Scientometrics (2025) 130:4915–4971
https://doi.org/10.1007/s11192-025-05410-2

Divided by discipline? A systematic literature review on the quantification of online sexism and misogyny using a semi-automated approach

Aditi Dutta[1,4] · **Susan Banducci**[1] · **Chico Q. Camargo**[1,2,3,4]

Received: 24 August 2024 / Accepted: 13 August 2025 / Published online: 7 October 2025

Abstract
Several computational tools have been developed to detect and identify sexism, misogyny, and gender-based hate speech, particularly on online platforms. These tools draw on insights from both social science and computer science. Given the increasing concern over gender-based discrimination in digital spaces, the contested definitions and measurements of sexism, and the rise of interdisciplinary efforts to understand its online manifestations, a systematic literature review is essential for capturing the current state and trajectory of this evolving field. In this review, we make four key contributions: (1) we synthesize the literature into five core themes—definitions of sexism and misogyny, disciplinary divergences, automated detection methods, associated challenges, and design-based interventions; (2) we adopt an interdisciplinary lens, bridging theoretical and methodological divides across social psychology, computer science, and gender studies; (3) we highlight critical gaps, including the need for intersectional approaches, the under-representation of non-Western languages and perspectives, and the limited focus on proactive design strategies beyond text classification; and (4) we offer a methodological contribution by applying a rigorous semi-automated systematic review process guided by PRISMA, establishing a replicable standard for future work in this domain. Our findings reveal a clear disciplinary divide in how sexism and misogyny are conceptualized and measured. Through an evidence-based synthesis, we examine how existing studies have attempted to bridge this gap through interdisciplinary collaboration. Drawing on both social science theories and computational modeling practices, we assess the strengths and limitations of current methodologies. Finally, we outline key challenges and future directions for advancing research on the detection and mitigation of online sexism and misogyny.

Keywords Systematic literature review · Online sexism and Misogyny · Semi-automated publication analysis · Applied natural language processing · Scientometrics

Extended author information available on the last page of the article

Introduction

> "*Millions of women and girls are affected by digital abuse and technology facilitated violence every year. Studies suggest that between 16 and 58 per cent of women have experienced this type of violence.*" (Secretary-General (2024))

The rapid growth of online spaces has been accompanied by increased online abuse targeting marginalized groups (Wilson & Land, 2020; FRA, 2023; Vidgen & Derczynski, 2020). Girls and women, in particular, have experienced hostility and harassment in online spaces and platforms (Jurasz & Barker, 2019; Nadim & Fladmoe, 2021; Vitak et al., 2017). Research shows that women are twice as likely as men to experience gender-based online harassment (Duggan, 2017), often resulting in self-censorship and withdrawal from digital spaces (Mantilla, 2013; FRA, 2023; International, 2017). This growing concern on the disproportionate impact of online hate speech towards girls and women has given rise to an active interest among the research community in countering online sexism and misogyny (Megarry, 2014; Guest et al., 2021; Citron, 2014), and an increase in research on quantifying the same using machine learning approaches (Vidgen & Derczynski, 2020).

However, these computational approaches face several limitations. They are difficult to implement effectively (Hewitt et al., 2016; Nozza et al., 2019; Samory et al., 2021), as they differ fundamentally in how sexism and misogyny are defined, measured, and operationalized—reflecting the complexity and multiplicity of the underlying concepts (Richardson-Self, 2018; Matsuda, 2018). While these approaches show impressive performance, they fail to identify and capture all forms of sexism or misogyny—especially overlooking the subtler forms of sexist discourse (Rodriguez-Sanchez et al., 2020; Rodríguez-Sánchez et al., 2021), and are often prone to erroneous classifications. This calls for the need to examine the current state of research in online sexism or misogyny, and identifying the current challenges arising due to disciplinary and methodological divide. Compounding this issue is the disciplinary divide: while social science research often adopts qualitative methods to explore sexism in rich, contextualized ways (Yasseri et al., 2016), computational work tends to rely on narrow, binary definitions, with limited integration of social science theories to analyze the immense amount of available online data on sexism and misogyny.

Despite growing work, there is no integrative, methodologically rigorous synthesis bridging theoretical frameworks from social science with practical automated computational approaches in computer science. This gap is particularly critical given that language is a form of social behavior that reflects identity and power structures (Dinan et al., 2020). Text analysis has been proven to be one of the established methods in mapping and analyzing hostility in online discourses, particularly for online gendered hate-speech (Jane, 2016). Thus, the need arises to apply a natural language processing (NLP) approach to analyze such data to advance both sociological understanding of the kind of sexism existing in online spaces, and methodological understanding of using and improving computational models to capture the same through detection and identification tasks.

Nonetheless, most of the earlier works have neglected or retrofitted the link between the data and sexism as a theoretical construct (Samory et al., 2021; Abburi et al., 2021; Parikh et al., 2019). Primarily, sexism and misogyny has been researched as a part of the hate-speech diaspora, disregarding the forms of sexism 'not involving hate' (Parikh et al., 2021), or other non-hostile forms that are subtle and often deceptive (Jha & Mamidi, 2017; Rodriguez-Sanchez et al., 2020; Rodríguez-Sánchez et al., 2021; Rodriguez-Sanchez et al.,

2022). While online communities now emphasize on the detection of sexism [or misogyny] (and other hate speech) more than ever before, automatic detection of these phenomena remains challenging, as most research focuses on using textual features to solve the issue (Das et al., 2023).

Recognizing this need, Fontanella et al. (2024) conducted a systematic literature review on the study of misogyny using computational methods, where they find a "limited connection between the areas of knowledge that are necessary to fully grasp this complex phenomenon". Through our research, we extend the review on articles beyond misogyny with an extensive discussion on the identified practices, along with their challenges and limitations in implementation, backed by social science literature. The review aims to consolidate, categorize, and critique existing approaches to quantifying online sexism and misogyny, identifying key research gaps and challenges through a theoretically informed, methodologically rigorous framework.

To address these gaps systematically, we developed a hybrid review pipeline that combines advanced NLP techniques with traditional PRISMA-based rigor—allowing for scalable yet methodologically transparent analysis of the literature. First, we applied transformer-based topic modeling (BERTopic) alongside UMAP dimensionality reduction and HDBSCAN clustering to group abstracts into coherent thematic clusters. Next, we validated and refined these clusters via a KeyBERT-driven keyword co-occurrence network, ensuring that selected topics truly reflected our research questions. Finally, we integrated these automated steps with manual title/abstract and full-text screening to enforce quality and consistency. This hybrid workflow not only accelerates large-scale literature reviews but also maintains the transparency and reproducibility demanded by systematic review standards.

In this paper, we use *quantification* to refer broadly to the identification, classification, or detection of sexism and misogyny in online texts. Our goal is to map how these constructs are operationalized across disciplines, assess the limitations of existing practices, and pave the way for future research that better integrate theory, method, and application.

