Decoding Bias: Exploring Sexism in Software Development Through Online Narratives and AI Analysis

Amanda Kolopanis

A Thesis

in

The Department

of

Computer Science & Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Software Engineering) at

Concordia University

Montréal, Québec, Canada

August 2024

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Amanda Kolopanis**

Entitled: **Decoding Bias: Exploring Sexism in Software Development Through Online Narratives and AI Analysis**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Software Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. René Witte

_____ Examiner
Dr. Jinqui Yang

_____ Examiner
Dr. René Witte

_____ Co-Supervisor
Dr. Tristan Glatard

_____ Co-Supervisor
Dr. Tanja Tajmel

Approved by _____
Joey Paquet, Chair
Department of Computer Science & Software Engineering

_____ 2024

_____
Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Decoding Bias: Exploring Sexism in Software Development Through Online Narratives and AI Analysis

Amanda Kolopanis

The persistent gender gap in Software Engineering (SE) and related software development fields necessitates a thorough examination to expose the root causes and advance women's engagement in technological innovation. This disparity presents both societal and technical challenges, perpetuating implicit gender biases in technology due to the limited representation of women. Online forums provide insight into women's experiences with sexism in technical environments, but the unstructured nature of this data complicates the extraction of such specific instances. Our research aims to address this issue by analyzing online narratives from women software developers illustrating their experiences with sexism in technical teams. We initiate this study by constructing a taxonomy to identify various forms of sexism. Subsequently, we apply conventional data extraction techniques, such as static keyword-matching, and advanced artificial intelligence (AI) methods, including semantic similarity, to identify sexist experiences in the online dataset. Lastly, we evaluate the AI model's effectiveness with Equity, Diversity, and Inclusion (EDI) experts to ensure alignment with nuanced human understandings of sexism. Our results reveal the development of a taxonomy encompassing four distinct classes of sexism, supported by definitions, anchor examples, and lexicons. We observe that while semantic similarity techniques are proficient in extracting narratives of sexist experiences, the model encounters difficulties in accurate classification. Furthermore, our results highlight the intricate challenges of trying to align AI systems with human interpretations of sexism as defined in our taxonomy. Additionally, our findings reveal three previously overlooked instances of sexism. Based on our outcomes, we propose a code of conduct for practitioners to mitigate sexism within technical teams, enhancing women's participation in SE and technological innovation.

Statement of Originality

I, Amanda Kolopanis, hereby declare that I am the sole author of this thesis. All ideas and inventions attributed to others have been properly referenced. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners. I understand that my thesis may be made electronically available to the public.

Acknowledgments

- I wish to express my profound gratitude to my family, especially my mother, for their unwavering support and encouragement throughout my academic journey. Their constant guidance and belief in me have been crucial, and I would not have reached this milestone without their ongoing support over the years.

- My sincere thanks go to my exceptional supervisors, Dr. Tanja Tajmel and Dr. Tristan Glatard. Their dedication, insightful guidance, and continuous support have been vital to the completion of this Master's program. Their mentorship has profoundly shaped my academic experience and success.

- I am deeply grateful to Dr. Gita Ghiasi, who initially inspired me as an exceptional professor for ENGR 392 and became a remarkable mentor. Her guidance and continuous support through conferences, workshops, my thesis, and my NSERC application, has been invaluable to my academic journey.

- I am also profoundly grateful to the Software Engineering for Artificial Intelligence (SE4AI) program for offering me the opportunity to enhance my technical skills in a field I am deeply passionate about. This program has been instrumental in my development and growth.

- Additionally, I wish to acknowledge the Equity, Diversity, and Inclusion (EDI) lab at Concordia University for their valuable participation and expertise in validating the results of this study. Their constructive feedback and support have been immensely appreciated.

- Finally, I would like to express my heartfelt thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC) for their financial support, which has been crucial to

the advancement of this research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Software engineering (SE) and related fields, such as computer science (CS) and information technology (IT) that encompass the software development domain, are exhibiting an ongoing gender gap. For example, a 2010 report indicated that women represented about 25% to 30% of the IT industry workforce Hill, Corbett, and St. Rose (2010). However, in 2022, the representation decreased as "women represented less than 24% of employees in the software development industry" Trinkenreich, Britto, Gerosa, and Steinmacher (2022). Although there have been various attempts to promote women in software development, such as creating inclusive communities and events specifically aiming towards women's involvement in technology, as well as showcasing prominent women role models in software development González González et al. (2018), the representation remains stagnant and on the verge of decreasing over time.

Consequently, the software development field is male-dominated and could be exhibiting inhospitable characteristics, like sexism, that discourage women from either entering or persisting within the domain. Numerous studies and articles document women's shared negative experiences while collaborating in their development teams Daub (2021); Faulkner (2000); Guzmán, Fischer, and Kok (2023). For instance, there are reports illustrating women's experiences of sexism by feeling ostracized, objectified, and harassed throughout their involvement in software development Doyle (2020). Furthermore, these individual experiences unite and propagate to industrial-level issues where employees organize company walkouts to protest the hostile environment towards women Wakabayashi, Griffith, Tsang, and Conger (2018) and undergo million-dollar lawsuits to combat

harassment Bond (2019).

The lack of women programmers produces a diversity gap in technological development through the absence of valuable knowledge that could assist in building quality applications. For example, the United Nations Educational, Scientific and Cultural Organization (UNESCO) reported in 2019 that popular virtual assistants such as Siri, Alexa, Cortana, and Google Assistant, implicitly enable sexual harassment by responding in a flirtatious, comedic, or dismissive manner when conversing with male participants United Nations Educational (2019). On the contrary, virtual assistants respond defensively when female participants vocalize the same statements of sexual harassment United Nations Educational (2019). Therefore, increasing women's involvement in software development is considered necessary to detect and avoid implicit misogyny and sexism before releasing the technology to the general public to propagate these negative social behaviours.

Accordingly, online platforms, such as Reddit, offer dedicated forums that concentrate on women's involvement in the software development domain. These online outlets enable women to share their personal experiences and suggestions for navigating the intricacies of the field in a supportive environment and can aid in identifying the possible true causes of women practitioners experiencing sexism. However, there is a significant challenge in extracting context-specific content from online resources as the data is often unstructured and contains an ample amount of noise. Thus, applying artificial intelligence (AI) approaches, such as semantic similarity, provides an avenue to efficiently filter relevant data pertinent to different types of sexism.

This thesis aims to investigate the various forms of sexism experienced by women software developers through an integrated approach that applies knowledge of sexism and misogyny to SE practices and AI. Initially, we explore research on sexism and misogyny to select a pertinent feminist theoretical resource to construct a taxonomy, complete with lexicons, that identifies the most common types of sexism women encounter in male-dominated fields. Next, we conduct a case study on the experiences of women software developers by extracting relevant Reddit content using these keywords and applying semantic similarity techniques to determine whether these categories reflect the online narratives. The model's effectiveness is then evaluated in collaboration with Concordia University's Equity, Diversity, and Inclusion (EDI) lab to assess its accuracy in classifying

instances of sexism. Finally, we present the discovered online experiences pertinent to our taxonomy and develop a code of conduct for practitioners and researchers to use to help reduce sexism within software development teams, thereby fostering a more inclusive and collaborative work environment.

1.1 Thesis Overview

The structure of the thesis progresses with the following content. Chapter 2 offers a literature review of the related works in our area of research, such as the concept of sexism, prior and current experiences of sexism in SE and related software development fields, and recent advancements in sexism classification systems. The section includes detailed analyses of the datasets following the utilized taxonomies and the models implemented in these classification systems. In Chapter 3, we illustrate the overview of our implemented semantic similarity approach by explaining concepts such as transformer models, embeddings, and cosine similarity scores. Chapter 4 details the methodology of our research, including the establishment of the taxonomy with the corresponding lexicons, the data extraction process, the semantic similarity analysis, and the model evaluation using a panel of four EDI experts. In Chapter 5, we present our main results in three sections, each encapsulating the motivation, approach, and outcomes as outlined below:

- Chapter 5.1 addresses our first research question **"What categories of sexism are women most likely to experience in the software development field?"** Through our literature research on sexism and misogyny, we identify that Kate Manne's book *"Down Girl: The Logic of Misogyny"* aligns closely with our research objectives. From this foundation, we develop a preliminary taxonomy, which we identify as *Sexism in Software Development*, containing four distinct classes that define different manifestations of sexism. Additionally, we create corresponding lexicons for the taxonomy and use them in the data extraction process.

- Chapter 5.2 tackles our second research question **"To what extent can semantic similarity effectively extract the experiences of women in software development that align with the constructed taxonomy?"** Through three distinct tasks, we discover that fine-tuning the model for extracting personal narratives of sexism from women software developers yields

promising results. However, obstacles arise when attempting to accurately classify narratives into the appropriate taxonomy class.

- Chapter 5.3 strives to answer our third research question **"What are the challenges in aligning AI systems with human interpretations of sexism?"** Through a comparative analysis of EDI experts' performance versus a zero-shot classification model in classifying twenty manually-selected examples from the dataset, we reveal challenges in interpreting the model's responses to different definitions with their anchor examples and the time-consuming nature of locating specific examples for the training dataset.

Chapter 6 provides a detailed examination of our findings by exploring the discovered experiences of sexism pertinent to our taxonomy and offers a code of conduct for practitioners and researchers to address sexism in software development teams while helping to narrow the gender gap in the field. Finally, Chapter 7 concludes the thesis by evaluating potential limitations -— such as internal, external, and construct validity issues —- assessing the research's societal impact, and outlining directions for future investigation.

1.2 Thesis Contributions

This thesis provides five main contributions to practitioners and the software development research community:

- We present a simplified taxonomy delineating the various forms of sexism commonly encountered in software development, including lexicons for each category.
- We assess and compare the constraints of semantic similarity in identifying and classifying sexism within unstructured online discourses by evaluating the model's efficacy alongside a team of EDI experts.
- We reflect on both established and newly uncovered experiences of sexism within the software development field while offering intriguing insights from our experimental findings.

- We establish a code of conduct for software development teams to reference, with the intention of raising awareness of sexism toward their female colleagues.
- We address the United Nations Sustainable Development Goals 5, 8, and 10, emphasizing gender equality, decent work and economic growth, and reduced inequalities. Our contribution to advancing these critical objectives involves promoting the inclusion of women in SE and related software development fields and tackling sexism within the domain.

Chapter 2

Sexism in Software Development

In this chapter, we conduct a literature review in the context of sexism in software development and highlight our contributions to the field. First, we define the notion of sexism to contextualize our study within this framework. Subsequently, we explore documented experiences of sexism in the software development domain to provide information on the pre-defined phenomena and to support our findings. Finally, we scrutinize recent research on the classification of sexism to ascertain how SE and related communities define and categorize these issues while also exploring the technologies employed to achieve this task.

2.1 Sexism

Sexism often refers to prejudice or discrimination that disadvantages individuals based on their gender, leading to phenomena such as systemic inequality and social injustice Leaper and Robnett (2016). A related term, misogyny, describes the hatred explicitly directed towards women and girls by encompassing violence and extreme hostility against them Ussher (2016). Therefore, this study focuses on sexism due to its broader scope of potential impacts that could be prominent in systemic and social environments such as technical teams. Moreover, our preliminary analysis of the Reddit dataset reveals a higher prevalence of sexist experiences compared to instances of misogyny, which reinforces the relevance of our chosen focus.

Furthermore, we recognize that gender is a socially constructed category that may not necessarily align with an individual's sex assigned at birth. While we fully acknowledge that sexism affects LGBTQIA+ individuals, we have chosen to focus in this study on sexism experienced by individuals who identify as women. We believe their unique challenges deserve dedicated research to ensure their voices are heard and address their specific discrimination experiences. We encourage future researchers to adjust our approach by focusing on the sexism experienced by LGBTQIA+ individuals. This shift will help illuminate their specific challenges and contribute to creating a more inclusive and diverse environment in software development. Therefore, in this subsection, we aim to thoroughly explore the complex notion of sexism defined throughout existing research to better understand the phenomenon and its implications, specifically against women.

One foundational psychology study exploring different types of sexism is Glick and Fiske (1997) in their article *Ambivalent Sexism Theory*. This study shows how stereotypic attitudes and beliefs about women sustain gender inequality. The authors differentiate between two prominent types of sexism: benevolent and hostile. Benevolent sexism involves subjective positive attitudes towards women who conform to traditional gender roles. Furthermore, hostile sexism encompasses negative attitudes, including dominative paternalism, derogatory beliefs, and heterosexual hostility. Both dimensions of sexism, according to the study, contribute to maintaining gender hierarchies, restricting women's opportunities, and perpetuating stereotypes. This research provides a valuable foundation for our literature review on sexism by helping us identify and validate potential categories and issues relevant to software development domains.

Moreover, a notable piece of feminist scholarship is Sara Mills' book *"Language and Sexism"* Mills (2008), where she offers a thorough critique of existing research on linguistic sexism and provides a critical analysis of feminist linguistics. Mills argues that previous studies have often concentrated on easily identifiable forms of overt sexism in language, thereby neglecting the subtler, context-dependent instances of indirect sexism. She suggests moving beyond a simplistic understanding of feminist linguistic concepts, such as direct and indirect sexism, political correctness, language reform, femininity, and masculinity Mills (2008). Instead, she suggests reinterpreting these notions within their cultural and linguistic contexts. Furthermore, Mills highlights that analyzing language in isolation can lead to significant oversimplifications. Consequently, she advocates

for a more nuanced approach that examines the context in which sexist expressions occur. This perspective is particularly pertinent to our research as it aligns with our objective to investigate the specific context of sexism faced by women software developers within their technical teams.

Another influential work in feminist literature that explores concepts of misogyny and sexism in various public spheres, such as the workplace and politics, is Kate Manne's *"Down Girl: The Logic of Misogyny"* Manne (2018). In this book, Manne presents the "first book-length exploration of misogyny," offering a clear, dictionary-style definition of the terms Manne (2018). Manne dissects misogyny as a mechanism for upholding patriarchy, manifesting through the control, policing, punishment, and exclusion of women who resist male dominance. Additionally, she examines sexism as an ideology that supports and rationalizes the patriarchal social order. One of the most engaging aspects of Manne's work is her integration of previous research, such as, for instance, the article by Glick and Fiske (1997), and her provision of accessible definitions for various types of sexism and misogyny, accompanied by a helpful list of key terms. For instance, Manne defines a segment of the concept of *Testimonial Injustice* as follows:

> "Testimonial injustice arises due to systematic biases in the 'economy of credibility,' as Fricker (2007) aptly calls it. It afflicts members of a certain social group, most notably when the group has historically been and to some extent remains unjustly socially subordinate. Testimonial injustice then paradigmatically consists in subordinate group members tending to be regarded as less credible when they make claims about certain matters, or against certain people, hence being denied the epistemic status of knowers, in a way that is explained by their subordinate group membership." Manne (2018)

Manne further elaborates on preliminary terms related to testimonial injustice, describing them as instances where individuals are, for example, "accused, impugned, convicted, corrected, diminished, or, alternatively, simply outperformed by those who have historically held dominance" Manne (2018). This approach distinguishes her work from other literature, offering a more accessible framework for a technical analysis to identify and extract relevant data from online resources.

Finally, the paper by Wrisley (2021) critiques existing feminist theories for their inadequate treatment of misogyny and its real-world implications. Wrisley argues that while these theories

effectively identify misogyny in society, they often fail to address its practical consequences and emotional dimensions. The paper identifies three primary issues with current feminist analyses of misogyny: the conflation of sexism and misogyny into a single category, the neglect of emotional aspects in feminist evaluations, and the merging of misogyny with violence against women. Furthermore, the author illustrates the distinction between misogyny and sexism by defining misogyny as a negative emotional orientation towards women. In contrast, sexism represents the institutionalized form of this prejudice, manifesting in societal structures such as unequal wages and restricted access to healthcare. Given this distinction, this thesis emphasizes sexism over misogyny to encompass a broader range of experiences. This approach is integral to our taxonomy design, which aims to address potential prejudices women software developers encounter within their technical teams. By focusing on sexism, we can more effectively identify and analyze the various forms of gender-based biases and their impact on women's professional experiences rather than solely addressing the hostility represented by misogyny.

Additionally, the study critically engages with Kate Manne's *"Down Girl: The Logic of Misogyny"* Manne (2018), arguing that Manne's framework inadequately addresses the emotional dimensions of misogyny, thus offering a flawed approach to understanding and combating the complexities of misogyny. While we recognize and respect the author's critiques and contributions to the field, our focus remains on exploring sexism in a broader context of gender prejudice, specifically concerning women software developers. We acknowledge that opinions on complex issues like sexism and misogyny may vary. However, for this study, we are using a comprehensive literature review on both topics to inform and design our taxonomy. This approach aims to better understand and address women's diverse prejudices in technical environments.

