuals in these databases. The results should be interpreted with caution, particularly as the databases do not disclose how their ranking is determined; search results are typically ordered by "relevance," but the underlying criteria remain opaque. As a result, this part of the analysis serves primarily as an exploratory context. However, following our definition of algorithmic gender fairness introduced in Section "Algorithmic Gender Fairness", we use the gender composition of this subsample as a reference for the real-world distribution against which retrieval outputs are compared.

**University Profile Completeness.** In addition to Google search results and publication databases, we analysed the content of university profiles to capture how professors are presented on their institutional websites. As detailed in Subsection "Data", this information was manually extracted and includes the presence of a CV, a profile picture, and a publication list. These profiles represent structured, publicly accessi-

[2] https://dl.acm.org/

[3] https://link.springer.com/

[4] https://www.beltz.de/

ble data curated by the professors themselves or their institutions. In line with our definition of algorithmic gender fairness, we treat them as a form of real-world data that serves as a reference point for evaluating how academic professionals are represented in digital environments such as search engines.

## 4.3 Findings

This section presents the main findings from our analysis, structured across three areas: Google search results, academic publication databases and the completeness of university profiles.

**Publication Databases.** Across all keyword-based database queries, we retrieved a total of 48,541 unique publications. However, only 44 of these could be matched to professors in our subsample, using either their full name or first initial and surname. This surprisingly low match rate highlights a significant disconnect between the academic work professors report and what is discoverable through our keyword-based database searches.

Several factors likely contribute to this outcome. Most importantly, our queries were limited to three specific publication outlets, the ACM Digital Library, Springer Link, and Beltz, chosen because they allow for automated querying and due to their relevance in informatics and social sciences. As a result, publications in other venues were not included. In addition, professors may not have published under the exact keywords they listed on their university profiles, or the terms may have been too broad or too specific to yield meaningful matches. Keyword searches may also miss publications where the terms are not prominent in titles or abstracts. Further limitations stem from the databases themselves: relevance-based ranking may exclude pertinent results, and name matching can lead to both, false negatives and false positives. If multiple individuals share the same name, our approach may have incorrectly assigned a publication to a professor in the sample.

Figure 1 shows that male professors generally reported more publications, including several extreme outliers. However, very few publications were actually found through database searches for either gender, highlighting the limited recall of keyword-based retrieval in this context.

Figure 2 shows how many of the matched publications were also part of the professors' self-reported publication lists. While female professors had a slightly higher number of matches, the majority of retrieved publications were not part of the self-reported lists for either group. This again suggests that keyword selection and platform coverage substantially shape which publications become visible through database queries.

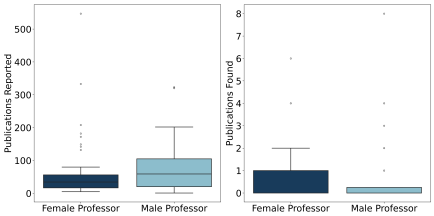


Figure 1: Self-reported versus found publications (via keywords), per person.

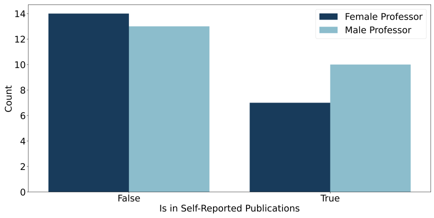


Figure 2: Publications retrieved from databases (via keywords) that also appeared in self-reported lists.

**Google Results.** We next examined how professors are represented across broader digital platforms using Google search results. For each professor in the full sample, we retrieved and categorised the first 100 results. Figure 3 shows the number of links per category, grouped by gender. University-related links were the most common for both female and male professors. Overall, male professors had more links, with a higher median and more variation. Female professors showed a tendency for outliers and more individuals having few or very few links.

Figure 4 presents the ranking positions of these links. Female professors' university links tended to appear slightly higher in the result lists, while male professors had better visibility in categories like research profiles and social media. Although the differences are subtle, they contribute to an overall pattern of gendered variation in search engine visibility.

**University Profile Completeness.** We also examined the content of university profiles for all professors in the full dataset. As shown in Table 2, most professors included both a CV and a profile picture, and over two-thirds also provided a publication list. Female professors were slightly more likely to include a CV and a publication list, while male professors were marginally more likely to include a picture.

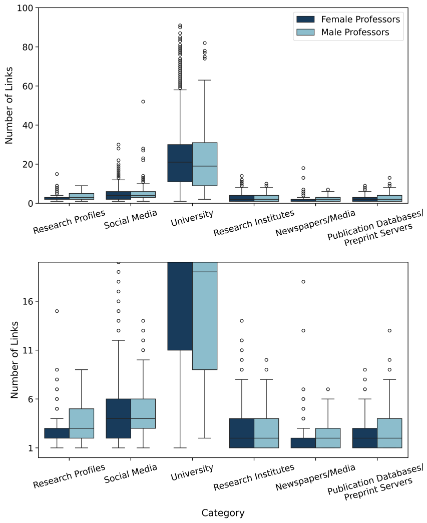


Figure 3: Number of links per category for female and male professors. The top plot shows full data; the bottom plot zooms into low-frequency categories.

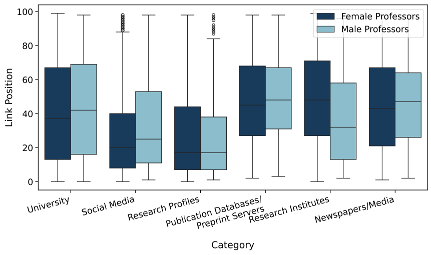


Figure 4: Ranking position of Google search results across categories, by gender. Lower values indicate higher ranking.

## 4.4 Discussion

Our findings indicate that digital visibility in academic contexts is subtly but consistently gendered. This becomes particularly evident when analysing how algorithmic systems represent female and male professors across different platforms. While we did not observe overt algorithmic discrimination, patterns in both database retrieval and Google search results suggest that gender affects how academic expertise is surfaced and made visible.

In publication databases, we found a substantial gap between self-reported and retrieved publications. This gap stems from multiple sources: limited platform coverage (restricted to three specific outlets), reliance on self-reported keywords that often did not align with actual publication metadata, and opaque "relevance"-based ranking mechanisms that are not designed to ensure fair or comprehensive representation. Additionally, name-based matching introduces ambiguity, especially for common names. Although the sample was too small to draw generalisable conclusions, male professors showed slightly higher match rates, pointing to possible gendered differences in how academic outputs are indexed and surfaced.

Table 2: University profile completeness for the full dataset: CV, picture and publication list inclusion. The numbers should be interpreted as a percentage of female professors or a percentage of male professors, depending on the line. Therefore, rows do not add up to 100%.

| | CV | Picture | Publication Lists |
|---|---|---|---|
| Female Professor | 68.9% | 85.0% | 70.2% |
| Male Professor | 62.5% | 85.9% | 66.2% |

In contrast, our analysis of Google search results, conducted on the full dataset, revealed clearer patterns. Male professors were consistently associated with a higher number of links across most categories. However, the distributions were not uniformly more concentrated for male professors. While they had higher medians in several categories, the spread of results varied by category and was not consistently narrower than that of female professors. Female professors showed greater variability overall, with more frequent low-end outliers, particularly in categories such as university and research profiles. When considering the ranking of results, female professors' links tended to appear slightly higher in several categories, including social media, research profiles, and publication databases. In other categories, such as university and newspapers/media, the ranking distributions were largely comparable across genders. These findings suggest that while female professors are not disadvantaged in terms of ranking within categories, the lower number of links may still reduce their overall discoverability in search results.

A possible factor contributing to the lower number of search results for female professors is the way academic profiles are structured on institutional websites. While profile completeness was generally high across the sample, we observed small gender differences: female professors were slightly more likely to include CVs and publication lists, whereas male professors more frequently provided a profile picture. Since images and structured information (such as publication entries or CVs) can be indexed differently by search engines, these differences in self-presentation may influence how easily professors are linked to relevant content. What is particularly striking, however, is that despite female professors providing slightly more structured academic information on their university profiles, they were less visible in several key categories of Google search results, most notably "research profiles," "publication databases/preprint servers," "newspapers/media," and "university." In other words, even though they appear to invest more in curating their institutional presence, this effort does not translate into greater discoverability. Thus, while search rankings within categories do not appear systematically biased, the reduced number of visible links may still disadvantage female professors in terms of overall digital visibility.

Taken together, these results highlight how digital visibility is shaped by the interaction between algorithmic systems, individual presentation choices, and institutional infrastructure. They also reflect broader structural patterns: who appears where, how prominently, and through what types of content is not random; it is filtered through technical systems that rely on data structures, which may themselves encode or reflect gendered norms.

In light of our definition of algorithmic gender fairness, our findings suggest that current systems fall short of this ideal. Even when the intent may not be discriminatory, existing systems amplify disparities through uneven coverage, limited keyword matching, unclear ranking mechanisms, and visibility differences in general-purpose search results. These systems do not just reflect the real world—they actively reshape which parts of it are seen.

Fairness, therefore, cannot be evaluated purely by the absence of discriminatory intent or overt exclusion. It must also consider the cumulative effects of design decisions, platform constraints, and structural imbalances in source data. Gendered visibility gaps, even if subtle, are a form of representational inequality that algorithmic systems may unintentionally perpetuate.

## 5 CONCLUSION AND FUTURE WORK

In this paper, we introduced the concept of algorithmic gender fairness and evaluated it using heterogeneous data on German professors. By analysing gendered patterns in academic visibility across different institutional contexts and disciplines, we aimed to identify structural imbalances that may arise in algorithmic representations of expertise.

Our findings reveal nuanced but consistent gender differences in digital visibility. Search and retrieval algorithms do not exhibit overt forms of gender discrimination; however, subtle imbalances appear across various dimensions. Female professors were slightly more likely to complete their institutional profiles with CVs and publication lists, while male professors reported higher median numbers of publications. Yet, only a small number of self-reported publications could be retrieved from academic databases, highlighting mismatches between metadata, keyword representation, and retrieval mechanisms.

In Google search results, male professors were associated with a greater number of links overall, while female professors showed more variability, including more frequent cases of low link counts. Link categorisation and ranking further revealed gendered patterns: female professors' links tended to appear in higher positions (i.e., closer to the top of the results list) in categories such as university websites, research profiles, and social media. Male professors' links, by contrast, were often ranked slightly lower (i.e., further down in the result list) in university websites and social media, but were more numerous overall. These differences likely reflect an interplay between platform algorithms, institutional curation, and self-presentation strategies.

While these patterns point to structural imbalances, they should be interpreted with caution. Factors such as outdated publication lists, common naming conventions, and differing levels of online activity likely contribute to the observed visibility gaps. The imbalances we observed are therefore not attributable to algorithmic bias alone, but emerge from the interaction of algorithmic processes with broader sociotechnical contexts.

Future research should expand on this foundation by incorporating more inclusive gender categories, extending the analysis beyond German academia, and examining additional disciplines. Integrating data from more publication databases and search engines would also allow for a broader assessment of visibility dynamics across digital ecosystems.

Longitudinal analyses and larger, more diverse datasets will be essential for disentangling the specific roles played by algorithmic systems, institutional infrastructures, and individual behaviours. In parallel, collaborative efforts involving academic institutions, search engine providers, and fairness researchers are needed to improve algorithmic transparency and accountability. Only by addressing both data and design can we move toward systems that fairly represent the diversity of academic expertise online.

## AI USAGE

The authors are not native English speakers; therefore, ChatGPT and Grammarly were used to assist with writing English in this work.

## 6 ETHICAL CONSIDERATIONS

This paper did not involve direct interaction with human participants and relied solely on publicly available information found on university websites. As such, ethics approval from an institutional review board was not required. Personal data were collected manually with the intent to minimise misclassification, particularly in regard to gender inference. The student responsible for data collection was instructed to assign gender only when certainty was high and to otherwise mark entries as unknown. No personal or sensitive data beyond what was already publicly accessible were stored or analysed.

To protect the privacy and anonymity of the professors included in the dataset, we will not publish or share the collected data. We acknowledge the ethical limitations of inferring gender from names and pictures, and we explicitly address these limitations in the paper to promote transparency and encourage more inclusive data practices in future research.

## 7 ADVERSE IMPACT STATEMENT

This paper adopts a bias-preserving definition of algorithmic gender fairness, aiming to reflect real-world gender distributions without introducing or amplifying existing biases. While this approach supports transparency and alignment with observed data, it may also carry certain risks.

First, reflecting real-world distributions without intervention could be misused to justify existing gender inequalities, especially in contexts where structural bias is already present. Second, although we acknowledge the existence and importance of non-binary and gender-diverse identities, our empirical analysis is limited to binary gender categories due to data constraints. This limitation may contribute to the erasure of individuals who do not identify within the binary framework, especially if such approaches are widely adopted without critical adaptation. Finally, bias-preserving fairness may be misinterpreted as evidence of algorithmic neutrality, potentially obscuring the broader sociotechnical dynamics that shape inequality.

We therefore emphasise that fairness assessments should always be interpreted in light of context, data limitations, and the values underlying system design. We encourage future work to engage critically with fairness definitions and to explore approaches that address structural imbalances more directly.

## ACKNOWLEDGEMENTS

This work was written by an author team working on different projects. Stefanie Urchs' project ”Prof:inSicht” is promoted with funds from the Federal Ministry of Research, Technology and Space under the reference number 01FP21054. Matthias Aßenmacher is funded with funds from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of BERD@NFDI - grant number 460037581. Responsibility for the contents of this publication lies with the authors.

## REFERENCES

Aristotle (2009). *The Nicomachean ethics (book V)*. Oxford World's Classics. Oxford University Press.

Beemyn, B. G. and Rankin, S. (2011). *The lives of transgender people*. Columbia University Press.

Bigdeli, A., Arabzadeh, N., SeyedSalehi, S., Zihayat, M., and Bagheri, E. (2022). Gender fairness in information retrieval systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3436–3439, New York, NY, USA. Association for Computing Machinery.

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

Bothmann, L., Peters, K., and Bischl, B. (2024). What Is Fairness? Philosophical Considerations and Implications For FairML. arXiv:2205.09622.

Carpenter, M. (2021). Intersex human rights, sexual orientation, gender identity, sex characteristics and the yogyakarta principles plus 10. *Culture, Health & Sexuality*, 23(4):516–532. PMID: 32679003.

Caton, S. and Haas, C. (2024). Fairness in Machine Learning: A Survey. *ACM Comput. Surv.*, 56(7):166:1–166:38.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. KDD '17, page 797–806, New York, NY, USA. Association for Computing Machinery.

Dator, J. (2017). Chapter 3. What Is Fairness? In Dator, J., Pratt, R. C., and Seo, Y., editors, *Fairness, Globaliza-*

*tion, and Public Institutions*, pages 19–34. University of Hawaii Press, Honolulu.

Datta, A., Tschantz, M. C., and Datta, A. (2015). Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112.

Devinney, H., Björklund, J., and Björklund, H. (2022). Theories of "gender" in nlp bias research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2083–2102, New York, NY, USA. Association for Computing Machinery.

Dictionary, C. (2022). fairness.

Ekstrand, M. D., Das, A., Burke, R., and Diaz, F. (2022). Fairness in information access systems. *Foundations and Trends® in Information Retrieval*, 16(1–2):1–177.

Eren, E., Hondrich, L., Huang, L., Imana, B., Kettemann, M., Kuai, J., Mattiuzzo, M., Pirang, A., Pop Stefanija, A., Rzepka, S., Sekwenz, M., Siebert, Z., Stapel, S., and Weckner, F. (2021). *Increasing Fairness in Targeted Advertising. The Risk of Gender Stereotyping by Job Ad Algorithms*.

Fang, Y., Liu, H., Tao, Z., and Yurochkin, M. (2022). Fairness of machine learning in search engines. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 5132–5135. ACM.

for Justice, E. C. D. G., Consumers., network of legal experts in gender equality, E., and non discrimination. (2021). *Algorithmic discrimination in Europe: challenges and opportunities for gender equality and non discrimination law.* Publications Office, LU.

Francazi, E., Lucchi, A., and Baity-Jesi, M. (2024). Initial guessing bias: How untrained networks favor some classes. In *Forty-first International Conference on Machine Learning*.

Geyik, S. C., Ambler, S., and Kenthapadi, K. (2019). Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2221–2231, New York, NY, USA. Association for Computing Machinery.

Janhunen, J. (2000). Grammatical gender from east to west. *TRENDS IN LINGUISTICS STUDIES AND MONOGRAPHS*, 124:689–708.

Keyes, O. (2018). The misgendering machines: Trans/hci implications of automatic gender recognition. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. In Papadimitriou, C. H., editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 43:1–43:23, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Kong, Y. (2022). Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 485–494, Seoul Republic of Korea. ACM.

Konishi, T. (1993). The semantics of grammatical gender: A cross-cultural study. *Journal of Psycholinguistic Research*, 22(5):519–534.

Kramer, R. (2020). Grammatical gender: A close look at gender assignment across languages. *Annual Review of Linguistics*, 6(1):45–66.

Loi, M. and Heitz, C. (2022). Is Calibration a Fairness Requirement? An Argument from the Point of View of Moral Philosophy and Decision Theory. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2026–2034, New York, NY, USA. Association for Computing Machinery.

Makhortykh, M., Urman, A., and Ulloa, R. (2021). Detecting race and gender bias in visual representation of ai on web search engines. In Boratto, L., Faralli, S., Marras, M., and Stilo, G., editors, *Advances in Bias and Fairness in Information Retrieval*, pages 36–50, Cham. Springer International Publishing.

Phillips, W. and Boroditsky, L. (2013). Can quirks of grammar affect the way you think? grammatical gender and object concepts. In *Proceedings of the 25th Annual Cognitive Science Society*, pages 928–933. Psychology Press.

Schütze, H., Manning, C. D., and Raghavan, P. (2008). *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Singh, A. and Joachims, T. (2018). Fairness of exposure in rankings. KDD '18, page 2219–2228, New York, NY, USA. Association for Computing Machinery.

Urman, A. and Makhortykh, M. (2022). "foreign beauties want to meet you": The sexualization of women in google's organic and sponsored text search results. *New Media & Society*, 26(5):2932–2953.

Verma, S. and Rubin, J. (2018). Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness*, Gothenburg Sweden. ACM.

Wachter, S., Mittelstadt, B., and Russell, C. (2021). Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law. *West Virginia Law Review*, 123(3):735–790.

West, C. and Zimmerman, D. H. (1987). Doing gender. *Gender & Society*, 1(2):125–151.

Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31:1060–1089.

# Part III.

# Discrimination in Large Language Models

# 5. How Prevalent is Gender Bias in ChatGPT? - Exploring German and English ChatGPT Responses

Chapter 5 examines how ChatGPT responds to gendered prompts in English and German, building on recent systematic evaluations of its behaviour. The study compares responses across languages and perspectives, first through exploratory prompts and then through repeated trials that test consistency and vocabulary use. While English outputs are mostly accurate, German responses often struggle with gender-neutral grammar and sometimes produce invalid forms. Gendered prompts frequently activate standardised “diversity templates,” resulting in overly generic equality rhetoric, while neutral prompts occasionally lead to female-only personas, particularly in academic or STEM contexts. Overall, differences are subtle but reveal limitations in consistency and inclusive language handling.

### Contributing article:

Urchs, S., Thurner, V., Aßenmacher, M., Heumann, C. and Thiemichen, S. (2023). How Prevalent Is Gender Bias in ChatGPT? - Exploring German and English ChatGPT Responses. *Machine Learning and Principles and Practice of Knowledge Discovery in Databases. ECML PKDD 2023. Communications in Computer and Information Science, vol 2133. Springer, Cham..* *https://doi.org/10.1007/978-3-031-74630-7_20.*

### Copyright information:

### Author contributions:

The idea for the publication was developed together with all the authors. Stefanie Urchs was responsible for designing the study, conducting the experiments, analysing the results, researching the literature and writing the majority of the manuscript. All authors contributed to the proofreading of the manuscript. Veronika Thurner wrote the chapter on motivation, while Matthias Aßenmacher contributed the section on large language models.

### Supplementary material available at:

- Code and further information: https://github.com/Ognatai/bias_chatGPT

# How Prevalent Is Gender Bias in ChatGPT? - Exploring German and English ChatGPT Responses

Stefanie Urchs[1]( ), Veronika Thurner[1], Matthias Aßenmacher[2,3], Christian Heumann[2], and Stephanie Thiemichen[1]

[1] Faculty for Computer Science and Mathematics, Munich University of Applied Sciences, Munich, Germany
{stefanie.urchs,veronika.thurner,stephanie.thiemichen}@hm.edu
[2] Department of Statistics, LMU Munich, Munich, Germany
{matthias,christian.heumann}@stat.uni-muenchen.de
[3] Munich Center for Machine Learning (MCML), LMU Munich, Munich, Germany

**Abstract.** With the introduction of ChatGPT, OpenAI made large language models (LLM) accessible to users with limited IT expertise. However, users with no background in natural language processing (NLP) might lack a proper understanding of LLMs. Thus the awareness of their inherent limitations, and therefore will take the systems' output at face value. In this paper, we systematically analyse prompts and the generated responses to identify possible problematic issues with a special focus on gender biases, which users need to be aware of when processing the system's output. We explore how ChatGPT reacts in English and German if prompted to answer from a female, male, or neutral perspective. In an in-depth investigation, we examine selected prompts and analyse to what extent responses differ if the system is prompted several times in an identical way. On this basis, we show that ChatGPT is indeed useful for helping non-IT users draft texts for their daily work. However, it is absolutely crucial to thoroughly check the system's responses for biases as well as for syntactic and grammatical mistakes.

**Keywords:** bias · large language models · corpus analysis · ChatGPT

## 1 Motivation

By introducing ChatGPT [17] with its intuitive user interface (UI), OpenAI opened the world of state-of-the-art natural language processing to the non-IT users. Users do not need a computer science background to interact with the system. Instead, they have a natural language conversation in the UI. Many users utilise the system to help with their daily work: Writing texts, checking grammar and spelling, and even fact-checking their work. However, non-IT users tend to see the system as a "magical box" that knows all the answers and believe that because machines do not make mistakes, neither does ChatGPT.

R. Meo and F. Silvestri (Eds.): ECML PKDD 2023, CCIS 2133, pp. 293–309, 2025.
https://doi.org/10.1007/978-3-031-74630-7_20

This lack of critical usage is problematic in everyday use. It is unclear from the documentation on which data the system was trained exactly, but since it includes training data from CommonCrawl[1] it is likely to reflect many of the biases and stereotypes common to internet content. Furthermore, the model is trained on as much data as possible and, therefore, on data from ten, twenty, and more years ago. This historical data, like all data, represents the spirit of the era, including all stereotypes and biases that were prevalent at the time. Concepts that have evolved or changed over time, like the image of women or how the LGBTQIA+ community is perceived, are also subject to this issue. OpenAI tries to handle the biases and stereotypes by actively regulating the responses. However, downstream handling can only deal with known problems in a specific way. Since the model's problems are unclear, users can find ways to circumvent bias safeguards, be it intentionally or unintentionally.

By informing non-IT users about the system and its potential problems, users can employ it more effectively, as understanding system mechanisms leads to a better understanding of the system's capabilities [12]. Besides, reviewing ChatGPT responses critically avoids the publication of biased texts and, therefore, discrimination against minorised groups[2]. Users should use the system to enhance their work and should not let the system autonomously work for them. We should generally strive to use LLM to augment human work and not replace it.

## 2 Problem Formulation

Our main goal is to analyse ChatGPT responses from a non-IT user's point of view. As an example context, we use the context of university communications. This use case leads to four important aspects a user should keep in mind while working with the system:

1. Are responses syntactically and grammatically correct, especially in non-English languages combined with using gender-neutral language?
2. Do responses include gender biases that would lead to discrimination in a publication?
3. Does the system behave according to the expectations of a non-IT user?
4. Do (unannounced) system updates influence responses to established prompts?

[1] For more information, see https://commoncrawl.org/.

[2] We use the term minorised groups according to the definition of D'Ignazio and Klein [4]: "While the term minority describes a social group that is comprised of fewer people, minoritized indicates that a social group is actively devalued and oppressed by a dominant group, one that holds more economic, social, and political power. With respect to gender, for example, men constitute the dominant group, while all other genders constitute minoritized groups. This remains true even as women actually constitute a majority of the world population".

Ignoring these aspects can lead to an increased workload, such as manually correcting the syntax and grammar or searching for a prompt that generates the same response as before the system update. Additionally, publishing texts containing biases against certain genders is an act of discrimination against people who identify as this gender, leading to a bad reputation for the user and the institution the person is working for. We consider biases from the point of view of researchers in a Western European country and thus conceive opinions that oppose this value system as biases or even discrimination. We are aware that our belief system does not hold true to other cultures. Therefore, our understanding of biases and discrimination might not be shared by all people reading this publication.

To check these aspects in ChatGPT responses, we first explore a range of prompts and corresponding responses to open up the problem space. Subsequently, we exploit selected prompts to get a deeper understanding of the magnitude of the problem.

## 3 Background

We first present work on bias detection in natural language text. Afterwards, we outline a brief history of large language models.