To address these gaps, this paper makes four key contributions. First, it synthesizes the fragmented literature across social science and computer science into five central themes: definitions of sexism and misogyny, disciplinary divergences, computational detection methods, current challenges, and design-oriented interventions—addressing the lack of conceptual and methodological cohesion. Second, it bridges disciplinary silos by integrating insights from gender studies, social psychology, and NLP, offering an interdisciplinary perspective on detection models. Third, it highlights critical underexplored areas, such as the absence of intersectional approaches, the dominance of Western contexts and languages, and the limited focus on proactive system design beyond keyword matching. Finally, it provides a methodological contribution through a semi-automated systematic review pipeline, combining topic modeling, network analysis and PRISMA guidelines to establish a transparent and scalable standard for future research in this space.

Background

Defining sexism and misogyny: challenges and contexts

Understanding and detecting sexism and misogyny—particularly through automated means—requires careful attention to how these terms are defined. However, definitions of

both concepts often prove too narrow to capture their full complexity, especially in computational contexts where nuanced social phenomena must be rendered in operational terms. Despite the central role these concepts play in understanding gendered power relations, there is no cross-disciplinary consensus on their precise definitions. Sexism and misogyny are central concepts in understanding the status of women yet there is no consensus across disciplines on their definition. Wrisley (2023) highlights the difficulty of establishing operational definitions for sexism and misogyny, noting that both terms have evolved well beyond their original conceptual boundaries. This ambiguity is compounded in computational fields, where definitional clarity is often sacrificed for operational convenience. Consequently, many studies tailor their definitions to align with their specific research aims, particularly in the detection of hate speech that manifests as sexism, misogyny, or both.

Manne (2017) provides a foundational framework for distinguishing the two: sexism functions to "justify [patriarchal] norms, largely via an ideology of supposedly 'natural' differences between men and women concerning their strengths interests, proclivities, and appetites," whereas misogyny serves to "uphold the social norms of patriarchies by policing and patrolling them" indicating it is a systemic property embedded within social structures. While Richardson-Self (2018) distinguishes between sexist and misogynistic speech, arguing that while sexist speech can oppress without overt violence, misogynistic speech often exhibits key features of hate speech. Building on this systemic view, Srivastava et al. (2017) defines misogyny as a form of hatred or contempt for women, arising directly from patriarchal systems. Traditionally rooted in face-to-face social interactions, misogyny has also functioned historically as a political mechanism to domesticate women, control their sexuality, and undermine collective feminist solidarities (Anderson, 2014).

Some scholars treat sexism and misogyny as synonymous or closely related (Rahali et al., 2021; Bhattacharya et al., 2020; Abburi et al., 2021), while others conceptualize misogyny as a subset or intensified form of sexism (Butt et al., 2021; Rodriguez-Sanchez et al., 2020). Even when the distinction is acknowledged, the terms are frequently used interchangeably in computational studies due to their frequent co-occurrence (Frenda et al., 2018). Much of the computational literature adopts this interchangeable usage, informed by theoretical positions that treat misogyny as an extreme articulation of sexist ideology (Chiril et al., 2020; Zeinert et al., 2021; Kohli et al., 2021).

Given these definitional challenges, this study aims to synthesize the broad spectrum of existing definitions and examine how sexism and misogyny are conceptualized across both social science and computer research domains.

From offline harm to online hate: the rise of digital sexism and misogyny

Back in 2013, the World Health Organization, WHO (2013) identified violence against women as "a global health problem of epidemic proportions", primarily referring to offline violence, while also warning of its likely expansion into social media. Indeed, the Internet—particularly social media—has since become a key space for the perpetration of sexism and misogyny, where women are subjected to various forms of violence (Jurasz & Barker, 2019). Prior research has also highlighted the role of specific linguistic forms and categories—such as the generic masculine[1]—in reinforcing prejudices, sexist attitudes, and gender stereotypes (Sensales & Areni, 2017). These manifestations may take different forms but share a common aim: to discredit women's participation in public life and silence

[1] A gender-biased form used to indicate those of both masculine and feminine gender, reinforcing a hierarchy that privileges men.

their political voices (Jurasz & Barker, 2019). In recent years, systemic gender inequality has increasingly manifested in cyberspace through the proliferation of abusive content that is even more aggressive, prompting further research into this evolving form of online misogyny (Fontanella et al., 2024). As a result, online platforms have contributed to the erosion of boundaries between online and offline experiences (Megarry, 2014).

Gendered harassment and the silencing of women online

Even self-identifying as a woman online can significantly increase the risk of internet harassment. When gender identity is known, gender stereotyping and discrimination from the "real world" often carry over into digital spaces, contributing to a "gender asymmetry" in the dynamics of online abuse (Herring, 1999). Even seemingly neutral actions—such as the perceived tone of a post—can be enough to "trigger" misogynistic mockery. Speaking out against such behavior often invites further backlash, with responses that are both sexist and misogynistic in nature, and notably, these can come from both men and women. Those who deliberately derail online feminist spaces often do so to suppress the free speech of those communities (Bartow, 2009). Megarry (2014) contextualizes online abuse within discursive practices, arguing that such hostility seeks to silence women's voices on digital platforms and regulate their public behavior.

The overwhelming volume of gendered abuse online raises serious social concerns. While some victims have been celebrated for exposing abusers through acts of 'feminist digilantism' such responses risk reinforcing the notion that these issues should be addressed privately by individuals rather than collectively through systemic or public intervention (Jane, 2016). Crucially, the impact of misogyny extends beyond psychological harm; it also has material consequences, particularly in how resources and opportunities are distributed in society. Thus, understanding misogyny and gender-based violence in online contexts requires a deeper exploration of their complex entanglement with digital culture and technology—an understanding that is essential to shaping equitable digital gender politics for the future (Ging & Siapera, 2018, 2019).

Given its impact, online misogyny and sexism can be seen as "seeking to prevent women from participating in building the forthcoming technological future" (Ging & Siapera, 2018). It is therefore necessary to stop such proliferation in online spaces to promote gender equality, raise awareness and eliminate it at the earliest by detecting them through computational tools.

Research questions

The challenge of considering sexism and misogyny from a quantitative perspective, when considering their highly subjective nature, motivates our research questions:

RQ1: What are the main topics in the studies identified, and how do they differ by discipline and over time?
RQ2: How has the existing literature operationalised sexism and misogyny?
RQ3: What are the main challenges and opportunities of computational approaches to the study of sexism and misogyny? Which of the challenges do they address?

The main objective of this paper is to provide a comprehensive systematic literature review, drawn from the research landscape of sexism and misogyny, studied over the years of 2012–2022. The aim is not to focus on specifics from any individual paper but to provide a general overview of the existing literature and draw conclusions from their study designs and research outputs. These observations are to inspire researchers on best working practices and approaches, while also contributing to future research objectives.

Our systematic literature review is divided into two stages: (a) Identifying the relevant studies through multiple steps by performing a semi-automated selection flowchart as illustrated in the PRISMA flowchart (Fig. 1) in Sect. 4, (b) Conducting an in-depth analysis of the selected study results in Sect. 5. While stage 1 is expected to answer the first research question, stage 2 will answer the second and third research questions.