In examining popular studies, we aimed to enhance our understanding of sexism to develop a comprehensive taxonomy that details its various forms. Our review revealed that while many papers provided nuanced analyses of misogyny and sexism from numerous perspectives, some variations in terminology and definitions could cause ambiguities. This divergence highlighted the challenges in establishing a standardized categorization as conflicting interpretations across research papers complicated our efforts. Moreover, we encountered difficulties sourcing references that offered clear definitions and illustrative examples of sexism suitable for our technical application. Among

the notable contributions, Kate Manne's book *"Down Girl: The Logic of Misogyny"* Manne (2018) stands out for its detailed content and extensive prior literature support. Consequently, we chose to base our taxonomy on Manne's work due to its taxonomy-based approach, which aligns well with our research objectives.

2.2 Experiences of Sexism in Software Development

Recent studies illustrate prevalent forms of sexism experienced by women software developers. As a general overview, women programmers often encounter various challenges in the field, ranging from explicit gender-targeted actions to managing internalized emotions. For instance, Oliveira et al. (2023) conducted a cross-sectional survey using convenience and snowball sampling. Their approach involved 42 women participants in the software engineering field to discuss their perspectives, challenges, and support tactics throughout their journey from academia to industry. Their research revealed that many women software engineers express concerns about being perceived as a *diversity hire*. This label undermines women's technical abilities and accomplishments while superficially addressing gender gaps in development teams. Furthermore, the study found that women software engineers commonly face hostile work environments, unequal opportunities, invisibility in technical contributions, and inadequate support compared to their male counterparts.

Additionally, Guenes, Tomaz, Kalinowski, Baldassarre, and Storey (2023) also performed an online survey using convenience and snowball sampling among software project managers and developers to investigate the prevalence and impact of the imposter phenomenon (IP) within the software engineering field. Originating from Clance's psychotherapy research (1978), IP defines the "experience of intellectual phoniness perceived by high-achieving professionals" Clance and Imes (1978). Guenes et al. (2023) found that 60.64% of women participants suffer from IP compared to men, with 48.82%, a factor that can contribute to mental disorders such as depression or burnout.

Sultana, Cavaletto, and Bosu (2021) adopted a methodology similar to that of Oliveira et al. (2023) and Guenes et al. (2023) to employ an online survey sent anonymously to self-identified software development practitioners. Their study aimed to explore the current status quo of gender

bias towards women across four dimensions in contemporary computing organizations: task selection, sexual harassment, gender harassment, and career progression. They received 78 completed responses, revealing that women software developers share that their teammates often implicitly pressure them to perform administrative tasks, such as scheduling meetings and organizing content for group discussions, as the team exhibits dissonance regarding their competence towards individually completing challenging, technical assignments. Moreover, the study highlights instances where women software developers experience unwanted sexual attention from male colleagues, which sometimes escalates into sexual harassment following rejection.

Trinkenreich et al. (2022) study aimed to investigate the challenges women practitioners face in global development teams within the software development sector to better understand the gender gap and propose mitigation strategies to retain women in the field. The research used a case study approach by distributing an online survey within Ericsson, a leading telecommunications company, and received 94 responses from women software developers. Their findings highlight prevalent issues, including experiences of benevolent and hostile sexism, lack of peer recognition and parity, impostor syndrome (i.e., IP), the effects of the glass ceiling bias, the "prove-it-again" phenomenon, and the maternal wall. For context, glass ceiling bias refers to systemic barriers that prevent qualified individuals, such as women, from advancing to higher positions within an organization despite their technical skills and achievements. Furthermore, the "prove-it-again" phenomenon illustrates individuals having to repeatedly provide evidence to demonstrate their competence to the team. Lastly, the maternal wall describes the systemic discrimination against mothers in software development due to assumptions made about their competence, commitment, and productivity after having children.

Finally, Guzmán et al. (2023) focused on observing micro-inequities and barriers experienced by software professionals in technical roles. Micro-inequities refer to subtle actions or behaviours that convey bias, such as purposeful conversation interruptions, lack of eye contact, and assigning menial tasks to teammates. Their research utilizes a purposive sampling approach by advertising a survey targeting software professionals through personal networks, industry contacts, and postings in online communities such as Reddit, LinkedIn and Facebook. They received 177 responses from women participants in technical roles. According to their findings, women experience more

micro-inequities compared to their male participants, along with encountering external and internal barriers. External barriers include witnessing or experiencing sexism and harassment in the workplace and having limited authority to make necessary decisions in their work. Internal barriers include feeling less valued and less recognized by teammates and receiving significantly less support than male colleagues.

While existing research extensively documents prevalent instances of sexism within the relative software development domains, many of the methodologies rely on traditional approaches such as surveys or interviews. In contrast, this thesis proposes an innovative strategy by examining narratives sourced from lesser-explored subreddits dedicated to women's engagement in programming and technology. Chapter 4 elaborates on this methodology in further detail by offering novel perspectives for understanding gender dynamics in tech communities.

2.3 Sexism Classification

This section aims to present the latest advancements in sexism classification research. We begin by highlighting the prevalent use of datasets in this field by aiming to reveal their underlying taxonomies. Additionally, we critically evaluate existing models deployed in this domain by examining their respective outcomes, encountered obstacles, and gained advantages. This comparative analysis helps us determine the most effective approach to address our research questions.

2.3.1 Datasets

In the field of sexism classification research, several popular datasets have been formulated based on prior studies containing large corpora of labelled data. However, a closer analysis reveals that many studies often overlook significant shortcomings. For example, the research by Kirk, Yin, Vidgen, and Röttger (2023) introduces the SemEval 2023 online sexism detection dataset, developed in conjunction with the annual International Workshop on Semantic Evaluation (SemEval) SemEval (2023). The workshop aims to evaluate advancements in semantic analysis systems across twelve distinct tasks. Task 10 specifically addresses Explainable Detection of Online Sexism (EDOS), employing a taxonomy grounded in prior theoretical and empirical research composed of eleven

fine-grained classes: (1) threats of harm, (2) incitement and encouragement of harm, (3) descriptive attacks, (4) aggressive and emotive attacks, (5) dehumanizing attacks and overt sexual objectification, (6) casual use of gendered slurs, profanities, and insults, (7) immutable gender differences and gender stereotypes, (8) backhanded gendered compliments, (9) condescending explanations or unwelcome advice, (10) supporting mistreatment of individual women, (11) supporting systemic discrimination against women as a group. However, a critical review of the literature used to construct this taxonomy reveals that many prior studies draw their classifications from broader research contexts rather than solely on sexism, as exemplified by works such as Waseem, Davidson, Warmsley, and Weber (2017) and Farrell, Fernandez, Novotny, and Alani (2019). Moreover, these taxonomies often lack updated insights from feminist studies, such as those presented in Jha and Mamidi (2017) and Samory, Sen, Kohne, Flöck, and Wagner (2021).

Quantitatively, the dataset consists of 20,000 labelled entries sourced evenly from Reddit and Gab, as detailed by Kirk et al. (2023). Additionally, the study provides a more extensive, unlabeled dataset for model adaptation. A notable aspect of the dataset is that it exhibits a high imbalance with 4,854 entries labelled as *sexist* and 15,146 as *not sexist*. This distribution raises concerns as it may not adequately support the nuanced multi-tiered taxonomy of eleven classes proposed by SemEval, mainly due to the limited data points under the *sexist* category. Moreover, upon closer examination of the dataset, it becomes evident that some data points do not align with the context of the research on sexism against women in online communities. For instance, the text "+1 You were acting like a complete douchebag" Kirk et al. (2023) is labelled as *sexist* and classified under the taxonomy category of *casual use of gendered slurs, profanities, and insults*. Although the term *douchebag* typically refers to a male, this classification appears inconsistent and could compromise the validity of the dataset. Additionally, another entry in the SemEval dataset that raises concerns is the text "We must require annual physicals to include a mental check. Something is seriously wrong with that woman. She's totally unhinged!" Kirk et al. (2023). This example is labelled as *sexist* under the categories of *descriptive attacks* and *aggressive and emotive attacks*. However, in the absence of contextual information, this comment could be interpreted as a critique rather than a clear instance of sexism directed towards women.

Another prominent dataset utilized in sexism classification research is the sEXism Identification

in Social neTworks (EXIST), associated with a series of scientific events and shared tasks focusing on sexism identification and classification in social networks Plaza et al. (2024). As of 2023, the dataset consists of approximately 9,400 tweets from Twitter in each language, English and Spanish, ranging from September 2021 to September 2022. The EXIST dataset categorizes tweets into five classes for the multi-class sexism classification task: (1) ideological and inequality, (2) stereotyping and dominance, (3) objectification, (4) sexual violence, and (5) misogyny and non-sexual violence. However, upon closer examination, the researchers explain that they developed this taxonomy through empirical observation of the scraped data in their original research paper Rodríguez-Sánchez, de Albornoz, and Plaza (2020). They recognize that the original dataset and taxonomy construction were solely based on Spanish tweets, potentially resulting in omitting categories in English tweets. Additionally, their official website cites prior research, such as Donoso-Vázquez and Rebollo-Catalan (2018) and Anzovino, Fersini, and Rosso (2018), and references to feminist literature like Manne (2018) in their general description, but lacks detailed information supporting their taxonomy. These insights determine significant concerns within the sexism classification research field, particularly regarding validating the constructed taxonomies and datasets.

As a result, Kalra and Zubiaga (2021) utilized the EXIST 2021 dataset in their sexism classification research and identified several issues with its labelling. They noted instances where specific labels appeared inappropriate from their perspective. Detailed analysis revealed numerous examples that they believed were inaccurately labelled. For example, they highlight the following excerpt: "kaliati says it is unfortunate that a day hardly passes without hearing of a case of rape defilement and other violence against girls women and children." Despite being categorized as *sexist* in the binary classification task of the EXIST 2021 dataset, the researchers argue that it does not constitute a sexist remark. Another instance they discussed was a passage labelled as *sexual-violence* in the dataset: "leaders stop normalizing sexual harassment it's not okay do not call it fine or normal it is unacceptable." In contrast, the researchers contend that this text does not adequately reflect a sexist statement. These examples present some deficiencies in the widely used EXIST dataset, as the labelling lacks clarity or justification, which could introduce noise and undermine the quality of studies in sexism classification research.

Finally, the work by Guest et al. (2021) introduces the Expert Annotated Dataset for the Detection of Online Misogyny, created by scraping posts and related comments from 34 misogynistic subreddits including *masculism*, *TheRedPill*, and *badwomensanatomy*. They developed a novel hierarchical taxonomy for online misogyny, annotating 6,567 labels with robust annotation guidelines. At its lowest level, the taxonomy comprises twenty classifications of misogyny such as *threatening language*, *privacy*, *controlling*, and *manipulation*. Unfortunately, the researchers note that 88.6% of their dataset labels as *None of the categories*. They further explain that their taxonomy relies on prior research, such as Vidgen et al. (2019) and Anzovino et al. (2018), which do not exclusively focus on supported research of sexism or misogyny. This limitation suggests that their categories may not fully capture the complexities of sexism and misogyny as understood within feminist discourse.

The dataset exposes several ambiguities and inconsistencies in its classification, thereby raising significant concerns. For instance, the entry containing the text "They want our money and peace of mind. I say, come and get it!" is classified as *misogynistic* under the *moral inferiority* category. This classification appears to be based on the phrase "*they want our money*". However, the absence of contextual information that directly associates this comment with women renders the misogynistic label questionable. Similarly, another entry, "Telling them to hit the treadmill is better," is categorized as *misogynistic* under *sexual or physical limitations*. The lack of contextual background for this text further complicates the application of the misogynistic label, rendering it ambiguous and problematic.

A pitfall of recent literature in sexism classification is the lack of reference to research that adequately dissects sexism in established taxonomies. Moreover, popular datasets in this field often need more expert scrutiny to exhibit correct labels and class imbalances. Therefore, in this study, we review the literature on sexism and sexism classification to construct a taxonomy representative of the types of sexism defined by experts in the field. Additionally, we focus on specific data points related to the context of women software developers' experiences found in the online narratives rather than concentrating on general forms of sexism and misogyny, which are present in most current text-based datasets.

Furthermore, our approach diverges from existing practices by focusing specifically on narratives illustrated by self-identified women software developers on Reddit. These narratives provide

contextual details about events and emotional responses related to sexism in software development. Unlike many current datasets that predominantly capture overtly misogynistic remarks, our methodology aims to capture nuanced experiences that encompass to the broader impact of sexism in software development. This shift in focus allows us to explore and categorize sexism within a specific professional context, offering insights that can potentially enhance the understanding and mitigation of sexism in SE and related software development fields.

2.3.2 Models

In the field of sexism classification, numerous recent studies employ a combination of machine learning approaches and language models in their methodology. For instance, Karthikeyan, Sundarraj, Sampathkumar, Mouthami, and Yuvaraj (2023) initiated a binary and multi-class sexism classification task using the English dataset from the EXIST 2022, which includes 10,210 entries for the training dataset and 1,135 values for the testing dataset. They individually evaluate four machine learning models: Logistic Regression, Linear SVC, Multinomial Naive Bayes, and Random Forest. Furthermore, their research uses feature representations such as Bag-of-Words, Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, and Bidirectional Encoder Representations from Transformer (BERT). Their findings revealed that the optimal model for the binary classification task is the Logistic Regression model, achieving an F1-score of 0.74. Moreover, the Linear SVC model achieved the highest F1-score of 0.50 for the multi-class classification task. Therefore, the results in this and related research papers using traditional machine learning approaches determines the performance limitations and denotes potential enhancements of leveraging advanced technologies, such as language models, to more effectively address the intricacies of sexism classification.

Additionally, Das et al. (2023) employed machine learning and natural language processing techniques, like sentence-BERT (sBERT) and word2vec, to integrate user gender information with textual features for tasks such as binary and multi-class classification of sexism in social media content. Furthermore, they express that they pre-trained their model using SemEval 2023 to perform the binary and multi-class classification tasks described in Section 2.3.1. Their dataset includes 6,796 posts for training, 972 for validation, and 1,940 for testing data. The study also utilizes the

PAN 2015 gender prediction English dataset to incorporate a gender feature into the model's training phase. Their results present an optimal binary classification accuracy of 80.97% with gender-inclusive embeddings and 79.38% without, while multi-class tasks achieved 64.43% and 63.60%, respectively, under similar conditions.

Butt, Ashraf, Sidorov, and Gelbukh (2021) conducted a study on sexism classification using the multilingual dataset EXIST 2021, comprising English and Spanish tweets. They performed binary and multi-class classification tasks based on the taxonomy provided by the dataset. The researchers augmented the original dataset of 6,977 texts, resulting in 13,954 data points. They evaluated their models using ten-fold cross-validation and compared the performance of various classifiers: Logistic Regression, Multilayer Perceptron, Random Forest, Support Vector Machine, 1D Convolutional Neural Network (1D-CNN), Long Short-Term Memory (LSTM), and BERT. Furthermore, they explored different feature representations such as word n-grams, character n-grams, and GloVe pre-trained embeddings. Their best results showed an F1-score of 78.02% for sexism identification and 49.08% for sexism classification with BERT when augmented with additional data.

Kalra and Zubiaga (2021) performed a comparative study of various model architectures, including Bag of Words (BOW), GLoVE embeddings, Long Short-Term Memory (LSTMs), Bidirectional LSTMs, Convolutional Neural Networks (CNNs), as well as BERT and DistilBERT models with additional data augmentation, to classify sexism in tweets and gab posts using the EXIST 2021 dataset. Their training dataset comprises 3,436 tweets, while the test set includes 1,716 tweets and 492 gab posts. Apart from the mentioned feature representation techniques, the authors do not report any additional features beyond those present in the dataset (i.e., the full tweet and gab content and the corresponding task labels). Their results indicated that combining BERT with a multi-filter CNN achieved the highest performance, with an F1-score of 0.760 for binary classification and 0.519 for multi-class classification with data augmentation.

Finally, recent studies adopting methodologies similar to ours in the sense of leveraging Reddit entries (i.e., posts and comments) primarily focus on identifying and categorizing misogynistic content. For instance, as discussed in Section 2.3.1, Guest et al. (2021) explored misogynistic subreddits to develop a novel hierarchical taxonomy for online misogyny by building on prior research and undiscovered Reddit content. Their dataset comprises 6,567 labelled data points annotated

by trained annotators following robust guidelines. Notably, their dataset is composed of 89.4% of non-misogynistic content, with only a tiny representation per taxonomy classification. Their work executes a comparison between logistic regression, BERT, and BERT with class weights that emphasize the dataset's minority class to adjust for the dataset imbalance. They pre-trained their models using a stratified 80/20 train/test split of the dataset. The model considers features such as the full online text, a *span* feature that includes specific text describing the type of sexism relevant to their taxonomy, and the corresponding labels to perform the predictions. Based on their experiments, the results identified that the optimal model was the weighted BERT model, achieving an F1-score of 0.43. Although such related studies could help determine misogynistic content resonating from toxic masculinity, they do not consider sexism experienced from females' perspectives, thus ignoring the social implications imposed on the victims of misogyny.