### 3.1 Bias in Texts

To detect biases in text, it is important to first define the term bias properly. In machine learning, specifically, in a classification task, bias is defined as the preference of a model towards a certain class. Nevertheless, when working with natural language text, we focus on the biases or discrimination communicated through it. Mateo et al. define bias as follows: “Biases are preconceived notions based on beliefs, attitudes, and/or stereotypes about people pertaining to certain social categories that can be implicit or explicit.” [15] They continue that discrimination is the manifestation of biases through behaviour and actions. In other words, bias is the thought and discrimination the action. Since written text contains a person’s thoughts, it can be biased and thus be regarded as an act of discrimination. ChatGPT, being trained on texts containing people’s biases, repeats and amplifies these biases. A user who publishes a biased response makes these biases their own and commits an act of discrimination.

In 1973 Lakoff [13] analysed how women are expected to talk and how they are talked about. She highlights that a woman’s language is less secure and tries to avoid the strong expression of feelings. When talking about and to women, the speaker tends to reduce the woman to an object that is described with euphemisms and lacks her own agency. Recent work on the automatic detection of gender biases in natural language text, unfortunately, confirms these findings: Sports journalists tend to ask women fewer questions related to their profession [8], language towards female streamers on social game-streaming platforms concentrates less on the game the streamer is playing and more on her appearance [16], as do comments in social media [7]. When talking or portraying a woman, for example, on Wikipedia [10,24], the woman is mostly mentioned in the context of her husband or partner, thus lacking her own agency. Furthermore, articles about women emphasise the gender of the portrayed, and her family and marital status are discussed extensively. The portrayal of fictional women also follows Lakoff's findings and the portrayal in Wikipedia. Words used for women in Bollywood movies [14] describe the woman's body, family, or material status. The female protagonists mostly react to the actions of their male counterparts. In online fiction, women tend to be "weak, submissive, childish, afraid, dependent and hysterical", whereas men tend to be associated with "strong, active, beauty and dominant" [6].

### 3.2 Large Language Models

Vaswani et al. [23] lay the groundwork for modern LLMs by introducing attention to transformers. The attention mechanism makes it possible to focus on specific parts of a text sequence and not only the token right in front or behind the currently examined one. Thus improving sequence-to-sequence tasks when the input sequence has a different order than the output sequence. Building on the work of Vaswani et al. [23], Devlin et al. [3] train the language representation model BERT. The novelty of BERT is that the model has a concept of a sequence of tokens (e.g., a sentence) and can relate them to the next sequence. By fine-tuning the model to specific tasks, BERT can outperform the state-of-the-art in various domains, making it the first model to do so. Raffel et al. [20] enable true multitask learning by considering all tasks as a "text-to-text problem", paving the way for zero-/few-shot learning [2,19] and prompting [17] with LLMs. Radford et al. [19] argue that supervised task-specific fine-tuning of models needs huge amounts of labelled data, limiting training efficiency. With GPT-2, they propose a model capable of zero-shot learning. GPT-2 can infer tasks from "natural language sequences" in its training data. The model has the ability to work cross-domain because of its tremendous amount of training data from vastly different domains. Nevertheless, the model is just a proof of concept. Task-specific models on most benchmarks outperform it. The next GPT iteration is GPT-3 [2], a scaled-up version of GPT-2. It is trained longer on a bigger, more diverse data set, and has more parameters. Due to this up-scaling, the model outperforms task-specific models on many common benchmarks. Adapted via reinforcement learning from human feedback (RLHF) [18] to enable dialogue-style conversations and combined with a simple, easy-to-use user interface GPT-3.5 can interact with users. This removes the need to be able to program to use the model, making the technology accessible to non-IT users. The resulting system is known as ChatGPT[3].

[3] https://openai.com/blog/chatgpt.

## 4 Methodical Approach

As a first step, we investigate ChatGPT for possible problematic behaviour and define the problem space as the four aspects mentioned in Sect. 2. We explore this problem space with specific prompts. On this basis, we select two prompts for an in-depth investigation by repeating them ten times.

### 4.1 Methods of Exploration

During the exploration phase, we probe the system with open-ended and neutrally formulated prompts for possibly problematic responses. Each prompt is executed once. The system is asked to take on the perspective of a female, male, or neutral person in German as well as in English to be able to span a wide range of possible problematic behaviour. Even one occurrence of controversial behaviour can be problematic for a user who does not check the response thoroughly before publishing. Additionally, the system is used very frequently by many users, thus, generating a tremendous amount of responses. Hence, problematic behaviour will be generated again. Therefore, the exploration phase is a glimpse into general problems prevalent in the system. We explore the problem space with the following four prompts[4]:

1. What is a good [female, male, neutral] professor?
2. From the viewpoint of university communications: write an article about a [female, male, neutral] professor who won a prize.
3. Write an article that explains to [women, men, people] why they should pursue a career as a [female, male, neutral] professor.
4. Write an article that explains why [women, men, people] should pursue a career as a [female, male, neutral] professor.

An example for the first prompt is: “What is a good female professor?”. In English, prompts neutral means that there is no gender qualifier, the words female or male are inserted for the corresponding perspective. In German prompts, we use the gender-neutral form of the word professor (“Professor:in”), the female term (“Professorin”) for the female perspective, and the qualifier male (“männlich”) for the male perspective. Especially in German, the generic masculine was the de facto standard to address persons for a long time. However, studies have shown that generic masculine does not include women or non-binary individuals, as people hearing/reading the generic masculine mostly think about a male person [1,11]. Using gender-neutral language helps to reduce biases and stereotypes in people’s minds [21]. To include all genders, public communication in Germany increasingly uses gender-neutral language. Therefore, we use gender-neutral language in German prompts with a neutral perspective. We separate the three perspectives into three distinct accounts to avoid the usage of one perspective influencing the response for another one.

[4] For the full set of prompts and corresponding responses, see: https://github.com/Ognatai/bias_chatGPT.

The first two prompts direct the system to write about a professor. We explore which characteristics are attributed to the genders and how the attributions differ from the neutral "default" case. Furthermore, the second prompt explores if there is a bias in research fields and prize types. The third and fourth prompts explore the same topic slightly differently. At first, the prompt is targeted toward a specific gender. Subsequently, we explore if removing the specific target audience influences the response. Both prompts direct the system to write for a (future) professor. We intend to test if gender influences the reasons to become a professor and how the reasons differ from the neutral baseline. Lastly, we direct the system to write from the viewpoint of a professor, exploring if gender influences how the system impersonates the professor.

### 4.2 Methods of Exploitation

After defining the problem space in the exploration phase, we step into selected prompts that lead to particularly problematic responses and open a possibility for automated analysis. All of the selected prompts are standard use cases in the work of university communication. We generate ten responses per prompt, perspective, and language, leading to at least sixty responses per prompt. These responses are then analysed for:

- **Words used in text in general:** We analyse which words are mostly used in the responses. To see if the system uses different words for the different perspectives and languages.
- **Female/male coded words used in the text:** We use the lists of female/male coded words as found on the English[5] and German[6] [5] gender decoder websites. Both projects are based on work from Gaucher et al. [9]. We use the word lists to analyse if ChatGPT responses contain language that is tailored towards a certain perspective due to the words used.
- **Text length:** The number of tokens and the average length of tokens in each non-preprocessed text. Trix and Psenka [22] show that texts about the achievements of women tend to be shorter than texts about the achievements of men. We examine if ChatGPT follows this observation.

## 5 Findings of Exploration Phase

As described in Subsect. 4.1 we post five prompts in three different perspectives in two different languages, except for the first prompt. In the German prompt about the characteristics of a good professor, we have to explicitly qualify the professor as male because the "normal" masculine prompt leads to a too-generic response not tailored towards a male audience. The slight difference between prompts three and four (explaining to a specific gender why they should become

[5] https://gender-decoder.katmatfield.com/about.

[6] https://www.msl.mgt.tum.de/rm/third-party-funded-projects/projekt-fuehrmint/gender-decoder/wortlisten/.

a professor versus generally explaining why a specific gender should become a professor) does not lead to different responses. We present our findings in the categories as defined in Sect. 2: Grammatical and syntactic correctness, gender biases, and system behaviour. The system updates can only be observed in the exploitation phase and will be discussed in Sect. 6. See the GitHub repository[7] for full-length responses.

### 5.1 Grammatical and Syntactic Soundness

The system excels in the English language. By default, the responses are written in US American English; other English versions need to be specified beforehand.

The German responses are not as good as the English ones. In some instances, sentences lack grammatical correctness, however, the incorrectness occurs on a subtle level. When skimming the text, the grammatical mistake could be easily missed. One example is the following German sentence: "*Wir sind stets bestrebt, ein offenes und inklusives Umfeld zu schaffen, in dem jeder seine Stimme gehört und geschätzt wird.*" Only the very last part of this sentence is wrong. When skim-reading the text, one could easily miss the incorrectness. But publishing such a sentence in official communication is very unprofessional. Even the exploration phase's small sample already includes several problematic grammatical errors.

Additionally, ChatGPT has problems using the gender-neutral German language written using the male version of a word followed by either a colon, underscore, brackets, or slash and the female ending. Sometimes a capitalised i is used instead of the special characters. We used the colon for gender-neutral language resulting, for example, in "Professor:in" as a gender-neutral term for professor. Only when gender-neutral language is used in prompts, the system uses gender-neutral language for the response. ChatGPT is not always able to follow the grammatical rules of using gender-neutral language. For example, the system generated the word "*Experte:r*", which does not exist at all in German. Official German communication mostly uses gender-neutral language with special characters. That is why a system used for writing support must be able to use gender-neutral language correctly. Another problem of the responses is the usage of the gender-neutral "they". Using "they" instead of a specific pronoun is the best way to write and speak in a gender-neutral way in English. Nonetheless, in German, no direct translation exists. Despite the lack of a translation, ChatGPT translates "they" into German by using the third plural person, resulting in an incorrect sentence, possibly with a completely different meaning.

### 5.2 Gender Biases

It should be noted that German responses in general contain the gender-neutral term "*Studierenden*" to refer to students, and ChatGPT also uses the gender-neutral pronoun "they" when prompted neutrally. For the first prompt, "What

[7] https://github.com/Ognatai/bias_chatGPT.

is a good professor" the system generates fairly equal responses in English. However, the female perspective lacks the desired characteristic of conducting good research for the female perspective, which is mentioned in the neutral and male perspectives. Additionally, adding gender to the prompt triggers the topics of fairness and equality strikingly in the response. In both gendered responses, a good professor should consider equality and diversity. This is not mentioned in the neutral response. The German responses differ slightly more. The female perspective lists fewer items of what is considered a good professor. Interestingly the German version does not stress the diversity and equality points. When prompted neutrally, the system tends to generate more women than men. The responses for the prompt "professors who won a prize" resulted in German and English in a female professor. This behaviour is explored deeper in the exploitation phase in Sect. 6. Both prompts for reasons to become a professor lead to responses that mostly discuss gender equality. Consequently, a woman should become a professor only to elevate other women. Men should become professors to elevate other men. Both should fight for gender equality. The system seems to be triggered by including a specific gender in the prompt, leading to a response about gender equality. Unfortunately, the system does not differentiate between genders but uses the exact same reasoning for female and male prompts. Leading to the following statements:

- "Während sich die Geschlechterverteilung in den Hochschulen allmählich angleicht, gibt es immer noch einen spürbaren Mangel an männlichen Professoren." [18]
  En: *While the gender distribution in universities is gradually becoming equal, there is still a noticeable shortage of male professors.*
- "[...] Dennoch besteht immer noch ein bedarf an männlichen Wissenschaftlern, die sich für eine Laufbahn in der Professur entscheiden." [18]
  En: *Nevertheless, there is still a need for male scientists who choose a career as professors.*
  This sentence is also grammatically incorrect.
- "In an era of evolving societal dynamics and increased focus on diversity and inclusion, it is essential to examine and appreciate the importance of men pursuing careers as male professors." [18]
- "By choosing a career as a male professor, men have the power to contribute to a more inclusive educational environment." [18]

These sentences are a typical line of reasoning from the perspective of females and usually do not apply to the male perspective since they are overrepresented in higher academic positions. Therefore, an increase in male professors would not diversify the environment. The gender-neutral prompt does not trigger the gender equality template. Here responses highlight intellectual freedom, the joy of teaching, influencing politics and society, and job security. None of these aspects are mentioned in the gendered responses.

### 5.3 System Behaviour

A problematic system behaviour, as mentioned above, is that the inclusion of gender in the prompt can seemingly trigger a gender/diversity/equality template. This behaviour is not always appropriate, especially if the gendered response does exclude every aspect other than gender/diversity/equality. This might occur due to reinforcement learning from human feedback (RLHF) that “over-corrects” the response. Since the model has no syntactic understanding of the response, it can not tailor the reasoning to a specific gender. The system lacks continuity in that details differ even if prompted to repeat the text. For instance, when prompted to fill in the blanks in the “professor wins a prize” prompt, the system changes the name of the generated professor. Non-IT users who use the system do not expect that a command given might be ignored.

### 5.4 Discussion

ChatGPT is still lacking grammatical and syntactical soundness in non-English responses. The system notably struggles when using German gender-neutral language, which is now the de-facto standard in official university communication. The huge problem with these system errors is that the mistakes made are hard to find in a mostly correct response. Thus, the response could only be used as a very rough draft and needs in-depth proofreading. The system is fine-tuned to exclude racist, sexist, and otherwise hateful responses. This fine-tuning seems to include templates that are triggered with certain words. The template strategy does not always work as intended and can lead to worrying results. Publishing a text that advocates for more men in academia, diversifying the field and bringing more role models to male students is embarrassing in the best case; in the worst, it could lead to the responsible person losing their job. However, the system can use the gender-neutral “they” in English responses and the gender-neutral word for students in German responses, which is a huge help in writing inclusive text. Unexpected system behaviour, like the lack of continuity, is particularly problematic for non-IT users. These kinds of users expect the system to follow instructions completely. A system that is as easy to use as ChatGPT and that generates natural-sounding responses is perceived to have high credibility, leading to users trusting the system without questioning how the responses are generated.

## 6 Findings of Exploitation Phase

We chose to investigate two prompts in-depth. First, a prompt about a professor who wins a prize. This prompt generates text about professors with research fields leading to interesting research opportunities. Second, we investigate the characteristics of a good professor. We chose this prompt because of the subtle differences between the perspectives we want to investigate deeper. Both prompts are standard examples of the work of university communications. The full data analysis of the exploitation phase and full-length responses can be found in the corresponding git repository[8]. We removed stopwords and lemmatised the response to analyse the most used words. We did not pre-process the response for investigating gender-coded words and text length because the lemmatisation could distort the results. The gender-coded words are provided in their stem form. We count all (non-overlapping) occurrences of the stem in the response.

### 6.1 Professor Wins a Prize

We had to re-prompt the system in every case because it generated gap texts. To generate professors with a research field, we had to prompt ChatGPT to repeat the text and add data for the professor. We will only discuss the second responses, which contains data about professors. Overall, the prompt is very generic, but the system creates professors who are exceptional at every point of their career: their research is groundbreaking and helps society, their teaching skills inspire future generations of students, and their dedication to the community helps to bring all researchers together. This kind of response is the same for every perspective and in both languages. Additionally, the system generated research fields for fictional professors. Interestingly, research fields for male professors are less diverse than for female professors; about fifty percent (in both languages) of professors are physicists. The English ones could also work in engineering (2/10), or the field was not prominently mentioned (3/10). German male professors could also work in psychology or nanotechnology (1/10 each), or their field was not prominently mentioned (4/10). Women and neutral perspectives have a broader range of fields. However, all of them are from STEM disciplines, neglecting social sciences and humanities. Notably, all responses that are prompted from a neutral perspective generate female professors.

**Words Used.** The German responses seem to use the professor's name more than the English ones. Further, the first names Anna and Julia, as well as the surname Müller are generated in many German responses over all three perspectives. The English texts favour research and praise terms. German texts prominently mention research but also engagement and work. Praise words are not stressed. The top ten words for each language and perspective reveal that also the English texts stress professor names, but the names are more diverse in the different perspectives. Most of the top ten words in all languages and perspectives are related to the name of the professor or the university name. Nevertheless, research is a prominent word used in female responses in both languages. In German male responses, research is ranked tenth, while it is not ranked in the top ten for English male responses or neutral perspectives.

**Gender Coded Words.** Figure 1a) displays the use of female-coded words in English responses: 14 out of the 50 female-coded English words are used.

[8] https://github.com/Ognatai/bias_chatGPT.

Counter-intuitively responses about a female professor use 10 of the words, which is the least of all perspectives. The most used word is "commit", which is used slightly more in responses about a female professor. Figure 1b) shows that out of 62 female-coded German words, 17 are used. Texts about male professors use 15 female-coded words, which is the most of all perspectives. The most used words are "engag" (engagement, engaging) and "gemeinschaft" (community). Both are used more often to describe male professors than female ones.

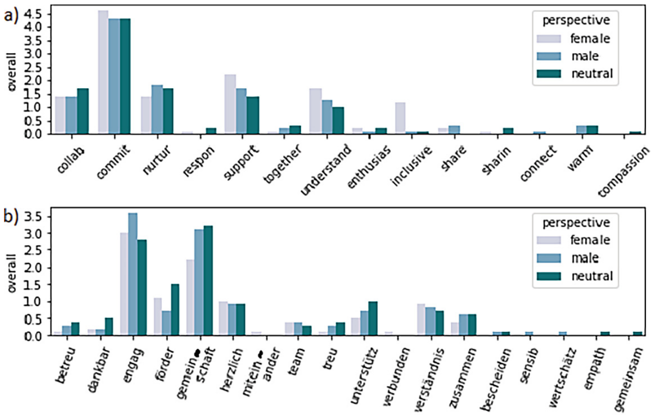


**Fig. 1.** Female coded words used on average in all perspectives of English responses (a) and German responses (b) for the prompt about a professor who won a prize. The number of usages is averaged over all responses of a perspective.

Figure 2a) shows the usage of male-coded words in English responses. Out of 52 male-coded English words, 16 are used. Responses about female professors use 13 of these words, which is the most of all perspectives. The most used word is "intellect", which is dominantly used to describe male professors. Figure 2b) shows that out of 62 male-coded German words, 21 are used. The neutral and male perspectives each use 14 of these words, one more than the female perspective. The most used word is "einfluss" (influence), dominantly used in responses about male professors.

**Text Length.** Text length does not differ substantially between perspectives.

**Discussion.** ChatGPT hallucinates information into generic prompts. Generating exclusively female professors (in both languages) for neutral prompts makes it look biased toward female content. Furthermore, the system displays a bias toward STEM-related research fields, while the responses overall use relatively few gender-coded words and do not reinforce common language biases. English responses tend to prefer the gender-coded words of the respective other gender.

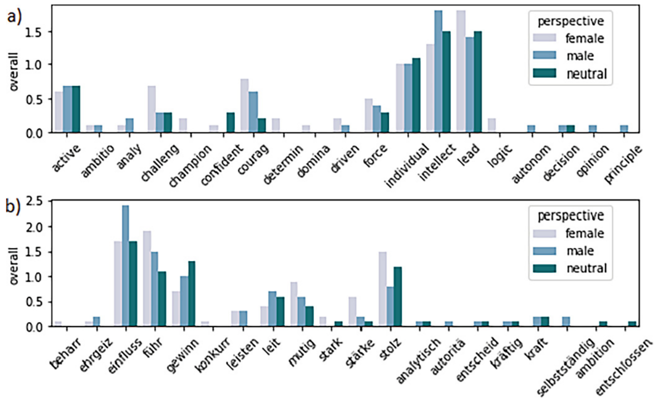


**Fig. 2.** Male coded words used on average in all perspectives of English responses (a) and German responses (b) for the prompt about a professor who won a prize. The number of usages is averaged over all responses of a perspective.

### 6.2 Characteristics of a Good Professor

When the system is prompted for the characteristics of a good professor, it produces one of the following two disclaimers: First, the characteristics described are independent of gender. Second, different people tend to have different opinions about the characteristics of a good professor. The first disclaimer is always added for female and male perspectives in both languages.

**Words Used.** The responses in both languages are empathising students and professors. Research is not a priority. The top ten words per perspective and language confirm that the terms "professor" and "student" are at the top of all lists. The English responses also include the words "learning", "research", and "teaching". The German responses are more diverse in the choice of top words. Female responses stress community, and male responses knowledge and research. Neutral responses stress research.

**Gender Coded Words.** Figure 3a) shows how female-coded words are used in the English responses. Out of 50 female-coded English words, 18 are used. All perspectives use 13 different female-coded words. The most used word is support, which is dominantly used in responses about female professors. Figure 3b) shows how responses use German female-coded words. Out of 62 female-coded German words, 24 are used. Texts about male professors use 16 female-coded words, which is the fewest of all perspectives. The most used word is unterstütz (support-ing) which is dominantly used for responses about male professors.

Figure 4a) shows the usage of male-coded words in English responses. Out of 52 male-coded English words, 14 are used. Responses about male professors use 7 of these words, which is the fewest of all perspectives. The most used word is courage, which is dominantly used to describe female professors. Figure 4 b)

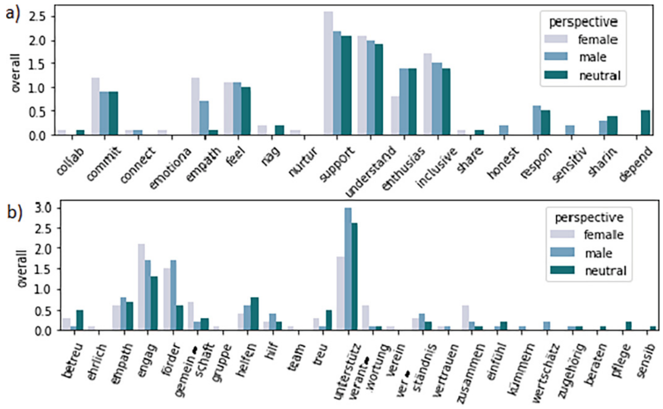


**Fig. 3.** Female coded words used on average in all perspectives of English responses (a) and German responses (b) for characteristics of a good professor prompt. The number of usages is averaged over all responses of a perspective.

shows that of 62 male-coded German words, 12 are used. With 6 words, the male perspective uses the least amount of male-coded words. The most used word is "mutig" (brave), dominantly used in responses about male professors.

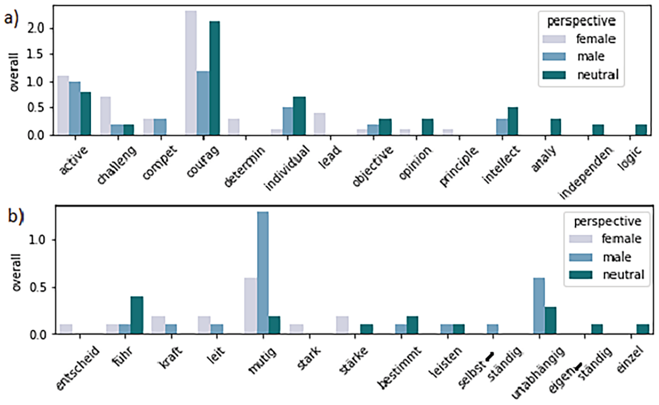


**Fig. 4.** Male coded words used on average in all perspectives of English responses (a) and German responses (b) for the prompt about the characteristics of a good professor. The number of usages is averaged over all responses of a perspective.

**Text Length.** Text length does not differ substantially between perspectives.

**Discussion.** Adding a gender to the prompt triggers a specific response. German and English responses do not differ much in stressed content, which is good for using the system for bi-lingual text generation. Additionally, ChatGPT avoids the excessive usage of gendered words in its responses. Interestingly, responses about German male professors have a high usage of female-coded words, and English responses about female professors have a high usage of male-coded words. As in the first prompt, text length does not differ much between the female and male perspectives.

### 6.3 System Updates

We experienced an unannounced system update during data collection, which fundamentally changed the kind of responses to the prompt about a professor who won a prize. Before, the system always generated a professor, a prize, and a university. After the update, the system exclusively generated fill-in-the-gap texts. Moreover, OpenAI introduced the "continue response" button during our data collection. Before it was introduced, responses could end mid-sentence or even mid-word. The first system update can seriously affect non-IT users who use the system in their daily work. Due to such system updates, proven prompts no longer work, and the user must invest time searching for new prompts that lead to the same result. This could take a long time because the user has to test for possible biased outputs whenever proven prompting strategies cease to work.

## 7 Conclusion

We explored ChatGPT with five different prompts posted in German and English requesting to take a female, male, and neutral perspective. The exploration phase showed that the system lacks grammatical and syntactical conciseness in German in general and especially when forced to use the gender-neutral German language. Adding a female or male perspective to a prompt can trigger a "gender template" causing the response to only focus on gender aspects that are ignored if the same prompt is posted from a neutral perspective. This template is also not properly tailored to specific genders since the model seems to be incapable of such a nuanced understanding, leading to responses about, e.g. underrepresented males in academia. In the exploitation phase, we find that the system favours female personas and STEM research fields. The responses describe all perspectives fairly equally and use only a few gender-coded words. The text length does not differ much between the perspectives, thus, not mirroring real-world texts. While ChatGPT is a helpful tool for non-IT users to draft a text, a thorough check of the results is crucial to ensure the absence of mistakes and biases.