Identifying relevant studies

Search strategy

We searched six databases—Google Scholar, ArXiv, Elsevier, Scopus, Semantic Scholar, and Web of Science—using a closely related set of keywords that operationalized our review criteria of 'quantifying' sexism and misogyny. This returned a comfortable number of results that were useful for performing the quantitative analysis. Search results were implemented such that the range of year of publication lay between 2012 and 2022. All of the articles should be in English, containing the full abstracts and titles for each of them. The reporting strategy follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), presented in Fig. 1,[2] which uses a checklist approach to systematic literature reviews.

This research was conducted to review papers with three main characteristics, namely:

(1) The papers study sexism and/or misogyny.
(2) The papers ideally study the propagation in social media platforms or other broadcast (preferably text-based) media.
(3) The papers use various methodologies for measuring or quantifying sexism and misogyny (e.g., scales, models, etc.).

ArXiv and Web of Science were chosen to collect studies from the fields of CS and SS, using the search criteria as shown in Table 2. As the Fig. 1 illustrates, the systematic screening and selection process used in this review. An initial total of 1745 records were retrieved from the databases, comprising 1511 results from Web of Science and 234 from arXiv. An additional 72 records were identified through external sources during the first stage of the review. Following the removal of 104 duplicates, 1691 records met the inclusion criteria and were retained, while 22 partially eligible records underwent further assessment and were subsequently excluded. Automated exclusion techniques—comprising text pre-processing, topic modeling, and network analysis—were then applied, resulting in a subset of 84 records selected for title and abstract screening.

[2] "The flow diagram depicts the flow of information through the different phases of a systematic review. It maps out the number of records identified, included, and excluded, and the reasons for exclusions. Different templates are available depending on the type of review (new or updated) and sources used to identify studies" (Takkouche and Norman, 2011).

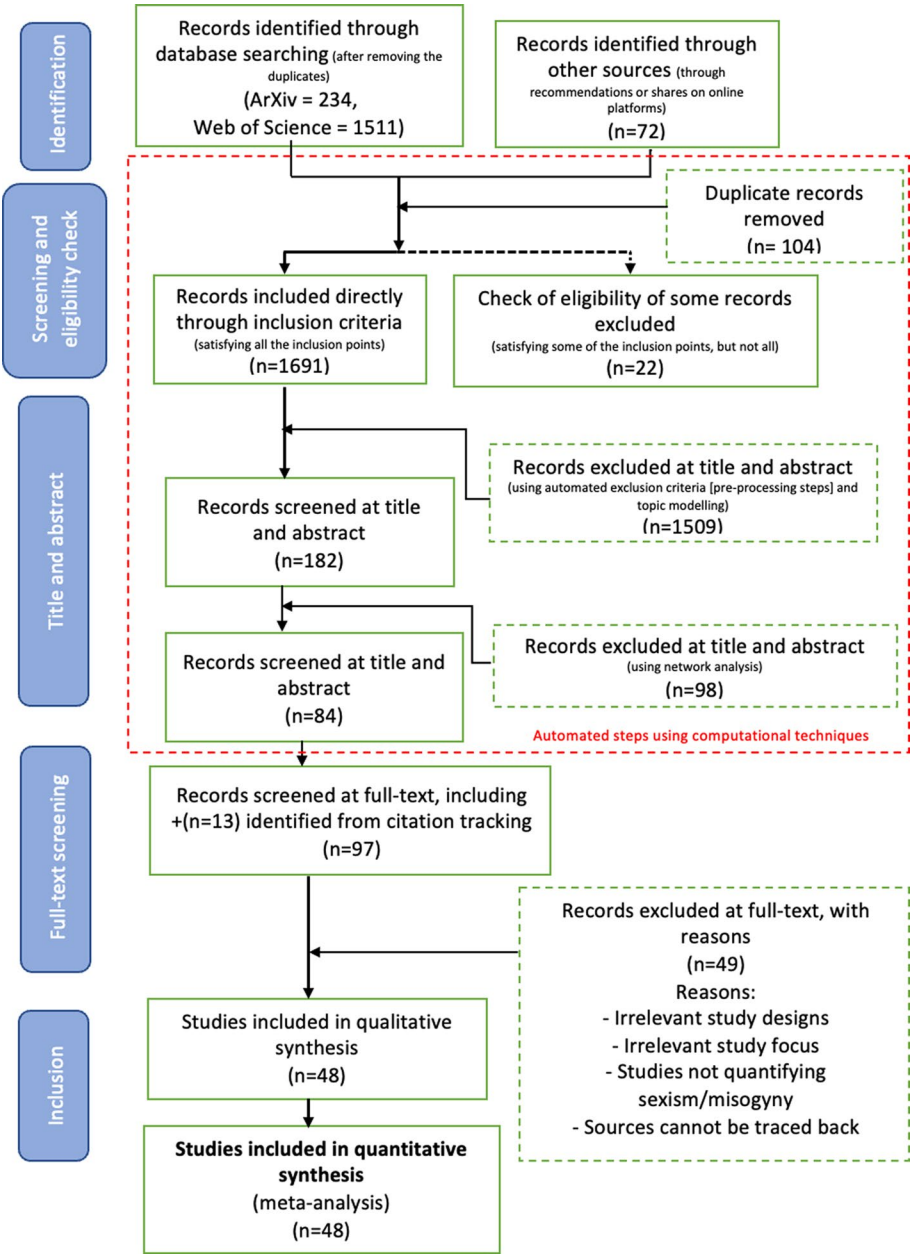

Fig. 1 PRISMA flowchart diagram ((Takkouche and Norman, 2011)) for this research. Each step shows the number of studies included and eliminated at that point of the research

Subsequently, 97 articles (including 13 from citation tracking) were reviewed at the full-text level. After excluding 49 articles due to irrelevance or lack of methodological clarity, a final set of 48 studies was included for both qualitative and quantitative synthesis.

This rigorous process incorporated both manual and computational methods to ensure a focused and comprehensive literature base.

The data collection method will be discussed in detail in the next section.

Data search and collection

In this section, we elaborate on the experimentation conducted with each of the citation databases, and the advantages and disadvantages encountered during the study. For this research, some fields of the search results, namely - title, abstract, year of publication, and the discipline of research for each of the search results were integral to the study. To perform the automated step of narrowing down our search results, some measures were taken to check the consistency and reliability of the data, which is shown in Table 2.

ArXiv is a platform that offers researchers to e-publish a draft version of their final work preceding a formal peer review and publication in a peer-reviewed scholarly or scientific journal, also referred to as 'pre-prints'.[3] Due to the popularity of ArXiv among CS researchers, its API was used with the expectation of returning unpublished or pre-published works for all disciplines. However, it was found that only the areas of CS vary widely. "In theoretical computer science and machine learning, over 60% of published papers are on ArXiv, while other areas are essentially zero" Sutton and Gong (2017). We opted to use advanced search queries to narrow down the results, as simpler queries were expected to return more irrelevant results, that had to be removed before analysis. Though the API returned only a limited number of papers, most of them were found to be relevant. Hence, we took it for analysis but did not use it as our only source due to its skewed disciplinary variety.