Due to the complexity of the task, sexism identification and classification pose significant challenges in achieving high-performance scores. Prior research highlights these difficulties by showing that achieving high performance in sexism classification remains elusive even with extensive fine-tuning on large datasets. This underscores the requirement for novel approaches to improve the predictive performance of the model. In response to these hurdles, our methodology takes a fresh approach by deliberately avoiding using pre-established taxonomies and datasets, such as SemEval and EXIST. Instead, we develop our own simplified taxonomy grounded in the literature on sexism and misogyny and utilize online Reddit resources that have not been previously explored in this field of study. This approach allows us to construct a more tailored framework and to analyze the unexplored data sources.

Finally, it is crucial to recognize the limitations of AI approaches that this software-focused literature review does not fully address. Numerous studies and books explore these limitations from a social science perspective Crawford (2021); Eubanks (2019); Noble (2018). The critical insight from this research is that current AI technologies cannot fully replace human judgment in categorizing various forms of sexism. The literature highlights that social scientists struggle to agree on definitive classifications of such complex phenomena and that we should not expect technology alone to resolve these issues. Therefore, it is expressed that AI should be used as a supplementary tool, working alongside experts to analyze results and make informed decisions based on the insights

provided by the technology.

Chapter 3

The Concept of Semantic Similarity

This chapter strives to provide background information regarding the operations behind the semantic similarity approach employed in this research. Semantic similarity is a technique that aims to quantify the relative meanings and contexts of text elements such as words, phrases, sentences, or documents. While traditional approaches like n-grams and syntactic parsing concentrate on syntactic structures, semantic similarity investigates the deeper, underlying meanings of the provided texts. As such, semantic similarity has a variety of applications, including natural language processing, information retrieval, and recommendation systems.

In this thesis, we conduct the process of semantic similarity in three sequential steps: use a transformer model to generate the embeddings for each piece of input text, incorporate a pooling technique to render the embeddings into fixed-length representations, and calculate the cosine similarity score between the two input texts to determine their relevancy. This approach follows a similar implementation to Reimers and Gurevych (2019) with the adjustment of including a pooling layer for further data refinement. Figure 3.1 displays a simplified diagram detailing the explained chain of events.

3.1 Transformers

Google introduced transformers as a deep learning architecture in their 2017 paper, *"Attention Is All You Need"* Vaswani et al. (2017). A *transformer* uses a semi-supervised learning approach

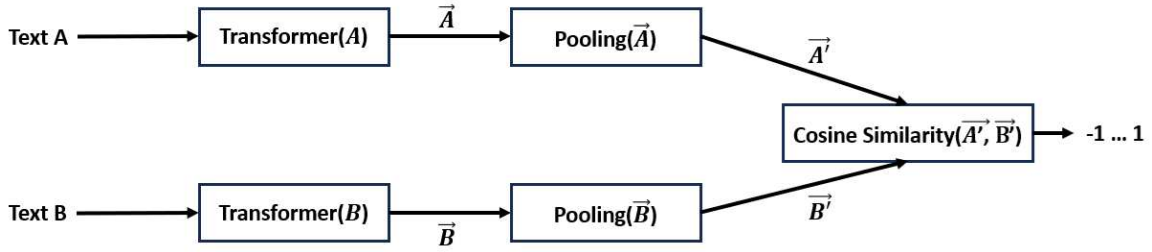

Figure 3.1: An overview of the implemented semantic similarity approach.

relying on pre-trained knowledge from a large corpus of data which can then be fine-tuned for specific tasks. The advantages of the architecture are that it utilizes parallelization to process the input tokens in a non-sequential manner and exhibits a self-attention mechanism that enables the model to consider each word's context in the supplied input text.

A transformer model uses a given number of encoders and decoders of equal quantity. The encoder processes the input sequence (e.g., a sentence containing ordered words) into a set of vectors that capture the context of the input. The encoder's result is supplied to the decoder, which generates output sequences based on the encoder's learned representations. The model repeats this operation until it has processed the entire input.

There are several variations in the implementation of the transformer architecture. For this thesis, we employ the Bidirectional Encoder Representations from Transformers (BERT) model produced by Google in 2018 to generate the vector representations of the input text for the semantic similarity task Devlin, Chang, Lee, and Toutanova (2018). Thus, the BERT model only utilizes the encoder component to output the vector representations.

Specifically, for this study, we implement the sentence-BERT (sBERT) model proposed by Reimers and Gurevych (2019) to convert large input texts containing numerous sentences to adequately resemble the structure of the detailed online narrations. As shown in Figure 3.1, this description entails the original texts, denoted as $\mathbf{TextA}$ and $\mathbf{TextB}$, are inputs to the transformer model. The model then refines these texts into vector representations, also called embeddings.

3.2 Embeddings

In the context of this thesis, an embedding is the output of the transformer model that numerically represents textual data using vectors in a continuous space. The placement of the vectors in a continuous space dictates the semantic relevancy of the data points through the observation of distances. As such, if several vectors are near each other, there is a high possibility that the corresponding texts will discuss a similar topic. This approach is a crucial component in semantic similarity as it depends on the selected model's ability to generate adequate vector representations to determine the contextual relevance between the data points. Therefore, concerning Figure 3.1, this explanation corresponds to the *Transformer*'s output embeddings $\vec{\mathbf{A}}$ and $\vec{\mathbf{B}}$ which the pooling layer then uses as inputs.

3.3 Pooling

The pooling operation aims to aggregate the information from the generated embeddings into a fixed-length vector representation per input text. The pooling layer performs a necessary function by ensuring that the vectors have equal cardinalities, which allows for effective similarity comparisons. Several types of pooling operations can also help influence the semantic similarity between input texts. For instance, *average pooling* is a technique that calculates the average of all embeddings present in a given vector to better help identify the general context of the provided input texts. Furthermore, *max pooling* is used to extract the highest values in the embeddings to capture the most prominent topics of the input texts. In this thesis, we use max pooling because our preliminary analysis indicates that the target online content often consists of lengthy texts covering a broad range of topics. Max pooling is preferable to average pooling in this context, as it accounts for numerous topics and helps identify texts containing narrations related to women software developers' shared experiences with sexism. As presented in Figure 3.1, the pooling layer accepts the embeddings mentioned above $\vec{\mathbf{A}}$ and $\vec{\mathbf{B}}$ to perform the required modifications based on the selected type of pooling and generates the vectors $\vec{\mathbf{A}}'$ and $\vec{\mathbf{B}}'$, which are ready to be evaluated by the cosine similarity metric.

3.4 Cosine Similarity

The final step of the semantic similarity process involves computing the cosine similarity score between two input texts to gauge their relevance. This score measures the cosine of the angle between the two embeddings (i.e., vectors), indicating the extent to which they align within a similar direction in the multi-dimensional continuous space. We calculate the cosine similarity score using the following equation:

$$\text{cosine similarity}(\vec{\mathbf{A}}, \vec{\mathbf{B}}) = \frac{\vec{\mathbf{A}} \cdot \vec{\mathbf{B}}}{\|\vec{\mathbf{A}}\|\|\vec{\mathbf{B}}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

The result is a floating-point value ranging between -1 and 1, where a positive value signifies a semantic resemblance between the two vectors. Moreover, a negative value suggests dissimilarity in the embeddings' semantic content. As such, Figure 3.1 illustrates this process, where the computation accepts the pooling layer's generated vectors $\vec{\mathbf{A}}'$ and $\vec{\mathbf{B}}'$ and outputs the cosine similarity score.

Chapter 4

Methodology

In this chapter, we divide the methodology into four sequential subsections. First, we explore research on sexism and misogyny to select a pertinent feminist theoretical resource to construct our taxonomy with related lexicons that detail various types of sexism potentially prominent in SE and associated software development domains. We then use these keywords to extract Reddit content containing at least one of them. Next, we analyze the raw dataset with semantic similarity tasks to filter for content relevant to women software developers' experiences of sexism while collaborating in the field. Finally, four members of Concordia University's EDI lab evaluate the model's performance to ensure the technology's classification abilities match reality through human-AI interactions. Figure 4.1 provides a visual overview of these steps.

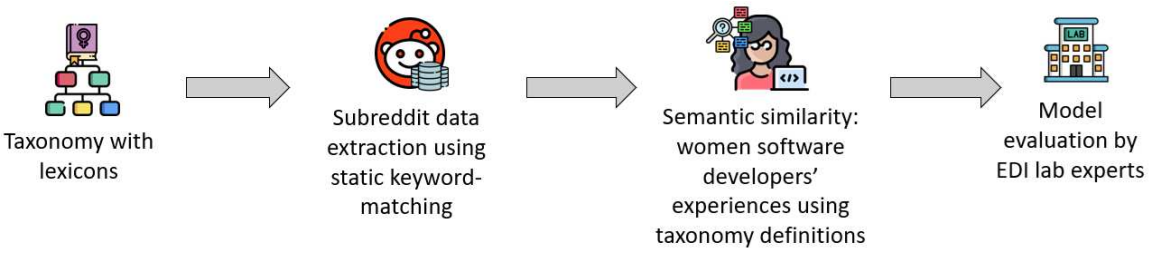

Figure 4.1: Methodology overview

4.1 Taxonomy with Lexicons

The study begins by utilizing foundational concepts from Kate Manne's *"Down Girl: The Logic of Misogyny"* Manne (2018) to develop our taxonomy. As outlined in Section 2.1, Manne's work is distinguished by its comprehensive definitions and relevant keywords, which are well-supported by prior research. Her exploration of misogyny and sexism across various public spheres, including workplace environments, makes her theories particularly pertinent to our research objectives. Consequently, Manne's framework is a primary resource for constructing our taxonomy. As such, in this study, we adopt the concept of sexism for its broader applicability in capturing the diverse experiences of women practitioners in software development. Additionally, we chose to focus on sexism rather than misogyny, given its more prevalent documentation in the field of software development, as illustrated in Chapter 1. To our benefit, Manne's book offers a list of key terms associated with each definition to elucidate the nature of these concepts. For instance, Manne elaborates on the concept of *feminine-coded goods and services* using the following text:

> "feminine-coded goods and services include simple *respect, love, acceptance, nurturing, safety, security,* and *safe haven*. There is *kindness* and *compassion*, *mortal attention, care, concern,* and *soothing*." Manne (2018)

Furthermore, we expand on the provided lists of keywords by incorporating relevant synonyms that match the context provided by Manne using three online thesauruses: Thesaurus.com Thesaurus.com (2024), Collins Dictionary HarperCollins (2024), and Webster Dictionary Merriam-Webster (2024). The objective of using various resources is to uniquely select as many synonyms per category that follow the context of the definitions without causing an overlap of words. For example, when examining the potential synonyms for the word *acceptance*, Thesaurus.com describes it as "related to the belief in goodness of something" Thesaurus.com (2024). However, Merriam-Webster Dictionary illustrates acceptance as "a form of obedience" Merriam-Webster (2024).

4.2 Data Extraction

We extract the initial dataset by referencing subreddits that specifically focus on the context of women in technology and using an extended list of keywords per taxonomy definition. We utilize eleven subreddits where the virtual community forums explicitly mention in their description the context of women in software development: GirlsGoneWired GirlsGoneWired (2024), WomenInTech WomenInTech (2024), XXSTEM XXSTEM (2024), CSWomen CSWomen (2024), LesbianCoders LesbianCoders (2024), WomenWhoCode WomenWhoCode (2024), PyLadies PyLadies (2024), LadyDevs LadyDevs (2024), ChicksWhoCode ChicksWhoCode (2024), LaunchCoderGirl LaunchCoderGirl (2024), LadyCoders LadyCoders (2024). Then, we use the Python Reddit API Wrapper (PRAW) Boe (2023) library to retrieve the 2,000 *hot*, *top*, and *new* posts and comments from 2018 to 2023 using the lexicons to perform static keyword-matching to extract the initial dataset for further examination. We establish this step as the baseline in our methodology to implement fundamental data extraction techniques, such as static keyword-matching, for capturing a wide range of topics from online content while potentially including our target narrations. This dataset enables us to evaluate the effectiveness of semantic similarity in identifying relevant data points that illustrate women software developers' experiences of sexism while filtering out irrelevant content.

4.3 Semantic Similarity Analysis

As a general overview, we divide our semantic similarity analysis into two tasks. Task 1 involves fine-tuning a model using a small dataset to identify content that closely aligns with a generalized definition of negative experiences among women software developers. Subsequently, we utilize these results to select the most pertinent subset of data for Task 2, which applies a zero-shot classification model that evaluates the online content with the taxonomy's four definitions and anchor examples. We opted for a zero-shot classification approach due to the absence of taxonomy class examples that precisely match the contextual nuances in the Reddit narratives. Furthermore, we equip our models with comprehensive inputs that include context-related keywords and influential adjectives to specifically concentrate on content relative to women software developers experiencing sexism. This enhancement aims to facilitate the discovery of content that precisely meets our

specific requirements.

The raw dataset undergoes pre-processing by removing special characters and syncategorematic words (i.e., is, a, the) along with tokenization. Each data point represents the full extracted text from a post or comment that includes at least one keyword from the established lexicons. As the dataset originates from online discourse and contains a variety of contexts, we initiate Task 1 by applying semantic similarity to expose data points that describe women software developers' experiences of sexism. We use the BERT sentence transformer (sBERT) with the *all-mpnet-base-v2* model Reimers and Gurevych (2019) coupled with a max pooling layer. As illustrated in Section 3.3, we select max pooling because it enables us to identify online content containing topics related to the constructed definitions. Furthermore, we train the model using 200 manually-labelled data points, where 120 are true instances of women software developers' experiences of sexism, and 80 data points do not illustrate such events. These data points are obtained from the mentioned subreddits and removed from the input dataset to avoid data leakage. The model interprets the dataset by analyzing features related to the full text of the Reddit content, followed by the associated binary label. We fine-tune the model using sBERT's *CosineSimilarityLoss* and *EmbeddingSimilarityEvaluator* functions over ten epochs while using three warm-up steps and five evaluation steps. Then, we generate sentence embeddings for the input dataset and a generalized definition of women software developer experiences to calculate the cosine similarity scores. The authors of this research manually construct the generalized definition and describe it as a first-person narrative while highlighting negative sentiment, similar to the target data points found in the input dataset, as follows:

> "As a female software engineer, woman in tech, and woman software developer, I've experienced challenging situations while collaborating with colleagues in my teams. These include encountering sexism and navigating a hostile environment."

The generalized definition purposely mentions various ways to describe a woman software developer so that the model can better grasp the context used in the online narratives. Furthermore, the generalized definition concentrates on negative experiences to prioritize potential areas of sexism instead of including positive experiences. Continuing, we calculate the cosine distance distributions

with the embeddings of the cleaned dataset and the generalized definition to set a suitable threshold to extract the portion of the dataset exhibiting the lowest cosine distances. This process aims to retrieve data points that most likely describe women software developer's experiences of sexism. Next, we conduct Task 2 by using the derived dataset and a separate sBERT with the *all-mpnet-base-v2* model Reimers and Gurevych (2019) to perform a zero-shot classification task by evaluating the cosine similarity score of the sentence embeddings with each of the rendered taxonomy definitions to match the context of women in software development and the corresponding anchor examples constructed by the authors. Moreover, the model interprets the subset of data by analyzing features related to the full text of the Reddit content, followed by the corresponding taxonomy label. The objective of this task is to identify specific instances of sexism reported by women software developers that align with the forms of sexism outlined in the taxonomy. The implemented source code for the baseline, Task 1, and Task 2, along with instructions and the utilized datasets, are available for further analysis in the referenced public GitHub repository [1].

Finally, to showcase the performance of each task, we calculate the model's precision in the top 10, 50, and 100 highest cosine similarity scored data points through manual evaluation. The rationale of using a metric that evaluates the model's precision in the top k value is that we are not aiming to extract every possible experience of sexism in the dataset but rather attempting to gather a suitable number of similar experiences to coincide with the taxonomy. Hence, we cannot use metrics such as recall and F1-score in this particular case, as it would not be feasible to manually determine all true positives in a substantially large dataset.

4.4 Model Evaluation

We select five potential data points pertinent to each illustrated definition, summing to twenty narrations, and evaluate the analysis with experts from Concordia University's EDI lab, consisting of of four women with diverse academic backgrounds and levels, using the Delphi technique Christie and Barela (2005). The Delphi technique is a method that collects experts' perspectives on

[1] Kolopanis, (2024). Decoding-Bias-Exploring-Sexism-in-Software-Development-through-Online-Narratives-and-AI-Analysis. GitHub. https://github.com/Amanda-Kolopa/Decoding-Bias-Exploring-Sexism-in-Software-Development-through-Online-Narratives-and-AI-Analysis/tree/main

a specific subject. For this particular instance, the method involves a series of structured examples extracted from the initial dataset, where the experts can provide their opinion on which classes of sexism are most relevant to the examples. This iterative process typically involves detailed discussions to aim for a group consensus on the appropriate classification Christie and Barela (2005). As such, for this thesis, we supply the iterators with a Google Form containing the twenty data points and options to select one of the four taxonomy categories. Additionally, we include an option called *Other* that enables the iterators to provide an alternative perspective if they believe the taxonomy does not meet the criteria. We evaluate the inter-rate agreement of the four participants to denote the percent agreement of exact matches between their classifications. Furthermore, using the labeling outcomes of the Delphi technique, we compare the average F1-score with the corresponding standard deviation of each iterator and the model to assess the alignment between humans and technology's understanding of sexism. Finally, we calculate the p-value for the predictions provided by each iterator and the model using scikit-learn permutation test at a significance level of $p = 0.05$ Gramfort and Liu (2024). For context, this test evaluates the extremity of the observed results by comparing them to a distribution of outcomes generated through random data permutations Mayo and Hand (2022). A p-value of 0.05 implies a 5% probability that the observed results could arise by random chance alone. If the p-value is below 0.05, it indicates that the results are statistically significant, which suggests that they are unlikely to be attributed to random chance.