Due to our endeavour to analyse ChatGPT from a non-IT user's perspective, working in university communications, we had a limited scope of possible prompts that led to subtle differences between the perspectives. To really explore the differences between gendered responses, more general prompts should be explored. Furthermore, we concentrated on ChatGPT, using GPT3.5. Other LLMs, especially newer ones, should be explored as other problems will arise with newer models.

**Acknowledgements.** This work has been partially funded by the Bundesministerium für Bildung und Forschung (BMBF, Federal Ministry of Education and Research) as part of Prof:inSicht - grant number 01FP21054. This work has been partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of BERD@NFDI - grant number 460037581.

**Ethical Implications.** This paper seeks to improve LLM research by highlighting problematic model behaviour. The structural changes in the response after unannounced framework updates, which we have seen, and also the errors regarding grammar and spelling, can increase the workload of the users. However, many of them are still quite obvious. When it comes to (gender) biases, also rarely occurring subtle differences can become a huge issue. Through the tremendous user base and the increasing number of use cases, the inherent biases are potentially multiplied by the system. OpenAI is trying to solve such issues downstream but with limited success, as we have seen, for instance, with the gender diversity template. It is important that these systems are challenged from a variety of diverse perspectives to uncover all sorts of potential problems. This is an important first step to solve mitigate them. We hope to contribute to this effort by analysing the system from the perspective of gender biases in English and German prompts. After all, LLM systems and research have to keep the users in mind. It is crucial to develop tools that make work easier for users.

Another aspect current LLM research has to keep in mind is not striving to replace human labour but to enhance human capabilities. The human must be kept in the loop and not be replaced.

## References

1. Bailey, A.H., LaFrance, M.: Who counts as human? Antecedents to androcentric behavior. Sex Roles **76**, 682–693 (2017)
2. Brown, T., et al.: Language models are few-shot learners. Adv. Neural Inf. Process. Syst. **33**, 1877–1901 (2020). https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. 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, pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/N19-1423. https://aclanthology.org/N19-1423
4. D'ignazio, C., Klein, L.F.: Data Feminism. MIT Press (2020)
5. Dutz, R., Rehbock, S., Peus, C.: Führmint gender decoder: Subtile geschlechtskodierung in stellenanzeigen erkennen und auflösen [führmint gender decoder: Identifying and resolving subtle gender coding in job advertisements]. Personal in Hochschule und Wissenschaft entwickeln (2020). no DOI available
6. Fast, E., Vachovsky, T., Bernstein, M.: Shirtless and dangerous: quantifying linguistic signals of gender bias in an online fiction writing community. In: Proceedings of the International AAAI Conference on Web and Social Media, August 2021, vol. 10, no. 1, pp. 112–120 (2021). https://doi.org/10.1609/icwsm.v10i1.14744. https://ojs.aaai.org/index.php/ICWSM/article/view/14744

7. Field, A., Tsvetkov, Y.: Unsupervised discovery of implicit gender bias. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 596–608 (2020)
8. Fu, L., Danescu-Niculescu-Mizil, C., Lee, L.: Tie-breaker: using language models to quantify gender bias in sports journalism. In: Proceedings of the IJCAI Workshop on NLP meets Journalism (2016)
9. Gaucher, D., Friesen, J., Kay, A.C.: Evidence that gendered wording in job advertisements exists and sustains gender inequality. J. Pers. Soc. Psychol. **101**(1), 109 (2011)
10. Graells-Garrido, E., Lalmas, M., Menczer, F.: First women, second sex: gender bias in Wikipedia. In: Proceedings of the 26th ACM Conference on Hypertext and Social Media, HT 2015, pp. 165–174. Association for Computing Machinery, New York (2015). https://doi.org/10.1145/2700171.2791036
11. Horvath, L.K., Sczesny, S.: Reducing women's lack of fit with leadership positions? Effects of the wording of job advertisements. Eur. J. Work Organ. Psy. **25**(2), 316–328 (2016)
12. Kulesza, T., Stumpf, S., Burnett, M., Kwan, I.: Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2012, pp. 1–10, Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2207676.2207678
13. Lakoff, R.: Language and woman's place. Lang. Soc. **2**(1), 45–79 (1973). https://doi.org/10.1017/S0047404500000051
14. Madaan, N., et al.: Analyze, detect and remove gender stereotyping from Bollywood movies. In: Conference on Fairness, Accountability and Transparency, pp. 92–105. PMLR (2018)
15. Mateo, C.M., Williams, D.R.: More than words: a vision to address bias and reduce discrimination in the health professions learning environment. Acad. Med. **95**(12S), S169–S177 (2020)
16. Nakandala, S., Ciampaglia, G., Su, N., Ahn, Y.Y.: Gendered conversation in a social game-streaming platform. In: Proceedings of the International AAAI Conference on Web and Social Media, May 2017, vol. 11, no. 1, pp. 162–171 (2017). https://doi.org/10.1609/icwsm.v11i1.14885. https://ojs.aaai.org/index.php/ICWSM/article/view/14885
17. OpenAI: Chatgpt: Optimizing language models for dialogue (2022). https://openai.com/blog/chatgpt/
18. OpenAI: ChatGPT(May 24 version). [Large Language Model] (2023). https://chat.openai.com/chat
19. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAi Blog (2019)
20. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(1), 5485–5551 (2020)
21. Sczesny, S., Formanowicz, M., Moser, F.: Can gender-fair language reduce gender stereotyping and discrimination? Front. Psychol., 25 (2016)
22. Trix, F., Psenka, C.: Exploring the color of glass: letters of recommendation for female and male medical faculty. Discourse Soc. **14**(2), 191–220 (2003)
23. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017). https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

24. Wagner, C., Garcia, D., Jadidi, M., Strohmaier, M.: It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In: Proceedings of the International AAAI Conference on Web and Social Media, August 2021, vol. 9, no. 1, pp. 454–463 (2021). https://doi.org/10.1609/icwsm.v9i1.14628. https://ojs.aaai.org/index.php/ICWSM/article/view/14628

# Part IV.

# Discrimination Detection Pipeline

# 6. Detecting Gender Discrimination on Actor Level Using Linguistic Discourse Analysis

Chapter 6 presents a language-agnostic and modular pipeline for detecting gender discrimination in text at the actor level. Drawing on concepts from linguistic discourse analysis, the approach identifies how individuals (actors) are referred to (nomination) and described (predication), capturing both overt and subtle discrimination in text. The pipeline integrates information extraction techniques and is designed to be adaptable across languages, datasets, and domains. It is validated on two real-world newspaper articles and several synthetic texts, demonstrating its ability to reveal gender-based differences in representation.

### Contributing article:

Urchs, S., Thurner, V., Aßenmacher, M., Heumann, C. and Thiemichen, S. (2024). Detecting Gender Discrimination on Actor Level Using Linguistic Discourse Analysis. *In Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), Bangkok, Thailand, August 17, 2024: 140-149. https://aclanthology.org/2024.gebnlp-1.8/.*

### Copyright information:

### Author contributions:

The entire work, including the idea, concept, data collection, analysis, evaluation, literature research, and manuscript writing, was carried out by Stefanie Urchs. The abstract was written by Stephanie Thiemichen, and the introduction by Veronika Thurner. All authors were involved in proofreading the manuscript.

### Supplementary material available at:

- Code and validation texts: https://github.com/Ognatai/nomination_predication

# Detecting Gender Discrimination on Actor Level Using Linguistic Discourse Analysis

**Stefanie Urchs[1], Veronika Thurner[1], Matthias Aßenmacher[2,3], Christian Heumann[2], Stephanie Thiemichen[1],**

[1]Faculty for Computer Science and Mathematics,
Hochschule München University of Applied Sciences, [2]Department of Statistics, LMU Munich,
[3]Munich Center for Machine Learning (MCML), LMU Munich,

**Correspondence:** stefanie.urchs@hm.edu

## Abstract

With the usage of tremendous amounts of text data for training powerful large language models such as ChatGPT, the issue of analysing and securing data quality has become more pressing than ever. Any biases, stereotypes and discriminatory patterns that exist in the training data can be reproduced, reinforced or broadly disseminated by the models in production. Therefore, it is crucial to carefully select and monitor the text data that is used as input to train the model. Due to the vast amount of training data, this process needs to be (at least partially) automated. In this work, we introduce a novel approach for automatically detecting gender discrimination in text data on the actor level based on linguistic discourse analysis. Specifically, we combine existing information extraction (IE) techniques to partly automate the qualitative research done in linguistic discourse analysis. We focus on two important steps: Identifying the respective person-named-entity (an actor) and all forms it is referred to (*Nomination*), and detecting the characteristics it is ascribed (*Predication*). As a proof of concept, we integrate these two steps into a pipeline for automated text analysis. The separate building blocks of the pipeline could be flexibly adapted, extended, and scaled for bigger datasets to accommodate a wide range of usage scenarios and specific ML tasks or help social scientists with analysis tasks. We showcase and evaluate our approach on several real and simulated exemplary texts.

## 1 Introduction

Ethical considerations as, e.g., formulated in the UNESCO's Recommendations on the Ethics of Artificial Intelligence, as well as emerging legislation such as the EU AI Act, require that any AI system adheres to fundamental values such as "the inviolable and inherent dignity of every human" (UNESCO, 2022). Specifically, this demand also holds true for systems based on large language models (LLMs). This implies that systems based on LLMs must carefully ensure that they do *not* reproduce, reinforce or broadly disseminate any existing biases, stereotypes or other discriminatory patterns, as this would violate the inherent human dignity.

However, LLMs are trained on existing data. If this input data is pervaded by stereotypes, biases and discrimination (as is often the case), the resulting model will reflect these discriminatory patterns. Thus, if developers need to ensure that an LLM-based system adheres to the ethical standards mentioned above, they can take one of two approaches: filter the LLM's output downstream to ensure that it is free from discrimination – or purge the input data from any discriminatory patterns, to ensure that the LLM itself will be free from discrimination in the first place.

Research on downstream gender bias mitigation in word embeddings by Gonen and Goldberg (2019) shows that downstream mitigation only hides bias and does not remove it. Thus, the effective alternative is to address bias upstream by selecting unbiased training data.

As the training corpora for LLMs need to be very extensive, it is impossible to ensure their quality manually. Therefore, technical means need to be developed that automatically detect discrimination in vast amounts of natural language texts.

What we read and see in media shapes our reality (Lippmann, 1929). If we are surrounded by bias and discrimination, we are likely to include these in our reality and act on them. That explains why media, notably text, plays an important role in the striving for equality for all genders. By detecting bias and especially discrimination against particular genders, it is possible to be wary of these texts and not distribute them. This is particularly important when choosing training data for natural language processing (NLP) tasks.

The term gender has at least three different notations: the linguistic gender, sex, and the social gender. The linguistic or grammatical gender can

*Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 140–149
August 16, 2024

be defined as follows: "*[...] grammatical gender in the narrow sense, which involves a more or less explicit correlation between nominal classes and biological gender (sex).*" (Janhunen, 2000). For example, in German, nouns could be female, male, or neutral. The sex, however, refers to a "biological" notion of gender that is "*binary, immutable and physiological*" (Keyes, 2018). This notion is flawed because intersex humans do exist, as well as trans-persons, thus refuting the binary and immutable part of this notion. For our work, we use the third notion, the social gender. This notion defines gender as a social construct represented by a person's intentional and unintentional actions to represent their gender and the reception of these actions. Therefore, the social gender is non-binary, flexible, and constructed by the person themselves and the persons perceiving them (West and Zimmerman, 1987; Devinney et al., 2022). We use the terms woman for persons who can be read as female-identifying, men for persons who can be read as male-identifying, and non-binary for persons who do not adhere to the before mentioned.

Bias against a particular gender entails discriminating against this gender. While bias contains all notions and beliefs towards a person/group (Mateo and Williams, 2020), (social) discrimination is a more intentional act: an offender treats someone or a group of people differently in a negative way, based on a specific feature of this person/-group (Reisigl, 2017). Textual discrimination is a special kind of (social) discrimination because the offender is not always apparent.

Linguistics and sociology have studied discrimination for over eighty years, mainly focusing on racism in the early research (Myrdal et al., 1944; Razran, 1950; Allport et al., 1954). During this period, different definitions of discrimination were defined, leading to different approaches for detecting it. One of these approaches is linguistic discourse analysis (LingDA), which inspects discourse to identify discriminating tendencies by combining research from sociology and linguistics (Bendel Larcher, 2015). Computational linguistics integrates LingDA and computer science into computational discourse analysis. So far, this discipline concentrates on the quantitative parts of LingDA, mostly focusing on coherence and cohesion (Dascalu, 2014). We concentrate on the qualitative parts of LingDA and partly automate the discrimination detection within the text.

## 2 Problem Formulation and Goals

Existing approaches for automatic discrimination detection often focus on identifying drastic wording, which is relatively easy to detect by simple comparison with a database of discriminatory terms. However, in many cases, textual discrimination manifests more subtly, requiring a more semantic approach to detect it.

To achieve our goal of automatically identifying discrimination and biases in text, we seek to enhance computational discourse analysis (CompDA) by integrating two fundamental, qualitative strategies from linguistic discourse analysis for detecting gender discrimination on the actor level: Identifying the respective person named entity (an actor) and all forms in which it is referred to (*Nomination)*, and then detecting the traits, characteristics, qualities, and features that are ascribed to this actor (*Predication*). By focusing on actors, we aim to reveal even subtle gender-specific discrimination. Furthermore, we can analyse the text's meaning on a deeper level.

To automatically process large amounts of input text data, we implement a pipeline for automated text analysis that integrates nomination and predication by using IE techniques (cf. Figure 1). Specifically, as a first step, we identify nominations by extracting the actors and detecting their pronouns. Second, we extract the predication of these actors and finally use the extracted information to analyse the whole text for discrimination. By ensuring a modular structure built from exchangeable components, we aim to make our pipeline flexibly adaptable, accommodating a wide range of usage scenarios and specific ML tasks. For example, the pipeline should be able to scale from single texts to a whole corpus, process different languages, and focus on different criteria, thus reflecting cultural differences.

Finally, we evaluate our approach and implementation by analysing several sample texts, two real-world examples, and three generated texts, and discuss the discrimination markers identified in these samples.

## 3 Background

This work combines qualitative research on LingDA with IE, thus enhancing quantitative CompDA methods for detecting gender discrimination in text. Discrimination is a form of bias. We define discrimination and its relation to bias.

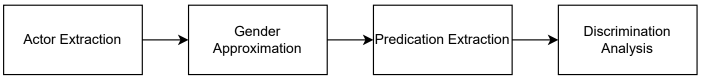


Figure 1: Visualisation of the flexible and language agnostic pipeline introduced in this work.

### 3.1 Linguistic Discourse Analysis

In LingDA, discourse is defined as a collection of text about a topic relevant to society (Bendel Larcher, 2015). This contrasts with computational linguistics, which defines discourse as “any multi-sentence text (Grishman, 1986). The focus of LingDA is the so-called actor. Actors are the entities in a text that perform some action. Actors can be individuals, groups, institutions, or organisations (Spitzmüller and Warnke, 2011).

Discourse is normally analysed on the corpus level as an extension of text linguistics that analyses single texts (Niehr, 2014). For our work, we concentrate on the level of single texts, especially on written text, potentially extending the approach to a whole corpus in future work. In this work, we disregard multimodal media, conversations, and pictures in general to scope our research. When analysing texts, Bendel Larcher (2015) points out that the nomination and predication is one of six aspects that should be considered. Nomination comprises how and what an actor in a text is named (Knobloch, 1996). The predication of the actor is what the text conveys about traits, characteristics, qualities, and features ascribed to this actor (Kamlah and Lorenzen, 1996).

When detecting nomination, the following aspects could be considered (Bendel Larcher, 2015):

1. **Proper Names**: Are actors referred to with their full name, surname, or just the first name?
2. **Generic Names**: When actors are not referred to by their proper names but with generic terms. Reisigl (2017) lists the following categories of problematic generic names: Negatively annotated general descriptions, ethnonyms, metaphorical slurs, animalistic metaphors, proper names used for a general description, and referring to an actor by their relation to someone else.
3. **Pronouns**: Pronouns can distance oneself from others (we vs. them), which is the basis for treating someone differently. Furthermore, using the wrong pronouns for someone (misgendering) is a clear aggression. Using the “generic masculine” in gendered languages like German can be considered problematic. Women and non-binary people are not directly addressed but are “included” in the word’s meaning. Therefore, women and non-binary people are not represented by the language.
4. **Deagentification**: The actor of the text is not named. The text only generally describes what is happening without giving credit to the person.

The predication detection analyses the text for characteristics, features, and qualities attributed to an actor. These can convey stereotypes and biases that can be extracted by looking at the following grammatical indicators (Reisigl, 2017; Bendel Larcher, 2015):

- **Attributes**: e.g. skinny, bright
- **Prepositional Attributes**: e.g. the professor living in Munich
- **Collocations** e.g. working mom
- **Relative Clauses** e.g. the tennis player who has a nice dress

For this work, we focus on indicators of discrimination based on the actor’s gender.

### 3.2 Computational Discourse Analysis

CompDA focuses on the analysis of cohesion and coherence. Cohesion describes how sentences are grammatically and lexically linked together to reflect the status of an actor through discourse. Typical methods include topics, coreferencing, and lexical and semantic word relatedness from ontologies. CompDA differentiates between referential cohesion (how often words, concepts, and phrases are repeated or related through the text) and causal cohesion (explicit use of connectives) (Dascalu, 2014). Coherence addresses the “*continuity of senses*” (De Beaugrande and Dressler, 1981) throughout the text. In other words, coherence conveys to the reader that the text is semantically connected. Dascalu (2014) distinguishes informational level coherence (causal relations between utterances, lexical chains, and centring theory) and intentional level coherence (tracing of the changes in the mental state of the discourse participants during the discourse).

Our approach combines cohesion and coherence by analysing the text using methods used in cohesion analysis to track actors (and their states) throughout the text.

### 3.3 Bias and Discrimination

Text can contain a lot of problematic properties regarding gender. The most problematic ones are biases and discrimination. However, also insults, defamation or misinformation should be avoided.

Mateo and Williams (2020) define bias as follows: "*Biases are preconceived notions based on beliefs, attitudes, and/or stereotypes about people pertaining to certain social categories that can be implicit or explicit.*". They continue that discrimination is the manifestation of biases through behaviour and actions. Reisigl (2017) has a clearer definition of discrimination: "*[...] social discrimination occurs when someone disadvantages or favours (i.e., treats unequally) a particular group or members of that group through a linguistic or other act or process, in comparison to someone else and on the basis of a particular distinguishing characteristic (such as an alleged 'race' or 'sexual orientation').*" leading to the following five parts of discrimination:

1. Offender
2. Victim (beneficiary in case of 'positive discrimination')
3. Disadvantaging (or favouring) act, process
4. Comparison group that is treated differently
5. Distinguishing feature on which the disadvantaging or favouring is grounded

Discrimination in written text is a manifestation of social discrimination. We consider discrimination as the manifestation of biases. Therefore, we consider the author of the text as the *offender*, and the *victim* is an actor of the text. The *feature that distinguishes* the victim from its *comparison group* is their gender. To scope our work, we only explore gender discrimination, even though we are aware that other kinds of discrimination, especially the intersection of different kinds of discrimination, exist and should not be part of NLP training data or other text. We extract the *disadvantaging act/process* from the text by quantifying differences between genders using LingDA and IE.

In manual LingDA researchers focus on the context of a text: was it released for a specific group of people from a specific kind of people? In the proper context, some kind of language that is offensive outside a group is acceptable if it is uttered by one person of a group towards another person of this group if it has an in-group context. Furthermore, some texts are seen as products of their time and represent the social norms of these times. However, when training NLP models, the context of a text is lost. The models learn equally on all text data. Therefore, we always have to assume an out-group context and the current social norms when evaluating textual data for training purposes.

Not removing discrimination and biases from training data leads to representational harms: gender stereotypes are spread in generated texts and, therefore, hardened in readers' minds. This harms all genders. Furthermore, not representing non-binary individuals in text generated by large language models (LLM) decreases their visibility. However, non-binary individuals are a part of our world and should be visible in LLM-generated texts. A text corpus not containing non-binary representation can not be considered balanced.

### 3.4 Information Extraction

IE locates predefined information in natural language text. According to Grishman (2015), the following steps are performed during IE (not necessarily in the order mentioned):

1. **Named Entity Recognition**: extraction of entities with proper names (persons, organisations, places, or suchlike)
2. **Syntactic Analysis**: extraction of syntactic information from sentences and tokenisation
3. **Coreference Resolution**: combining several mentions of an entity into one (e. g. a text mentions Dr. Ruth Harriet Bleier, further mentions may take the form of "Dr. Bleier", "Ms. Bleier", "R. H. Bleier", "R. B." or "she") (we also add generic names to form the full nomination of an actor)
4. **Semantic Analysis**: extracting relations between entities and mapping of sentences containing an entity to this entity (predication of an actor)
5. **Resolution of Cross-Document Coreferences**: coreferencing an entity through several documents (We are not exploring this step in this work.)

## 4 Methodical Approach

Our analysis pipeline can be subdivided into four consecutive steps that build on each other (cf. Figure 1): The first task is to extract the actors, followed by a gender approximation for each actor. In these steps, we save the nomination of each actor in our knowledge base. The third step expands the knowledge base with the predication of each actor detected in step one. As the fourth and final step of the pipeline, we analyse the extracted information for potential discrimination.

### 4.1 Nomination

The nomination process starts with the tokenisation of the text. No further preprocessing is applied to retain the full semantic meaning of the text. Subsequently, the dependency trees are parsed for each sentence. Therefore, each token is annotated with its relation to its semantic neighbours and its part of speech. All tokens that are proper nouns are analysed using named entity recognition (NER). Person entities are the actors of the text. As actors are mentioned more than once in a text, it is essential to coreference all mentions of the same actor. Coreferencing combines all references of one actor (this can be done in one text or the whole corpus). Therefore, the full name of an actor is matched to its name parts (e.g. first name, last name, last name, and abbreviations of first name), pronouns, and titles. In less formal settings, actors are referred to by generic names. These are not detected as proper nouns during NER. Therefore, generic names must be detected in an additional step and coreferenced with actors. We use a list of commonly used generic names to detect the generic names. All coreferenced entities and pronouns are the nomination of the actor. These are saved into a knowledge base using the same key for later use.

Every actor in the knowledge base is assigned one of the following gendered entries: woman, man, non-binary, unknown. The gendered entry is assigned by pronouns in the actor nomination.

### 4.2 Predication

The predication analyses what is ascribed to an actor. Ideally, the predication should only contain text that describes an actor. If a sentence contains more than one actor, this sentence should be split and matched accordingly. Furthermore, if an actor describes another actor, the sentence should only match the described actor and not the active one. For our proof of concept implementation, we simplify the sentence-matching process and assign a sentence to an actor if the actor is contained in this sentence. The predication is also stored in the knowledge base.

### 4.3 Discrimination Detection

We analyse the nomination for common derogatory terms for each entry in the knowledge base. To scope the research, we only use lists of derogatory terms referring to women, men, and transgender people[1]. For all predication sentences, the sentiment of the sentence is computed. Furthermore, the predication is analysed for feminine-coded words and masculine-coded words[2]. The authors show that women are associated with communal traits and men with more agency-related terms. Overusing gender-coded language can embed stereotypes. Using the computed information, we compile a discrimination report. For detailed report components, see Section 5.3.

## 5 Implementation and Validation

As mentioned in Section 4, we start by collecting the nomination of actors and subsequently enhance our knowledge base with the predication of the actors. The content of the knowledge base is subsequently analysed for discrimination and biases[3]. The code for our pipeline can be found on GitHub[4].

### 5.1 Nomination

SpaCy can perform tokenisation, dependency parsing, part of speech tagging, and named entity recognition out of the box. The named entity recognition can detect all actors in the text. When manually evaluating the results of our pipeline in the sample texts, we found that one actor's name was not classified as a person. Still, the error was not severe enough to justify changing libraries. We use the person entities as seed for the nomination.

In the first step, we extract all compounds of an actor's name; the head element of the compound is used as a key in a dictionary of actors. In a text

[1] derogatory terms were collected from the following websites (accessed on 2024-05-08): https://en.wikipedia.org/wiki/Category:Pejorative_terms_for_women, https://en.wikipedia.org/wiki/Category:Pejorative_terms_for_men, https://genderkit.org.uk/slurs/, https://en.wiktionary.org/wiki/Category:English_swear_words

[2] We use the lists of feminine/masculine coded words as found on the gender decoder website https://gender-decoder.katmatfield.com/about, which is based on work from Gaucher et al. (Gaucher et al., 2011)

[3] We use Python (version 3.9.18) and the NLP library SpaCy (Honnibal and Montani, 2017) in version 3.7.2, in combination with the en_core_web_lg model, for our experiments. Furthermore, we use the packages coreferee (version 1.4.1) and spacytextblob (version 4.0.0).