Web of Science and *Scopus* showed results in retrieving studies from CS. Though Zhang (2014) found that Scopus retrieved "significantly" more studies in CS as compared to the Web of Science, with all of the kinds of document types - conference proceedings, journal articles, reviews, and editorials; yet for our search type, more relevant works were found in Web of Science. As Fiala and Tutoky (2017) mentioned in their work, CS has a greater reliance on conference proceedings as compared to other disciplines. To some extent, these conference proceedings papers are also indexed in Web of Science in the Conference Proceedings Citation Index, which makes it possible to carry out scientometric studies of CS based on the data from Web of Science (Fiala & Tutoky, 2017).

For *Google Scholar*, we used two external APIs like SerpAPI for scraping the data, as well as a software named '*Publish or Perish*' (Harzing, 2007) to collect the search results. Both of the methods were rejected because of their disadvantages. Such as, Publish or Perish could only extract 1000 results at a time for each search query. While this drawback was overcome by searching for documents with a shorter range of years to stay within the limit, it lacked some of the fields that were needed for this study—abstract and discipline. Alternatively, SerpAPI[4] worked similar to a web scrapping tool and could only scrape the results as the search engine demonstrates, i.e., it only scrapes what Google shows on their Google Scholar pages, nothing more. Even though the fields we got through this API were relevant, they did not contain the full information we needed for the analysis. For example, the full

[3] A preprint is a full draft research paper that is shared publicly before it has been peer-reviewed. Most preprints are given a digital object identifier (DOI) so they can be cited in other research papers. A preprint is a full draft of a research paper that is shared publicly before it has been peer-reviewed (Mudrak, 2018).

[4] https://serpapi.com/.

text in the title and abstract was missing and was instead indicated with dotted extensions in the beginning and end of the text. For the remaining tested citation databases—*Elsevier* and *Semantic Scholar*, the possible search queries were either too simple (consequently giving back a lot of irrelevant studies), did not give back enough studies on our topic, or lacked some of the essential fields (e.g., abstract) that were integral to this study, especially for the automated search strategy used to eliminate non-relevant studies.

Therefore, we found empirical evidence indicating that the research outputs we got from ArXiv and Web of Science were ideal for our work. Alongside the search queries, we augmented the dataset with manually added papers that satisfied the selection criteria (Sect. A.2). This data from external sources included studies shared in the social platforms Twitter (or X) and LinkedIn, recommendations of other researchers in the field, and following the references of the reviewed papers (i.e., citation tracking).

Final methodology selection criteria

Observing the pros and cons of all the citation databases, it was decided to use the Web of Science API to collect data based on the individual areas of discipline—SS and CS, as the primary data source. Since many of the relevant computational papers were seen to be published in ArXiv within the given period, those papers were also considered as part of the data collection. It was done to ensure that we get full coverage of both published and unpublished works (pre-prints), relevant to the study of sexism and misogyny during the 11 years. As discussed before, we also included the publications that were informed through external sources. While the Web of Science was taken as the main source for published works, ArXiv was taken as a source for unpublished works. We then combine the selected search results for the next Sect. 4.3, before removing the duplicates.

Data extraction and synthesis

In this section, we first provide an overview of the collected data from the previous Sect. 4.2, and then use automated approaches for the data extraction stage. The analyses are performed before the application of the selection criteria (See Sect. A.2). For each of the following subsections, the fields considered were:

- Title of the paper,
- Abstract (Multiple abstracts of the same paper were replaced with the first abstract),
- Year of publication (or pre-printing),
- Language of the paper.

Figure 2 shows a steep rise in the study and publication of research on sexism and misogyny, in both the fields of CS and SS. While SS studies always dominated research on the topic, CS works also showed admirable improvement, with a lot of the papers getting published in 2022 alone. The study of online sexism and misogyny has grown significantly since 2014, with a notable rise in scholarly attention from 2018 onwards. Figure 2 illustrates the yearly distribution of selected publications, revealing a marked upward trajectory that peaks in 2021.

As we had discussed in the Sect. 4.2, there has been a rising trend of pre-prints in CS (see Fig. 8), many of which were later published and indexed in citation databases. Studies researching social media platforms like Facebook, Twitter, and Instagram were

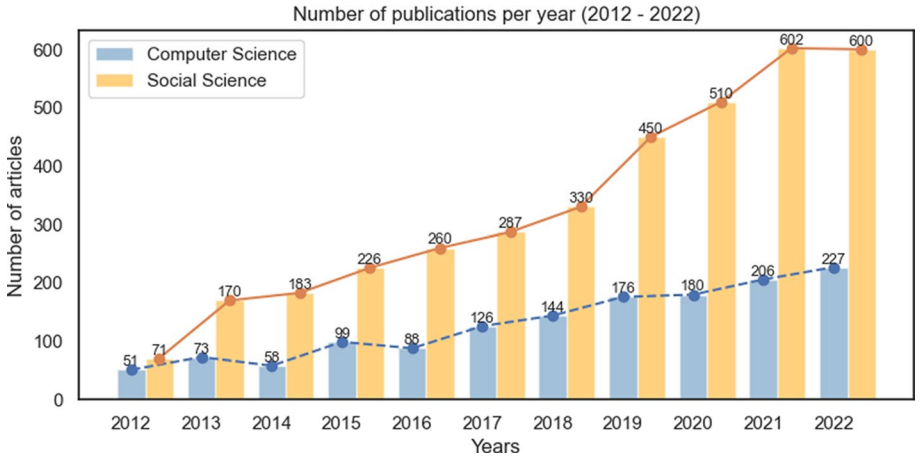

Fig. 2 Number of publications per year. The blue bars reflect the research articles on Computer Science, while the yellow bars reflect the research articles on Social Science, between the years of 2012 and 2022. (Color figure online)

seen to be limited, with less than 100 works dedicated to research on sexism and misogyny in these online platforms. While almost all of the returned results indicated that works were published majority in English, among the other languages—Spanish and Portuguese followed through, though separated by huge margins.

Pre-processing of the text was done to drop duplicates and remove characters in the text that could hinder the automated selection of the studies based on the titles and abstracts. Studies containing no abstracts at this stage were removed as they could not be added for automated selection criteria. Given that the count of such papers was only 13, the abstracts were looked up in Google Scholar and later manually checked, if they satisfied the selection criteria for this research.

For the automated extraction stage, we perform two steps in chronological order: **topic modeling** and **keyword co-occurrence network** to narrow down our search.

Topic modeling

Topic modeling[5] was used with the pre-processed data containing the abstracts and titles from both disciplines, to generate clusters of topics based on the documents (i.e., the collection of studies containing abstracts and topics). Among all the topic modeling techniques experimented with, BERTopic (Grootendorst, 2022) proved to be the best choice for the task. It is because BERTopic "leverages transformers and c-TF–IDF to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions" (Grootendorst, 2022), hence enhancing the topic recognition ability by the model.[6]

[5] "Topic modeling is a machine learning technique that automatically analyzes text data to determine cluster words for a set of documents. It leverages 'unsupervised' machine learning to analyze and identify clusters or groups of similar words within a body of text" (Pykes, 2023).

[6] More information on the techniques used in this methodology is explained in the supplementary section: J.1.