Chapter 5

Results

This thesis aims to investigate the diverse forms of sexism women encounter in SE and affiliated software development domains. Our research endeavours not only to bridge the gender gap by identifying factors contributing to sexism against women in the field but also to evaluate the limitations of AI-based methods for detecting and classifying sexism, such as those relying on semantic similarity with advanced language models. We also examine the challenges in aligning AI systems with human interpretations for such a nuanced task. To accomplish these goals, we have formulated three research questions (RQ1, RQ2, and RQ3) that serve as the framework for our inquiry.

RQ1 aims to identify the predominant categories of sexism prevalent within the software development field by serving as an initial step toward uncovering additional potential categories. This investigation forms the basis for our comprehension of sexism categories by enhancing our ability to recognize the likely types of sexism in online content. It also allows us to draw upon relevant research, such as the studies referenced in Section 2.2, that documents experiences of sexism in the software development domain while aiding us in identifying the most probable categories.

RQ2 seeks to assess the efficacy of semantic similarity in identifying and categorizing narratives of sexism as recounted by women in software development. Our approach diverges from conventional semantic similarity methods in sexism classification research. We utilize unstructured textual data extracted from diverse subreddits by encompassing discussions related to women's personal involvement in technology.

RQ3 aims to explore the complexities of aligning AI systems with the diverse interpretations

of sexism outlined in our taxonomy, specifically in the context of experiences narrated by women software developers. Our goal is to inform practitioners of the challenges in identifying the most influential aspects that contributed to our final results.

Therefore, the following sections provide insights into the motivation behind each research question, the approach taken to address each question, and the corresponding results that answer each question comprehensively.

5.1 RQ1: What categories of sexism are women most likely to experience in the software development field?

Our motivation behind RQ1 arises from the necessity to address the prevalent gender gap in the software development sector. As explained in Chapter 1, the lack of women in software development fosters negative societal perceptions about women's contributions to technological innovation and leads to significant technical repercussions. The absence of women's perspectives can result in technology that is embedded with gender biases, which in turn impedes model advancement and reinforces a cycle of sexism that affects end users. Therefore, our goal is to investigate identified forms of sexism in existing software development research, in conjunction with relevant literature on the notion of sexism, to gain deeper insights and contribute to bridging the gender gap.

Our approach to investigating RQ1 begins with a comprehensive review of documented experiences of sexism in SE and related fields to identify existing challenges, as detailed in Section 2.2. We then build on this foundation and review prominent literature on sexism and misogyny to grasp a better understanding of the notion of sexism, as presented in Section 2.1. We use *"Down Girl: The Logic of Misogyny"* Manne (2018) as our primary resource to develop our taxonomy that focuses on a limited set of categories central to our methodology. We believe Manne's work meets our research objectives by providing structured definitions and lists of keywords supported by prior research. Therefore, this taxonomy and associated lexicons are instrumental in guiding our data extraction process. Moreover, we aim to assess how well the selected Reddit data aligns with our theoretical framework of sexism.

Our results produced a taxonomy identified as *Sexism in Software Development* that describes

four distinct classifications from our identified primary source and affiliated research. The definitions provided by Manne are rendered to suit the context of the target data of women in software development. Furthermore, the corresponding anchor examples that the authors establish are detailed as follows:

- **Feminine-Coded Goods and Services**: The characteristics that women are expected to naturally provide to men because they are entitled to receive the benefits of women's goods and services. Moreover, these characteristics are used to reinforce traditional gender roles. For example, care-mongering is when women are disproportionately required to be caring and are expected to develop personal relationships with individuals Manne (2018).

    - *Anchor example*: I am the only woman in our dev team and I am always implicitly expected to do the administrative tasks during our meetings. When I confront my team about this, they explain that my organization and note-taking abilities are a natural talent that benefits the team.

- **Gendered Split Perception**: Women are judged more harshly when performing the same actions as their male counterparts even though they have done nothing wrong in moral and social reality. Women may be subject to moral suspicion and consternation for violating edits of the patriarchal rule book Manne (2018).

    - *Anchor example*: As a female software engineer, I feel like my source code is heavily scrutinized by my male teammates. When I submit similar work as my male co-workers, I tend to receive more critiques compared to my colleagues despite our work being identical in logic and performance.

- **Testimonial Injustice**: Arises due to systematic biases that afflict women as a social group that has historically been and to some extent remains unjustly socially subordinate. The group members tend to be regarded as less credible when making claims about certain matters, or against certain people, hence being denied the epistemic status of knowers Manne (2018).

    - *Anchor example*: I am a woman software developer. I find that when I present an idea to

my development team, they often ignore my input. However, when my male colleague repeats the same idea in a follow-up meeting, the team almost immediately accepts them.

- **Social Dominance Penalty**: People are (often unwittingly) motivated to maintain gender hierarchies by applying social penalties to women who compete for, or otherwise threaten to advance to, high-status, masculine-coded positions. This is demonstrated when women in such positions who are agentic (i.e., competent, confident, assertive) are perceived as extreme in masculine-coded traits like being arrogant and aggressive Manne (2018).

    - *Anchor example*: As one of the female programmers in our team, I sometimes experience a sense of hostility when I provide constructive criticism or potential improvements to my male counterparts' source code. I give the same type of feedback to my female colleagues and receive praises.

It is important to note that the authors constructed the anchor examples using data reviewed during the manual analysis process. While they may not encompass every type of incident or experience within each category, they serve as a foundational starting point for understanding various forms of sexism. Moreover, they play a crucial role in our semantic similarity approach by highlighting keywords essential for retrieving contexts relevant to software development and women programmers.

Based on the established taxonomy, the associated lexicons from Manne's work yield 100 primary keywords, each corresponding to a specific category within the taxonomy. The detailed distribution of the collected keywords and corresponding synonyms is displayed in Figure 5.1. Essentially, the *feminine-coded goods and services* classification has a total of 184 keywords, the *gendered split perception* category contains 85 keywords, *testimonial injustice* assigns 201 keywords, and *social dominance penalty* possesses 191 keywords. Therefore, the list comprises 661 keywords to extract the initial dataset. Note that the *gendered split perception* category contains the least relative keywords, consequently influencing the succeeding results compared to the other classes. All identified lexicons are available in Appendix A for a comprehensive review.

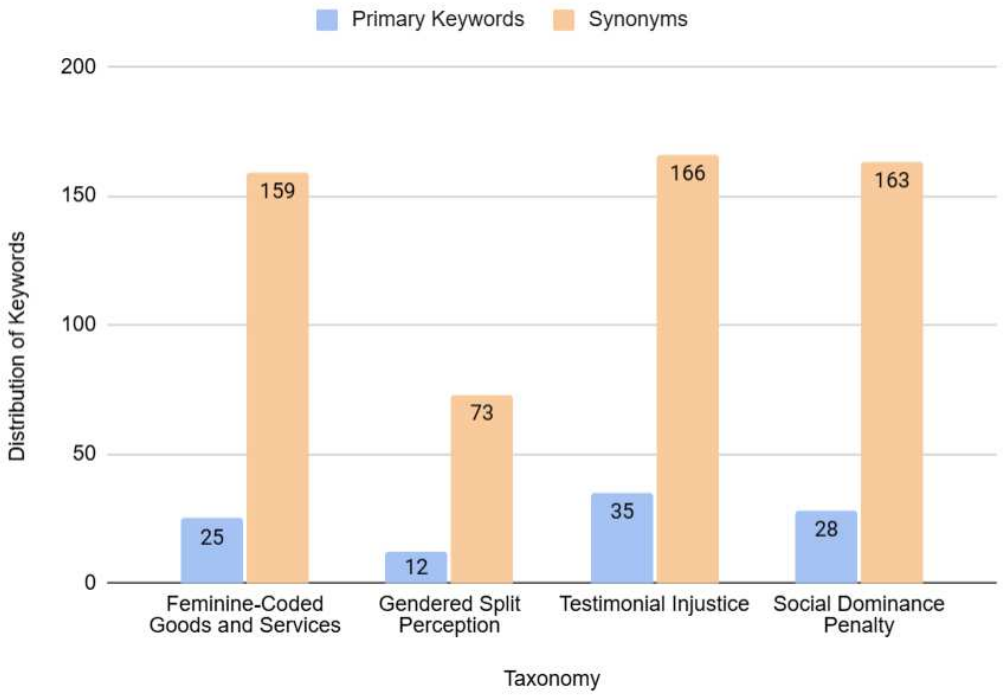

Figure 5.1: The distribution of primary keywords with associated synonyms per taxonomy category.

5.2 RQ2: To what extent can semantic similarity effectively extract the experiences of women in software development that align with the constructed taxonomy?

Our motivation for RQ2 revolves around evaluating the effectiveness of semantic similarity in identifying and categorizing instances of sexism experienced by women in software development from a vast corpus of unstructured online data. Drawing from our literature review on sexism classification in Section 2.3, we observed that many researchers have turned to advanced language models, like BERT, to achieve such tasks with adequate results. The benefits of such models demonstrate superior capability in understanding textual data, offering the potential for fine-tuning on labelled datasets and outperforming traditional machine learning approaches. However, due to the unavailability of a suitable pre-labelled dataset for our specific research objectives, we decided to employ semantic similarity techniques to a large, unlabeled dataset to identify content likely to contain first-person narratives from women software developers detailing their encounters with sexism in the

domain.

In addressing RQ2, we divide our approach into three distinct tasks. Initially, the baseline aims to evaluate the efficacy of a non-AI method, such as static keyword-matching, to extract potential data points from eleven selected subreddits. Using the Python Reddit API Wrapper (PRAW) Boe (2023) library, we retrieve the 2,000 *hot, top,* and *new* posts and comments from 2018 to 2023. Then, we extract data containing at least one term from our constructed lexicons. We deliberately employ this rudimentary keyword-matching extraction technique to encompass a broad spectrum of topics that may not directly relate to sexism. This intentional inclusion aims to generate a large dataset that allows us to thoroughly assess the data filtration capabilities of semantic similarity. Although alternative extraction methods could align more closely with taxonomy definitions, RQ2 prioritizes evaluating semantic similarity with a complex, unlabeled dataset that requires minimal fine-tuning. However, future researchers may explore enhancements in this area as needed.

Following this, Task 1 involves identifying potential data points related to sexist experiences narrated by women software developers. We fine-tune the sBERT model using the *CosineSimilarityLoss* and *EmbeddingSimilarityEvaluator* functions over ten epochs while using three warm-up steps and five evaluation steps. We train the model on 200 manually-labelled data points, where binary labels indicate whether the text narrates an experience of sexism by a self-identified woman software developer. Therefore, in reference to Figure 3.1, $\mathbf{TextA}$ represents the manually-constructed generalized definition of women software developer experiences, as illustrated in Section 4.3, while $\mathbf{TextB}$ represents a data point from the outcome of the baseline approach. Subsequently, using a cosine distance threshold (0.40 in this case), we extract a subset of the data most relevant to our study. Finally, Task 2 employs a zero-shot classification sBERT model on the subset of data, which aims to assign the text to the most suitable taxonomy category based on the highest cosine similarity score. The model incorporates the provided taxonomy definitions and anchor examples for comparison within the subset. Therefore, in reference to Figure 3.1, $\mathbf{TextA}$ represents a taxonomy definition with its corresponding anchor example, as illustrated in Section 4.3, while $\mathbf{TextB}$ represents a data point from the outcome of Task 1. We use the top k precision metric to assess the results of each task through manual evaluation to ensure the models align with the thesis goals. Additionally, we use permutation testing to determine the statistical significance of the results to

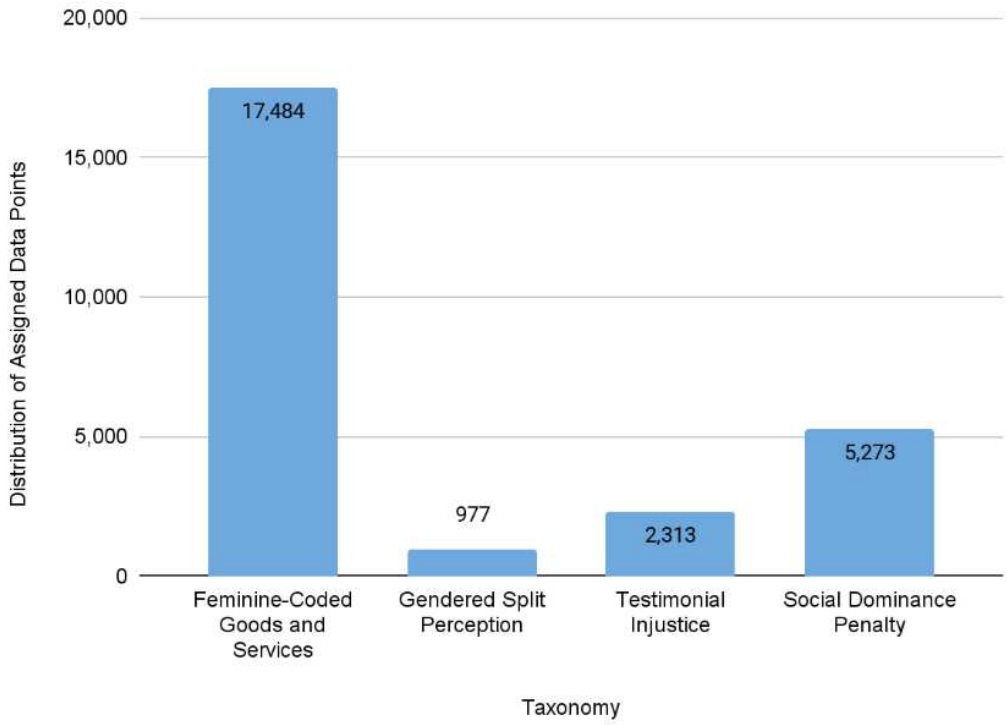

Figure 5.2: The results of the baseline approach that describe the distribution of classified data points using the static keyword-matching approach on eleven subreddits.

distinguish genuine findings from chance occurrences.

5.2.1 Baseline: Static Keyword-Matching Approach

Our data collection results using a static keyword-matching approach generate the taxonomy distributions presented in Figure 5.2, where the total number of extracted posts and comments is 26,047. As a reflection, each data point represents the full extracted text from a post or comment that includes at least one keyword from our lexicons. Additionally, Table 5.1 presents the results of the top 10, 50, and 100 precision scores when applying the static keyword-matching approach to the online content. The outcomes in the table dictate that the static keyword-matching approach helps extract potentially relevant data instead of solely utilizing semantic similarity over an abundance of online content. However, the results denote that the approach is insufficient for detecting content relative to women software developers' experiences pertinent to the constructed taxonomy.

A notable effect of applying static keyword-matching identifies that over half of the dataset

classifies under the *feminine-coded goods and services* category. However, this distribution does not accurately represent the dataset's reality because the category includes keywords such as *natural, healthy,* and *loyal*, which can loosely apply to other contexts. Upon further inspection, it is evident that the extracted online content contains a variety of topics aside from personal experiences of sexism that do not correspond to the definition of *feminine-coded goods and services*. These topics include, but are not limited to, technical discussions, conversations of upcoming events, and generally positive feedback to narratives. As mentioned in the prior section, the *gendered split perception* class obtained the least amount of assigned data points due to its list of keywords, consequently containing the lowest distribution among the taxonomy. Moreover, these results help to compare our findings with a more traditional approach that does not apply AI practices (i.e., static keyword-matching) and use it as a baseline to determine the extent to which semantic similarity can extract women software developers' experiences in the dataset.

Approach	**Top 10 Precision Score**	**Top 50 Precision Score**	**Top 100 Precision Score**
Baseline	0.10	0.14	0.16
Task 1	0.70	0.40	0.45

Table 5.1: The top k precision scores of the baseline approach compared to Task 1 in detecting women software developers' experiences.

5.2.2 Task 1: Semantic Similarity to Extract Women Software Developers' Experiences

As for the results of Task 1, Table 5.1 illustrates the top 10, 50, and 100 precision scores when using the fine-tuned sBERT model to locate data points describing narratives from women software developers of sexist experiences. The outcome determines an improvement in performance as the top 10 precision score increased by 0.60 while the top 50 and top 100 precision scores rose by 0.26 and 0.29, respectively. Therefore, our results showcase that the semantic similarity approach outperforms the static keyword-matching approach for our research case study. Furthermore, this helps to determine how semantic similarity can efficiently identify women software developers' experiences using a generalized representation of the desired textual data.