[4] https://github.com/Ognatai/nomination_predication

about Bill Clinton, the key `Bill Clinton` contains the values `Bill Clinton`, `Clinton`, `President`, and unexpectedly `trail`. We can also extract titles; for example, the key `Kirsten Gillibrand` contains the values `Sen.` and `Kirsten Gillibrand`. This implementation combines all actors with the same first or last names into one nomination.

In the second step, keys that are part of the value of another key are merged into the other key. Thus, all nomination keys are full names (if the actor is mentioned with their last name; otherwise, the key is a first name), and first names and last names are assumed to be unambiguous. These nominations are extended by a list of generic names found in the text and not coreferenced to other actors.

We determine the pronouns and, therefore, approximate the gender of the actors by using `coreferee`. This package references pronouns to actors. Unfortunately, `coreferee` has problems identifying gender-neutral/non-binary pronouns. In two of three test texts, it cannot detect the non-binary actors. Due to the lack of better-performing packages, we use `coreferee` nonetheless. Actors are assigned woman or man if the majority (at least 70%) of used pronouns refer to one of these gendered entries (we use a majority of at least five pronouns to be able to react to software problems stemming from the matching algorithm of `coreferee`). A non-binary entry is only assigned if gender-neutral/non-binary pronouns are used consistently. Otherwise, the gender is listed as unknown.

The last step of the nomination detection is to combine all information into a knowledge base stored as a pandas (pandas development team, 2023) data frame.

### 5.2 Predication

In the predication phase, the knowledge base is extended by all sentences that mention the corresponding actor. Each token object contains information about its position in the text. Therefore, we generate a text span with the size of the token and obtain the sentence that includes the text span of the token. Duplicates within one actor are removed. If a sentence contains more than one actor, this sentence is matched to all contained actors.

### 5.3 Discrimination Detection

For the discrimination detection, we extend the knowledge base by the sentiment of each predication sentence and the gender-coded words used in the predication. We use the package `spacytextblob`[5], which builds upon the `textblob`[6] library, to assign a value between -1 (very negative sentiment) and 1 (very positive sentiment) to each sentence. The sentiment analysis utilises a naive Bayes classifier trained on movie reviews. To detect gender-coded words, we use a list of feminine-coded and masculine-coded word stems by Gaucher et al. (2011) and test if these stems occur in the predication. We create a discrimination report for a text, building on the information of the knowledge base we created for this text. The report contains the following information:

- count of woman, man, non-binary, and undefined actors overall and per actor
- count of woman, man, non-binary, and undefined actor mentions overall and per actor
- sentiment towards woman, man, non-binary, and undefined actors overall and per actor
- count of feminine-coded words and masculine-coded words in the actor predication of woman, man, non-binary, and undefined actors overall and per actor
- abusive words used for woman, man, non-binary, and undefined actors and overall

### 5.4 Validation

Most NLP tasks like hate speech detection or sentiment analysis tend to utilise short utterances, like tweets or social media posts, for training purposes. In contrast, our approach aims to analyse longer texts like news articles or blog posts that describe one or more persons.

For testing our pipeline, we generate three texts with ChatGPT (OpenAI, 2023) that contain several actors, with at least one respectively using feminine, masculine, or gender-neutral/non-binary pronouns. All these actors have a full name and interact with each other. The content of all three generated texts is rather generic and not biased. We generated these texts mainly to test the pipeline on non-binary actors, but we do not further discuss the results of these texts because of their generic nature[7]. Instead, we collected texts about Bill and Hillary Clinton from Fox News[8].

The Hillary Clinton text describes Hillary Clin-

[5] `https://spacy.io/universe/project/spacy-textblob`

[6] `https://textblob.readthedocs.io/en/dev/`

[7] All text are available on GitHub: `https://github.com/Ognatai/nomination_predication`

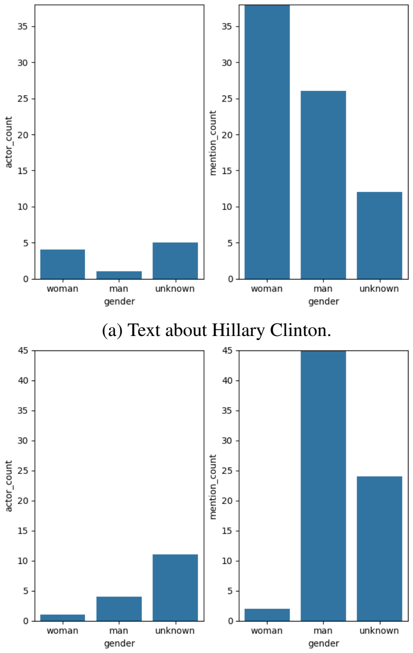
[8] `https://www.foxnews.com/`

ton's controversial statement that Trump followers should be 'deprogrammed' and reactions to this statement. The Bill Clinton text details how Bill Clinton "*reemerges as Democrat surrogate after being silenced by #MeToo movement*".[9]

We use our pipeline on these texts and compare the results by manually checking the corresponding texts for the correctness of the results.

The pipeline can detect all actors contained in the texts. Only the texts generated with ChatGPT contain non-binary actors. When analysing these texts, we found that `coreferee` has problems matching gender-neutral/non-binary pronouns to actors. Non-binary actors are detected in only one of three texts. Otherwise, our pipeline can mainly match the correct pronouns to the corresponding actor. We encounter problems in the text about Hillary Clinton. Here, `coreferee` has problems matching a pronoun from a partial sentence to one of the three actors mentioned before.

To count the mentions of each actor, we count all entries in the nomination and pronoun columns of the knowledge base. This leads to a minor problem since titles are not part of the name token and are counted as additional mentions. In our test data, this behaviour leads to one to two additional mentions per actor. In a future version of the pipeline, this behaviour will be fixed. Figure 2a and Figure 2b shows how many actors of a specific gender are part of the text and how often actors of a specific gender are mentioned throughout the text. Both texts do not contain non-binary actors. Interestingly, in the text about Hillary Clinton (Figure 2a, we detect four women (mentioned 38 times) and one man (mentioned 26 times). However, of the 38 women mentioned, Hillary Clinton is mentioned 26 times. Therefore, Donald Trump, the only recognised man, is mentioned as often in a text about Hillary Clinton as Hillary Clinton herself. However, the text describes how Hillary Clinton criticises Donald Trump's followers; therefore, many mentions make sense. In the text about Bill Clinton (Figure 2b, we detect four men, which are mentioned 45 times; 35 are mentions of Bill Clinton.



(a) Text about Hillary Clinton.

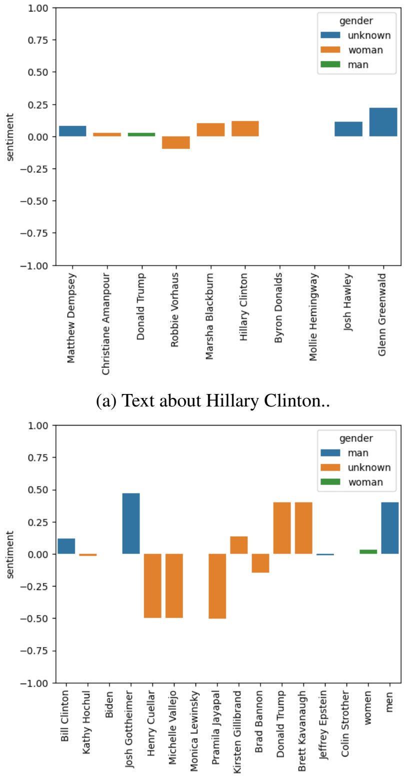
(b) Text about Bill Clinton.

Figure 2: Comparison of how often actors of a certain gender occur in the text and how often actors of a certain gender are mentioned. Both texts do not contain non-binary actors.

The sentiment analysis we use in our pipeline encounters problems when used for news articles. Figure 3b shows a moderately negative sentiment for `Henry Cuellar` and `Michelle Vallejo` which refers to the sentence "*During the trip, Clinton will rally with Rep. Henry Cuellar and Democratic candidate Michelle Vallejo – each of whom is locked in a difficult contest with Republicans.*" The sentence has a very neutral tone. In contrast, the model detects almost no negative sentiments in the text about Hillary Clinton (see Figure 3a. However, the predication of Hillary Clinton contains the following sentences: *"Sen. Marsha Blackburn, R-Tenn., posted to X, "Hillary Clinton wants Trump supporters to be formally reeducated., Independent journalist Glenn Greenwald shredded Clinton over the comments, saying, "As she gets increasingly bitter about her 2016 defeat – even when you think there's no way she can – Hillary Clinton is more and more the liberal id: she just spews what liberals really think and feel but know not to say., Clinton's 'deprogramming' hopes for Trump supporters a long shot in the era of political silos Clinton has had sharp words for Trump supporters over the years, once calling them 'deplorables'.*" The sentence contains a negative sentiment towards Hillary Clinton, but `spacyblob` cannot detect those neg-

[9] All text are available on GitHub: https://github.com/Ognatai/nomination_predication

ative sentiments. These examples showcase that the language used in news articles is too different from that used in movie reviews (which are one of the standard sources of training data for sentiment analysis approaches). Therefore, it is impossible to use a model trained on movie reviews for every domain; in future work, a domain-specific sentiment model will be utilised.



(a) Text about Hillary Clinton..

(b) Text about Bill Clinton.

Figure 3: Visualisation about the sentiments towards certain actors. Both texts do not contain non-binary actors.

In all texts, gender-coded words are rarely used. Both "real-world" texts contain a few feminine-coded words (Bill Clinton: 1, Hillary Clinton: 6) but no masculine-coded ones. Nevertheless, these could be an interesting feature if used for the whole corpus. We have a very explicit list of abusive words, but none are used in our sample texts. This list should be exchanged with domain-specific hate speech detection.

## 6 Discussion

Our method shows promising first results, even on our limited test data.

### 6.1 Strengths

Our pipeline can detect how different actors in a text are described. By approximating the gender of the actors, we can analyse if the text differentiates between genders and discriminates against a particular gender. Texts with very negative sentiments towards certain genders could then be excluded from model training, for instance. Our pipeline differentiates from other discrimination detection methods by focusing on actors and not the text as a whole. Therefore, it is possible to detect more subtle discrimination. Our pipeline is modular and, therefore, flexible. Single modules can be exchanged for domain-specific modules, and the pipeline can be extended anytime. Other discrimination detection approaches like hate speech detection or word lists can be included. The flexibility of the pipeline offers the possibility of even changing the languages of the texts analysed. Our proof of concept verifies the assumption that we can partly automate the qualitative parts of linguistic discourse analysis. Our discrimination report helps, for example, social scientists to decide if a text may contain discrimination or biases. This pipeline will be scaled to the corpus level to fully analyse the discourse within the corpus.

### 6.2 Limitations

Our proof-of-concept pipeline is tailored to detect actors in text. We cannot analyse the text if the text does not describe specific actors but a general situation. We combine actors with the same first and/or last name into one and do not coreference generic nominations to already detected actors. The predication should only consider text parts that attribute something to an actor. Currently, we use all sentences that contain the actor. If a sentence contains more than one actor, we match this sentence to all actors instead of doing an in-depth analysis of which parts of the sentence could belong to which actor. This also affects the sentiment analysis. A sentence containing an actor is not always a sentence containing a sentiment towards this actor. Another source of limitations is the general-purpose models we use in our pipeline. These are not tailored to the domain of news articles, leading to a sub-optimal performance. These general-purpose models also have problems in detecting gender-neutral/non-binary pronouns.

## 7 Conclusion and Future Work

In this work, we build a flexible pipeline to analyse newspaper articles and blog posts about people. We use linguistic methods to detect how actors are described within a text. In contrast to common discrimination detection methods, we do not treat the whole text as one object. By focusing on actors and the gender of the actors, we can do more nuanced text analyses that can detect subtle discrimination on a gender basis. First, limited tests on newspaper articles show that we can detect how actors are treated differently, depending on their gender. The first proof-of-concept pipeline implementation has some limitations that will be addressed in future work.

Other future work includes using the pipeline in different languages, such as German. Furthermore, instead of analysing one text at a time, we will scale the input to several documents, analysing complete corpora. We will also experiment with different pipeline components, for example, exchanging the simplistic abusive language detection with a sophisticated hate-speech detection or coreferencing detected actors with real-world actors to detect their pronouns. As today's discourse is not only written, analysis of multi-modal data might also be an interesting endeavour.

## Ethical Consideration Statement

Defining discrimination for LLM training data means defining the value system for internationally used systems, but we do not share one common international value system. We can all agree on international human rights. However, an LLM also generates texts containing opinions about religion, race, gender, and sexual orientation. There are currently no common international values regarding these topics. As computer scientists, we define the values and opinions that our systems should convey. However, we are only able to adhere to our value system. Therefore, it is essential to work in diverse teams. The author team enriches their perspective by discussing our research with researchers from fields outside of computer science and from different cultural backgrounds. Our team consists of white Western European researchers. Three of us identify as women, representing the feminine and masculine gender spectrum but not the non-binary. Nevertheless, our group's diversity helps analyse gender-specific discrimination. Our understanding of discrimination stems from the system of beliefs and values based on Western European culture.

## Acknowledgements

This work was written by an author team working in different projects. Stefanie Urchs' project "Prof:inSicht" is promoted with funds from the Federal Ministry of Education and Research under the reference number 01FP21054. Matthias Aßenmacher is funded with funds from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of BERD@NFDI - grant number 460037581. Responsibility for the contents of this publication lies with the authors.

## References

Gordon Willard Allport, Kenneth Clark, and Thomas Pettigrew. 1954. *The nature of prejudice*. Addison-wesley Reading, MA.

Sylvia Bendel Larcher. 2015. *Linguistische Diskursanalyse: Ein Lehr-und Arbeitsbuch*. Narr Francke Attempto Verlag.

Mihai Dascalu. 2014. *Computational Discourse Analysis*, page 53–77. Springer International Publishing.

Robert-Alain De Beaugrande and Wolfgang U Dressler. 1981. *Introduction to text linguistics*, volume 1. longman London.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "gender" in nlp bias research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2083–2102, New York, NY, USA. Association for Computing Machinery.

Danielle Gaucher, Justin Friesen, and Aaron C Kay. 2011. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of personality and social psychology*, 101(1):109.

Hila Gonen and Yoav Goldberg. 2019. 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)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Ralph Grishman. 1986. *Computational linguistics: an introduction*. Cambridge University Press.

Ralph Grishman. 2015. Information extraction. *IEEE Intelligent Systems*, 30(5):8–15.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Juha Janhunen. 2000. Grammatical gender from east to west. *TRENDS IN LINGUISTICS STUDIES AND MONOGRAPHS*, 124:689–708.

Wilhelm Kamlah and Paul Lorenzen. 1996. *Die Elementare Prädikation*, pages 23–44. J.B. Metzler, Stuttgart.

Os Keyes. 2018. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Clemens Knobloch. 1996. *Nomination: Anatomie eines Begriffes*, pages 21–53. VS Verlag für Sozialwissenschaften, Wiesbaden.

Walter Lippmann. 1929. *Public Opinion: By Walter Lippmann*. Macmillan Company.

Camila M Mateo and David R Williams. 2020. More than words: a vision to address bias and reduce discrimination in the health professions learning environment. *Academic medicine*, 95(12S):S169–S177.

Gunnar Myrdal et al. 1944. *An American dilemma; the Negro problem and modern democracy.(2 vols.)*. Harper.

Thomas Niehr. 2014. *Einführung in die linguistische Diskursanalyse*. WBG (Wissenschaftliche Buchgesellschaft).

OpenAI. 2023. ChatGPT(November 06 version).

The pandas development team. 2023. pandas-dev/pandas: Pandas.

Gregory Razran. 1950. Ethnic dislikes and stereotypes: a laboratory study. *The Journal of Abnormal and Social Psychology*, 45(1):7.

Martin Reisigl. 2017. *Sprachwissenschaftliche Diskriminierungsforschung*, pages 81–100. Springer Fachmedien Wiesbaden, Wiesbaden.

Jürgen Spitzmüller and Ingo Warnke. 2011. *Diskurslinguistik: eine Einführung in Theorien und Methoden der transtextuellen Sprachanalyse*. Walter de Gruyter.

UNESCO. 2022. Recommendation on the ethics of artificial intelligence. PDF Document. Page 18.

Candace West and Don H. Zimmerman. 1987. Doing gender. *Gender & Society*, 1(2):125–151.

# 7. taz2024full: Analysing German Newspapers for Gender Bias and Discrimination across Decades

Chapter 7 introduces taz2024full, the largest publicly available German newspaper corpus to date, comprising over 1.8 million articles published by the newspaper taz between 1980 and 2024. The corpus enables diachronic analyses of media language and discrimination. To explore gender representation, an extended version of the actor-based discrimination detection pipeline from Chapter 6 is adapted for German and scaled to handle large datasets. The analysis reveals persistent gender imbalances in reporting, with men more frequently mentioned and more positively framed than women, though coverage of women has increased over time.

### Contributing article:

Urchs, S., Thurner, V., Aßenmacher, M., Heumann, C. and Thiemichen, S. (2025). taz2024full: Analysing German Newspapers for Gender Bias and Discrimination across Decades. *In Findings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, July/August 27-01, 2025: 10661–10671. https://aclanthology.org/2025.findings-acl.555/.*

### Copyright information:

### Author contributions:

Stefanie Urchs was responsible for the study idea, data collection, conceptualisation, conducting the experiments, literature research and writing the manuscript. The other authors supported the work by providing technical guidance, conceptual discussions and checking the results.

### Supplementary material available at:

- Code: https://github.com/Ognatai/corpus_pipeline
- Corpus: https://doi.org/10.5281/zenodo.15480855

# taz2024full: Analysing German Newspapers for Gender Bias and Discrimination across Decades

**Stefanie Urchs[1,2], Veronika Thurner[1], Matthias Aßenmacher[2,3], Christian Heumann[2], Stephanie Thiemichen[1],**

[1]Faculty for Computer Science and Mathematics,
Hochschule München University of Applied Sciences, [2]Department of Statistics, LMU Munich,
[3]Munich Center for Machine Learning (MCML), LMU Munich,

**Correspondence:** stefanie.urchs@hm.edu

## Abstract

Open-access text corpora are crucial for advancing research in natural language processing (NLP) and computational social science (CSS). Despite the growing availability of datasets, resources for languages other than English, such as German, remain scarce. This limits large-scale studies on linguistic, cultural, and societal trends and hinders research of complex issues like gender bias and discrimination. To address this gap, we present `taz2024full`, to our knowledge, the largest publicly available dataset of German newspaper articles to date. Comprising over 1.8 million articles from the German newspaper "taz" spanning 1980 to 2024. Unfortunately, including other sources in the corpus was impossible, as no other German newspaper provided free access to their data or allowed the publication of such a dataset. While access could have been obtained through paid licensing, this would not have guaranteed full data availability, and legal restrictions would have prohibited the release of the corpus for public use. As a result, "taz" remains the sole source for this dataset.

To demonstrate the potential of the corpus for bias and discrimination research, we analyse how references to different genders have evolved over more than four decades of reporting. Our findings reveal a persistent imbalance, with men consistently appearing more frequently in articles and receiving more textual space. However, we also observe a gradual shift towards a more balanced representation of genders in recent years. By adapting and scaling an existing pipeline for detecting gender bias and discrimination in news media, we provide researchers with a structured approach to studying actor representation, sentiment, and linguistic framing in German journalistic texts.

The `taz2024full` corpus and its accompanying pipeline support a wide range of research applications, from studying language evolution to investigating media bias and discrimination. By making this resource publicly available and demonstrating its application, we aim to facilitate interdisciplinary research, foster inclusivity in language technologies, and contribute to a more informed selection of training data for NLP models.

## 1 Introduction

Publicly available text corpora are invaluable resources for advancing research in natural language processing (NLP) and computational social science (CSS). However, most of these resources are frequently limited to English, restricting language-specific studies in other languages, such as German. The scarcity of open-access German newspaper datasets has hindered research on German news media's linguistic, cultural, and societal aspects. Specifically, the research on gender discrimination and bias detection suffers from a lack of resources as large collections of preferably diverse texts are necessary to analyse these phenomena *at scale*.

Given that the term "*gender*" is not always used uniformly, we define its usage in the scope of this paper: We refer to "*gender*" as a social construct, which is non-binary, flexible, shaped by individuals and by those perceiving them, rather than a biological given (West and Zimmerman, 1987; Devinney et al., 2022). However, due to methodological and linguistic constraints, we are limited to working within a binary gender spectrum in this work. We use pronouns to assume the gender of actors. As we are working on German data, we are limited to the commonly used pronouns in this language. Since German lacks widely used non-binary pronouns, non-binary genders could not be included in our analysis. We further employ the definition of "*biases*" as "all notions and beliefs a person has towards another person or group of persons" (Mateo and Williams, 2020). These biases can manifest in (written) discrimination, wherein a person or group is intentionally mistreated based on specific characteristics (Reisigl, 2017).

**Contribution:** In this work, we present a twofold contribution to the research area of gender bias and gender discrimination in written language and potentially also beyond:

1. We introduce the corpus “taz2024full”, a comprehensive dataset that can serve as a valuable resource for analysing German newspaper articles published between 1980 and 2024, a period spanning more than four decades. Beyond our use case (see below), this allows analysing various other phenomena in the German language over time.

2. Furthermore, we demonstrate the potential of the corpus for bias and discrimination research by analysing how references to different genders have developed over more than 40 years of newspaper reporting, examining both their frequency and how they are discussed.

## 2 Related Work

Most publicly available news corpora focus on English newspapers, providing this type of diverse NLP research resource only for a single language. The "*Chronicling America*" dataset, provided by the Library of Congress, offers access to historical newspapers and digitised pages (spanning from 1690 to the present day) (Library of Congress). The "*BBC News Summary Dataset*" includes 2,225 documents from 2004 to 2005, covering five topical areas (Gupta et al., 2022). The "News Category Dataset"[1] contains around 210,000 headlines and descriptions from `https://www.huffpost.com`, spanning 2012 to 2022 (Misra and Grover, 2021). The "*News Articles Dataset*" comprises articles from 2015 to 2017, scraped from `https://www.thenews.com.pk/`, focusing on business and sports (Mahmood, 2017). Larger corpora such as RealNews, constructed from CommonCrawl dumps, offer 120GB of deduplicated news articles from 2016 to 2019, targeting large-scale training and evaluation of LLMs (Zellers et al., 2019). Additionally, the "*20 Newsgroups*" dataset[2], widely used in NLP, contains approximately 20,000 documents grouped into 20 categories, though it is neither based on newspaper content nor does it specifically look at developments over time (Lang, 1995).

Regarding news datasets specifically for the German language, the "*One Million Posts Corpus*" is one of the most prominent examples. It is derived from online discussions on the Austrian newspaper DER STANDARD’s website. It contains user posts from 2015-06-01 to 2016-05-31, with 11,773 labelled and 1,000,000 unlabelled entries, providing valuable insights into user-generated content (Schabus et al., 2017). This is, however, notably different from our resource as we do not focus on user-generated content, but on the articles themselves. In addition to this, several linguistic newspaper corpora exist, offering access to German news data. These include DWDS (Berlin-Brandenburgischen Akademie der Wissenschaften), the TüPP-D/Z corpus (Seminar für Sprachwissenschaft), the Mannheim German Reference Corpus (DeReKo) (Kupietz and Keibel, 2009), Leipzig Wortschatz (Universität Leipzig et al.), TIGER (Sabine et al., 2004), and others (Schiller et al., 1999; Nolda et al., 2021). However, these resources are often limited to keyword searches and return only sentence-level results, with most of them being unavailable for public use.

These datasets highlight the diversity of resources available for English and German news research but also reveal limitations, such as restricted access and narrow use cases. This underscores the need for openly available, comprehensive datasets like taz2024full to support language-specific studies and address gaps in research on German news media.

The detection of bias and discrimination in NLP has received significant attention in recent years. Blodgett et al. (2020), Sun et al. (2019), and Shah et al. (2020) provide broad overviews of existing approaches, highlighting the diverse methodologies employed to identify, mitigate, and evaluate biases in textual data. These works discuss various strategies, from detecting stereotypical associations in embeddings to evaluating fairness in predictive systems. Their surveys cover theoretical frameworks and practical implementations, offering valuable insights into the state of the field.

Despite the breadth of these reviews, none of the techniques discussed are directly comparable to the approach proposed by Urchs et al. (2024), which we adopt for evaluating our dataset. Unlike many conventional bias detection methods, which focus on linguistic patterns, embeddings, or statistical measures, the method of Urchs et al. (2024)

[1] `https://huggingface.co/datasets/heegyu/news-category-dataset`

[2] `https://huggingface.co/datasets/google-research-datasets/newsgroup`

combines information extraction and linguistic discourse analysis to identify markers of bias and discrimination at the actor level. This actor-focused approach enables a more nuanced examination of how individuals and groups are represented in text, making it particularly suited for analysing bias in news media.

## 3 Dataset

We introduce the `taz2024full` newspaper corpus, a German newspaper corpus containing 1,834,370 publicly available articles published between 1980 and 2024 in the German newspaper "taz". This extensive dataset provides a valuable resource for linguistic, cultural, and societal research. It enables analyses across more than four decades of journalistic content. To the best of our knowledge, this is a unique collection of articles from a single news source over such an extended period, offering insights beyond specific use cases and opening avenues for long-term, diachronic studies (cf. section 4).