We applied the BERTopic algorithm to the collections of CS and SS papers separately to capture the topics of research between these two disciplines, and to check the differences in the themes of sexism between them. Among all the experimentation conducted—including setting different ranges of parameters to get the best representative models from there, we further employed fine-tuning of the model to improve on that, by using multiple representations from the model. For our work, we used these different representations from keywords and phrases to summarize and custom labels. Figures 3 and 4 indicate the topics recognized by the model. Using the aforementioned parameters, we used the BERTopic model to groups documents into topic clusters, identified by their keywords and keyphrases. It uses clustering to define topics and hence does not assign more than one topic to each document. In the figures, each point corresponds to each document in their respective disciplines. BERTopic uses HDBSCAN by default for clustering, and it does not force all the data points to be a part of any of the recognized clusters or topics. Simultaneously, BERTopic uses UMAP to perform dimensionality reduction. We then used further customization of the UMAP by setting the parameter 'n_components' to 2, to 'pre-reduce' embeddings for visually depicting our model results in the two figures. For those topics that do not form a part of any groups (also termed as "outliers"), the points are marked in grey in the figures. The colored points in both the figures indicate topics, and each color represents a unique topic for the sets of documents, which have further been marked correspondingly with labels of the same color boxes. Larger clusters represent more densely populated or well-defined research areas, indicating a richer or more mature body of literature. The algorithm itself exhibits strong local clustering to group similar topic categories together, to which we also controlled the balance between the local and the final structure to efficiently distinguish between each topic.

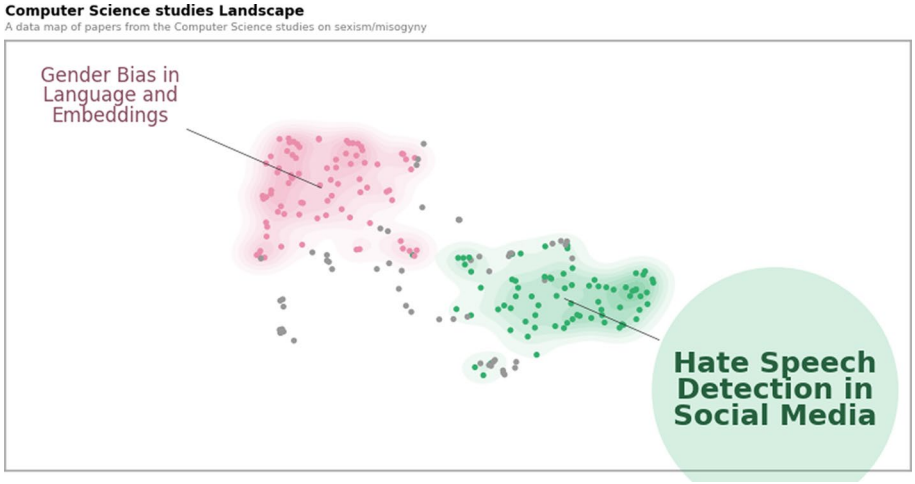

Fig. 3 This figure shows a UMAP scatterplot, where each point represents one document. The unique colors in the figure represent a different topic in computer science centering around sexism and misogyny between 2012 and 2022. Through topic modeling, usually each document gets assigned a set of key words as themes within the paper, which are then grouped together with a unique color, representing the same topic with similar sets of keywords found across all the documents. When grouped, each topic is described by their topic name in the same color. The grey points represent outliers (documents which did not get any assigned topic). The highlighted topic name indicates more relevance to our research objectives. (Color figure online)

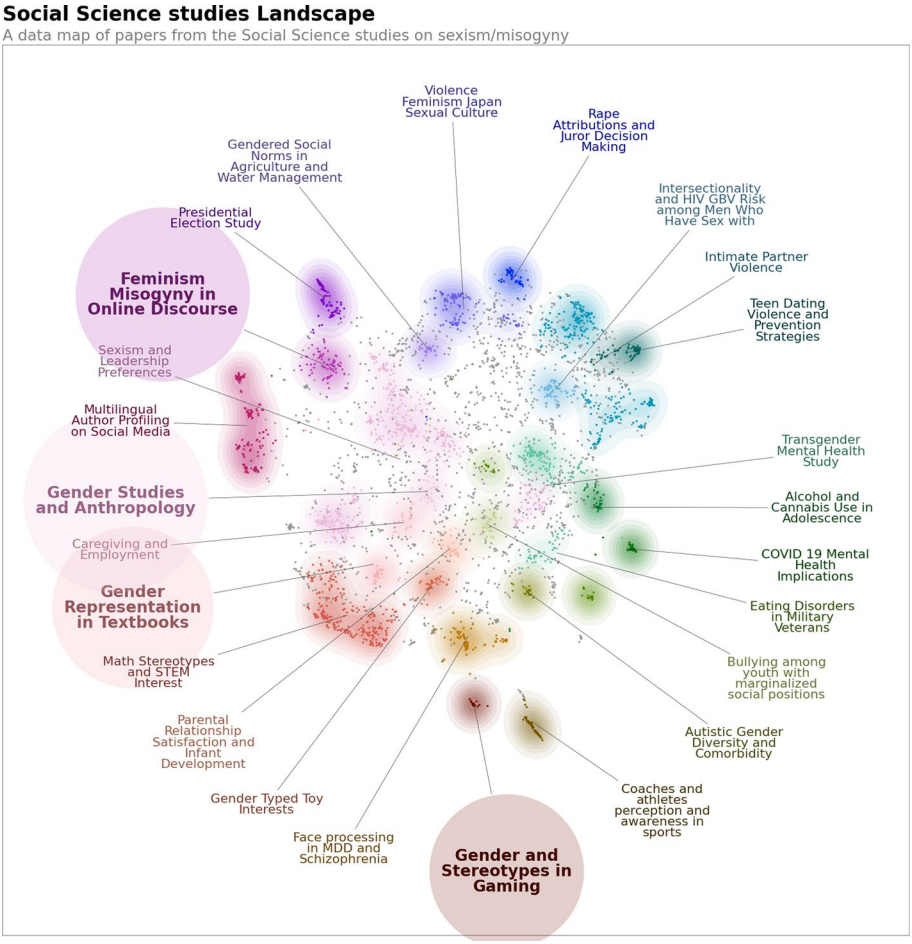

Fig. 4 Similar to Fig. 3, this figure shows a UMAP scatterplot where each unique color represents a different topic in social science centering around sexism and misogyny between 2012 and 2022. The highlighted topic indicates more relevance to our research objectives. (Color figure online)

It uses a light hue of the same colors encircling each topic to indicate the cluster belonging to the respective topic. In Figs. 4 and 3, different colors indicate different cluster of topics, as identified by the model. Usually in topic modeling, the models analyze "bags" or groups of words together to capture the meaning of the words. However, in our approach, we used the BERTopic framework using a Large Language model (LLM)—Mistral for better inference and contextual understanding of the abstracts and generation of logical topics (indicated by each color). While some points in the same cluster may look further away than the points from another cluster, it is due to its projection in 2D-dimensional space which we did for better visualization; hence the points within the same clusters are closer in a multi-dimensional space.

Usually in topic modeling, models analyze "bags" or groups of words to capture word co-occurrence patterns. However, in our approach, we used the BERTopic framework in combination with a Large Language Model (LLM)—specifically, the

OpenHermes-2.5-Mistral-7B (Q3_K_M.gguf) model—to enhance contextual understanding and generate more coherent, semantically meaningful topics based on the abstracts. This allows the model to generate more logical and semantically coherent topics, reflected by each color in Figs. 3 and 4. The model leverages multiple representations from keywords and phrases to summarize and create custom labels, improving the quality of topic inference beyond traditional bag-of-words methods.