We use a permutation test to evaluate the statistical significance of the results from our fine-tuned model. Based on 5,000 permutations, the fine-tuned model's F1-scores range from 0.26 to 0.34. The model exhibits an F1-score of 0.30 on the original dataset with a p-value of 0.539, which represents that the statistical significance of these improvements is marginal. This suggests that while the semantic similarity approach is promising and more effective than static keyword-matching, the evidence cannot conclusively assert its superiority. Further research with enhanced methods or larger datasets is recommended to validate and potentially strengthen these findings.

Furthermore, Figure 5.3 displays the distribution of cosine distances when applying semantic similarity to locate the potential data points explaining personal narratives of sexism from women software developers. As depicted from the red rectangle in Figure 5.3, we select a cosine distance threshold of 0.4 — corresponding to the first mode in the distribution of cosine distances — to perform further examination on the most relevant subset of the data for the succeeding classification task. Therefore, the subset of the data comprises 4,781 records, which eliminates approximately 81.6% (i.e., 21,266 data points) of the dataset potentially containing irrelevant online content for further analysis.

5.2.3 Task 2: Semantic Similarity for Taxonomy Classification

The objective of Task 2 is to utilize the rendered taxonomy definitions and anchor examples composed by the authors as inputs for the zero-shot classification model to identify the most pertinent data points for each category based on the highest cosine similarity scores. Figure 5.4 presents the resulting distribution of data points per category. Once again, the *feminine-coded goods and services* category contains the highest amount of potential data points by accounting for approximately half of the dataset. However, upon further inspection, the distribution contains many data points with the lowest cosine similarity scores across the taxonomy. This outcome is due to the model generally classifying any remaining irrelevant content under the *feminine-coded goods and services* category because of its broader sense of wording compared to the other definitions. Also, the *gendered split perception* class contains the lowest amount of potential data points due to the consequences, as mentioned above, of having the least amount of keywords compared to the rest of the taxonomy. Finally, *testimonial injustice* comprises the second highest distribution across

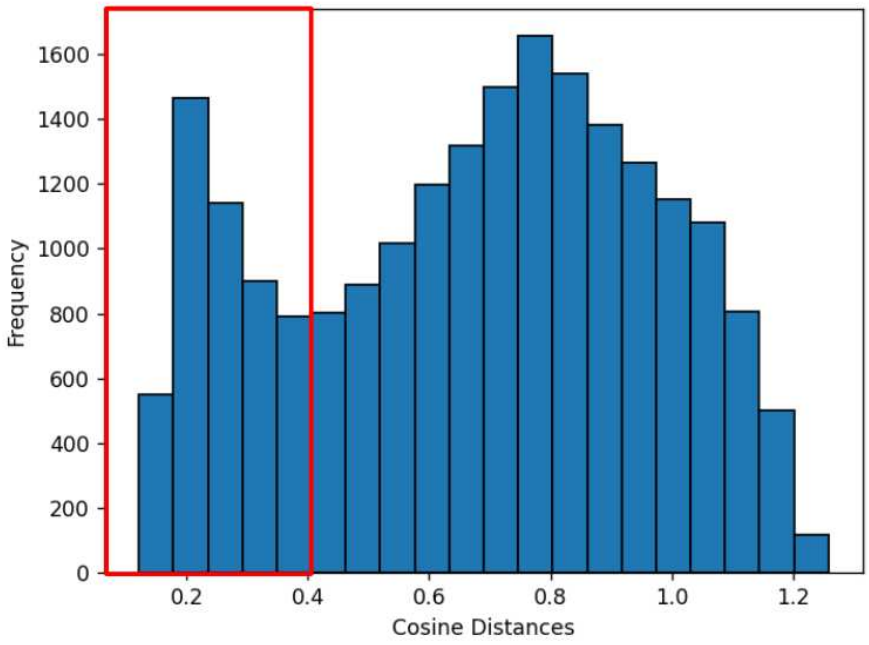

Figure 5.3: The results of Task 1 describing the distribution of cosine distances depicting semantic similarity for experiences of women software developers. The red rectangle highlights the subset of extracted data meeting the threshold of 0.40, which is further analyzed in the subsequent phase of the semantic similarity approach.

the taxonomy. In contrast, the *social dominance penalty* ranks third, contrasting with the results observed during the static keyword-matching portion of the study.

Taxonomy	**Top 10 Precision Score**	**Top 50 Precision Score**	**Top 100 Precision Score**
Feminine-Coded Goods and Services	0.40	0.14	0.14
Gendered Split Perception	0.40	0.10	0.05
Testimonial Injustice	0.20	0.12	0.10
Social Dominance Penalty	0.20	0.12	0.14

Table 5.2: The top k precision scores of Task 2 when applying semantic similarity with the taxonomy classifications.

Additionally, Table 5.2 describes the results for each category and the distribution of the dataset. The *feminine-coded goods and services* and the *gendered split perception* classes exhibited optimal

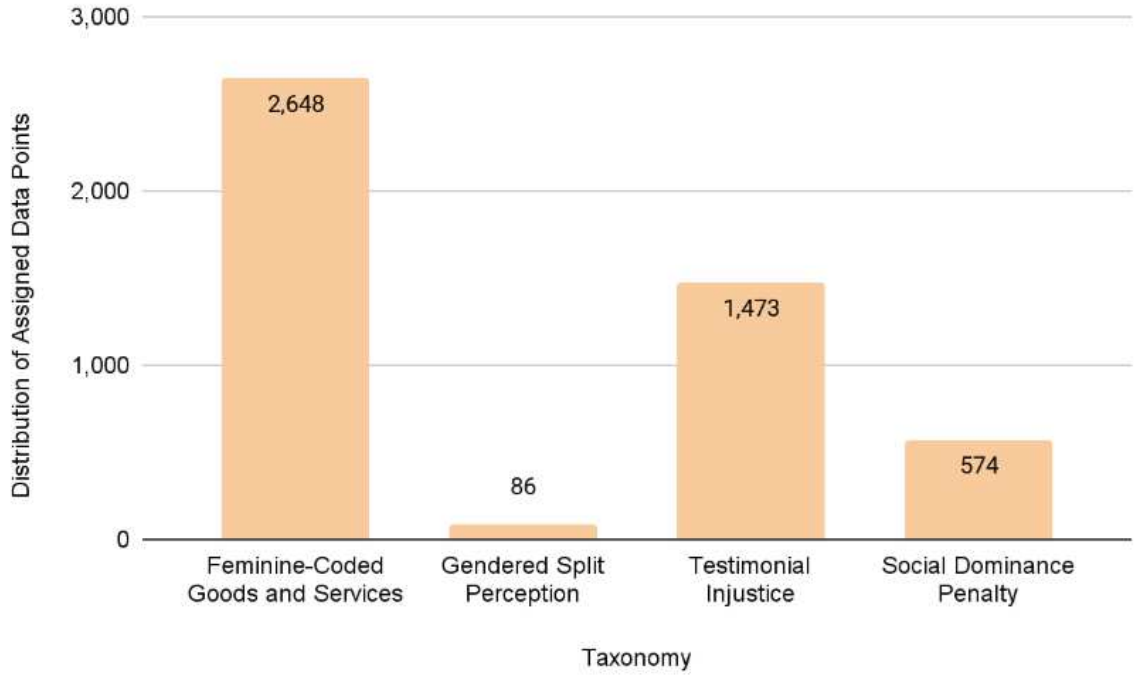

Figure 5.4: The results of Task 2 present the distribution of assigned data points per taxonomy classification using semantic similarity.

performances in the top 10 precision scores. However, *the gendered split perception*'s results drastically reduced when observing the top 50 and top 100 precision scores due to the class only containing 86 data points. Furthermore, the performance of the *testimonial injustice* and *social dominance penalty* categories are seen as the lowest in the taxonomy as the model's prediction is often confused between the two definitions and anchor examples. In the subsequent research question, we further elaborate on the results and details of Task 2 by presenting the permutation test outcomes from the zero-shot classification model in comparison with the findings from the human evaluation process.

5.3 RQ3: What are the challenges in aligning AI systems with human interpretations of sexism?

Our motivation for RQ3 is to gain a deeper understanding of how different elements in the provided input texts influence the model's ability to produce results that correspond with human interpretations of various types of sexism as outlined in our taxonomy. Given the intricate nature

of the social phenomenon addressed in this thesis, which is not easily discernible by humans alone, we aim to explore the challenges of aligning AI systems with human comprehension of the diverse forms of sexism. This investigation seeks to uncover the extent to which AI can effectively meet human standards in understanding and interpreting complex social issues.

Our approach to answering RQ3 involves evaluating the model's ability to classify various forms of sexism compared to Concordia University's EDI lab members. We consulted with four members of the EDI lab, all of whom identify as women and come from varied backgrounds in terms of race, academic status (graduate and undergraduate students), and disciplines, including software engineering, mechanical engineering, and social sciences. The model and the participants are equipped with the authors' constructed taxonomy definitions and anchor examples to guide their classifications. Moreover, we provide the participants with a Google Form containing twenty manually-selected examples from the extracted dataset to be preferably classified in one of the categories [1]. As the model can intentionally classify certain examples using a negative cosine similarity score to indicate that the texts do not fit into any predefined category, we provided an *Other* option for EDI experts to note any uncertainties or offer alternative interpretations. This approach allows us to discern ambiguities in texts and gain insights into different perspectives on sexism. Furthermore, due to time constraints preventing us from employing the complete Delphi technique to achieve unanimous decisions, we relied on majority voting among experts as the ground truth for evaluation. While multiple discussions would have been preferable, we address this limitation in detail in Section 7.1.1. Lastly, we assess the model's performance alongside the EDI experts by comparing their responses using average F1-score, standard deviation, and p-value metrics. Additionally, we present the results of permutation tests to rigorously evaluate the significance of our findings against chance occurrences.

Our survey results in Table 5.3 dictate that the Delphi technique's first iteration could not produce a unanimous decision for most examples. The *feminine-coded goods and services* category received the most votes as the participants classified eight of the twenty examples under this definition. Moreover, the succeeding category, *gendered split perception*, was selected as the dominant

[1] Kolopanis, (2024). Decoding Bias Taxonomy: Manual Classification of Online Narratives. Google Forms. https://docs.google.com/forms/d/e/1FAIpQLSf0_II5BnLUkuGMQBwErHne-f64sAYOmp0SmNTUQFwT9sM2aQ/viewform

option for six of the provided examples. Thus, the experts identified four examples relevant to *testimonial injustice*, while one example identifies with the label *a social dominance penalty*. Also, five examples contained the *Other* option where the participants struggled to decide between two or more classifications and stated which parts in the examples influenced their decision based on the taxonomy definitions. Therefore, as the inter-rater agreement using percentage agreement produces a relatively low outcome of 5%, it denotes that the experts struggle to consistently classify the examples, which could be attributed to ambiguities in the rendered definitions and anchor examples or complexities of multiple classes within the online data. Details regarding the first iteration of the Delphi technique are available for further analysis in Appendix B.

Example ID	**Model Prediction**	**Majority Vote**	**Number of Votes in Agreement with Model**
1	GSP	FCGS	0
2	SDP	GSP	1
3	TI	TI	3
4	GSP	GSP	2
5	FCGS	FCGS	2
6	FCGS	GSP	0
7	GSP	FCGS	0
8	FCGS	GSP	1
9	FCGS	TI	1
10	FCGS	FCGS	3
11	TI	SDP	1
12	TI	TI	3
13	SDP	TI	0
14	TI	FCGS	0
15	SDP	FCGS	0
16	TI	FCGS	0
17	TI	GSP	1
18	SDP	SDP	3
19	TI	GSP	0
20	TI	FCGS	0

Table 5.3: The comparison between the Task 2 model's prediction, the highest voted class by the EDI experts, and the number of votes in agreement with the model's prediction. The abbreviations are reflected as follows: *Feminine-Coded Goods and Services (FCGS), Gendered Split Perception (GSP), Testimonial Injustice (TI), and Social Dominance Penalty (SDP).*

Subsequently, we further assess the model's performance against the EDI experts' comprehension of the taxonomy by conducting permutation tests on their predictions. We evaluate each participant's responses, calculating their average F1-score and standard deviation, mainly focusing on instances where they used the *Other* option. Additionally, we re-examine the p-values to discern if their assessments lean towards random chance or hold substantive value. Furthermore, we compare the zero-shot classification model's performance with the EDI experts to provide qualitative insights into the challenges of aligning AI technologies to human interpretation of sexism. As illustrated in Figure 5.5, the four human participants achieved average F1-scores ranging from 0.62 to 0.80 with minimal standard deviation. Each participant's p-values were below the 0.05 significance level, indicating statistically significant predictions not attributable to chance. In contrast, the zero-shot classification model attained an average F1-score of 0.30 and exhibited a high standard deviation. This variability suggests that the model's predictions may be somewhat influenced by chance, as reflected in the distribution of F1-scores, including a score of 0.49. While the model's F1-score may seem modest, it underscores the inherent complexity in classifying different forms of sexism and the limitations of AI systems. As detailed in Section 2.3.2, recent studies explain the persistent challenges in classifying sexism with AI by revealing that high performance remains elusive despite extensive fine-tuning on large datasets. Additionally, the task is notably challenging when applying a zero-shot classification approach to novel datasets.

Lastly, our study highlights the challenges we encountered in attempting to align AI systems with human interpretations of sexism, particularly in the context of the experiences of women software developers. We struggled to understand the reason behind the model's classifications of certain texts due to its lack of transparency and inability to suggest improvements. Moreover, translating the human perspective on sexism into the input texts was difficult as we needed to accurately reflect the nuanced experiences of women in technical teams. Given that our research examines the experiences of women software developers facing sexism, it was imperative to embed this contextual detail into our input texts to ensure the model better comprehends the intended topic of sexism. This approach involved creating multiple versions of generalized definitions and anchor examples of sexism, which was time-consuming and required the manual evaluation of the top k precision scores to determine the optimal results. Additionally, selecting specific examples from the dataset

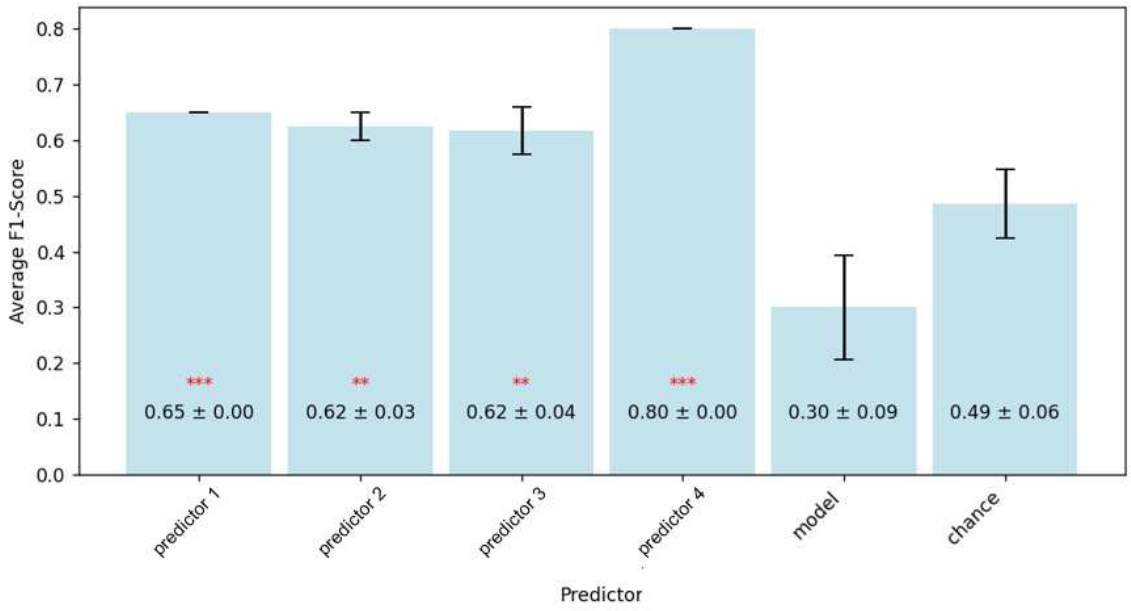

Figure 5.5: The results of the permutation test on the human predictors and the Task 2 zero-shot classification model. The graph presents the average F1-scores with the corresponding standard deviations and p-values per human predictor, model, and chance. The p-values are represented as follows: * indicates $p < 0.001$, ** indicates $p < 0.01$, and *** indicates $p < 0.05$.

while avoiding data leakage proved challenging, as we had to locate lengthy and detailed instances that genuinely represented the experiences of sexism narrated by women software developers.