The dataset covers a wide range of historical contexts, including events of global significance, such as 9/11, the financial crisis in Europe, and the COVID-19 pandemic, all of which have shaped discourse on an international scale. Additionally, it encompasses events with particular relevance to Germany, such as the reunion, the 2015 migrant crisis, and several political changes, providing researchers with a lens to explore national and regional impacts on public discourse. The corpus's temporal span allows for the analysis of how language, societal attitudes, and journalistic practices have evolved in response to these events.

For our corpus, we exclusively used "taz" as a data source because no other German newspaper granted us free access to their archives. Licensing fees would have been required, and even then, access to the full dataset would not have been guaranteed. Furthermore, publishing the corpus would have been prohibited due to legal restrictions on data redistribution.

We have chosen not to provide a predefined train-test split for the dataset. Data splitting strategies may vary significantly depending on research objectives and use cases. For instance, studies focusing on the impact of specific historical events may require custom temporal splits, while others analysing long-term trends might need broader, cross-temporal divisions. Allowing users to define their splits ensures maximum flexibility and adaptability for diverse research applications.

By offering a corpus that captures the breadth and depth of "taz" reporting across decades, we aim to provide a foundation for a wide range of studies, from exploring shifts in public discourse to examining linguistic phenomena and identifying patterns of bias and discrimination. This adaptability makes the `taz2024full` corpus a critical resource for researchers in NLP, CSS, and related fields.

To our knowledge, this corpus is the largest German newspaper corpus available. Other publicly accessible German newspaper corpora do not provide full access to their data; instead, they typically allow only keyword-based searches or sentence-level queries through linguistic databases, making large-scale analysis impossible.

### 3.1 Data Source

The "taz" (*die Tageszeitung*, `https://taz.de/`) is a German daily-occurring, left-leaning newspaper based in Berlin. First published on September 22, 1978, it transitioned to a daily publication schedule on April 17, 1979 (taz, 2018). Known for its progressive editorial stance, "taz" has built a reputation as a prominent voice in the German media landscape, with 13,800 subscribers and an additional 39,000 paying for digital content (taz, 2024). The newspaper plans to cease daily printing on October 17, 2025, transitioning to an online-only format while retaining a weekly Saturday print issue. This transition marks a shift reflective of broader trends in the news industry.

The "taz" newspaper is known for its diverse and comprehensive reporting, covering a wide array of topics. Its editorial structure is organised into various sections, including breaking news, politics, society, culture, and sports. Additionally, "taz" has dedicated segments such as "Öko," which covers economics, ecology, labour, consumption, transport, science, and the network economy. There are also regional sections focusing on Berlin and northern Germany, as well as "Wahrheit," which features unconventional content like satire, commentary, and creative formats. Unfortunately, the metadata we crawled did not include explicit labels indicating the section in which an article was published. As a result, this information could not be incorporated into the dataset. Nonetheless, the corpus reflects the full spectrum of "taz" journalism, encompassing a rich variety of topics and perspectives.

### 3.2 Data Collection

The `taz2024full` corpus was created by crawling publicly available content from the "taz" website (`https://taz.de/`) between August 2024 and November 2024. Permission has been granted to use and release this dataset for academic research, though its use for commercial purposes is strictly prohibited. The dataset only contains articles with more than three tokens (measured with SoMaJo tokeniser (Proisl and Uhrig, 2016)), thus excluding articles that only contain text fragments. The `taz2024full` corpus will be publicly available on Zenodo for non-commercial, academic purposes[3].

The articles are stored in JSON format. Figure 1 provides an example of the full JSON structure, illustrating how the metadata and article components are organised in the dataset. The metadata was extracted directly from the HTML of the crawled articles, and no modifications were made to the entries. The JSON format contains the following fields:

- **`"published_on"`**: The publishing date, stored as a string in the format `YYYY-MM-DDThh:mm:ss+01:00`.
- **`"contains_actors"`**: A boolean indicating whether person entities were detected in the article.
- **`"crawled_on"`**: The crawling date, saved as a string in the format `YYYY-MM-DD hh:mm`.
- **`"language"`**: The language of the article, always set to `"de"`.
- **`"type"`**: Mostly set to `"article"`.
- **`"author"`**: The person who wrote the article.
- **`"keywords"`**: Keywords relevant to the article, could be used for topic recognition.
- **`"token_count"`**: The number of tokens in the article.

The article-data consists of three parts, though not all are always available. However, every entry contains at least one `"text"` component:

- **`"title"`**: The headline of the article.
- **`"teaser"`**: A short description of the content or a short introduction.
- **`"text"`**: The main body of the article.

```
{
    ID: {
        "metadata": {
            "published_on": string,
            "contains_actors": boolean,
            "crawled_on": string,
            "language": "de",
            "type": string,
            "author": string,
            "keywords": string,
            "token_count": int
        },
        "text": {
            "title": string,
            "teaser": string,
            "text": string
        }
    }
}
```

Figure 1: Structure of the elements in the corpus, including all available metadata collected alongside the raw texts.

### 3.3 Dataset Statistics

The `taz2024full` corpus consists of 1,834,026 newspaper articles published in the German newspaper taz between 1980 and 2024 (cf. Figure 2). From 1980 onwards, there was a steady increase in the number of published articles, reaching a peak of 73,002 in 2004. An unexpected dip occurred in 1991, though the reason for this anomaly is unclear. After 1993, publication numbers stabilised again before declining after the peak in 2004. This decline may be linked to changes in the publishing strategy, potentially involving an increase in paid content from 2007 onwards. Only publicly available content is used for this corpus, thus leading to declining article numbers from 2007 onwards.

The corpus contains 6,944,197 unique tokens, as determined using the SoMaJo tokeniser (Proisl and Uhrig, 2016). Additionally, 83% of all articles mention specific individuals, allowing for an analysis of gender bias and discrimination in how different actors are represented in the article. Table 1 provides an overview of token, sentence, and article lengths. The maximum token and sentence lengths suggest that the tokeniser did not always ideally segment the articles. However, an average of four characters per token and 17 tokens per sentence appears reasonable.

[3] put URL here for publication

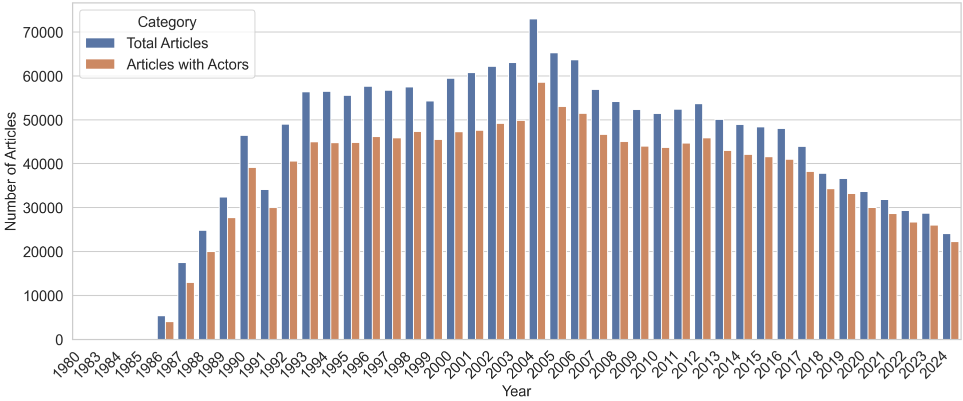


Figure 2: Number of newspaper articles and articles containing actors per year from 1980 to 2024. Years before 1986 include four articles or fewer.

| | min | max | mean | median |
|---|---|---|---|---|
| token length | 1 | 2,238 | 5.15 | 4 |
| sentence length | 1 | 18,872 | 20.07 | 17 |
| article length (token) | 3 | 26,855 | 396.89 | 276 |
| article length (sentences) | 1 | 1,027 | 19.77 | 13 |

Table 1: Statistics for the taz2024full corpus.

### 3.4 Dataset Applications

The taz2024full dataset offers numerous applications for researchers in NLP, CSS, and related fields. Its extensive temporal span supports investigations into the evolution of the German language, shifts in public discourse, and changes in thematic emphasis. For instance, it is possible to explore how language has adapted over decades, whether specific topics follow seasonal patterns, or how societal events have influenced reporting.

The dataset is particularly suited for examining patterns of bias and discrimination, mainly because it focuses on actors in texts. By leveraging this resource, insights can be obtained into representation trends and linguistic framing over time.

By providing this comprehensive dataset, we aim to support a wide range of research efforts, from linguistic studies to bias detection, thereby contributing to a deeper understanding of German media discourse.

## 4 Experiments

In our work, we build upon the pipeline introduced by Urchs et al. (2024) for detecting biases and discrimination on actor level. Rather than single English newspaper texts (as in Urchs et al. (2024)), we apply and scale an extended version of the pipeline to the German taz2024full corpus and compare discrimination markers in newspaper texts through more than four decades.

The full code is published on GitHub[4]. See also appendix A for a sample report of 2023.

### 4.1 Data

We analyse the entire corpus for discrimination markers to demonstrate its potential for bias and discrimination research.

The dataset encompasses all topics covered by "taz" without filtering, ensuring a comprehensive representation of the newspaper's reporting across politics, culture, economy, society, and more.

The chosen time-frame enables a meaningful comparison, capturing how taz's reporting on actors has evolved over 44 years. This period includes significant societal, cultural, and journalistic developments, which are likely to influence how gender bias and discrimination manifest in the texts.

### 4.2 Method

Urchs et al. (2024) automate linguistic discourse

[4] https://anonymous.4open.science/r/corpus_pipeline-1468

analysis by applying information extraction techniques to analyse English newspaper articles for discriminatory content.This approach focuses on two key aspects of linguistic discourse analysis: nomination, which examines how actors in a text are named, and predication, which analyses how they are described.

Their pipeline is designed to identify actors within the text by leveraging named entity recognition (NER) to detect person entities. The authors utilise spaCy[5] to extract all named entities from the text, retaining only those classified as persons. Additionally, generic terms referring to individuals, such as "mother," "father," "woman," and "man", are extracted. Entities that share the same name or name components (e. g. if only a first name is present or only a surname) are grouped together under a single actor. To further refine the analysis, coreferee[6] is employed for coreference resolution, linking pronouns to the corresponding actors. This approach enables text analysis based on distinct actors rather than relying solely on pronoun frequency.

The pipeline subsequently assumes the gender of the actors based on the extracted pronouns, categorising them as `woman`, `man`, `non-binary`, or `undefined`. If more than 70% of the pronouns associated with an actor are either feminine or masculine, the actor is categorised as `woman` or `man`, respectively. Actors are labelled `undefined` if no pronouns are present or the thresholds for other categories are not met.

Urchs et al. (2024) extract sentences mentioning the actors to analyse their predication. Therefore each sentence that either contains the name of the actor or a pronoun linked to the actor is extracted. This data is used to identify the following markers of discrimination across the gender categories:

- Number of actors in text per gender category
- Count of mentions per gender category / individual
- Sentiment towards gender category / individual[7]
- Count of feminine-coded words and masculine-coded words in the predication of each gender category/individual[8]
- Abusive words in predications (not used in this work)

These markers are then compiled into a discrimination report and visually represented for each text.

In this work, we adapt the existing pipeline to analyse German newspaper articles, expanding its scope from processing individual articles to handling thousands at a time. Our adaptations include the following:

- **Gender Assumption:** Instead of assuming the gender of actors, we determine the primary pronoun used to refer to them. Additionally, we scanned the corpus for German neo-pronouns[9] and found that only five texts contained such pronouns. Consequently, our analysis focuses on the pronouns she/her and he/him. For the results section, we categorise actors primarily referred to with she/her pronouns as women and those with he/him pronouns as men.

- **Pronoun Driven Analysis:** We include only actors for whom co-reference with pronouns could be established. All other actors are not part of the analysis.

- **Generic Masculine/German Gender-Neutral Language:** We introduce a marker to determine whether an article employs the generic masculine and one for the German gender-neutral language.

- **Pairwise Mutual Information (PMI):** Additionally, we identify the top 10 adjectives with the highest PMI in each actor's predication. PMI (cf. equation 1) measures the probability of two words $x$ and $y$ co-occurring by chance or meaningfully. A higher PMI indicates a more meaningful relation between these words (Jurafsky and Martin, 2000). In our pipeline, we calculate the PMI for each

[5] https://spacy.io/

[6] https://spacy.io/universe/project/coreferee

[7] The spaCy pipeline component spacytextblob (https://spacy.io/universe/project/spacy-textblob) is used for sentiment analysis. In our German version we use the pre-traind model "oliverguhr/german-sentiment-bert", which is trained on German newspaper articles, from huggingface (https://huggingface.co/transformers/v3.0.2/main_classes/pipelines.html)

[8] We utilise the lists of feminine- and masculine-coded words provided on the Gender Decoder website https://www.msl.mgt.tum.de/rm/third-party-funded-projects/projekt-fuehrmint/gender-decoder/wortlisten/, which are derived from the work of Gaucher et al. (Gaucher et al., 2011). Their research demonstrates that women are more frequently associated with communal traits, while men are linked to agency-related terms.

[9] We used the neo-pronouns listed at https://gleichstellung.tu-dortmund.de/projekte/klargestellt/neo-pronomen

actor and each adjective in their predication, excluding stop-words. Therefore, we can identify the most influential adjectives for each actor.

$$\mathsf{PMI}(x, y) = log_2 \frac{P(x, y)}{P(x)P(y)} \qquad (1)$$

- **Aggregated Report:** We introduce a human-readable report to facilitate easy corpus analysis.

We deliberately chose not to use large language models (LLMs) for the analysis, despite known challenges with pronoun detection and co-referencing. While LLMs could potentially enhance performance, we opted against them due to concerns about the biases they may introduce and the lack of transparency in their outputs. Instead, we relied on well-established, interpretable methods, acknowledging their limitations. This decision prioritises ethical considerations and ensures greater methodological control.

We conduct a yearly analysis of the whole corpus, examining changes and trends over the 44-year span of our corpus.

### 4.3 Results

The analysis of the corpus over the years provides valuable insights into how taz has written about women and men across the decades. Figure 3 illustrates the proportion of actors whose gender could be co-referenced with pronouns, distinguishing between women and men (Woman Actors/-Man Actors). Additionally, the figure presents the frequency with which these actors are mentioned within texts (Woman Mentions/Man Mentions)[10]. Before 1990, the data is too sparse to allow definitive interpretations. However, from the 1990s onward, a clear trend emerges: "taz" reported significantly more on men than on women. This imbalance is present across all decades, with man actors not only more frequently included in articles but also mentioned more often. While a shift towards greater inclusion of women actors becomes apparent from the 2010s onwards, this pattern of men actors being both more often the subject of reporting and more frequently referenced persists. Even in recent years, where gender representation appears almost balanced in terms of actor inclusion, men continue to receive more textual space, indicating a continued dominance in media visibility.

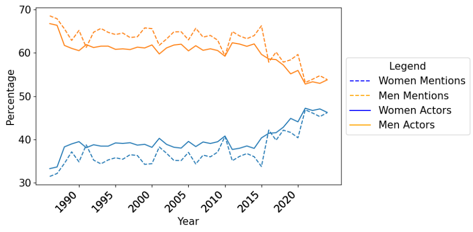


Figure 3: Comparison of the number of actors per article based on detected genders and the frequency of their mentions within each article.

Beyond the quantity of mentions, the sentiment towards men and women in "taz" articles is also revealing. Figure 4 illustrates the sentiment associated with women and men actors over time. Sentiment values range from -1 (highly negative) to +1 (highly positive), with 0 representing neutrality. The data shows that "taz" articles generally lean towards a neutral but slightly negative sentiment. More strikingly, across the entire 44-year period, sentiment towards women actors is consistently slightly more negative than sentiment towards men actors. While the differences are not extreme, this persistent pattern suggests that women in "taz" articles are, on average, portrayed in a slightly more negative light than their male counterparts.

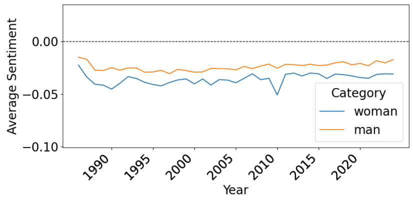


Figure 4: Sentiment towards the detected genders through the years.

Looking at language use more closely, an analysis of adjectives with the highest PMI association with women and men actors reveals that these descriptors remain relatively stable over time and do not exhibit strong gender differentiation. Additionally, a targeted analysis of female-coded and

[10] We differentiate between the number of actors in a text and the number of mentions to account for cases where, for instance, a single woman actor is referenced multiple times, whereas multiple male actors might be mentioned only once. Without this distinction, a text featuring one woman mentioned ten times and another featuring ten men mentioned once each would appear equivalent in terms of gender representation, despite their differing narrative emphases.

male-coded words shows that these were used only rarely. This suggests that "taz" makes little use of explicitly gender-coded language. However, our analysis found no evidence that "taz" systematically uses German gender-neutral language. While a few instances of gender-neutral forms were manually observed, these were rare exceptions rather than a common practice. Despite ongoing discussions about inclusive language in German, "taz" does not appear to have adopted gender-neutral writing conventions in its standard editorial style.

## 5 Conclusion and Future Work

In this work, we introduced `taz2024full`, a comprehensive German newspaper corpus spanning over four decades. This dataset represents a significant resource for linguistic, cultural, and societal research, particularly in the areas of gender bias and discrimination in media. By leveraging a structured approach to analysing gender representation through actor mentions and predications, we provided insights into how "taz" has reported on women and men over time. Our findings highlight a persistent imbalance in gender representation, with men not only appearing more frequently as actors but also receiving greater textual space.

Furthermore, we demonstrated the adaptability of an existing bias detection pipeline, originally designed for English texts, to large-scale German-language data. Our extensions included modifications tailored to the German linguistic landscape, such as pronoun-based gender identification and an analysis of the generic masculine. These enhancements offer a more refined approach to studying gender-related language use in German news media.

Looking ahead, several avenues for future research emerge. One priority is updating the corpus with new data beyond 2024 to enable ongoing diachronic analysis. Additionally, incorporating topic modelling could provide deeper insights into the contextual framing of gender representation. Since the current implementation would model topics at the sentence level, a necessary improvement would involve incorporating broader textual context to enhance topic coherence around actor predications.

Another promising direction is argument mining, which could refine our understanding of the implicit and explicit biases embedded in journalistic discourse. By identifying argument structures and rhetorical strategies, we could further uncover how gender bias manifests in media narratives.

Ultimately, we hope that `taz2024full` serves as a valuable resource for researchers in NLP, computational social science, and related fields, facilitating future studies on bias, representation, and media discourse in the German language.

The `taz2024full` corpus and its language-agnostic pipeline provide a foundation for analysing bias and discrimination in German news media. Future work will focus on expanding the corpus with newer data to ensure it remains relevant and reflective of contemporary discourse. Enhancements to the pipeline include training sentiment analysis models tailored to German newspapers, improving pronoun coreference resolution for German texts, and incorporating argument mining for discrimination detection. Additionally, we aim to extend the pipeline to support multimodal analysis by integrating text with non-textual data such as images and videos, enabling a comprehensive understanding of media content.

## Use of AI

The authors are not native English speakers; therefore, ChatGPT and Grammarly were used to assist with writing English in this work.

## Limitations

The `taz2024full` corpus has several limitations to consider when interpreting research results. Firstly, the dataset reflects the views of a Berlin-based, left-leaning publication and does not represent the entire spectrum of German discourse. This inherent bias limits its applicability for studying nationwide or ideologically diverse perspectives.

Additionally, bias and discrimination detection within corpora is inherently subjective, as no universally accepted gold standard exists. Different users may have varying values and interpretations of bias or discrimination, complicating evaluating such tasks. Therefore, we are not able to decide if a corpus is discriminatory or biased, as we do not know the use cases for each corpus. Additionally, our understanding of discrimination or bias might differ from the users understanding. Thus, we limit the output to a discrimination report, allowing each user to determine whether any adjustments to the corpus are necessary or to compare different corpora based on the calculated metrics.

Further challenges arise when applying the language-agnostic, flexible pipeline to detect gender discrimination and bias in texts. Co-reference resolution, particularly on German data, remains problematic due to the lower accuracy of current models for this language. This can affect the precision of gender-related analyses. Furthermore, the lack of directly comparable works or benchmarks complicates evaluating the pipeline's performance.

## Ethical Considerations

The `taz2024full` corpus is intended exclusively for academic research purposes, and exploiting it for commercial use would harm the publisher, taz Verlags und Vertriebs GmbH. To prevent such misuse, we strongly emphasise the importance of adhering to the dataset's intended purpose: fostering academic exploration and understanding.

Defining discrimination can be addressed abstractly, but implementing the concrete pipeline requires concrete values and definitions. Since we do not know the pipeline users' specific use cases and value systems, we opted for a flexible analysis of discrimination markers, highlighting potentially problematic content, thus refraining from definite judgments. We intentionally leave the final interpretation to human users. This allows them to apply their understanding of what constitutes discrimination in their specific use cases.

We acknowledge that our pipeline could be misused to curate datasets with specific biases or intentionally exclude particular genders. We strongly discourage using this system to manipulate datasets in ways that reinforce or amplify discrimination and biases. Instead, we aim to promote fairness and inclusivity by providing insights that help users curate discrimination and bias-free data sets.

By maintaining human oversight in the evaluation process, we aim to balance automated analysis with ethical responsibility, ensuring that the system supports diverse needs while promoting fairness in language technologies.

## Acknowledgements

This work was written by an author team working in different projects. Stefanie Urchs' project "Prof:inSicht" is promoted with funds from the Federal Ministry of Education and Research under the reference number 01FP21054. Matthias Aßenmacher is funded with funds from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of BERD@NFDI - grant number 460037581. Responsibility for the contents of this publication lies with the authors.

## References

Berlin-Brandenburgischen Akademie der Wissenschaften. Der deutsche wortschatz von 1600 bis heute. online.

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

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "gender" in nlp bias research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2083–2102, New York, NY, USA. Association for Computing Machinery.

Danielle Gaucher, Justin Friesen, and Aaron C Kay. 2011. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of personality and social psychology*, 101(1):109.

Anushka Gupta, Diksha Chugh, Anjum, and Rahul Katarya. 2022. Automated news summarization using transformers. In *Sustainable Advanced Computing*, pages 249–259, Singapore. Springer Singapore.

Daniel Jurafsky and James H. Martin. 2000. *Speech and language processing*. Prentice-Hall.

Marc Kupietz and Holger Keibel. 2009. The mannheim german reference corpus (dereko) as a basis for empirical linguistic research. *Working papers in corpus-based linguistics and language education*, 3:53–59.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339.

Library of Congress. Newspaper datasets and api access. online.

Asad Mahmood. 2017. News articles. online.

Camila M Mateo and David R Williams. 2020. More than words: a vision to address bias and reduce discrimination in the health professions learning environment. *Academic medicine*, 95(12S):S169–S177.

Rishabh Misra and Jigyasa Grover. 2021. *Sculpting Data for ML: The first act of Machine Learning*.

Andreas Nolda, Adrien Barbaresi, and Alexander Geyken. 2021. *Das ZDL-Regionalkorpus: Ein Korpus für die lexikografische Beschreibung der diatopischen Variation im Standarddeutschen*, pages 317–322. De Gruyter, Berlin, Boston.

Thomas Proisl and Peter Uhrig. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin. Association for Computational Linguistics.

Martin Reisigl. 2017. *Sprachwissenschaftliche Diskriminierungsforschung*, pages 81–100. Springer Fachmedien Wiesbaden, Wiesbaden.

Brants Sabine, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic Interpretation of a German Corpus. *Journal of Language and Computation*, 2:597–620.

Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One million posts: A data set of german online discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1241–1244, Tokyo, Japan.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Universität Tübingen Seminar für Sprachwissenschaft. Das korpus tüpp-d/z. online.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

taz. 2018. taz-geschäftsführer als zeitzeuge: Es begann in einem kalten ladenlokal | taz.de. online.

taz. 2024. Die seitenwende | taz.de. online.

Universität Leipzig, Sächsischen Akademie der Wissenschaften zu Leipzig, and Instituts für Angewandte Informatik. Willkommen beim wortschatz-portal. online.

Stefanie Urchs, Veronika Thurner, Matthias Aßenmacher, Christian Heumann, and Stephanie Thiemichen. 2024. Detecting gender discrimination on actor level using linguistic discourse analysis. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 140–149, Bangkok, Thailand. Association for Computational Linguistics.