The highlighted circles represent the most relevant thematic areas for our study. Figure 4 illustrates key domains within social science research, including digital feminism and online misogyny, gender representation in education and media, health-related concerns such as mental health and gender-based violence, and legal–cultural perspectives on sexual violence. The left side of the map emphasizes discourse and identity, while the right side is more focused on public health and psychological issues. The central regions reflect interdisciplinary intersections, suggesting potential for integrating discourse analysis with clinical and educational research to deepen understanding of gendered bias. Figure 3 illustrates the computer science research landscape on sexism and misogyny, revealing two dominant and distinct thematic clusters. The first, "Gender Bias in Language and Embeddings", centers on studies that investigate algorithmic bias in NLP models, focusing on how word embeddings and language representations encode and perpetuate gender stereotypes. The second cluster, "Hate Speech Detection in Social Media", includes work on building and evaluating systems for identifying misogynistic and abusive language online. Unlike the diverse and interdisciplinary structure observed in social science literature, the computer science landscape appears more siloed, with limited overlap between technical fairness research and applied detection systems. This highlights a relatively narrow scope and methodological uniformity in the field's engagement with gendered bias. We observe that social science studies generally lack research focused on the automated detection or identification of online sexism and misogyny, which is the focal area of this study. Therefore, we chose to exclude the social science articles from further analysis and concentrate exclusively on those from computer science.

Online sexism and misogyny are subsets of online hate speech. Therefore, in this study, we focus on the topic within Computer Science that specifically addresses the quantification of such content, rather than broader analyses of gender bias in its various forms (which was one of the two topics identified through topic modeling). To quantify hate speech, it is essential to review research that emphasizes detection tasks. Therefore, at this stage, we selected the topic **'Hate Speech Detection in Social Media'** from among the various categories and disciplines considered.

However, while topic modeling provided a structured overview of thematic clusters, it does not inherently guarantee that these clusters are semantically coherent or directly aligned with our review's focus. To strengthen the reliability of our topic selection, we conducted a keyword co-occurrence network analysis in the next section. This complementary step enabled us to assess the internal consistency of key concepts within the selected cluster, ensuring that the documents truly reflected the computational study of online sexism and misogyny. The analysis helped us validate our selection and informed the subsequent full-text review.

Keyword co-occurrence network

To validate if the topics captured from the automated selection of topics from each discipline in the previous section were representative of the corresponding documents, we

navigated the disciplines and each topic, alongside their respective keywords. To obtain the most frequent keywords in the set of documents, we use KeyBERT[7] (Grootendorst, 2020) to extract embeddings with a BERT model to get a document-level representation from our abstracts and titles. From each document, we used KeyBERT to identify key phrases that would provide with a more accurate summary of the documents, rather than simple keywords. KeyBERT works by creating an embedding of document texts, from which BERT key phrase embeddings of a pre-defined word *n*-gram range length of 1–2 words[8] were created. Consequently, cosine similarities between the document and their respective keyphrase embeddings are calculated to extract the top 10 keyphrases that best describe that document. These selected keyphrases per document are then compared against the whole set of documents. We chose to look into the 100 most common keywords in the documents taken at both discipline level (CS and SS), as well as the topic level (each topic based on the topics we generated in Sect. 4.3). This was done to check the relevance of the keywords, and consequently the set of documents that would best represent our research objective of performing a literature review on the quantification of online sexism and misogyny.

On comparing the keywords present in Computer Science and Social Science, we found that the contents of the papers (from their title and abstract) focused on different kinds of sexism and misogyny—both indicating similarity in topics, but contribution at different capacities. They reveal a complementary yet fragmented landscape in how sexism and misogyny are studied across disciplines. In Computer Science (see Fig. 11), research tends to cluster tightly around specific psychological constructs—such as gender stereotypes, hostile and benevolent sexism, ambivalent sexism, general sexist attitudes, and the broader term "sexism" itself—reflecting a strong focus on operationalizing and detecting discrete forms of bias in online text. In contrast, Social Science scholarship (see Fig. 12) embraces a broader, more intersectional view: central themes include gender stereotypes and roles, but extend into areas such as gender equality, identity (including transgender experiences), domestic and sexual violence, critical discourse, and systemic discrimination.

This divergence suggests that computational approaches risk oversimplifying the complexity of sexism by concentrating on narrowly defined categories, while social-theoretic work may lack the granular tools needed for large-scale detection. Bridging these perspectives would offer a path forward: embedding rich, contextualized constructs from social science themes of power, identity, and structural inequality into algorithmic models to enhance their sensitivity to subtle, intersectional manifestations of sexism and misogyny. Conversely, employing automated keyword and topic-modeling techniques to capture emergent patterns (e.g., discourse around political discussions) can help social scientists process vast corpora and refine theoretical frameworks. Ultimately, a synergistic integration—where computational precision is being guided by theoretic depth from social science will yield more robust, fair, and actionable tools for understanding and combating online sexism. We will discuss more on this in Sect. 5.3.

On further analysis, we selected the most relevant topic among all of the highlighted topics from both disciplines and performed a keyword search on each of them. Figure 5 corresponds to the topic 'Hate Speech Detection in Social Media', and it showed the most promising result of containing the necessary keywords needed for this study.

[7] KeyBERT is a minimal and easy-to-use keyword extraction technique that leverages BERT embeddings to create keywords and key phrases that are most similar to a document.

[8] Word *n*-gram range lets users decide the length of sequence of consecutive words that should be extracted from a given text.