Therefore, our findings illustrate that semantic similarity effectively identifies generalized contexts, such as the experiences of sexism encountered by women software developers, even when trained on a relatively small dataset of high-quality texts. Nevertheless, semantic similarity struggles with multi-class classification of sexism. Future research should explore alternative AI techniques to address these limitations. Additionally, it is essential to consider critiques from social science researchers, such as Crawford (2021), Eubanks (2019), and Noble (2018), who emphasize that AI technology experiences limitations in fully comprehending the nuanced and complex nature of social issues like sexism.

Chapter 6

Discussion and Code of Conduct

This chapter presents two key subsections derived from our results. First, we review the primary findings of sexism experiences from our Reddit case study, which is organized by taxonomy class. Next, we delineate the code of conduct we have devised, grounded in established guidelines aligned with our *Sexism in Software Development* taxonomy. This framework provides actionable recommendations for practitioners to effectively counteract sexism within software development teams and cultivate a more inclusive environment for female colleagues.

6.1 Reddit Insights: Sexism Experiences

The extracted subset of data revealed a variety of sexist experiences described by women software developers that align with the established taxonomy and prior research. We elaborate on these discovered experiences as follows:

- **Feminine-Coded Goods and Services**: Women software developers illustrate that their teammates often implicitly expect them to manage organizational tasks such as note-taking and scheduling meetings due to their gender. For example, Reddit user *CieloBlueStars* expresses that her colleagues consider her as the "team go-to secretary to do all administrative tasks like notetaking, documentation, meeting facilitation, which all end up invisible work that just adds on to overwork and not recognized as valuable promotable work" CieloBlueStars (2023).

Sultana et al. (2021) illustrates this type of experience in their research, where colleagues expect women software developers to perform administrative duties for the team. Furthermore, women software developers communicate that people often expect them to perform tasks perceived as less technical and more creative while maintaining a social balance within their teams. For instance, Reddit user *prettypangolin* explains that she is "being pushed towards projects that require strong soft skills (but maybe are less technical), feeling obligated to participate in diversity activities (which can be a lot of extra work!), being "the representative" in the room, being the de facto emotional support person..." patriotn8 (2019). The research by Sultana et al. (2021) and Guzmán et al. (2023) both highlight a common theme that women software developers often express feeling pressured to handle tasks that require minimal technical expertise to meet their team's expectations. This narration also supports previous findings from Jimenez et al. (2019), explaining that women and members from minority groups are more frequently assigned diversity activities compared to their peers. However, to our knowledge, there is a lack of studies documenting the experiences of women software developers being implicitly tasked with providing emotional support to the entire development team.

- **Gendered Split Perception**: Women software developers explain that their technical contributions are often heavily scrutinized through pull requests or general public feedback compared to their male counterparts. For example, Reddit user *queenannechick* illustrates her encounters with male software developers when publicizing her technical projects online for others to utilize for their benefit:

  > "I have and have had for a decade now an array of small, ultra-niche software products that I make and sell online. The first couple I attached my actual name too. I'd endlessly get harassing emails from men trying to "help" that were just berating tirades saying I seem like a ditz, floozy, incompetent moron and that I should let them help. They attached their names, titles, and companies to these emails! They actually thought they were helping. Now I use a made-up male name and I literally get emails saying I'm awesome (which I never got for female-faced

products)" queenannechick (2023).

From our comprehensive review of the literature on sexism in the software development field, we identified no prior studies documenting the phenomenon where women intentionally alter their online identities to male-oriented or gender-neutral ones, thereby reducing hostility from their remote colleagues. Moreover, women software developers mention that they experience harsher conditions when performing their work than their male colleagues. For instance, Reddit user *crowleyscot* explains, "I actually have this problem with a male supervisor where he is more lenient with his male student than me (female). In fact, he'll even do work for the male student (write code/come up with research ideas etc) whereas I will not only have to do everything myself, but I'll also be drafted extra duties such as teaching/mentoring etc." WayRoundTheWorld (2018). A parallel example of this experience is depicted in the research by Oliveira et al. (2023), explaining that women software developers frequently receive unequal support compared to their male teammates.

- **Testimonial Injustice**: Women software developers share that their colleagues in male-dominant teams often try to take ownership of their ideas and are then re-explained the same concepts in a simplified manner. For example, Reddit user *marmotte25* denotes that:

  > "Even today, one of my male colleagues stated that he'd never witnessed real discrimination against women at work yet, he's new and is already trying to take ownership of my project and credit for everything I've done and spent 30 minutes mansplaining to me this morning how to manage a technical team" kaiso_gunkan (2022).

This situation resembles the "prove-it-again" phenomenon described in the research by Trinkenreich et al. (2022), where women software developers are often required to demonstrate their technical competence repeatedly. However, to our knowledge, a lack of prior research in the field of software development indicates that males commonly attempt to take credit for their female colleagues' work. Furthermore, women software developers explain that they are often ignored or disregarded for their input while participating in technical discussions with

their male-dominant teams. For instance, Reddit user *schwarzekatze999* describes that "I'm the only female on the team, get ignored and talked over all the time, and guys make sexist jokes without thinking" imLissy (2023). Instances of this nature are highlighted in the work of Guzmán et al. (2023), where women software developers are shown to encounter micro-inequities, such as feeling overlooked during discussions, in contrast to their male colleagues.

- **Social Dominance Penalty**: Women software developers sometimes feel hostility when requesting clarification on technical aspects while participating in male-dominant teams. For example, Reddit user *user983763876* states that her colleagues respond unprofessionally to her questions as follows:

> "During that job, I was left questioning my own abilities and sometimes my own sanity. It was common for me to ask specific questions about architectural design or simply versions of components. These questions were answered either dismissively or in an overly aggressive manner." eggo14 (2022)

These types of responses are discussed in the research by Oliveira et al. (2023), which details how women software developers frequently encounter hostile responses within their male-centric team environments. Lastly, women software developers explain their various encounters with harassment and derogatory statements from their male counterparts. For instance, Reddit user *Vaqu3ra13* describes her unfortunate incidents of such behaviour during her career:

> "Throughout my career, I've been sexually harassed, underpaid, assaulted (both verbally and physically), and overlooked by male coworkers. I've been asked for "favors." I've been called a "bitch" when really, I was no more "assertive" than my male counterparts." LaikaBauss31 (2020).

Unfortunately, these documented experiences of encountering resentment and derogatory terms are prevalent among women in the field of software development, as supported by studies like those conducted by Sultana et al. (2021), Trinkenreich et al. (2022), and Oliveira et al. (2023).

6.2 Code of Conduct

To address sexism and misogyny in software development and counteract ongoing issues, we propose that software development teams adopt a code of conduct and receive training on its principles. This initiative aims to raise awareness and promote proactive measures against discriminatory practices. Although this code of conduct does not derive directly from our findings, it draws on best practices from organizational and managerial recommendations Ontario Ministry of Labour (2016); Sonke Gender Justice (2024), as well as the Council of Europe's guidelines against sexism Council of Europe (2019). It is also grounded in the established taxonomy. Consequently, the code of conduct is not based on empirical evidence from this thesis.

6.2.1 Feminine-Coded Goods and Services

- **Assign a Scrum Master**: As highlighted by Council of Europe (2019), portraying women in stereotypical roles reinforces gender biases, such as the assumption that women are expected to handle administrative tasks in team meetings. In agile software development, the scrum master plays a crucial role in team efficiency by documenting the meeting minutes and scheduling follow-up sessions while also managing the team's overall progress. Therefore, it is ideal to designate the most appropriate team member as the scrum master through team discussions. Furthermore, the team should consider other potential colleagues in case the candidates refuse the proposition. *As such, this contributes to mitigating the presumption that the women software developers are implicitly responsible for initiating the administrative work during team meetings.*

- **Equitable Diversity Efforts**: Team members should avoid assuming stereotypical roles, such as women colleagues are responsible for managing diversity-related tasks or acting as the sole representative for the group's efforts Council of Europe (2019). These responsibilities include attending team-building workshops, engaging in collaborative training sessions, and participating in broader Equity, Diversity, and Inclusion (EDI) initiatives. It is imperative that every team member actively engages in these efforts to collectively build and uphold an inclusive

and equitable work environment. *Therefore, this initiative prevents the disproportionate assignment of non-technical tasks to women software developers and promotes fair distribution of responsibilities, ultimately fostering a more balanced and supportive workplace environment.*

- **Address Sexist Humour**: While some jokes might seem innocuous and often stem from outdated cultural norms, sexist humour can intimidate and silence individuals while trivializing unacceptable behaviour Council of Europe (2019). Such remarks undermine a respectful work environment and reinforce harmful stereotypes. Therefore, Council of Europe (2019) suggests that teammates should avoid jokes that reference traditional gender roles or offensive stereotypes. If such comments arise, the team, including superiors, must address the issue collectively to correct and prevent such behaviour. *Evidently, this helps mitigate the risk of teammates inadvertently making offensive jokes that single out women in software development.*

6.2.2 Gendered Split Perception

- **Genderless Team Communication**: To cultivate an inclusive and respectful work environment, all team members should use genderless language when referring to colleagues' professional contributions and achievements. As stated by Council of Europe (2019), it is recommended to use gender-neutral forms of titles and pronouns to raise awareness and prevent sexist behaviour. For example, rather than labelling teammates as "our female user interface (UI) designer" or "the woman back-end developer," which diverts attention from their professional competencies to their gender, one should utilize terminology that highlights their role and accomplishments. *This practice ensures that women software developers' technical expertise receives proper recognition without being diminished by gender biases, thereby advancing a fair and equitable workplace for everyone.*

- **Objective Review Process**: When evaluating a teammate's work, it is imperative to approach the review with impartiality toward the author. For instance, when tasked with reviewing a

pull request, start by thoroughly reading the provided description and then proceed with analyzing the code itself. Once that is completed, the reviewer could continue addressing any remaining aspects of the pull request, such as contacting the author for further discussions. Sonke Gender Justice (2024) expresses that fostering a culture of open dialogue and collaborative learning is essential for team growth. *This method ensures that feedback remains objective, thereby diminishing the risk of disproportionate scrutiny being directed at women software developers during code reviews.*

- **Draft and Record Ideas**: Before sharing ideas in team discussions, consider informally documenting them for clarity and ownership. It is best practice to send notes to yourself via email or chat to effectively refine and outline the details. According to Council of Europe (2019), workplace sexism -—such as taking credit for women's contributions or claiming their ideas as one's own -— can have systematic and damaging effects. Systematically, it can limit professional opportunities and hinder career advancement for women. Internally, it can contribute to heightened anxiety and depression, undermining overall well-being and job satisfaction. *This practice supports women software developers in asserting ownership of their ideas through traceability.* If it is preferred to receive initial feedback, select a trusted colleague to review your final draft before the formal presentation.

6.2.3 Testimonial Injustice

- **Enable Individual Input**: During team meetings, it is important to create an environment where all members feel encouraged to share their input Sonke Gender Justice (2024). This approach ensures that everyone can contribute innovative ideas and fosters a supportive work atmosphere while also allowing individuals to opt-out if they have no additional insights. By promoting inclusive discussions, the team can address overlooked issues more effectively, taking into account each colleague's professional experience. *This approach is particularly valuable for women software developers, providing them with a platform to express their perspectives without interruptions.*

- **Iterative Task Reflection**: In agile software development, holding a retrospective at the end

of each sprint is crucial for improving team efficiency. This practice encourages documenting both accomplishments and challenges related to individual tasks. Over time, these records become valuable references for identifying recurring issues and areas for improvement. According to Sonke Gender Justice (2024), it is advantageous for the team when members strive to remain flexible and learn from previous challenges in their work practices. *Therefore, this helps enable women software developers to explain their experiences and prevent teammates from forming assumptions against their technical skills.* For instance, a teammate could have encountered delays in task completion due to factors such as system environments or faulty hardware issues which is beyond their immediate control.

- **Preferred Task Assignment**: To enhance task assignment, gather each team member's preferences and aim for a fair distribution of tasks across sprints. Allowing individuals to work on tasks they are passionate about can boost team morale, improve productivity, and foster mutual trust. When team members see that others are making an effort to engage in tasks they care about, it strengthens confidence in each other's commitment and capabilities Sonke Gender Justice (2024). *For women software developers, this approach provides the opportunity to take on meaningful technical work and showcase their abilities to the team.*

6.2.4 Social Dominance Penalty

- **Zero-Tolerance Harassment Policy**: As described by Ontario Ministry of Labour (2016), team members should establish a clear zero-tolerance harassment policy that outlines rules and procedures to address unprofessional conduct among colleagues. Therefore, new members should review this policy to ensure a shared understanding and respect among all team members, potentially signing it for accountability and clarity. *This approach supports women software developers by demonstrating a commitment to maintaining clear boundaries and promptly addressing any issues that may arise.*

- **Support and Report**: In the event of a team member violating the harassment policy, it is crucial that the victim feels fully supported when reporting the incident to either another team member or a superior, such as a manager, supervisor, or professor Council of Europe

(2019). The team must establish a secure and confidential reporting mechanism that ensures individuals can communicate incidents without fear of retaliation. *This initiative is important for women software developers as it provides reassurance that they can address situations without facing immediate judgment from their peers.*

- **Misconduct Resolution**: The team assembles – excluding the individual responsible for the harassment – to engage in a comprehensive discussion of the reported issue and determine appropriate disciplinary actions against the offender Council of Europe (2019). Additionally, external resources may be consulted for guidance if the situation's complexity warrants it. *This proactive approach demonstrates to women software developers that the team is committed to upholding the zero-tolerance harassment policy, thereby fostering a safe and supportive work environment for all team members.*

Chapter 7

Conclusion, Impact on Society, and Future Work

The concluding chapter provides a comprehensive thesis summary highlighting our main findings. Furthermore, we discuss our study's threats to validity, social impacts and contributions to Software Engineering (SE) and affiliated software development fields, such as computer science (CS) and information technology (IT). Finally, we explore future research prospects and suggest avenues for further investigation.

7.1 Conclusion

The gender gap in SE and affiliated domains is a prominent concern that requires further attention as the lack of women in technological advancement leads to embedded biases, which propagate into societal actions and enable sexist behaviours. In numerous studies and articles, women software developers express various types of experiences of sexism while collaborating in male-dominant teams, such as feeling ostracized, objectified, and harassed. Moreover, advancing technologies, such as artificial intelligence (AI), exhibit implicit prejudice towards females due to the absence of knowledge from women programmers to aid in identifying relative issues before publicizing the software to mitigate societal concerns. Therefore, in this research, we aim to contribute to the identification of sexism in SE and affiliated domains to help bridge the gender gap in technological

innovation.

In our study, we initiate a literature review of research on sexism and misogyny to select and construct a taxonomy to identify various prominent forms of sexism. Subsequently, we apply static keyword-matching and semantic similarity to identify narrations of sexist experiences from eleven subreddits between the years 2018 and 2023. Lastly, we evaluate the AI model's effectiveness with four Equity, Diversity, and Inclusion (EDI) experts to ensure its alignment with nuanced human understandings of sexism. Our research focuses on investigating (1) the most common categories of sexism that women software developers are likely to experience during their participation in the field, (2) the extent to which semantic similarity can effectively extract experiences of women in software development that align with the constructed taxonomy, and (3) the challenges in aligning AI systems with the human interpretation of sexism.

Our results present a comprehensive taxonomy titled *Sexism in Software Development*, which features four distinct categories of sexism, complete with definitions, anchor examples, and associated lexicons. Moreover, while semantic similarity methods effectively capture narratives of sexist experiences from women software developers, the model struggles with precise classification. Furthermore, our findings reveal the complex challenges of adapting AI systems to match human interpretations of sexism by addressing the technical aspects of influential input texts and manually selecting the most appropriate data points to train the model. Therefore, based on our results, we denote that AI requires significant human guidance to effectively identify and classify sexism. Continuing, we identify previously documented experiences of sexism while highlighting three novel reports of sexism in software development. In response to these insights, we recommend a code of conduct designed to help practitioners and researchers reduce sexism within technical teams, thereby fostering greater participation of women in SE and technological innovation.

Finally, it is essential to recognize that although the classifiers in Tasks 1 and 2 are currently in their preliminary stages, their underlying concepts possess the potential for expansion into various applications. For example, integrating these models into work-oriented communication platforms like Slack or Microsoft Teams can enhance the detection and classification of sexist language, promoting a more respectful and inclusive digital work environment. Additionally, educators can use

these classifiers as interactive tools to provide real-time feedback on communication patterns, to potentially foster greater awareness and understanding of gender bias. These applications might help mitigate sexist behaviour and possibly support broader initiatives to advance workplace diversity, educational equity, and societal change. However, it is important to note that AI can only play an assisting role in detecting and countering sexism.

7.1.1 Threats to Validity

In this thesis, we determine three main threats to the validity of our outcomes, denoted as internal, external, and constructive. Internal validity refers to factors within the experiment that could influence the results. Our study's internal validity may be compromised by its static keyword-matching approach, which might miss relevant content due to its limited keyword set, despite manual attempts to expand the lexicons. Therefore, advanced techniques like the near-neighbor approach could better capture community-specific language. Additionally, reliance on initial Delphi technique results and majority votes, constrained by time, could affect model evaluation accuracy. However, assessing the iterators' F1-scores and statistical measures helps mitigate this issue. Overall, while there are potential limitations, the initial dataset consists of 26,047, and our evaluations suggest minimal impact on the findings.