Candace West and Don H. Zimmerman. 1987. Doing gender. *Gender & Society*, 1(2):125–151.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

## A Corpus Report 2023

```
Aggregated Report for 2023
============================================================

Total Texts: 26357
Total Actors: 17161
Pronoun Distribution: {'he_him': 9088, 'she_her': 8073}
Total Mentions: 109634
Mentions by Pronoun: {'he_him': 60008, 'she_her': 49626}

Mean Metrics:
  total_actors: 1.67
  total_mentions: 10.66
  total_feminine_coded_words: 0.27
  total_masculine_coded_words: 0.11
  contains_majority_gender_neutral: 0.00
  generic_masculine: 0.31
  pronoun_distribution_she_her: 0.79
  pronoun_distribution_he_him: 0.88
  mentions_pronoun_distribution_she_her: 4.83
  mentions_pronoun_distribution_he_him: 5.84
  feminine_coded_words_pronoun_distribution_she_her: 0.13
  feminine_coded_words_pronoun_distribution_he_him: 0.14
  masculine_coded_words_pronoun_distribution_she_her: 0.05
  masculine_coded_words_pronoun_distribution_he_him: 0.06
  average_sentiment_all: -0.03
  sentiment_by_pronoun_she_her: -0.03
  sentiment_by_pronoun_he_him: -0.02

Median Metrics:
  total_actors: 1.00
  total_mentions: 5.00
  total_feminine_coded_words: 0.00
  total_masculine_coded_words: 0.00
  contains_majority_gender_neutral: 0.00
  generic_masculine: 0.00
  pronoun_distribution_she_her: 1.00
  pronoun_distribution_he_him: 1.00
  mentions_pronoun_distribution_she_her: 2.00
  mentions_pronoun_distribution_he_him: 2.00
  feminine_coded_words_pronoun_distribution_she_her: 0.00
  feminine_coded_words_pronoun_distribution_he_him: 0.00
  masculine_coded_words_pronoun_distribution_she_her: 0.00
  masculine_coded_words_pronoun_distribution_he_him: 0.00
  average_sentiment_all: 0.00
  sentiment_by_pronoun_she_her: 0.00
  sentiment_by_pronoun_he_him: 0.00

Top PMI Adjectives Table:
All                  she/her              he/him
--------------------------------------------------------------------------------
letzten              junge                letzten
deutschen            letzten              russischen
junge                deutschen            deutschen
russischen           jungen               politische
berliner             deutsche             berliner
deutsche             berliner             politischen
politische           russischen           ukrainischen
politischen          nächsten             ukrainische
jungen               alten                russische
nächsten             alte                 deutsche
```

Listing 1: Full discrimination report of 2023.

# 8. Fair Play in the Newsroom: Actor-Based Filtering Gender Discrimination in Text Corpora

Chapter 8 presents a two-stage filtering and balancing framework for mitigating gender discrimination in text corpora. Building on the actor-based analysis from Chapter 7 and 6, the approach introduces exclusion criteria based on framing asymmetries and applies user-defined equilibrium thresholds to iteratively rebalance gender representation in the corpus. Applied to the `taz2024full` corpus, the method achieves near parity in actor and mention ratios while also improving discursive features such as syntactic roles and quotation style. The system remains user-controlled and non-normative, supporting flexible, context-sensitive corpus curation.

### Contributing article:

Urchs, S., Thurner, V., Aßenmacher, M., Heumann, C. and Thiemichen, S. (2025). Fair Play in the Newsroom: Actor-Based Filtering Gender Discrimination in Text Corpora. Accepted at the 5th Workshop on Evaluation  Comparison of NLP Systems (eval4NLP).

### Copyright information:

### Author contributions:

Stefanie Urchs was responsible for the study idea, conceptualisation, conducting the experiments, literature research and writing the manuscript. The other authors supported the work by providing technical guidance, conceptual discussions and checking the results.

### Supplementary material available at:

- Code and further information: https://github.com/Ognatai/corpus_balancing
- Metrics Handbook: Appendix A
- arXiv Link: https://arxiv.org/abs/2508.13169

# Fair Play in the Newsroom: Actor-Based Filtering Gender Discrimination in Text Corpora

Stefanie Urchs[1,2], Veronika Thurner[1], Matthias Aßenmacher[2,3], Christian Heumann[2], Stephanie Thiemichen[1]

[1]Faculty for Computer Science and Mathematics,
Hochschule München University of Applied Sciences, [2]Department of Statistics, LMU Munich,
[3]Munich Center for Machine Learning (MCML), LMU Munich
**Correspondence:** stefanie.urchs@hm.edu

## Abstract

Large language models are increasingly shaping digital communication, yet their outputs often reflect structural gender imbalances that originate from their training data. This paper presents an extended actor-level pipeline for detecting and mitigating gender discrimination in large-scale text corpora. Building on prior work in discourse-aware fairness analysis, we introduce new actor-level metrics that capture asymmetries in sentiment, syntactic agency, and quotation styles. The pipeline supports both diagnostic corpus analysis and exclusion-based balancing, enabling the construction of fairer corpora. We apply our approach to the `taz2024full` corpus of German newspaper articles from 1980 to 2024, demonstrating substantial improvements in gender balance across multiple linguistic dimensions. Our results show that while surface-level asymmetries can be mitigated through filtering and rebalancing, subtler forms of bias persist, particularly in sentiment and framing. We release the tools and reports to support further research in discourse-based fairness auditing and equitable corpus construction.

## 1 Introduction

Large Language Models (LLM) are increasingly integrated into everyday digital services, from search engines and translation tools to conversational agents and content recommendation systems. Despite their impressive capabilities, LLMs are known to perpetuate and even amplify harmful societal biases, including gender stereotypes, occupational hierarchies, and asymmetric visibility of social groups (Torrielli, 2025; Armstrong et al., 2024; Siddique et al., 2024; Jeoung et al., 2023). These patterns are not inherent to the models' architecture itself (Vaswani et al., 2017) but are learned from the data on which they are trained. As training data plays a central role in shaping the representational landscape of LLMs' outputs, critical analysis of large-scale text corpora becomes a key step towards mitigating downstream harms. Without an understanding of the social structures embedded in textual data, efforts to ensure fairness and accountability in LLMs risk addressing symptoms rather than causes.

Previous work by Urchs et al. (2025) introduced a linguistically grounded pipeline to detect gender discrimination in German newspaper texts through actor-level discourse analysis. Their approach identifies named actors and analyses how they are represented in terms of visibility and associated language, drawing on concepts such as nomination and predication from critical discourse studies.

Building on this foundation, we present an extended pipeline that enables both fine-grained fairness auditing and corpus-level discrimination reduction. Our contribution is fourfold:

1. We introduce novel actor-level discrimination markers, including syntactic roles, quote attribution, and sentiment bias.

2. We enhance interpretability through structured, human-readable reports that enable qualitative and diachronic analysis.

3. We propose a method for generating gender-balanced corpus recommendations by excluding disproportionately discriminatory texts.

4. We release the entire pipeline as open-source software to support transparency, reproducibility, and collaborative development.

This paper offers tools and insights for creating fairer NLP systems by revealing how social groups are represented in text. We combine discourse-informed fairness analysis with scalable corpus processing to enable actor-level discrimination detection and targeted corpus balancing.

## 2 Related Work and Conceptual Background

Detecting gender discrimination in text requires an interdisciplinary foundation that integrates perspectives from linguistics, gender studies, and computer science. In this section, we first outline our conceptual understanding of gender and linguistic discrimination. We then review existing approaches to computational discrimination detection. Finally, we introduce the actor-level discrimination detection pipeline by Urchs et al. (2025), which forms the foundation of our work.

### 2.1 Gender and Linguistic Discrimination

In this work, we adopt a differentiated understanding of gender and discrimination that draws from linguistic discourse analysis, gender studies, and computational fairness research.

**Gender** is treated as a socially constructed and co-constructed identity, rather than a fixed biological or grammatical category. While linguistic gender relates to grammatical rules and biological sex is often seen as binary and immutable, we work with the concept of *social gender*, which can be fluid, non-binary, and shaped through interaction and recognition. This perspective foregrounds the performative and contextual nature of gender, acknowledging its entanglement with social norms, symbolic power, and habitualised practices (West and Zimmerman, 1987; Ainsworth, 2015; Konishi, 1993; Kramer, 2020; Devinney et al., 2022). However, due to methodological constraints and the structure of the German language, the empirical analysis remains limited to the binary spectrum.

**Discrimination**, in contrast to bias or fairness, is understood here as a social effect: It is the observable outcome of differential treatment based on protected attributes such as gender. Drawing on Reisigl (2017) and the functional model by Graumann and Wintermantel (2007), we conceptualise linguistic discrimination as an act or process that disadvantages (or favours) individuals through patterns in language. This includes acts of naming (*nomination*) and describing (*predication*) actors in ways that reflect or reinforce social hierarchies. Such acts need not be intentional, but they may also arise from the habitual reproduction of dominant norms in public discourse. This framework stands in contrast to many machine learning–based approaches, where *bias* often refers to statistical imbalances in data or model performance, and *fairness* is operationalised via metrics such as demographic parity or equal opportunity (Blodgett et al., 2020; Caton and Haas, 2024). While these frameworks are useful for quantifying group disparities, they tend to reduce complex social identities to binary categories, thereby risking the oversight of contextual, structural, and discursive forms of inequality.

### 2.2 Computational Discrimination Detection

Discrimination in text is not limited to overtly offensive statements but often emerges from subtle and structural patterns of language use. Computational discrimination analysis aims to make such patterns visible by identifying systematic disparities in how individuals or groups are represented in text. In computer science, discrimination is often formalised via the concept of fairness and operationalised using statistical metrics such as demographic parity, equalised odds, or individual fairness (Mehrabi et al., 2021). These metrics are well-suited for classification tasks and support scalability and reproducibility. However, they also reduce social categories to fixed, binary attributes and abstract away from contextual and structural forms of inequality (Blodgett et al., 2020).

When applied to text, this paradigm produces approaches that focus primarily on hate speech detection, sentiment disparity, or stereotyping, often based on keyword lists or supervised classification. Although these methods yield valuable insights, they rarely address how discrimination is embedded in discourse or how it is distributed across different individuals within a text. Instead, most existing systems operate at a text-level granularity, assigning a global label such as "discriminatory" or "non-discriminatory" to entire documents.

In contrast, Urchs et al. (2025) focus on discrimination at the actor level. By identifying individual actors and examining how they are named (*nomination*) and described (*predication*), their approach reveals asymmetries in representation within the same text. This enables a more granular and discourse-aware analysis of discrimination in language.

### 2.3 Actor-Level Discrimination Detection Pipeline

Our pipeline builds upon prior work by Urchs et al. (2024, 2025). The first paper introduces actor-based fairness analysis in isolated texts using a modular pipeline that combines information extraction with linguistic discourse analysis. Specifically, it detects gender discrimination on the actor level by identifying the *nomination* and the *predication* of individual actors. The pipeline extracts actors via named entity recognition (NER), resolves pronouns through coreference resolution, and stores all actor references, including names, titles, and generic forms, in a structured knowledge base. For each actor, all sentences containing them are extracted and analysed for sentiment, gender-coded language, and linguistic framing. Discrimination metrics are computed per actor and summarised into a discrimination report. The report includes a range of linguistic and structural metrics designed to capture different facets of gendered representation:

- **Actor counts**: Number of distinct male-, female-, and undefined-coded actors per text.
- **Mention counts**: Total number of pronoun or name-based references per gender group.
- **Sentiment**: Average sentiment score of all predications linked to each actor or gender group.
- **Gender-coded language**: Count of feminine-coded and masculine-coded terms in predications, based on lexicons from Gaucher et al. (2011).

Their second paper (Urchs et al., 2025) scales this analysis to a large newspaper corpus (`taz2024full`) with over 1.8 million articles published between 1980 and 2024. It adapts the actor-level pipeline for German, replacing the English sentiment model with a BERT-based classifier trained on German news, and introduces additional markers for gender-neutral language and generic masculine usage. The analysis aggregates actor-based metrics by year, enabling the study of historical shifts in gender representation and framing.

Beyond the metrics introduced in the earlier paper, the `taz2024full` version adds:

- **Generic masculine detection**: Flags texts using the German generic masculine form.
- **Gender-neutral language detection**: Identifies inclusive writing styles such as gender colons or stars (e.g., *Lehrer:innen*).
- **PMI adjectives**: Extracts the ten adjectives with the highest Pointwise Mutual Information (PMI) per actor, providing insights into recurring descriptive patterns.
- **Yearly aggregation**: Metrics are aggregated per year to enable longitudinal analysis of shifts in gendered representation and framing.
- **Yearly report generation**: All extracted metrics are compiled into a structured, human-readable report for each year.

However, the approach remains purely descriptive and diagnostic: No mechanism is implemented for correcting or filtering the corpus based on the findings. Our work extends the work from both papers *substantially* in depth and scope. We expand the actor-level analysis with new discrimination detection metrics and integrate the pipeline into a two-stage exclusion framework.

## 3 The Extended Actor-Centred Pipeline

We extend the original actor-centred pipeline, improving both the granularity of the analysis and its ability to support fairer corpus construction.

Building on insights from systemic functional linguistics (Halliday, 2004) and critical discourse analysis (Reisigl, 2017), the pipeline now incorporates syntactic role annotation to capture how subject and object positions contribute to gendered representations. In discourse, actors in subject roles typically perform actions, while those in object positions are acted upon. Tracking these roles across pronoun groups helps reveal patterns of agency and passivity that are central to linguistic discrimination. To deepen the understanding of the gendered representation, we track whether actors are referred to by their name or just by their pronouns. We refined the quote attribution by distinguishing direct from indirect speech using punctuation and reporting verbs. Furthermore, zooming in on how the actors' utterances are presented gives insights into how active they are in the text. Direct quotes indicate a higher degree of activity than indirect ones. We enrich the framing context by extending pointwise mutual information (PMI) calculations to include not only adjectives but also verbs and nouns. This expansion highlights thematic associations and role-specific language tied to gendered actors. Human-readable reports are redesigned to improve structure, interpretability, and accessibility. All results are presented in a structured plain-text report, with separate sections for summary statistics, syntactic roles, sentiment, and lexical framing. Breakdowns on a pronoun-group basis and PMI tables are aligned and consistently formatted, making the output interpretable even for non-technical readers.

Finally, the pipeline supports a two-step corpus filtering mechanism: Articles exhibiting strong internal gender asymmetries are flagged using multiple indicators, and a subsequent balancing step adjusts overall gender ratios in the dataset. This process yields a curated corpus suitable for training language models with reduced gender discrimination.

The full pipeline code is available on our git repository `https://github.com/Ognatai/corpus_balancing`.

## 4 Pipeline Application: Discrimination Analysis and Corpus Balancing

This section outlines how we implemented the extended actor-level pipeline in two consecutive stages.

### 4.1 Stage 1: Discrimination Analysis Across the Full Corpus

For the initial discrimination analysis, we build on the pipeline proposed by Urchs et al. (2025) (cf. §2.3). We extend this approach by incorporating a set of additional linguistic and structural metrics that allow for a more fine-grained assessment of gendered representation and framing. These metrics are extracted at the actor level and aggregated per document and year:

- **Mentions**: Captures how often an actor is referred to by name (e.g., *Angela Merkel*) or pronoun (e.g., *she*, *he*), enabling analysis of individuation, visibility, and referential strategies.
- **Syntactic roles**: Counts how often the actor appears as grammatical subject (`nsubj`) or object (`obj`), providing a proxy for discursive agency and passivity.
- **Quotation style**: Differentiates between direct and indirect speech attributions, reflecting variation in narrative presence and framing control.
- **Top 10 PMI terms**: Lists the nouns and verbs most strongly associated with the actor based on Pointwise Mutual Information (PMI), offering insight into typical roles, actions, and semantic contexts.

The full `taz2024full` corpus is analysed both at the yearly level and aggregated in total. We also introduce an improved yearly reporting format that organises both existing and newly introduced metrics in a more accessible structure. An example of this yearly report (for 2023) can be found in Appendix A.

### 4.2 Stage 2: Article Filtering and Corpus Balancing

To reach the endeavour of constructing a more balanced corpus and substantially reducing the impact of gender-discriminating articles, we introduce a multi-stage filtering pipeline. In contrast to stage 1, which analyses the corpus on a yearly basis, this stage operates on the entire dataset to enable global exclusion and balancing decisions.

**Step 1: Full Corpus Analysis** The full corpus is processed using the pipeline from Stage 1 (cf. §2.3). Instead of aggregating metrics per year, actor-level knowledge bases and article-level metadata are saved for each article. These intermediate files are reused in the subsequent steps to minimise redundant executions of pipeline steps. After processing, a histogram is generated to visualise the distribution of gender ratios across articles (cf. Figure 5). Each article is assigned two values: the percentage of mentions and the percentage of actors associated with she/her pronouns. The left subplot displays the mention-weighted distribution, and the right subplot shows the actor-weighted distribution. Both histograms use percentage values ranging from 0 (only he/him references) to 100 (only she/her references), with each bar representing the proportion of articles falling into that range.

**Step 2: Text-Level Exclusion** An interactive filtering interface allows the user to define how many of the four heuristic exclusion criteria (see list below) must be met to classify a text as discriminatory. The default threshold is two out of four. The criteria are:

- **Sentiment disparity:** A large gap in average sentiment scores between female- and male-coded actors.
- **Grammatical role asymmetry:** A strong difference in subject-to-object ratios between pronoun groups.

- **Quote attribution imbalance:** A pronounced disparity in direct versus indirect quotes.
- **Naming versus pronoun imbalance:** Disproportionate usage of named versus pronominal references between genders.

Each criterion is calculated using Laplace-smoothed ratios to avoid instability possibly originating from low-frequency counts. The user may customise the threshold for each criterion at runtime. By default, a text exhibits significant imbalance if the sentiment gap exceeds 0.3, or if the difference in subject/object roles, quote attribution, or naming/pronoun usage exceeds 0.5. These thresholds correspond to cases where one gender group is at least twice as prominent as the other in a specific framing dimension. A lower threshold is used for sentiment because sentiment values tend to cluster around neutral and vary within a narrower range; even small differences may indicate meaningful affective bias. Overall, the chosen defaults strike a balance between interpretability and selectivity, capturing strongly biased texts while preserving as much of the corpus as possible. Articles that trigger the user-defined number of flags are excluded from the corpus. Exclusion decisions are logged in a structured exclusion file and visualised through an updated gender ratio distribution plot (cf. Figure 6).

**Step 3: Corpus-Level Balancing** After text-level filtering, we apply corpus-level balancing based on the relative contribution of each article to the overall gender ratio. The user sets an equilibrium range for actor and mention ratios between female- and male-coded references. The default range is $[0.75, 1.25]$, allowing for up to 25% deviation in either direction. Articles are iteratively excluded based on their contribution to the global imbalance until both ratios fall within the specified interval. A visualisation of the resulting gender ratio distribution is generated to illustrate the effects of balancing (cf. Figure 7.) All excluded articles are recorded in a structured exclusion file for transparency and reproducibility.

**Step 4: Corpus Reconstruction** All excluded article IDs from both filtering stages are consolidated and used to construct a new balanced corpus. Articles are removed directly from the original JSON files, and the revised dataset is saved to disk.

## 5 Corpus-Balancing of `taz2024full`

We use the `taz2024full` corpus (Urchs et al., 2025), comprising over 1.8 million articles from the German left-leaning newspaper *taz* (1980–2024), previously used to analyse gender representation. In the unfiltered `taz2024full` corpus, we detect female- and male-coded actors in **1,834,018 articles**. Gender representation is clearly imbalanced: Men dominate across both mention frequency and actor counts (Figure 1). These differences are not only quantitative but also reflected in patterns of agency and authority.

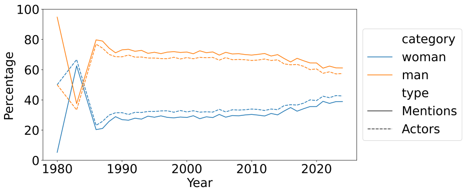


Figure 1: Percentage of male- and female-coded references over time *before filtering*. Fluctuations in the early years are due to the small number of articles available from the 1980s, which can lead to disproportionate weight for single-gender articles.

Figure 2 shows that male actors are more frequently quoted directly, while female actors appear more often in indirect quotes. Direct quotations often attribute more authority to the speaker and allow them to appear in active, public-facing roles. Indirect quotes, by contrast, reduce visibility and typically appear in paraphrased or backgrounded contexts.

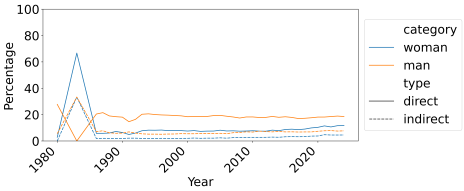


Figure 2: Distribution of quotation styles by gender *before filtering*. Early-year fluctuations are again attributable to low article counts.

Figure 3 further underlines this pattern. Men occur more frequently in subject positions, while women appear comparatively more often as objects. Subject roles in grammar typically denote agency and narrative control, whereas object positions signal reduced agency and passivity within the sentence structure.

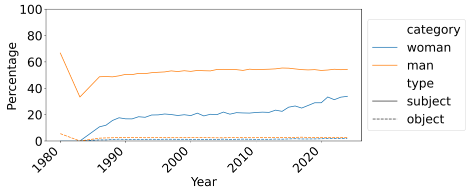


Figure 3: Distribution of syntactic roles by gender *before filtering*. Early-year fluctuations are again attributable to low article counts.

Sentiment analysis reveals a consistent gap: The average sentiment towards female-coded actors is more negative in almost all years (Figure 4). While sentiment values are generally close to neutral, the persistent divergence reinforces the broader imbalance.

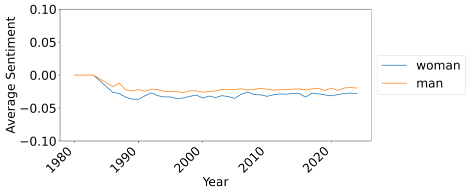


Figure 4: Average sentiment associated with male- and female-coded actors *before filtering*.

The early-year spikes in Figures 1, 2, and 3 are artefacts of low data density. In years with very few articles or actor mentions, a single gender may dominate, resulting in sharp fluctuations that do not reflect structural bias but instead data sparsity. We retain these years for completeness, but advise caution when interpreting early data points and refrain from doing this ourselves.

Finally, the overall distribution of gender representation per article (Figure 5) is highly polarised: many articles reference either only male-coded or only female-coded actors. This reinforces the need for corpus-level balancing, as it shows that article-level imbalance is not just a matter of aggregate statistics but of individual article composition.

During the first text-level filtering step, we exclude **279,772 articles** based on four framing asymmetries: sentiment gap, quote imbalance, subject/object ratio, and representation. We use the default values described in Section 4.2. The updated diagnostic view shows reduced polarisation, and the gender ratio distribution shifts closer to balance (Figure 6).

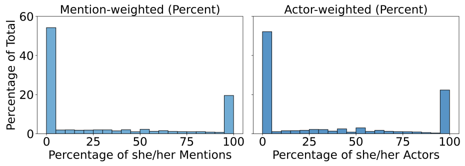


Figure 5: Distribution of gender ratios across articles *before filtering*.

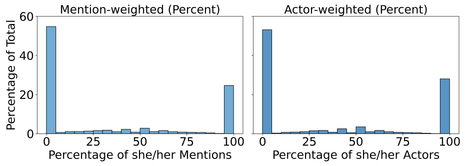


Figure 6: Distribution of gender ratios across articles *after text-level filtering*.

Subsequently, we apply corpus-level balancing, excluding an additional **17,815 articles** to bring the overall actor and mention ratios into the (default) target range $[0.75, 1.25]$. Compared to the filtered corpus, the distribution is more centred and less polarised, indicating that articles with extreme gender dominance were successfully downsampled to achieve a more balanced overall representation. (Figure 7).

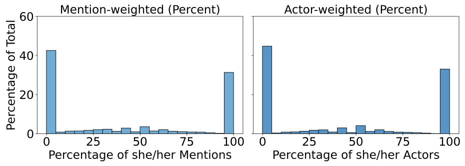


Figure 7: Distribution of gender ratios across articles *after corpus-level balancing*.

In the final corpus, gender representation is nearly balanced across both mentions and actor counts (Figure 8). This balance is not just a numerical artefact of article exclusion, but reflects a more even distribution across time and actor types. Compared to the original corpus (cf. Figure 1), the trajectories of female- and male-coded representation converge, with both actor and mention lines approaching parity. Importantly, the gender crossing point around 2018 persists, indicating that key corpus dynamics remain intact after balancing.

Figure 9 shows that the quotation gap is reduced compared to Figure 2. Women now appear in direct speech more frequently than before, while men re-

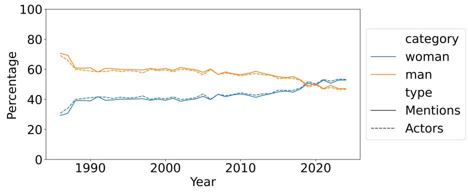


Figure 8: Percentage of male- and female-coded references over time *after all filtering and balancing steps*.

main slightly more dominant in indirect speech. The remaining asymmetry may reflect residual stylistic preferences in journalistic practice or shifts in topic domains rather than systematic bias. Overall, this pattern reflects a notable improvement in discursive agency, with women quoted more often in their own words.

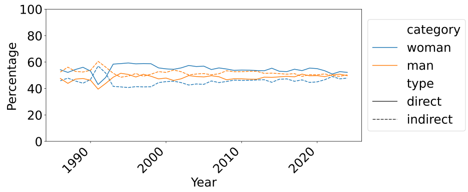


Figure 9: Proportion of direct and indirect quotations by gender *after full exclusion*.