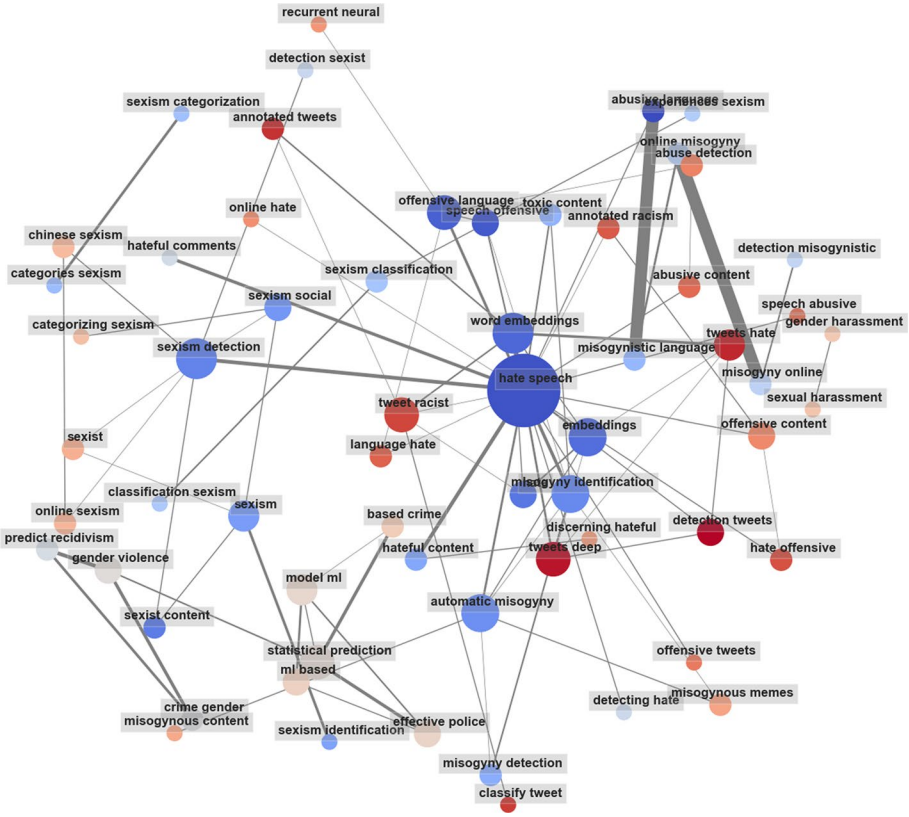

Fig. 5 Most frequent keywords gathered from the abstracts and titles of Computer Science studies in the topic of **'Hate Speech Detection using Deep Learning models'**

The figure consists of the 100 most frequent keyphrases in the topic. The size of each circle (node) indicates the weight of that particular keyphrase in the set of documents. The color of the nodes can be anything that is recognized by Matplotlib in colormap specified and could be randomly generated. The edge width of the edges joining the nodes indicates the number of associations between the two connected nodes—the more the number of connections, the thicker the edge width. The co-occurrence network for all the highlighted topics from the previous step showed us the most used keywords pertaining to the said topics. By studying the generated keywords, we could capture the general objectives of the respective papers under the same topic, as stated from their abstracts. We then selected the most relevant topic for our work, which focuses on computational approaches in detecting online sexism and misogyny (a part of the online hate speech discourse). Consequently, we decided to perform the full-text literature review on the articles which fell under the topic of 'Hate Speech Detection in Social Media'.

On completing this step, we included a few more of the studies that had been identified through citation tracking and did a full-text screening of all the collected texts.

Note: As we see from the plots, till here, all of the analyses largely relied on the information provided in the title or the abstract of the papers. Hence, it limits us in

providing a concise assessment of the exact models, methodologies, or datasets used by the corresponding papers. For which we would need a full-text assessment.

Final data selection

Screening

Following the selection of the citation database, the automatic filtering of papers using BERTopic, and the validation via topic keywords, we identified the necessary characteristics of the data to emphasize the findings which eventually led us to select the research articles based on the selection criteria as noted in Sect. A.2, by performing a manual screening. Eligible articles were divided into two categories at this stage. The first category is the data acquired from the automated stage, while the second category is the records identified from citation tracking.

Finally, we thoroughly reviewed all publications related to the quantification of sexism and misogyny in online social platforms to determine their focus and methodologies, by reading their full text.

Qualitative assessment of the selected studies

A total of 97 full-text articles were analyzed qualitatively, as shown in Fig. 1. We assessed them based on *four* criteria, namely:

(i) *Irrelevant study focus* whether they are focused on studying the propagation of sexism and misogyny (irrespective of whether they indicated such in the abstract). Many of the works focused on hate speech, but because we wanted to only review sexism and misogyny, they were eliminated.
(ii) *Irrelevant study designs* whether the intended outcome of the research was not about performing a computation analysis on the detection or identification of online sexism and misogyny.
(iii) *Studies not quantifying sexism/misogyny* whether the paper focused on a review of studies in the relevant topic; or contained a summary of the author's thoughts from multiple papers, such as opinion pieces around the same topic. We only focused on methodology papers which described their approach. Hence, summary of methods, shared task descriptions or dataset papers[9] are generally excluded from the list.
(iv) *Sources could not be traced back* in case the paper's paraphrased contents with citations were not reflective of the summary the original authors indicated in their study.

48 Out of the 97 research articles qualified from this step, from surpassing these exclusion criteria and as a result, were included in the final qualitative analysis.

The full text of all the papers was reviewed qualitatively and information about each was added to a summary table covering the following points:

- Forms of hate speech studied. It is because hate speech could encompass a lot of things, including sexism and misogyny.
- Definitions of sexism or misogyny (or both) used in the study.

[9] Our dataset analysis was based on the same set of papers selected during the screening process.

- Language(s) of the data used for the study.
- Data selection criteria. This could depend on the original data collection method, such as—using keywords, hashtags, public profiles by monitoring user's online activity, users identified as sexist/misogynist, tags of sexism, specific phrases; or even based on a particular timeline of interest.
- Datasets used and their types (external, API-generated, etc.).
- Dataset modifications, if done. This could be in the form of data augmentation, counterfactual examples, document expansion by adding semantically similar words, transliterating multilingual dataset to uniform language, and many more.
- Broadcast media or social media platform which is of interest for the study.
- Annotators used in the study, and their tasks. If each or group of annotators had different tasks, that was also recorded.
- Pertaining to the previous point, the κ values that are statistical measures used to measure inter-rater reliability, are also noted.
- Research bias addressed or acknowledged in the study. If acknowledged, it is posted as a limitation in the paper.
- Pre-processing or post-processing done on the data.
- Performance metric used.
- Embedding type used, since this could range from word-level to node-based.
- Classification or clustering type, and the respective models.
- Syntactic, linguistic, and semantic/lexical features.
- Prompt topics and intersectionality (if present).

Recently several SLR tools have incorporated semi-automation using Artificial Intelligence techniques, for supporting the screening and extraction (pre-screening) phases (Bolanos et al., 2024), like we did in our research. Of such tools, only a few use topic modeling for their work. Such as, RobotAnalyst[10] and SWIFT-Review[11] uses Latent Dirichlet Allocation (LDA) that assigns a topic to a paper based on the most recurrent terms shared by other papers using a generative probabilistic model, while Iris.ai[12] clusters the papers according to a two-level taxonomy of global topics and specific topics (Bolanos et al., 2024). The former two tools depend on the term frequencies while the later perform Named Entity Recognition (NER) and allow users to to customize entity extraction by letting them define their own set of categories beforehand. Even with its advantage of the superior language capabilities to produce one of the most advanced techniques in language topic modeling today (Briggs, 2023), BERTopic has remained unexplored for the same task. In our work, we use that potential alongside the promising result of Large Language Models (LLMs) in their information extraction capabilities, to cluster our topics before validating the results with network analysis and selecting the topic(s) more suited for our work. This proved to be particularly useful to us in the screening and qualitative assessment phase as empirical analysis of the topics generated to their corresponding papers showed that the approach accurately clustered similar papers together.

[10] https://www.nactem.ac.uk/robotanalyst/.
[11] https://www.sciome.comswift-review/.
[12] https://iris.ai/.

Results of the systematic literature review

Data statistics

Post data screening and qualitative assessment, we finally narrow down the number of manuscripts to 48, that satisfied the scope of our meta-analysis. In the first subsection, we provide a brief overview of the key statistics of the 48 papers. In the following subsections, we provide an overview of the existing computational approaches dedicated to quantifying sexism and misogyny. Beyond that, we discuss the challenges and limitations faced by the said approaches from the existing literature.