External validity examines factors beyond the experiment's control that may influence outcomes. This study utilizes Reddit content to uncover novel experiences of sexism in software development, but we cannot ensure that the data is exclusively from women developers since subreddits are open to everyone. Despite this, our approach remains valid and adaptable for future research using other datasets on sexism, such as those from the Automatic Detection of Sexist Statements Commonly Used at the Workplace Grosz and Conde-Cespedes (2020), "Call me sexist, but..." Samory, Sen, Kohne, Floeck, and Wagner (2021), and Explainable Detection of Online Sexism (EDOS) Kirk et al. (2023).

Lastly, construct validity is affected by limitations in our research approach, particularly the narrow scope of our taxonomy, which may not fully capture the diverse experiences of sexism in software development. Relying mainly on one resource restricts our perspective and could miss other significant forms of sexism. Furthermore, we relied on the non-empirical, but theoretically

predefined taxonomy of Manne (2018). Despite these limitations, this research provides a foundation for a new approach to sexism classification, encouraging future studies to enhance the taxonomy by incorporating broader insights from additional literature and feminist theories.

7.2 Impact on Society

The thesis presents significant contributions to SE and related software development fields by highlighting the neglected experiences of sexism narrated by women software developers on overlooked online platforms. Moreover, our study advances research in sexism detection and classification through the establishment of a taxonomy formulated using literature on sexism and misogyny. This approach surpasses arbitrary categorizations by providing a nuanced understanding of sexism that better reflects its complex realities. Additionally, our research provides a code of conduct for software development practitioners to recognize and effectively mitigate sexism directed towards their female colleagues by promoting a more inclusive and supportive work environment.

Lastly, this thesis helps progress the following three United Nations Sustainable Development Goals (SDG) United Nations Department of Economic and Social Affairs (2015):

- **SDG 5 - Gender Equality:** Focuses on eliminating all forms of discrimination and violence against women and girls while ensuring equal opportunities and empowerment. Accordingly, this thesis examines women software developers' experiences of sexism to raise awareness of the gender gap in software development and promote strategies for achieving gender equality for women.

- **SDG 8 - Decent Work and Economic Growth:** Aims to create the necessary conditions for inclusive and sustainable economic growth that promotes decent work for all. As such, our research supplies a code of conduct that software development teams can reference to ensure that women colleagues are provided with a supportive work environment to enable decent working conditions and help bridge the gender gap in software development.

- **SDG 10 - Reduced Inequality:** Concentrates on reducing inequality within and among countries. Therefore, this study aims to reduce gender inequality in the software development domain by sharing narratives of sexist experiences reported by women developers worldwide, alongside providing recommendations to software teams on mitigating sexism against their women counterparts.

The thesis aims to enhance inclusivity for women in software development by presenting their challenges of participating in a predominantly male field and offering guidelines for teams to mitigate sexism directed at them. Furthermore, our objective includes encouraging diversity within software development to reduce embedded implicit gender biases in technological innovation and promote women's involvement in technology. These efforts align with the United Nations' SDGs by contributing to the creation of a sustainable and equitable environment where all technological contributions are valued irrespective of gender.

7.3 Future Work

This study can be continued and enhanced in various areas for future work, as outlined in the following subsections.

7.3.1 Enhance Zero-Shot Classification Model

The current semantic similarity approach utilizes a zero-shot classification model to categorize online content relevant to women software developers' personal experiences based on the constructed taxonomy. As a starting point, future researchers could begin building the dataset with an equal distribution across taxonomy classes to fine-tuning the model using the examples mentioned in Section 6.1 and available on the GitHub repository stated in Section 4.3. It is advisable to carefully review, discuss, and document the logic behind the classifications of these examples with participants to achieve consensus-based categorization. Additionally, researchers may explore employing a model capable of processing full narratives to grasp the context of the text while performing a deeper analysis on a sentence-level basis to identify areas of classification overlap. Lastly, future researchers could consider employing other relevant models, such as Llama-3 Meta (2024) and

Universal Sentence Encoder (USE) Google (2024), or incorporating feature representations such as Doc2Vec Gensim (2024) to enhance performance efforts.

7.3.2 Refining Model Evaluation via Iterative Delphi Technique

As previously mentioned, due to time constraints, this thesis did not include iterative discussions with the EDI experts to achieve consensus on the classification of the model evaluation testing dataset. As such, future research could benefit from conducting model evaluations through multiple iterations of the Delphi technique. This approach would ensure that the ground truth values converge to a final result and foster a collective understanding of the categories. Moreover, future researchers may find it valuable to involve individuals with backgrounds in Women's and Gender Studies as they have a deeper understanding of sexism. Their insights could improve the model's performance by leveraging expertise in the relevant field.

7.3.3 Extend the Taxonomy and Lexicons

The thesis aims to serve as an initial step toward advancing sexism classification to align with supported literature on sexism and misogyny. Consequently, we recommend that future researchers develop this study further by incorporating additional categories and keywords related to sexism and misogyny derived from feminist research and relevant sources like psychological studies. For example, researchers could consider employing advanced techniques, such as WordNet University (2024) and NearestNeighbor Scikit-learn (2024), to extend the taxonomy lexicons and capture community-specific language. Moreover, future researchers could aim to find specific classes of sexism pertinent to the software development field to better align with the realities of the ongoing phenomenon. Instead of using theoretically predefined taxonomies, empirical methods, such as open coding could be employed. Lastly, we suggest that researchers try further refining the taxonomy definitions and anchor examples to enhance the relevance of the online data.

7.3.4 Analyze LGBTQIA+ Challenges in Software Development

Our current research is centred on the experiences of women software developers, as we found that narratives from the relevant subreddits predominantly reflect their perspectives. However, there

is an opportunity to broaden our approach to include the unique experiences of LGBTQIA+ individuals in SE and affiliated software development domains. During our data analysis, we encountered numerous accounts of sexist experiences from self-identified LGBTQIA+ members, which could inform future research. Expanding research in this direction would help create a more diverse and inclusive work environment, potentially driving technological innovation through a richer array of perspectives. However, it is essential to note that our current taxonomy does not adequately reflect the phenomenon experienced by LGBTQIA+ members and would require further refinement supported by additional literature.

7.3.5 Apply Methodology to Available Sexism Datasets

Future researchers can extend this thesis by applying our methodology to other text-based datasets that address sexism, such as Automatic Detection of Sexist Statements Commonly Used at the Workplace Grosz and Conde-Cespedes (2020), "Call me sexist, but..." Samory, Sen, Kohne, Floeck, and Wagner (2021), and Explainable Detection of Online Sexism (EDOS) Kirk et al. (2023). However, it is essential to note that many existing sexism datasets primarily consist of short texts that present sexist content without extensive contextual information, which could cause variations in the results.

Appendix A

Taxonomy Lexicons

In Chapter 5.1, we present the findings from our literature review along with the developed taxonomy and the associated lexicons for each category. This appendix provides an extended list of keywords from our taxonomy for those who wish to explore it further.

Keywords (25)	Synonyms (159)
cool	awesome, wonderful, lovely, great, excellent, beautiful, terrific, fantastic, fabulous, superb, hot, marvelous, stellar, fine, neat, prime, heavenly, calm
natural	genuine, unaffected, simple, honest, innocent, naïve, sincere, pure, raw, organic, wholesome, easy
healthy	strong, fit, hearty, active, lively
loyal	faithful, dependable, devoted, trustworthy, trusty, dedicated, reliable
good	pleasant, positive, favorable, valuable, noble, decent, ethical, mortal, auspicious, happy
affection	sentiment, liking
adoration	veneration
indulgence	blessing, privilege, courtesy, leniency, permissiveness
loving	admiring, affectionate, amiable, adoring, passionate
acceptance	approval, support, embracing, adoption
nurturing	female, feminine, matronly, womanly, parental
safety	protection, safeguards, safeness, guard
security	safekeeping, shield
safe haven	
kindness	goodwill, grace, kindliness, benevolence, gentleness, sweetness, kindheartedness, benignity
compassion	empathy, sympathy, mercy, pity, commiseration
mortal attention	
concern	worry, fear, anxiety, unease, concernment
soothing	relaxing, comforting, tranquilizing, calming, hypnotic, quieting, sedative, dreamy, peaceful, restful, reassuring
caring	compassionate, benevolent, helpful, sympathetic, thoughtful, generous, humane, kindly, warm, soft, sensitive, tender, responsive, receptive, considerate, warmhearted, tenderhearted, softhearted, nice
trust	confide, confidence, faith, assurance, entrustment, credence, depend on, count on
respect	admiration, regard, appreciation, praise, recognition, reverence
attentive	kind, respectful, solicitous, gracious, polite
relationship	connection, association, kinship, relation, linkage, affiliation, interaction, bond, communication, friendship
giving	bestowing, offering

Table A.1: Lexicon of Feminine-Coded Goods and Services

Keywords (12)	Synonyms (73)
duplicitous	deceitful, dishonest, fraudulent
vindictive	malicious, vengeful, vicious, revengeful, petty, spiteful, merciless, resentful
conniving	scheming, plotting, conspiring, collusive, shifty
untrustworthy	disloyal, unreliable, untrusty, devious, unfaithful
careless	thoughtless, reckless, sloppy, negligent, indifferent, unconcerned, absent-minded, unthinking, cursory, inconsiderate, unmindful, incautious, impetuous, unwary, mindless,
shady	dubious, questionable, unscrupulous, dodgy, suspect, fishy, disreputable
crooked	suborned, corrupt, dishonorable
rule-breaker	
dangerous	troubling, perilous, precarious, ugly, unsafe, unstable, alarming, menacing, insecure, irresponsible
suspicious	distrustful, skeptical, mistrustful, unusual, unbelieving, leery
risky	hazardous, threatening, dicey
deceptive	misleading, sneaky, spurious, ambiguous, delusive, fallacious, delusory, beguiling

Table A.2: Lexicon of Gendered Split Perception

Keywords (35)	Synonyms (166)
catcalling	jeering, hooting, snorting, sniffing, jibing, gibing, sneering, laughing, whistle, heckling, holler
trolling	
condescending	bossy, impudent, snooty
mansplain	
moralizing	lecturing, preaching
blaming	condemning, condemn, condemned, faulting, denouncing, knocking, attacking, slamming, censuring
silencing	suppressing, quelling, subduing, censor, muffling
lampooning	spoofing, burlesquing, mimicking, banter, bitterness, cynicism
satirizing	
sexualizing	
desexualizing	
belittling	minimizing, discounting, derogating, pejorative, contemptuous, contempt
caricaturing	deride, scoff, taunt, tease, parodying, imitating
exploiting	abuse, manipulate, misuse
erasing	eradicating, destroying, abolishing, obliterating
infantilizing	immaturity, ignorance, childishness
ridiculing	derisive, baiting, deriding, fooling
humiliating	mortifying, demeaning, embarrassing, degrading, ignominious, humbling
mocking	uncivil, sarcastic, satirical, disrespectful, sardonic, negativistic
slurring	disgrace, insinuate, affronting, blaspheming, cursing, berating
vilifying	insulting, offensive, rude, abusive, malign, smearing, libeling, slandering, defaming, discrediting
demonizing	diabolize, torment, affliction
shunning	avoidance, ostracism, exile, isolation, rejection, expulsion, evasion
shaming	disgracing, dishonoring, abasement, mortification, deceiving, groveling, grudging
patronizing	domineering, dominant, disdainful, authoritarian, snobbish
dismissive	
disparaging	dismissing, denigrating, bad-mouthing, derogative, defamatory, deprecatory
less credible	
less competent	incompetent, unskillful, helpless, inadequate, incapable, unqualified, useless, inept, unfit, inexperienced
accused	indicted, charged, blamed, prosecuted, censured
impugned	criticized, denounced, appealed, castigated, reprobate
convicted	guilty, culpable, punishable
corrected	rectified, amended, revised, culprit, imprison, rebuke, discipline, reprimand, chide, admonish, assessed
diminished	scorn, devalue, denigrate, decry, deprecate, depreciate, derogate
outperformed	beat, exceed, surpass, outdo, defeated, bested

Table A.3: Lexicon of Testimonial Injustice

Keywords (28)	Synonyms (163)
smother	overwhelm, stifle, repress, hold back, restrain, bottle up
intimidate	bully, frighten, scare, coerce, startle, browbeat, harass, bulldoze, pressure, terrify, hound, daunt, oppress, constrain, dishearten, dismay
powerful women	
powerful woman	
threatening	ominous, intimidatory, terrorizing, sinister
underestimate	underrate, undervalue, minimize
doubt	disbelief, hesitation, uncertainty, skepticism
victim blaming	
crazy	kooky, mad, nuts, nutty, silly, wacky, ridiculous, absurd, foolish, ludicrous, mental, irrational
hysterical	agitated, distraught, frantic, frenzied, neurotic, convulsive, upset
disliked	hatred, disgust, hostility, loath, disapproval, distaste, animosity, aversion, antagonism, displeasure, antipathy, enmity, animus, disinclination, repugnance, detestation, abhor, detest, execrated, despised
rejected	abandoned, deserted, disused, denied, disregarded, dumped, ditched, rebuff
hostile	antagonistic, mean, hateful, inhospitable, nasty, unfavorable, unfriendly, catty, sour, inimical, negative
abrasive	irritating, annoying, harsh, cruel, unpleasant, rough, unkind, frustrating, disturbing, aggravating, bothersome
manipulative	deceive, shrewd
arrogant	vain, smug, pompous, imperious, cocky, conceited, cavalier, bumptious, assumptive, pretentious
aggressive	belligerent, combative, destructive, intrusive, assertive, malevolent, pushy, pugnacious
ballbreaker	
castrating bitch	
punished	penalized, fined, sentenced, chastised, levied
real woman	
real women	
bitch	floozy, harlot, hussy, slut, tart, tramp, vamp, wench, whore, broad, hellion, termagant, vixen,
witch	hag, shrew
unfair	foul, shameful, biased, prejudiced, discriminatory
rigid	strict, rigorous, stern, stringent
cold	aloof, distant, frigid, apathetic, glacial
psychotic	demented, insane, unhinged, lunatic, paranoid, psycho, maniac

Table A.4: Lexicon of Social Dominance Penalty

Appendix B

First Iteration of Delphi Technique

In Chapter 5.3, we illustrate an overview of the results from the first iteration of the Delphi technique among the four EDI experts to compare with the model's performance in sexism classification of the *Sexism in Software Development* taxonomy. This appendix presents the complete responses from the participants to further explore additional details.

Example ID	Predictor 1	Predictor 2	Predictor 3	Predictor 4
1	TI	FCGS	FCGS	FCGS
2	SDP	GSP	Other: first part is TI/SDP and second part is GSP	GSP
3	TI	SDP	TI	TI
4	GSP	Other: GSP or SDP	GSP	SDP
5	FCGS	TI	GSP	FCGS
6	GSP	SDP	TI	GSP
7	SDP	FCGS	TI	FCGS
8	FCGS	GSP	GSP	GSP
9	TI	FCGS	TI	TI
10	FCGS	FCGS	FCGS	SDP
11	SDP	TI	GSP	SDP
12	TI	TI	TI	GSP
13	GSP	TI	TI	TI
14	SDP	FCGS	Other: TI and FCGS	FCGS
15	TI	FCGS	Other: TI and FCGS	FCGS
16	FCGS	FCGS	GSP	Other: ontology inferiority
17	GSP	TI	GSP	GSP
18	SDP	SDP	TI	SDP
19	GSP	FCGS	GSP	GSP
20	FCGS	FCGS	FCGS	FCGS

Table B.1: The results of completing the first iteration of the Delphi technique with four participants from Concordia University's Equity, Diversity, and Inclusion (EDI) lab. The results are pertinent to the distributed Google Form containing twenty manually-selected examples from the extracted Reddit content from the mentioned footnote in Section 5.3.