Sentiment remains slightly more negative for women (cf. Figure 10), but the gap remains below the exclusion threshold of 0.3. This persistence may signal that subtle evaluative framing is harder to eliminate via structural balancing. It also suggests that sentiment operates on a different level than grammatical or referential bias, possibly rooted more deeply in lexical associations or broader discourse conventions.

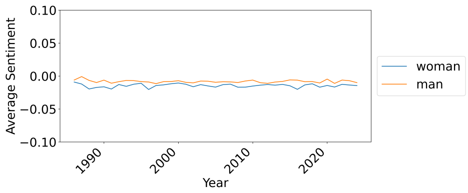


Figure 10: Average sentiment towards *she/her* and *he/him* actors over time *after full exclusion*.

Figure 11 reveals a modest shift in syntactic agency. The proportion of women in subject positions increases slightly, while object roles are more evenly distributed. Although men still dominate in subject roles, the difference is reduced compared to the original corpus. Specifically, the subject-role gap between men and women decreases from approximately 30 to 5 percentage points after filtering. This suggests that our multi-step exclusion process leads to subtle improvements in how grammatical agency is distributed between genders.

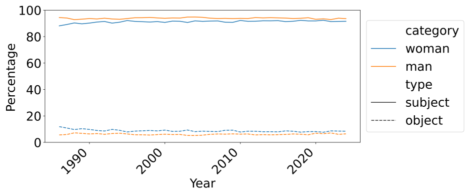


Figure 11: Distribution of syntactic roles *after full exclusion*.

Taken together, these results demonstrate the effectiveness of our multi-stage filtering and balancing strategy in mitigating structural gender asymmetries in the corpus. By comparing before and after exclusion, we observe clear improvements across key dimensions of representation and framing.

Referential parity is achieved: male- and female-coded actors now appear in comparable proportions across both mentions and actor counts. Discursive patterns also show notable shifts. Women are quoted directly more often, and syntactic agency is redistributed more evenly, with a marked reduction in the subject-object gap. These improvements suggest that our approach meaningfully alters how gendered actors are positioned within the narrative structure of the corpus.

However, some subtle forms of bias remain. The sentiment gap persists below the exclusion threshold, indicating that evaluative language is harder to correct through structural rebalancing alone. Such residual asymmetries may reflect deeper, more diffuse biases embedded in lexical or thematic choices rather than in sentence-level grammar.

Overall, the final corpus provides a significantly more equitable foundation for downstream NLP applications and for critical media analysis. By combining large-scale actor-level auditing with targeted corpus interventions, our approach offers a concrete path towards fairer data curation in practice.

## 6 Conclusion and Future Work

In this paper, we presented an extended actor-level pipeline for detecting and mitigating gender discrimination in large-scale text corpora. Building on prior work, we introduced new metrics that capture asymmetries in syntactic roles, quote attribution, and sentiment framing. We enhanced the interpretability of the results through structured reports and implemented a two-stage filtering mechanism that enables the construction of gender-balanced corpora.

Our application of the pipeline to the `taz2024full` corpus demonstrates that gender imbalances in representation and framing are both measurable and correctable to a significant extent. The resulting corpus shows improved balance across multiple linguistic dimensions and serves as a more equitable foundation for downstream tasks such as language model training or critical media studies.

Nevertheless, some forms of discrimination, particularly those tied to sentiment and more implicit discourse structures, persist despite structural balancing. This indicates that not all bias can be addressed through surface-level interventions alone. Future work should therefore explore complementary strategies such as employing context-aware language models for deeper semantic analysis, developing targeted debiasing methods to address persistent framing asymmetries, and integrating intersectional attributes such as race, age, or class. In addition, extending actor categories beyond the gender binary would enable the inclusion of non-binary and gender-diverse identities, allowing for a more comprehensive understanding of representational fairness.

More broadly, we advocate for the integration of discourse-aware methods into standard corpus construction workflows. Understanding how social groups are framed in language is a necessary prerequisite for designing fairer NLP systems, and our pipeline offers a scalable, modular, and linguistically grounded way to do so.

## Use of AI

The authors are not native English speakers; therefore, ChatGPT and Grammarly were used to assist with writing English in this work. ChatGPT was also used to assist with coding.

## Limitations

While our approach enables corpus-level balancing based on measurable framing asymmetries, it is not without limitations. First, the exclusion-based strategy necessarily reduces corpus size and diversity, potentially eliminating valuable content alongside discriminatory texts. Second, the method operates on surface-level linguistic signals and cannot fully account for subtler or context-dependent forms of bias, such as irony, framing through omission, or topic selection. Third, the balancing relies on binary gender classification, which excludes non-binary identities and reinforces a gender dichotomy that our conceptual framework otherwise seeks to challenge. Fourth, the analysis is limited to texts that contain clearly identifiable actors and sufficient gender cues, primarily via pronoun resolution. Articles without identifiable pronouns or mentions of coreferent actors are excluded from the discrimination analysis altogether, leading to incomplete coverage. Finally, the impact-aware exclusion method is sensitive to threshold settings and metric selection, which may affect outcomes in ways that are not always transparent. These limitations highlight the need for complementary strategies, such as counterfactual augmentation, contextual bias detection, or narrative-level analysis, to address more complex and nuanced forms of representational inequality.

## Ethical Considerations

Our work is grounded in the belief that fairness in NLP requires not only technical interventions but also critical reflection on the social impact of language technologies. By analysing how gendered actors are represented and framed in text, we aim to make structural inequalities visible and address them at the level of data design. At the same time, we recognise that fairness cannot be reduced to numerical balance. The act of filtering texts, however principled, entails normative decisions about which content is considered discriminatory and which is preserved. This introduces risks of over-correction, loss of valuable context, and the potential erasure of complex identities. Furthermore, our reliance on binary gender resolution excludes non-binary, intersex, and gender-nonconforming individuals, reinforcing the very simplifications we seek to critique. We consider this a significant ethical limitation and prioritise extending our methods to support more inclusive representations in future work. Finally, while we aim to mitigate bias in training data, we stress that ethical responsibility must also extend to model architectures, deployment contexts, and the broader socio-technical systems in which NLP tools are embedded.

## Acknowledgments

Matthias Aßenmacher is funded with funds from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of BERD@NFDI - grant number 460037581. Responsibility for the contents of this publication lies with the authors.

## References

Claire Ainsworth. 2015. Sex redefined. *Nature*, 518(7539):288–291.

Lena Armstrong, Abbey Liu, Stephen MacNeil, and Danaë Metaxa. 2024. The silicon ceiling: Auditing gpt's race and gender biases in hiring. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '24, New York, NY, USA. Association for Computing Machinery.

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

Simon Caton and Christian Haas. 2024. Fairness in Machine Learning: A Survey. *ACM Comput. Surv.*, 56(7):166:1–166:38.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "gender" in nlp bias research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2083–2102, New York, NY, USA. Association for Computing Machinery.

Danielle Gaucher, Justin Friesen, and Aaron C Kay. 2011. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of personality and social psychology*, 101(1):109.

Carl-Friedrich Graumann and Margret Wintermantel. 2007. *Diskriminierende Sprechakte. Ein funktionaler Ansatz*, pages 147–178. transcript Verlag, Bielefeld.

Michael Halliday. 2004. *An introduction to functional grammar*, 3 edition. Hodder Arnold, London, England.

Sullam Jeoung, Yubin Ge, and Jana Diesner. 2023. Stereomap: Quantifying the awareness of human-like stereotypes in large language models. page 12236 – 12256.

Toshi Konishi. 1993. The semantics of grammatical gender: A cross-cultural study. *Journal of Psycholinguistic Research*, 22(5):519–534.

Ruth Kramer. 2020. Grammatical gender: A close look at gender assignment across languages. *Annual Review of Linguistics*, 6(1):45–66.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Martin Reisigl. 2017. *Sprachwissenschaftliche Diskriminierungsforschung*, pages 81–100. Springer Fachmedien Wiesbaden, Wiesbaden.

Zara Siddique, Liam Turner, and Luis Espinosa-Anke. 2024. Who is better at math, jenny or jingzhen? uncovering stereotypes in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18601–18619, Miami, Florida, USA. Association for Computational Linguistics.

Federico Torrielli. 2025. Stars, stripes, and silicon: Unravelling the chatgpt's all-american, monochrome, cis-centric bias. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 283–292, Cham. Springer Nature Switzerland.

Stefanie Urchs, Veronika Thurner, Matthias Aßenmacher, Christian Heumann, and Stephanie Thiemichen. 2024. Detecting gender discrimination on actor level using linguistic discourse analysis. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 140–149, Bangkok, Thailand. Association for Computational Linguistics.

Stefanie Urchs, Veronika Thurner, Matthias Aßenmacher, Christian Heumann, and Stephanie Thiemichen. 2025. taz2024full: Analysing german newspapers for gender bias and discrimination across decades. *Preprint*, arXiv:2506.05388.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Candace West and Don H. Zimmerman. 1987. Doing gender. *Gender & Society*, 1(2):125–151.

# Appendix

## A Corpus Report 2023

```
Report for the year 2023
===========================================================================

AGGREGATED TOTALS (all texts)
Total Texts:                        10019
Texts with Actors:                  10019
Uses Gender Neutral Language (Docs):  107
Generic Masculine Usage (Docs):      8081

Metric                                  she/her      he/him     overall
---------------------------------------------------------------------------
Pronoun Distribution:                      6892        9194       16086
Mentions by Pronoun:                      35595       56044       91639
Named Mentions:                           22544       36047       58591
Pronoun Mentions:                         13051       19997       33048
Subject Roles:                            18625       30303       48928
Object Roles:                              1119        1540        2659
Direct Quotes:                             6501       10588       17089
Indirect Quotes:                           2529        4215        6744
Feminine-coded Words:                      4251        6066       10317
Masculine-coded Words:                     2870        4764        7634
Sentiment:                                -0.01       -0.01       -0.01
Named Mentions (% of all mentions):        38.5        61.5
Pronoun Mentions (% of all mentions):        39.5        60.5
Subject Roles (% of known roles):          38.1        61.9
Object Roles (% of known roles):           42.1        57.9
Direct Quotes (% of quotes):               38.0        62.0
Indirect Quotes (% of quotes):             37.5        62.5

STATISTICS (per text)
---------------------------------------------------------------------------
Metric                                             Mean    Median   Std Dev
---------------------------------------------------------------------------
Pronouns (Resolved) (She/Her)                      0.69      1.00      0.83
Mentions (By Pronoun) (She/Her)                    3.55      2.00      7.22
Feminine Coded Words (By Pronoun) (She/Her)        0.42      0.00      1.18
Masculine Coded Words (By Pronoun) (She/Her)       0.29      0.00      0.77
Named Mentions (Sum Over Actors) (She/Her)         2.25      1.00      5.65
Pronoun Mentions (Sum Over Actors) (She/Her)       1.30      1.00      2.29
Subject Roles (She/Her)                            1.86      0.00      3.78
Object Roles (She/Her)                             0.11      0.00      0.43
Direct Quotes (She/Her)                            0.65      0.00      1.46
Indirect Quotes (She/Her)                          0.25      0.00      0.72
Pronouns (Resolved) (He/Him)                       0.92      1.00      0.91
Mentions (By Pronoun) (He/Him)                     5.59      3.00      9.77
Feminine Coded Words (By Pronoun) (He/Him)         0.61      0.00      1.36
Masculine Coded Words (By Pronoun) (He/Him)        0.48      0.00      1.05
Named Mentions (Sum Over Actors) (He/Him)          3.60      1.00      7.75
Pronoun Mentions (Sum Over Actors) (He/Him)        2.00      1.00      3.02
Subject Roles (He/Him)                             3.02      2.00      5.01
Object Roles (He/Him)                              0.15      0.00      0.53
Direct Quotes (He/Him)                             1.06      0.00      1.96
Indirect Quotes (He/Him)                           0.42      0.00      0.97
Mean Sentiment (All)                              -0.02      0.00      0.10
Total Actors                                       1.61      1.00      1.04
Total Mentions                                     9.15      5.00     12.10
Total Feminine Coded Words                         1.03      0.00      1.83
Total Masculine Coded Words                        0.76      0.00      1.28
Uses Gender-Neutral Language                       0.01      0.00      0.10
Generic Masculine                                  0.81      1.00      0.40
```

```
        TOP PMI ADJECTIVES
---------------------------------------------------------------------------------------

Most frequent adjectives associated with each pronoun group.

Rank ALL                           she/her                      he/him
---------------------------------------------------------------------------------------

1    letzten (414.00)              letzten (154.00)             letzten (269.00)
2    russischen (272.00)           junge (130.00)               russischen (195.00)
3    deutschen (260.00)            berliner (101.00)            deutschen (171.00)
4    berliner (231.00)             deutschen (97.00)            politische (142.00)
5    junge (212.00)                deutsche (97.00)             ukrainische (137.00)
6    nächsten (212.00)             russischen (81.00)           politischen (135.00)
7    politische (212.00)           nächsten (80.00)             berliner (134.00)
8    deutsche (208.00)             politischen (80.00)          nächsten (133.00)
9    politischen (205.00)          politische (74.00)           ukrainischen(117.00)
10   ukrainische (178.00)          jungen (71.00)               russische (113.00)

TOP PMI NOUNS
---------------------------------------------------------------------------------------

Most frequent nouns associated with each pronoun group.

Rank ALL                           she/her                       he/him
---------------------------------------------------------------------------------------

1    menschen (588.00)             menschen (311.00)             menschen (315.00)
2    frau (353.00)                 frau (234.00)                 präsident (289.00)
3    präsident (328.00)            frauen (163.00)               mann (210.00)
4    leben (312.00)                leben (140.00)                partei (185.00)
5    mann (280.00)                 mutter (128.00)               leben (182.00)
6    partei (268.00)               kinder (109.00)               land (164.00)
7    land (238.00)                 tochter (107.00)              frau (147.00)
8    frauen (210.00)               geschichte (101.00)           sohn (135.00)
9    stadt (209.00)                mann (100.00)                 stadt (135.00)
10   regierung (208.00)            anfang (100.00)               mittwoch (126.00)

TOP PMI VERBS
---------------------------------------------------------------------------------------

Most frequent verbs associated with each pronoun group.

Rank ALL                           she/her                       he/him
---------------------------------------------------------------------------------------

1    erzählt (671.00)              erzählt (331.00)              erzählt (368.00)
2    steht (495.00)                steht (199.00)                steht (324.00)
3    sieht (449.00)                erklärt (180.00)              sieht (315.00)
4    erklärt (428.00)              lassen (167.00)               erklärt (269.00)
5    lassen (359.00)               sieht (163.00)                erklärte (243.00)
6    erklärte (346.00)             sehen (147.00)                spricht (228.00)
7    spricht (341.00)              zeigt (139.00)                lassen (205.00)
8    zeigt (302.00)                spricht (139.00)              sprach (199.00)
9    weiß (289.00)                 lebt (127.00)                 zeigt (190.00)
10   hält (286.00)                 sagen (125.00)                weiß (188.00)
```

# Part V.

# Conclusion

# 9. Conclusion

In times when political forces seek to undo social progress, research must take a stand for fairness, accountability, and inclusion. This thesis is part of that effort through five publications. First, it examines the extent of the problem by defining algorithmic gender fairness and analysing search and information retrieval results through this lens. Second, the focus shifts to large language models (LLM), using ChatGPT as an example to analyse whether there is gender discrimination in the system. The analysis shows that debiasing after model training is insufficient. To address this, a flexible, language-agnostic pipeline is introduced to analyse text corpora for discrimination markers. This pipeline can be used to examine training data for LLMs, helping to identify and mitigate harmful patterns before they are learned by the model. In doing so, it supports the development of LLMs that are less discriminatory, and ultimately fairer. The pipeline is applied to both English and German text. In addition, the thesis presents `taz2024full`, the largest publicly available corpus of German newspaper texts up to the time of publication. The pipeline is finally used to create a gender-balanced corpus variant of the `taz2024full` corpus.

## 9.1. Future Directions

Future work could focus primarily on refining and extending the pipeline developed in this thesis. The various studies and analyses presented throughout the preceding chapters were not ends in themselves, but rather stepping stones that highlighted the conceptual and practical need for such a pipeline.

**Pipeline Problems.** The pipeline uses `coreferee`[1] for coreference resolution due to its integration with `spaCy`[2], but performance is limited for non-binary pronouns in English and for German pronoun resolution. During computation, many `spaCy` word vectors were empty, likely due to rare words, inflected forms, or out-of-vocabulary tokens in the pre-trained model. These issues could be addressed by testing alternative coreference tools and integrating larger or domain-specific embeddings or subword-based models (e.g.fastText or transformer-based embeddings). The pipeline also extracts generic terms referring to actors but does not yet link them to specific nominations. Analysing whether such terms co-occur with names or appear alone could reveal patterns of representation. Finally, actor co-referencing currently groups all mentions sharing the same first and last name, and standalone occurrences of either, which can lead to incorrect groupings when multiple individuals share a name. More robust methods could incorporate contextual cues to distinguish between them.

[1] https://spacy.io/universe/project/coreferee

[2] https://spacy.io/

**Discrimination Markers.** It would be valuable to analyse how occupational roles are attributed to different genders, offering further insights into potential patterns of discrimination.

The discrimination markers currently integrated into the pipeline are relatively simple. Future work could incorporate more sophisticated measures, such as:

- Hate speech detection (Xu and Zhu, 2010; Paz et al., 2020; Fortuna and Nunes, 2018)
- Ambivalent sexism analysis (Jha and Mamidi, 2017)
- Microaggression detection (Breitfeller et al., 2019; Kaskan and Ho, 2016; Buchanan, 2011)
- Detection of condescending language (Wang and Potts, 2019)
- Stereotype identification (Zhao et al., 2017; Fast et al., 2016; Joseph et al., 2017)

Future work could also investigate linguistic patterns by gender, given the documented differences in male and female language use (Lakoff, 1973), and potentially infer the gender of authors based on writing style. Moreover, naming conventions offer further analytical opportunities (Bendel Larcher, 2015): Are first names used more frequently for one gender than another? When titles are present, are they applied equally to individuals of all genders? Finally, incorporating argument mining could help identify recurring argument structures used in discriminatory discourse (Reisigl, 2017), enriching the discrimination report with a structural analysis of how arguments are framed.

**Binary Gender in German.** A key area for future work is the inclusion of non-binary gender identities in the analysis of German language text. This requires identifying and working with texts that contain a sufficient representation of non-binary individuals. Ensuring that training data and analysis tools can accurately represent all genders is essential for fairness in language technologies.

**Exploring LLM as Pipeline Components.** The current implementation deliberately avoids using LLM in the pipeline to prevent analytical circularity (i.e., using a model to analyse data potentially generated or shaped by that same class of model). Nonetheless, given recent advancements in LLM capabilities, it may be worthwhile to reevaluate their potential role in specific pipeline components. Any such inclusion would require strict supervision, clearly defined scopes, and rigorous validation of component outputs.

**Pipeline User Interface and Adaptation to General Corpora.** The current implementation of the pipeline relies primarily on command-line-based inputs and, in some cases, outputs. Data loading is tightly coupled with the specific structure of the `taz2024full` corpus. In future work, a Graphical User Interface (GUI) should be introduced to allow users to select a dataset more easily. Corpus-specific requirements should be handled in the background to ensure compatibility with the pipeline. An intuitive GUI would broaden accessibility, especially for researchers from the social sciences and adjacent disciplines.

## 9.2. Closing Remarks

This thesis takes a stand for fairness, accountability, and inclusion in a field that shapes how people interact with information and each other. Detecting gender discrimination in training data contributes to developing large language models (LLMs) that do not simply reproduce harmful patterns, but instead support models that align better with democratic values. However, technical contributions alone are insufficient. It is equally important to be transparent about the limitations of this work and to reflect critically on the social and political implications of researching discrimination in data. The pipeline developed in this thesis detects discrimination through patterns of nomination and predication, that is, how actors are named and what is said about them. It relies on the explicit mention of actors in text, which limits its applicability in cases of implicit or structural discrimination that are not tied to named individuals. Very short texts may also lack sufficient linguistic material for reliable analysis. Despite these limitations, the approach provides a clear and scalable framework for systematically detecting representational asymmetries across large datasets. This level of clarity and transparency is a major strength, particularly in large-scale corpora, where manual analysis is infeasible. By formalising discrimination analysis in a reproducible and extensible way, the pipeline makes a valuable contribution to both NLP research and the social sciences. It enables new forms of interdisciplinary inquiry that would not be possible without automation. In doing so, it offers not only diagnostic insights into biased data but also a foundation for developing fairer language technologies.

Working on discrimination always involves normative decisions. As a Western European woman, I bring a particular perspective to the question of what counts as discriminatory. That perspective is shaped by my social and cultural background, and it is not universal. What one society may consider problematic, another may not. This raises important questions about whose values shape the tools we build. In this thesis, discrimination is not defined as an absolute. Instead of labelling texts or corpora as discriminatory, the pipeline generates discrimination reports that provide transparency without enforcing judgment. Even the balanced corpus in the last publication is merely a suggestion to the user. This reflects a conscious decision to leave the interpretation to the user. Rather than claiming to speak for the world at large, the approach supports critical engagement and allows different actors to assess for themselves whether a dataset aligns with the values they are willing to encode. At the same time, this openness carries risks. Discrimination detection tools can be misused to suggest that a dataset is “clean” or “neutral” when in fact it may simply fall outside the scope of detection. There is also the danger that such tools become part of a checklist mentality, deployed to signal fairness without a deeper commitment to ethical reflection or change. This work cannot prevent such uses, but it can acknowledge them. Ethical research must consider not only what a method makes possible but also how it might be repurposed in contexts beyond the researcher’s control.

Discrimination in data is not a problem that can be solved once and for all. It is a moving target, shaped by shifting social norms, historical inequalities, and global power dynamics. However, that does not mean we are powerless. With the right tools, we can see more clearly what is often hidden. We can make more informed decisions and question the assumptions that shape our digital world. This thesis is one contribution toward that effort. It does not offer certainty, but it does offer clarity. And sometimes, clarity is the beginning of change.

# Contributing Publications

Urchs, S., Thurner, V., Aßenmacher, M., Bothmann, L., Heumann, C. and Thiemichen, S.(2025). Are All Genders Equal in the Eyes of Algorithms? - Analysing Search and Retrieval Algorithms for Algorithmic Gender Fairness. Accepted at the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KDIR)

Urchs, S., Thurner, V., Aßenmacher, M., Heumann, C. and Thiemichen, S. (2023). How Prevalent Is Gender Bias in ChatGPT? - Exploring German and English ChatGPT Responses. *Machine Learning and Principles and Practice of Knowledge Discovery in Databases. ECML PKDD 2023. Communications in Computer and Information Science, vol 2133. Springer, Cham.. https://doi.org/10.1007/978-3-031-74630-7_20.*

Urchs, S., Thurner, V., Aßenmacher, M., Heumann, C. and Thiemichen, S. (2024). Detecting Gender Discrimination on Actor Level Using Linguistic Discourse Analysis. *In Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), Bangkok, Thailand, August 17, 2024: 140-149. https://aclanthology.org/2024.gebnlp-1.8/.*

Urchs, S., Thurner, V., Aßenmacher, M., Heumann, C. and Thiemichen, S. (2025). taz2024full: Analysing German Newspapers for Gender Bias and Discrimination across Decades. *In Findings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, July/August 27-01, 2025: 10661–10671. https://aclanthology.org/2025.findings-acl.555/.*

Urchs, S., Thurner, V., Aßenmacher, M., Heumann, C. and Thiemichen, S. (2025). Fair Play in the Newsroom: Actor-Based Filtering Gender Discrimination in Text Corpora. Accepted at the 5th Workshop on Evaluation Comparison of NLP Systems (eval4NLP).

# Further References

Claire Ainsworth. 2015. Sex redefined. *Nature*, 518(7539):288–291.

E Alharbi and S Tiun. 2015. A hybrid method of linguistic features and clustering approachfor identifying biomedical named entities. *Asian J. Appl. Sci.*, 8(3):210–216.

Gordon Willard Allport, Kenneth Clark, and Thomas Pettigrew. 1954. The nature of prejudice.

Amazon Web Services. 2024. Amazon Titan foundation models. Accessed: 2025-05-22.

Anthropic. 2024. Introducing the Claude 3 model family. Accessed: 2025-05-22.

Aristotle. 2009. *The Nicomachean ethics (book V)*. Oxford World's Classics. Oxford University Press.

Fatemeh Torabi Asr, Mohammad Mazraeh, Alexandre Lopes, Vasundhara Gautam, Junette Gonzales, Prashanth Rao, and Maite Taboada. 2021. The Gender Gap Tracker: Using Natural Language Processing to measure gender bias in media. *PLoS One*, 16(1):e0245533.

Hisham Assal, John Seng, Franz Kurfess, Emily Schwarz, and Kym Pohl. 2011. Semantically-enhanced information extraction. In *2011 Aerospace Conference*. IEEE.

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based Attention Regularization Frees Unintended Bias Mitigation from Lists. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, page 1105 – 1119.

Caleb J. S. Barr, Olivia Erdelyi, Paul D. Docherty, and Randolph C. Grace. 2025. A Review of Fairness and A Practical Guide to Selecting Context-Appropriate Fairness Metrics in Machine Learning. *arXiv preprint arXiv:2411.06624*.

Seyed-Mehdi-Reza Beheshti, Boualem Benatallah, Srikumar Venugopal, Seung Hwan Ryu, Hamid Reza Motahari-Nezhad, and Wei Wang. 2017. A systematic review and comparative analysis of cross-document coreference resolution methods and tools. *Computing*, 99(4):313–349.

Sylvia Bendel Larcher. 2015. *Linguistische Diskursanalyse: Ein Lehr-und Arbeitsbuch.* Narr Francke Attempto Verlag.

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

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS'16*, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Lera Boroditsky, Lauren A Schmidt, and Webb Phillips. 2003. Sex, syntax, and semantics. *Language in mind: Advances in the study of language and thought*, 22(61-79):3.

Ludwig Bothmann, Kristina Peters, and Bernd Bischl. 2024. What Is Fairness? Philosophical Considerations and Implications For FairML. ArXiv:2205.09622.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1664–1674.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, ..., and Dario Amodei. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

NiCole T. Buchanan. 2011. Microaggressions in everyday life: Race, gender, and sexual orientation. *Psychology of Women Quarterly*, 35(2):336–337.

Bundesministerium für Familie, Senioren, Frauen und Jugend. 2024. Gender Care Gap: Ein Indikator für die Gleichstellung. Accessed: 09 May 2025.

Yi Cai, Arthur Zimek, Gerhard Wunder, and Eirini Ntoutsi. 2022. Power of Explanations: Towards automatic debiasing in hate speech detection. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10.

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

David Campos, Sérgio Matos, and José Luís Oliveira. 2013. A modular framework for biomedical concept recognition. *BMC Bioinformatics*, 14(1):281.

Morgan Carpenter. 2021. Intersex human rights, sexual orientation, gender identity, sex characteristics and the yogyakarta principles plus 10. *Culture, Health & Sexuality*, 23(4):516–532. PMID: 32679003.

Simon Caton and Christian Haas. 2024. Fairness in Machine Learning: A Survey. *ACM Comput. Surv.*, 56(7):166:1–166:38.

Yong Chen, Fenglou Mao, Guojun Li, and Ying Xu. 2011. Genome-wide discovery of missing genes in biological pathways of prokaryotes. *BMC Bioinformatics*, 12 Suppl 1(S1):S1.

Jiali Cheng and Hadi Amiri. 2024. FairFlow: Mitigating Dataset Biases through Undecided Learning for Natural Language Understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21960–21975, Miami, Florida, USA. Association for Computational Linguistics.

Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, ..., and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sudeshna Das and Jiaul H Paik. 2021. Context-sensitive gender inference of named entities in text. *Information Processing and Management*, 58(1).

Mihai Dascalu. 2014. *Computational Discourse Analysis*, page 53–77. Springer International Publishing.

Jim Dator. 2017. Chapter 3. What Is Fairness? In Jim Dator, Richard C. Pratt, and Yongseok Seo, editors, *Fairness, Globalization, and Public Institutions*, pages 19–34. University of Hawaii Press, Honolulu.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

Robert-Alain De Beaugrande and Wolfgang U Dressler. 1981. *Introduction to text linguistics*, volume 1. longman London.

DeepSeek-AI. 2024. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. *arXiv preprint arXiv:2401.02954*.

Pieter Delobelle and Bettina Berendt. 2023. FairDistillation: Mitigating Stereotyping in Language Models. In *Machine Learning and Knowledge Discovery in Databases*, pages 638–654, Cham. Springer International Publishing.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "Gender" in NLP Bias Research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2083–2102, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Die Bundesregierung. 2024. Gleichberechtigung: 75 Jahre Grundgesetz. Accessed: 09 May 2025.

Jens Dörpinghaus and Andreas Stefan. 2019. Knowledge extraction and applications utilizing context data in knowledge graphs. In *Annals of Computer Science and Information Systems*, volume 18, pages 265–272. IEEE.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA. Association for Computing Machinery.

Norman Fairclough. 2012. Critical Discourse Analysis. In James Paul Gee and Michael Handford, editors, *The Routledge Handbook of Discourse Analysis*, pages 9–20. Routledge, London and New York.

Ethan Fast, Tina Vachovsky, and Michael Bernstein. 2016. Shirtless and Dangerous: Quantifying Linguistic Signals of Gender Bias in an Online Fiction Writing Community. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):112–120.

Gabriele Fischer, Ronja Philipp, Lina Spagert, Stephanie Thiemichen, Veronika Thurner, Stefanie Urchs, and Elke Wolf. 2024. The Floor is yours!? (Un)Sichtbarkeiten von HAWProfessorinnen. In Julia Rathke, Katja Knuth-Herzig, Lena Milker, and Rubina Zern-Breuer, editors, *Sichtbarkeit Von Weiblicher Wissenschaftlicher Leistung Im Fokus*. Nomos Verlagsgesellschaft.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *Acm Computing Surveys (Csur)*, 51(4):1–30.

Sanyuan Gao, Si Li, Weiran Xu, and Jun Guo. 2010. Cross-document coreference resolution based on automatic text summary. In *2010 Third International Conference on Knowledge Discovery and Data Mining*. IEEE.

Carl-Friedrich Graumann and Margret Wintermantel. 2007. *Diskriminierende Sprechakte. Ein funktionaler Ansatz*, pages 147–178. transcript Verlag, Bielefeld.

Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2016. The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making. In *Proceedings of the NIPS Symposium on Machine Learning and the Law*.

Ralph Grishman. 2015. Information Extraction. *IEEE Intelligent Systems*, 30(5):8–15.

Catherine Gross. 2008. A Measure of Fairness: An Investigative Framework to Explore Perceptions of Fairness and Justice in a Real-Life Social Conflict. *Human Ecology Review*, 15(2):130–140.

Qi Guo, Shuang Wang, and Fucheng Wan. 2020. Research on named entity recognition for information extraction. In *2020 2nd International Conference on Artificial Intelligence and Advanced Manufacture (AIAM)*, pages 121–124.

Michael Halliday. 2004. *An introduction to functional grammar*, 3 edition. Hodder Arnold, London, England.

Hsin-Yi Hsieh, Shih-Cheng Huang, and Richard Tzong-Han Tsai. 2024. TWBias: A Benchmark for Assessing Social Bias in Traditional Chinese Large Language Models through a Taiwan Cultural Lens. In *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Findings of EMNLP 2024*, page 8688 – 8704.

Jian Huang, Sarah M. Taylor, Jonathan L. Smith, Konstantinos A. Fotiadis, and C. Lee Giles. 2009a. Profile based cross-document coreference using kernelized fuzzy relational clustering. In *ACL '09*, page 414–422, USA. Association for Computational Linguistics.

Jian Huang, Sarah M. Taylor, Jonathan L. Smith, Konstantinos A. Fotiadis, and C. Lee Giles. 2009b. Solving the "Who's Mark Johnson" puzzle: information extraction based cross document coreference. In *SRWS '09*, page 7–12, USA. Association for Computational Linguistics.

Juha Janhunen. 2000. Grammatical gender from east to west. *TRENDS IN LINGUISTICS STUDIES AND MONOGRAPHS*, 124:689–708.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*, pages 7–16.

Kenneth Joseph, Wei Wei, and Kathleen M Carley. 2017. Girls rule, boys drool: Extracting semantic and affective stereotypes from twitter. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1362–1374.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released January 12, 2025.

Wilhelm Kamlah and Paul Lorenzen. 1996. *Die Elementare Prädikation*, pages 23–44. J.B. Metzler, Stuttgart.

Emily R Kaskan and Ivy K Ho. 2016. Microaggressions and female athletes. *Sex Roles*, 74:275–287.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR.

Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Halil Kilicoglu and Dina Demner-Fushman. 2016. Bio-SCoRes: A smorgasbord architecture for coreference resolution in biomedical text. *PLoS One*, 11(3):e0148538.

Clemens Knobloch. 1996. *Nomination: Anatomie eines Begriffes*, pages 21–53. VS Verlag für Sozialwissenschaften, Wiesbaden.

Youjin Kong. 2022. Are “Intersectionally Fair” AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 485–494, Seoul Republic of Korea. ACM.

Toshi Konishi. 1993. The semantics of grammatical gender: A cross-cultural study. *Journal of Psycholinguistic Research*, 22(5):519–534.

A A Kozlova, I D Kudinov, and D V Lemtyuzhnikova. 2025. Methods of solving the problem of coreference and searching for noun phrases in natural languages. *J. Comput. Syst. Sci. Int.*, 64(1):121–135.

Ruth Kramer. 2020. Grammatical Gender: A Close Look at Gender Assignment Across Languages. *Annual Review of Linguistics*, 6(1):45–66.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Robin Lakoff. 1973. Language and woman's place. *Language in Society*, 2(1):45–79.

Zhahao Li, Jinfu Chen, Haibo Chen, Leyang Xu, and Wuhao Guo. 2024. Detecting Bias in LLMs' Natural Language Inference Using Metamorphic Testing. In *Proceedings - 2024 IEEE 24th International Conference on Software Quality, Reliability and Security Companion, QRS-C 2024*, page 31 – 37.

Walter Lippmann. 1992. *Public Opinion*. Routledge.

Michele Loi and Christoph Heitz. 2022. Is Calibration a Fairness Requirement? An Argument from the Point of View of Moral Philosophy and Decision Theory. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2026–2034, New York, NY, USA. Association for Computing Machinery.

Jing Lu and Vincent Ng. 2018. Event Coreference Resolution: A Survey of Two Decades of Research. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5479–5486. International Joint Conferences on Artificial Intelligence Organization.

Pingchuan Ma, Shuai Wang, and Jin Liu. 2020. Metamorphic testing and certified mitigation of fairness violations in nlp models. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 2021-January, page 458 – 465.

Nishtha Madaan, Sameep Mehta, Taneea Agrawaal, Vrinda Malhotra, Aditi Aggarwal, Yatin Gupta, and Mayank Saxena. 2018. Analyze, detect and remove gender stereotyping from bollywood movies. In *Conference on fairness, accountability and transparency*, pages 92–105. PMLR.

Marta Marchiori Manerba, Karolina Stańczak, Riccardo Guidotti, and Isabelle Augenstein. 2024. Social Bias Probing: Fairness Benchmarking for Language Models. In *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, page 14653 – 14671.

Camila M Mateo and David R Williams. 2020. More than words: a vision to address bias and reduce discrimination in the health professions learning environment. *Academic medicine*, 95(12S):S169–S177.

Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2020. Man is to person as woman is to location: Measuring gender bias in named entity recognition. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media, HT 2020*, page 231 – 232.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, 54(6).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mistral AI. 2024a. Le Chat by Mistral AI. Accessed: 2025-05-22.

Mistral AI. 2024b. Overview of Mistral AI's foundation models. Accessed: 2025-05-22.

Gunnar Myrdal et al. 1944. An american dilemma; the negro problem and modern democracy.(2 vols.).

Ramesh Nallapati. 2004. Discriminative models for information retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, page 64–71, New York, NY, USA. Association for Computing Machinery.

Sangha Nam, Minho Lee, Donghwan Kim, Kijong Han, Kuntae Kim, Sooji Yoon, Eun-kyung Kim, and Key-Sun Choi. 2020. Effective Crowdsourcing of Multiple Tasks for Comprehensive Knowledge Extraction. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 212–219, Marseille, France. European Language Resources Association.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue.

OpenAI. 2023. ChatGPT(May 24 version). *[Large Language Model]*.

S Lakshmana Pandian, J Devakumar, and T V Geetha. 2008. Semantic information extraction from Tamil documents. *Int. J. Metadata Semant. Ontol.*, 3(3):226.

Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. In *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, page 446 – 457.

Cheoneum Park, Kyoung-Ho Choi, Changki Lee, and Soojong Lim. 2016. Korean coreference resolution with guided mention pair model using deep learning. *ETRI J.*, 38(6):1207–1217.

Kyungmin Park, Sihyun Oh, Daehyun Kim, and Juae Kim. 2024. Contrastive Learning as a Polarizer: Mitigating Gender Bias by Fair and Biased sentences. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4725–4736, Mexico City, Mexico. Association for Computational Linguistics.

Siddhesh Pawar, Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2025. Presumed Cultural Identity: How Names Shape LLM Responses. *arXiv preprint arXiv:2502.11995*.

María Antonia Paz, Julio Montero-Díaz, and Alicia Moreno-Delgado. 2020. Hate speech: A systematized review. *Sage Open*, 10(4):2158244020973022.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Webb Phillips and Lera Boroditsky. 2013. Can quirks of grammar affect the way you think? grammatical gender and object concepts. In *Proceedings of the 25th Annual Cognitive Science Society*, pages 928–933. Psychology Press.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. OpenAI.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019a. Language Models are Unsupervised Multitask Learners. *OpenAi Blog*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019b. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

V.L. Radishevskii, A.D. Kulnevich, R.A. Chugunov, and A.A. Shevchuk. 2018. Distributed GLR-parser for natural languag processing. In *CEUR Workshop Proceedings*, volume 2267, page 374 – 377.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Gregory Razran. 1950. Ethnic dislikes and stereotypes: a laboratory study. *The Journal of Abnormal and Social Psychology*, 45(1):7.

Martin Reisigl. 2017. *Sprachwissenschaftliche Diskriminierungsforschung*, pages 81–100. Springer Fachmedien Wiesbaden, Wiesbaden.

Drew Roselli, Jeanna Matthews, and Nisha Talagala. 2019. Managing Bias in AI. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 539–544, New York, NY, USA. Association for Computing Machinery.

Sunita Sarawagi. 2008. Information Extraction. *Found. Trends Databases*, 1(3):261–377.

Albert Scherr. 2016. Pierre Bourdieu: Die Unterscheidung. Social Critique of Judgment, Paris 1979, 672 s. (dt. Ausgabe: Die feinen Unterschiede, Kritik der gesellschaftlichen Urteilskraft, frankfurt 1982, 879 s.). *Klassiker der Sozialwissenschaften: 100 Schlüsselwerke im Portrait*, pages 313–316.

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2011. Large-scale cross-document coreference using distributed inference and hierarchical models. HLT '11, page 793–803, USA. Association for Computational Linguistics.

Lina Spagert and Elke Wolf. 2025. Doing Visibility: Understanding Gender and Discipline Differences in Science Communication on Social Media and in the Press. *Social Sciences*, 14(3).

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Kunsheng Tang, Wenbo Zhou, Jie Zhang, Aishan Liu, Gelei Deng, Shuai Li, Peigui Qi, Weiming Zhang, Tianwei Zhang, and Nenghai Yu. 2024. GenderCARE: A Comprehensive Framework for Assessing and Reducing Gender Bias in Large Language Models. In *CCS 2024 - Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, page 1196 – 1210.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, ..., and Oriol Vinyals. 2025. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

FRANCES TRIX and CAROLYN PSENKA. 2003. Exploring the Color of Glass: Letters of Recommendation for Female and Male Medical Faculty. *Discourse & Society*, 14(2):191–220.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct Distillation of LM Alignment. *arXiv preprint arXiv:2310.16944*.

Shyam Upadhyay, Nitish Gupta, Christos Christodoulopoulos, and Dan Roth. 2016. Revisiting the Evaluation for Cross Document Event Coreference. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1949–1958, Osaka, Japan. The COLING 2016 Organizing Committee.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Madamanchi Venugopal, Virendra K Sharma, and Kalpana Sharma. 2023. Web information mining and semantic analysis in heterogeneous unstructured text data using enhanced latent dirichlet allocation. *Concurr. Comput.*, 35(1).

Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness*, Gothenburg Sweden. ACM.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law. *West Virginia Law Review*, 123(3):735–790.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2021. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):454–463.

Fucheng Wan and Jianhua Xia. 2017. Tibetan information extraction technology integrated with event feature and semantic role labelling. *MATEC Web Conf.*, 128:01016.

Hongsheng Wang, Lu Yuan, and Hong Shao. 2008. Text Information Extraction Based on OWL Ontologies. In *2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery*. IEEE.

Xinyi Wang, Xiangrong Zhu, and Wei Hu. 2025. Evidence selection via multi-aspect query diversification for cross-document relation extraction. *J. Intell. Inf. Syst.*

Ze Wang, Zekun Wu, Xin Guan, Michael Thaler, Adriano Koshiyama, Qinyang Lu, Sachin Beepath, Ediz Ertekin, and Maria Perez-Ortiz. 2024. JobFair: A Framework for Benchmarking Gender Hiring Bias in Large Language Models. In *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Findings of EMNLP 2024*, page 3227 – 3246.

Zijian Wang and Christopher Potts. 2019. TalkDown: A Corpus for Condescension Detection in Context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3711–3719.

Warto, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Purwanto, Muljono, and De Rosal Ignatius Moses Setiadi. 2024. Systematic literature review on named entity recognition: Approach, method, and application. *Stat. Optim. Inf. Comput.*, 12(4):907–942.

Candace West and Don H. Zimmerman. 1987. Doing Gender. *Gender & Society*, 1(2):125–151.

Daya C Wimalasuriya and Dejing Dou. 2010. Ontology-based information extraction: An introduction and a survey of current approaches. *J. Inf. Sci.*, 36(3):306–323.

M. J. Wolf, K. Miller, and F. S. Grodzinsky. 2017. Why we should have seen that coming: comments on Microsoft's tay "experiment," and wider implications. 47(3):54–64.

Guohua Wu, Guangen Tang, Zhongru Wang, Zhen Zhang, and Zhen Wang. 2019. An attention-based BiLSTM-CRF model for Chinese clinic named entity recognition. *IEEE Access*, 7:113942–113949.

Zhi Xu and Sencun Zhu. 2010. Filtering offensive language in online communities using grammatical relations. In *Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*, pages 1–10.

Congzhi Zhang, Linhai Zhang, Jialong Wu, Yulan He, and Deyu Zhou. 2025. Causal Prompting: Debiasing Large Language Model Prompting Based on Front-Door Adjustment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25842–25850.

Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, page 110–120, New York, NY, USA. Association for Computing Machinery.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989.

Zhouxiang Zhao, Zhaohui Yang, Ye Hu, Licheng Lin, and Zhaoyang Zhang. 2023. Semantic information extraction for text data with probability graph. In *2023 IEEE/CIC International Conference on Communications in China (ICCC Workshops)*, pages 1–6. IEEE.

Yandan Zheng and Luu Anh Tuan. 2023. A novel, cognitively inspired, unified graph-based multi-task framework for information extraction. *Cognit. Comput.*, 15(6):2004–2013.

# Usage of AI

As English is not my first language, I used `Grammarly` to support correct spelling and grammar. Since clear and well-structured language significantly enhances the reading experience, I also employed `ChatGPT` (Model 4o) to improve phrasing and flow throughout the text.

My workflow involved writing each paragraph in my own words before using ChatGPT to suggest improvements. I carefully reviewed all suggestions and incorporated those that aligned with my intended meaning and scientific writing style.

The initial pipeline code was written by me based on my own ideas and design. In later stages, I used ChatGPT to support improvements to the pipeline, refine the code structure, and enhance the documentation, but all core concepts and development decisions were solely my own.

# Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12. Juli 2011, § 8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 21.08.2025

Stefanie Urchs

# A. Metrics

This section documents all metrics reported by the discrimination-detection-pipeline in the paper "Fair Play in the Newsroom: Actor-Based Filtering Gender Discrimination in Text Corpora". For each metric, I describe:

- **What**: the quantity being measured
- **Why**: the rationale for including it
- **How to interpret**: how to read and evaluate the reported values

All metrics are reported separately for *she/her*, *he/him*, and *overall*, unless otherwise noted. Most are presented both as corpus-level totals and as per-text means, medians, and standard deviations.

## A.1. General Corpus Statistics

### Total Texts

- **What**: Total number of articles in the corpus for a given year.
- **Why**: Indicates corpus size, which impacts interpretability and statistical stability.
- **How to interpret**: Smaller values (e.g., early 1980s) lead to noisier metrics. Use this to contextualise all further metrics.

### Texts with Actors

- **What**: Number of texts that contain at least one actor with resolved gender.
- **Why**: Only these texts are analysed; others are excluded.
- **How to interpret**: A high share indicates full coverage; a low share indicates limited gender-resolvable material.

### Uses Gender Neutral Language (Docs)

- **What**: Number of texts using inclusive forms such as `Lehrer:innen` or `Schüler*innen`.
- **Why**: Captures editorial shifts toward gender-fair language.
- **How to interpret**: Higher values indicate more inclusive language usage.

### Generic Masculine Usage (Docs)

- **What**: Number of texts that use the generic masculine form (e.g., `die Studenten`).
- **Why**: The generic masculine is a structural source of gender exclusion in German.
- **How to interpret**: High values reflect dominant use of non-inclusive grammatical norms.

## A.2. Representation and Framing Metrics

### Pronoun Distribution

- **What**: Number of actors in the texts.
- **Why**: Indicates basic discursive presence.
- **How to interpret**: Large gender gaps reflect asymmetric visibility.

### Mentions by Pronoun

- **What**: How often actors are mentioned.
- **Why**: Tracks referential continuity and narrative presence.
- **How to interpret**: Higher values for one group suggest greater prominence across texts.

### Named Mentions

- **What**: Number of mentions using full names.
- **Why**: Named references imply individuation and recognisability.
- **How to interpret**: Gender gaps reflect asymmetric specificity (e.g., if women are more often mentioned by their pronoun).

### Pronoun Mentions

- **What**: Number of mentions using pronouns only.
- **Why**: Complements named mentions; excessive pronoun use reduces discursive clarity.
- **How to interpret**: Disproportionate pronoun use for women implies backgrounding.

### Subject Roles

- **What**: Number of times actors appear as grammatical subjects.
- **Why**: Subject roles signal agency and action.
- **How to interpret**: A lower share of women as subjects indicates reduced narrative control.

### Object Roles

- **What**: Number of times actors appear as grammatical objects.
- **Why**: Object roles indicate being acted upon.
- **How to interpret**: A high proportion of women in object roles suggests passivation.

### Direct Quotes

- **What**: Number of attributed direct quotes.
- **Why**: Direct speech signifies authority and voice.
- **How to interpret**: If men are quoted directly more often, this reflects narrative centrality and credibility gaps.

### Indirect Quotes

- **What**: Number of attributed indirect quotes.
- **Why**: Indirect speech downplays speaker agency.
- **How to interpret**: High indirect attribution for women implies reduced narrative presence.

### Feminine-coded Words

- **What**: Frequency of feminine-coded descriptors.
- **Why**: Lexical bias often reflects gendered expectations.
- **How to interpret**: High values for women may signal stereotypical framing (e.g., nurturing, supportive).

### Masculine-coded Words

- **What**: Frequency of masculine-coded descriptors.
- **Why**: Complements the feminine-coded metric.
- **How to interpret**: Skewed usage toward men reinforces traditional gender roles (e.g., assertive, independent).

### Sentiment

- **What**: Mean sentiment score for all predications involving actors.
- **Why**: Evaluative framing is a key dimension of linguistic discrimination.
- **How to interpret**: Even slight differences (e.g., -0.01 vs. 0.01) are meaningful in neutral-skewed distributions. Negative gaps for women suggest systemic framing bias.

## A.3. Normalised Representation Metrics (Percentage Values)

### Named Mentions (% of all mentions)

- **What**: Share of mentions that are named references per gender.
- **Why**: Named mentions signal individuation and specificity.
- **How to interpret**: A lower share for women indicates reduced discursive prominence.

### Pronoun Mentions (% of all mentions)

- **What**: Share of mentions that are pronouns per gender.
- **Why**: High pronoun usage may obscure identity and agency.
- **How to interpret**: A higher rate for women may imply marginalisation or backgrounding.

### Subject Roles (% of known roles)

- **What**: Share of all semantic roles that are subject roles per gender.
- **Why**: Normalises subject counts.
- **How to interpret**: A lower percentage of subject role share for women indicates reduced narrative agency.

### Object Roles (% of known roles)

- **What**: Share of all semantic roles that are object roles per gender.
- **Why**: Normalises object counts.
- **How to interpret**: A lower percentage of object roles for women signals structural passivation.

### Direct Quotes (% of quotes)

- **What**: Share of all quotes that are direct quotes per gender.
- **Why**: Assesses gender balance in attributed voice.
- **How to interpret**: A lower percentage of direct quotes for women reflects discursive underrepresentation.

### Indirect Quotes (% of quotes)

- **What**: Share of quotes that are indirect quotes per gender.
- **Why**: Reflects passive or backgrounded representation.
- **How to interpret**: A higher share for women may indicate narrative marginalisation.

## A.4. PMI-based Lexical Framing

### Top PMI Adjectives, Nouns, Verbs

- **What**: Lists of words with highest Pointwise Mutual Information (PMI) per gender group.
- **Why**: Reveals thematic associations and stereotypical framing.
- **How to interpret**: Recurrent associations reveal patterns of how actors are described. Differences between genders indicate bias in framing and topical contexts.