References

Anzovino, M. E., Fersini, E., & Rosso, P. (2018). Automatic identification and classification of misogynistic language on twitter. In *International conference on applications of natural language to data bases.* Retrieved from https://api.semanticscholar.org/CorpusID:43958570

Boe, B. (2023). *PRAW 7.7.1 documentation.* Retrieved 2024-03-20, from https://praw.readthedocs.io/en/stable/index.html

Bond, S. (2019, December). Uber To Pay $4.4 Million To Employees Who Were Sexually Harassed At Work. *NPR*. Retrieved 2024-05-06, from https://www.npr.org/2019/12/19/789949239/uber-to-pay-4-4-million-to-employees-who-were-sexually-harassed-at-work

Butt, S., Ashraf, N., Sidorov, G., & Gelbukh, A. (2021). Sexism identification using bert and data augmentation - exist2021. *CEUR Workshop Proceedings*, *2943*, 381–389. (Publisher Copyright: © 2021 CEUR-WS. All rights reserved.; 2021 Iberian Languages Evaluation Forum, IberLEF 2021 ; Conference date: 21-09-2021)

ChicksWhoCode. (2024). *Chicks Who Code.* Retrieved 2024-03-20, from https://www.reddit.com/r/chickswhocode/

Christie, C. A., & Barela, E. (2005, March). The Delphi Technique as a Method for Increasing Inclusion in the Evaluation Process. *Canadian Journal of Program Evaluation*, *20*(1), 105–122. Retrieved 2024-05-06, from https://utpjournals.press/doi/10.3138/cjpe.020.005 doi: 10.3138/cjpe.020.005

CieloBlueStars. (2023, September). *Feel undervalued in my role* [Reddit Post]. Retrieved 2024-05-09, from www.reddit.com/r/girlsgonewired/comments/16fahan/feel_undervalued_in_my_role/

Clance, P. R., & Imes, S. A. (1978). The imposter phenomenon in high achieving women: Dynamics and therapeutic intervention. *Psychotherapy: Theory, Research & Practice*, *15*(3), 241–247. Retrieved 2024-07-06, from https://doi.apa.org/doi/10.1037/h0086006 doi: 10.1037/h0086006

Council of Europe. (2019). *Sexism: See it. Name it. Stop it.* Retrieved 2024-08-12, from https://human-rights-channel.coe.int/stop-sexism-en.html

Crawford, K. (2021). *The atlas of ai: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

CSWomen. (2024). *Women in Computer Science.* Retrieved 2024-03-20, from https://www.reddit.com/r/cswomen/

Das, A., Rahgouy, M., Zhang, Z., Bhattacharya, T., Dozier, G., & Seals, C. D. (2023, September). Online Sexism Detection and Classification by Injecting User Gender Information. In *2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings)* (pp. 1–5). Mount Pleasant, MI, USA: IEEE. Retrieved 2024-03-17, from https://ieeexplore.ieee.org/document/10292474/ doi: 10.1109/AIBThings58340.2023.10292474

Daub, A. (2021, April). How Sexism Is Coded Into the Tech Industry. *The Nation*. Retrieved 2024-07-29, from https://www.thenation.com/article/society/gender-silicon-valley/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018, May). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* arXiv. Retrieved 2024-06-24, from http://arxiv.org/abs/1810.04805 (arXiv:1810.04805 [cs])

Donoso-Vázquez, T., & Rebollo-Catalan, A. (2018). *Violencias de género en entornos virtuales/ Gender violence in virtual environments*.

Doyle, R. (2020, October). *Sexism in Tech: An Inconvenient Truth.* Retrieved 2024-05-06, from https://medium.com/swlh/sexism-in-tech-an-inconvenient

-truth-26df0329e39

eggo14. (2022, May). *Underestimated by Male Counterparts* [Reddit Post]. Retrieved 2024-05-09, from www.reddit.com/r/womenintech/comments/ugvg0n/underestimated_by_male_counterparts/

Eubanks, V. (2019). *Automating inequality: How high-tech tools profile, police, and punish the poor*. Picador.

Farrell, T., Fernandez, M., Novotny, J., & Alani, H. (2019). Exploring misogyny across the manosphere in reddit. In *Proceedings of the 10th acm conference on web science* (p. 87–96). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi-org.lib-ezproxy.concordia.ca/10.1145/3292522.3326045 doi: 10.1145/3292522.3326045

Faulkner, W. (2000). Dualisms, hierarchies and gender in engineering. *Social Studies of Science*, *30*(5), 759–792.

Gensim. (2024). *Doc2Vec Model.* Retrieved 2024-08-28, from https://radimrehurek.com/gensim/auto_examples/tutorials/run_doc2vec_lee.html

GirlsGoneWired. (2024). *Girls Gone Wired.* Retrieved 2024-03-20, from https://www.reddit.com/r/girlsgonewired/

Glick, P., & Fiske, S. T. (1997, March). Hostile and Benevolent Sexism: Measuring Ambivalent Sexist Attitudes Toward Women. *Psychology of Women Quarterly*, *21*(1), 119–135. Retrieved 2024-05-10, from http://journals.sagepub.com/doi/10.1111/j.1471-6402.1997.tb00104.x doi: 10.1111/j.1471-6402.1997.tb00104.x

González González, C., García-Holgado, A., Martinez-Estevez, M., Gil, M., Martin-Fernandez, A., Marcos, A., ... Gershon, T. (2018, April). Gender and engineering: Developing actions to encourage women in tech. In *Gender and engineering: Developing actions to encourage women in tech* (pp. 2082–2087). doi: 10.1109/EDUCON.2018.8363496

Google. (2024). *Universal Sentence Encoder | TensorFlow Hub.* Retrieved 2024-08-28, from https://www.tensorflow.org/hub/tutorials/semantic_similarity_with_tf_hub_universal_encoder

Gramfort, A., & Liu, L. (2024). *Test with permutations the significance of a classification score.*

Retrieved 2024-08-11, from https://scikit-learn/stable/auto_examples/model_selection/plot_permutation_tests_for_classification.html

Grosz, D., & Conde-Cespedes, P. (2020, May). Automatic Detection of Sexist Statements Commonly Used at the Workplace. In *Pacific Asian Conference on Knowledge Discovery and Data Mining (PAKDD), Wokshop (Learning Data Representation for Clustering) LDRC*. Singapour, Singapore. Retrieved from https://hal.science/hal-02573576

Guenes, P., Tomaz, R., Kalinowski, M., Baldassarre, M. T., & Storey, M.-A. (2023, October). Impostor Phenomenon in Software Engineers. *ICSE SEIS 2024*. Retrieved 2024-03-16, from https://zenodo.org/doi/10.5281/zenodo.8415205 doi: 10.5281/ZENODO.8415205

Guest, E., Vidgen, B., Mittos, A., Sastry, N., Tyson, G., & Margetts, H. (2021, April). An expert annotated dataset for the detection of online misogyny. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume* (pp. 1336–1350). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.eacl-main.114 doi: 10.18653/v1/2021.eacl-main.114

Guzmán, E., Fischer, R. A.-L., & Kok, J. (2023, May). Mind the gap: gender, micro-inequities and barriers in software development. *Emperical Software Engineering (2024)*, 36. doi: https://doi.org/10.1007/s10664-023-10379-8

HarperCollins. (2024). *Collins Online Dictionary | Definitions, Thesaurus and Translations.* Retrieved 2024-03-20, from https://www.collinsdictionary.com/

Hill, C., Corbett, C., & St. Rose, A. (2010). *Why so few? women in science, technology, engineering, and mathematics*. Washington, D.C: AAUW. (OCLC: ocn607105042)

imLissy. (2023, February). *Women in computer science* [Reddit Post]. Retrieved 2024-05-09, from www.reddit.com/r/girlsgonewired/comments/10xvaul/women_in_computer_science/

Jha, A., & Mamidi, R. (2017, August). When does a compliment become sexist? analysis and

classification of ambivalent sexism using twitter data. In D. Hovy et al. (Eds.), *Proceedings of the second workshop on NLP and computational social science* (pp. 7–16). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W17-2902 doi: 10.18653/v1/W17-2902

Jimenez, M. F., Laverty, T. M., Bombaci, S. P., Wilkins, K., Bennett, D. E., & Pejchar, L. (2019). Underrepresented faculty play a disproportionate role in advancing diversity and inclusion. *Nature Ecology & Evolution*, *3*, 1030–1033.

kaiso_gunkan. (2022, September). *How do you deal with being unequal to men?* [Reddit Post]. Retrieved 2024-05-09, from www.reddit.com/r/womenintech/comments/xnt7is/how_do_you_deal_with_being_unequal_to_men/

Kalra, A., & Zubiaga, A. (2021, November). *Sexism Identification in Tweets and Gabs using Deep Neural Networks.* arXiv. Retrieved 2024-05-10, from http://arxiv.org/abs/2111.03612 (arXiv:2111.03612 [cs]) doi: 10.48550/arXiv.2111.03612

Karthikeyan, B., Sundarraj, S., Sampathkumar, C., Mouthami, K., & Yuvaraj, N. (2023). Sexism Classification in Social Media Using Machine Learning Algorithms. In A. Abraham, T. Hanne, N. Gandhi, P. Manghirmalani Mishra, A. Bajaj, & P. Siarry (Eds.), *Proceedings of the 14th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2022)* (pp. 14–23). Cham: Springer Nature Switzerland. doi: 10.1007/978-3-031-27524-1_2

Kirk, H. R., Yin, W., Vidgen, B., & Röttger, P. (2023, May). *SemEval-2023 Task 10: Explainable Detection of Online Sexism.* arXiv. Retrieved 2024-05-10, from http://arxiv.org/abs/2303.04222 (arXiv:2303.04222 [cs]) doi: 10.48550/arXiv.2303.04222

LadyCoders. (2024). *LadyCoders.* Retrieved 2024-03-20, from https://www.reddit.com/r/LadyCoders/

LadyDevs. (2024). *Women in Web & Software Development.* Retrieved 2024-03-20, from https://www.reddit.com/r/ladydevs/

LaikaBauss31. (2020, March). *[Misc] Set the example of respecting women in the industry by being welcoming yourself.* [Reddit Post]. Retrieved 2024-05-09, from www.reddit.com/r/girlsgonewired/comments/fsifgg/misc_set_the_example_of_respecting_women_in_the/

LaunchCoderGirl. (2024). *LaunchCoderGirl Reddit!* Retrieved 2024-03-20, from https://www.reddit.com/r/LaunchCoderGirl/

Leaper, C., & Robnett, R. D. (2016). Sexism. In R. J. Levesque (Ed.), *Encyclopedia of adolescence* (pp. 1–10). Cham: Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-319-32132-5_226-2 doi: 10.1007/978-3-319-32132-5_226-2

LesbianCoders. (2024). *lesbiancoders.* Retrieved 2024-03-20, from https://www.reddit.com/r/lesbiancoders/

Manne, K. (2018). *Down Girl: The Logic of Misogyny*. Oxford University Press.

Mayo, D. G., & Hand, D. (2022, May). Statistical significance and its critics: practicing damaging science, or damaging scientific practice? *Synthese*, *200*(3), 220. Retrieved 2024-08-11, from https://doi.org/10.1007/s11229-022-03692-0 doi: 10.1007/s11229-022-03692-0

Merriam-Webster. (2024). *Merriam-Webster: America's Most Trusted Dictionary.* Retrieved 2024-03-20, from https://www.merriam-webster.com/

Meta. (2024). *Llama 3.1.* Retrieved 2024-08-28, from https://llama.meta.com/docs/overview

Mills, S. (2008, 01). Language and sexism. *Language and Sexism*, 1-178. doi: 10.1017/CBO9780511755033

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

Oliveira, T., Barcomb, A., Santos, R. d. S., Barros, H., Baldassarre, M. T., & França, C. (2023, December). Navigating the Path of Women in So ware Engineering: From Academia to Industry. *ICSE SEIS 2024*.

Ontario Ministry of Labour. (2016, August). Code of Practice to Address: Workplace Harassment Under Ontario's Occupational Health and Safety Act. *Queen's Printer for Ontario*.

patriotn8. (2019, September). *Female engineers of Reddit, how would you describe your experience in the engineering field? What would you tell someone younger to encourage them to pursue a career in engineering?* [Reddit Post]. Retrieved 2024-05-09, from www.reddit.com/r/AskWomen/comments/czkn73/female_engineers

_of_reddit_how_would_you_describe/

Plaza, L., Carrillo-de Albornoz, J., Amigó, E., Gonzalo, J., Morante, R., Rosso, P., ... Ruiz, V. (2024). Exist 2024: sexism identification in social networks and memes. In N. Goharian et al. (Eds.), *Advances in information retrieval* (pp. 498–504). Cham: Springer Nature Switzerland.

PyLadies. (2024). *PyLadies.* Retrieved 2024-03-20, from https://www.reddit.com/r/pyladies/

queenannechick. (2023, June). *Around 10 years in, we leave. Where did you go?* [Reddit Post]. Retrieved 2024-05-09, from www.reddit.com/r/womenintech/comments/1470lz1/around_10_years_in_we_leave_where_did_you_go/

Reimers, N., & Gurevych, I. (2019, 11). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing*. Association for Computational Linguistics. Retrieved from https://arxiv.org/abs/1908.10084

Rodríguez-Sánchez, F., de Albornoz, J. C., & Plaza, L. (2020). Automatic classification of sexism in social networks: An empirical study on twitter data. *IEEE Access*, *8*, 219563-219576. Retrieved from https://api.semanticscholar.org/CorpusID:229307824

Samory, M., Sen, I., Kohne, J., Floeck, F., & Wagner, C. (2021, June). *"Call me sexist, but...": Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples.* arXiv. Retrieved 2024-05-10, from http://arxiv.org/abs/2004.12764 (arXiv:2004.12764 [cs]) doi: 10.48550/arXiv.2004.12764

Samory, M., Sen, I., Kohne, J., Flöck, F., & Wagner, C. (2021, May). "call me sexist, but..." : Revisiting sexism detection using psychological scales and adversarial samples. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*(1), 573-584. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/18085 doi: 10.1609/icwsm.v15i1.18085

Scikit-learn. (2024). *NearestNeighbors.* Retrieved 2024-08-28, from https://scikit-learn/stable/modules/generated/sklearn.neighbors.NearestNeighbors.html

SemEval. (2023). *SemEval-2023 Tasks.* Retrieved 2024-07-14, from https://semeval.github.io/SemEval2023/tasks.html

Sonke Gender Justice. (2024). *Code of Conduct.* Retrieved 2024-08-12, from https://genderjustice.org.za/code-of-conduct/

Sultana, S., Cavaletto, L. A., & Bosu, A. (2021, July). *Identifying the Prevalence of Gender Biases among the Computing Organizations.* arXiv. Retrieved 2024-05-10, from http://arxiv.org/abs/2107.00212 (arXiv:2107.00212 [cs]) doi: 10.48550/arXiv.2107.00212

Thesaurus.com. (2024). *Synonyms and Antonyms of Words | Thesaurus.com.* Retrieved 2024-03-20, from https://www.thesaurus.com/

Trinkenreich, B., Britto, R., Gerosa, M. A., & Steinmacher, I. (2022, May). An empirical investigation on the challenges faced by women in the software industry: a case study. In *Proceedings of the 2022 ACM/IEEE 44th International Conference on Software Engineering: Software Engineering in Society* (pp. 24–35). Pittsburgh Pennsylvania: ACM. Retrieved 2024-03-20, from https://dl.acm.org/doi/10.1145/3510458.3513018 doi: 10.1145/3510458.3513018

United Nations Department of Economic and Social Affairs. (2015). *Sustainable Development Goals.* Retrieved from https://sdgs.un.org/goals

United Nations Educational, S. a. C. O. U. (2019). *I'd blush if I could: Closing Gender Divides in Digital Skills through Education.* EQUALS and UNESCO EQUALS. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000367416/PDF/367416eng.pdf.multi

University, P. (2024). *WordNet.* Retrieved 2024-08-28, from https://wordnet.princeton.edu/homepage

Ussher, J. (2016, 04). Misogyny. University of Western Sydney. doi: 10.1002/9781118663219.wbegss381

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. Retrieved 2024-06-24, from https://proceedings.neurips.cc/paper/2017/

hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. (2019, August). Challenges and frontiers in abusive content detection. In S. T. Roberts, J. Tetreault, V. Prabhakaran, & Z. Waseem (Eds.), *Proceedings of the third workshop on abusive language online* (pp. 80–93). Florence, Italy: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W19-3509 doi: 10.18653/v1/W19-3509

Wakabayashi, D., Griffith, E., Tsang, A., & Conger, K. (2018, November). Google Walkout: Employees Stage Protest Over Handling of Sexual Harassment. *The New York Times*. Retrieved 2024-05-06, from https://www.nytimes.com/2018/11/01/technology/google-walkout-sexual-harassment.html

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017, August). Understanding abuse: A typology of abusive language detection subtasks. In Z. Waseem, W. H. K. Chung, D. Hovy, & J. Tetreault (Eds.), *Proceedings of the first workshop on abusive language online* (pp. 78–84). Vancouver, BC, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W17-3012 doi: 10.18653/v1/W17-3012

WayRoundTheWorld. (2018, February). *I'm failing with female supervisors and don't know why (I'm a woman)* [Reddit Post]. Retrieved 2024-05-09, from www.reddit.com/r/cswomen/comments/7vdeqb/im_failing_with_female_supervisors_and_dont_know/

WomenInTech. (2024). *Women in Tech.* Retrieved 2024-03-20, from https://www.reddit.com/r/womenintech/

WomenWhoCode. (2024). *womenwhocode.* Retrieved 2024-03-20, from https://www.reddit.com/r/womenwhocode/

Wrisley, S. (2021, 10). Feminist theory and the problem of misogyny. *Feminist Theory*, *24*, 146470012110393. doi: 10.1177/14647001211039365

XXSTEM. (2024). *Women that STEM.* Retrieved 2024-03-20, from https://www.reddit.com/r/xxstem/