Author collaboration network

We provide an author collaboration network in Fig. 6, where the name of the researchers are nodes, their size and color indicating the number of relevant manuscripts they authored

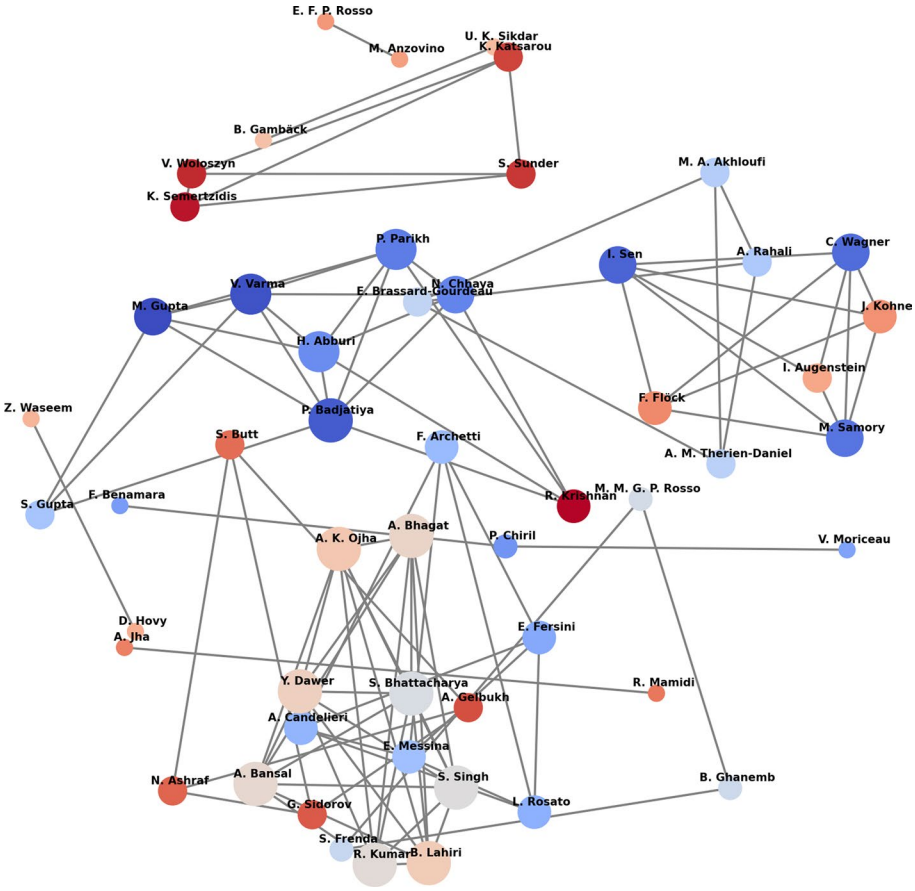

Fig. 6 Network displaying all the author collaborations between the 48 research articles screened for our literature review

Table 1 Summary table of some of the topmost categories of designs/methodologies in all the observed characteristics across the selected studies

Characteristics of study or research designs			
Characteristics	Count	Characteristics	Count
Benchmark Datasets used		**Languages of datasets**‡	
Waseem and Hovy (2016) and Waseem (2016)	14	English	36
Fersini et al. (2018) and Basile et al. (2020)	13	Spanish	7
Basile et al. (2019)	5	Italian	6
Bates (2015)	6	South Asian languages (e.g., Bangla, Hindi)	6
Rodríguez-Sánchez et al. (2021)	4	Other European languages	2
Fersini et al. (2018)	3	**Paradigms** 12*‡	
Machine Learning models		Perspective	40
Support Vector Machine (SVM)	16	Descriptive	5
Bidirectional Encoder Representations from Transformers (BERT)	15	Unsupervised (as per model features)	3
Long Short-Term Memory (LSTM)	15	**Evaluation type**‡	
Logistic Regression (LR)	10	Binary Classification	32
Convolutional Neural Networks (CNN)	9	Multi-class Classification	23
Naive Bayes (NB)	6	Multi-label Classification	4
Random Forest (RF)	6	Cluster Analysis	1
Decision Tree (DT)	4	Results per class	1
Multilayer Perceptron (MLP)	4	**Annotator types**	
XGBoost	3	External dataset	24
fastText	2	Experts	5
Platform of interest		Authors	4
Twitter (or X)	28	Amateurs/Crowdsourced external annotators	3
Sexism reported online (from Everyday Sexism Project) (Bates, 2015)	3	Students of linguistic, communication and gender	3
Facebook	2	Machine learning models	3
Reddit	2	Social Scientists	2
Gab	2	Annotator character not stated	2

*Unique features of a document, i.e., a document can only have either of the categories in the field, ‡the categories listed below are a part of the exhaustive list for that particular field

or co-authored. The connections between the authors are indicative of co-authorship on manuscripts, and their weighted edges imply the frequency of co-authorship.

Characteristics of study or research designs

Table 1 gives a summary of all the general characteristics we found in the full-text reviews of the selected studies. It provides a summary of the most used categories of each field (design/methodology) that the documents in our literature review have used. The other categories which were not featured in the table were mostly used by only one document. Each document could have multiple categories under the same field of design or methodology. For example, one document could be researching datasets of multiple languages in different platforms of interest and using multiple models, with different levels of classification at different stages. The categories that are uniquely present in a document are marked with an asterisk ( ) beside it, while the fields with their entire list of categories in the table are marked with an obelisk (†). In here, the asterisk ( ) symbol would not just indicate that the feature itself is unique to the field, but also all the documents should add up to the total number of literature listed.

Overview of the general methodologies

As per the Table 1, we do see the frequency of each source of online data and the machine learning models used in the 45 manuscripts we reviewed. The type of classification of sexism and misogyny used in the said studies are otherwise unknown and how they link between the sources of online data and the computational methodologies is an important source of information to indicate the multi-level connections between the variables, and consequently its impact on the quantification of sexism and misogyny. Figure 7 show the connection between nodes in each level of information (source of online data, classification type and model used), while the connections are the links between each level with their weights indicating the frequency of connections between each node type at different levels. This is a many-to-many mapping between the three levels, and show the flow of information between each of them. The colors of each node at different levels represent the unique relation between the linked nodes at each level (e.g., Twitter [level 1] => Misogyny (5 categories) [level 2] is different from Twitter [level 1] => Misogyny (binary) [level 2], and Reddit [level 1] => Misogyny (binary) [level 2] is different from Reddit [level 1] => Abuse/Aggression [level 2]). The abbreviations for the models in level 3 are found in Supplementary Sect. H. Overall, the Fig. 7 gives a clear evidence that Twitter was the most explored online data source to investigate different forms of sexism and misogyny.[13]

Overview of the existing computational approaches

General definitions and strategies used

Most computational works on sexism and misogyny took the automated identification problem as a binary classification task, i.e., deciding if the text in question is sexist/misogynistic or not. For defining the terms, researchers used different (non-standardized) forms for their work because of computational benefits, such as model performances. Though

[13] More on the paradigms (Rottger et al., 2022) in the Supplementary Sect. 3, in the table of terminologies.

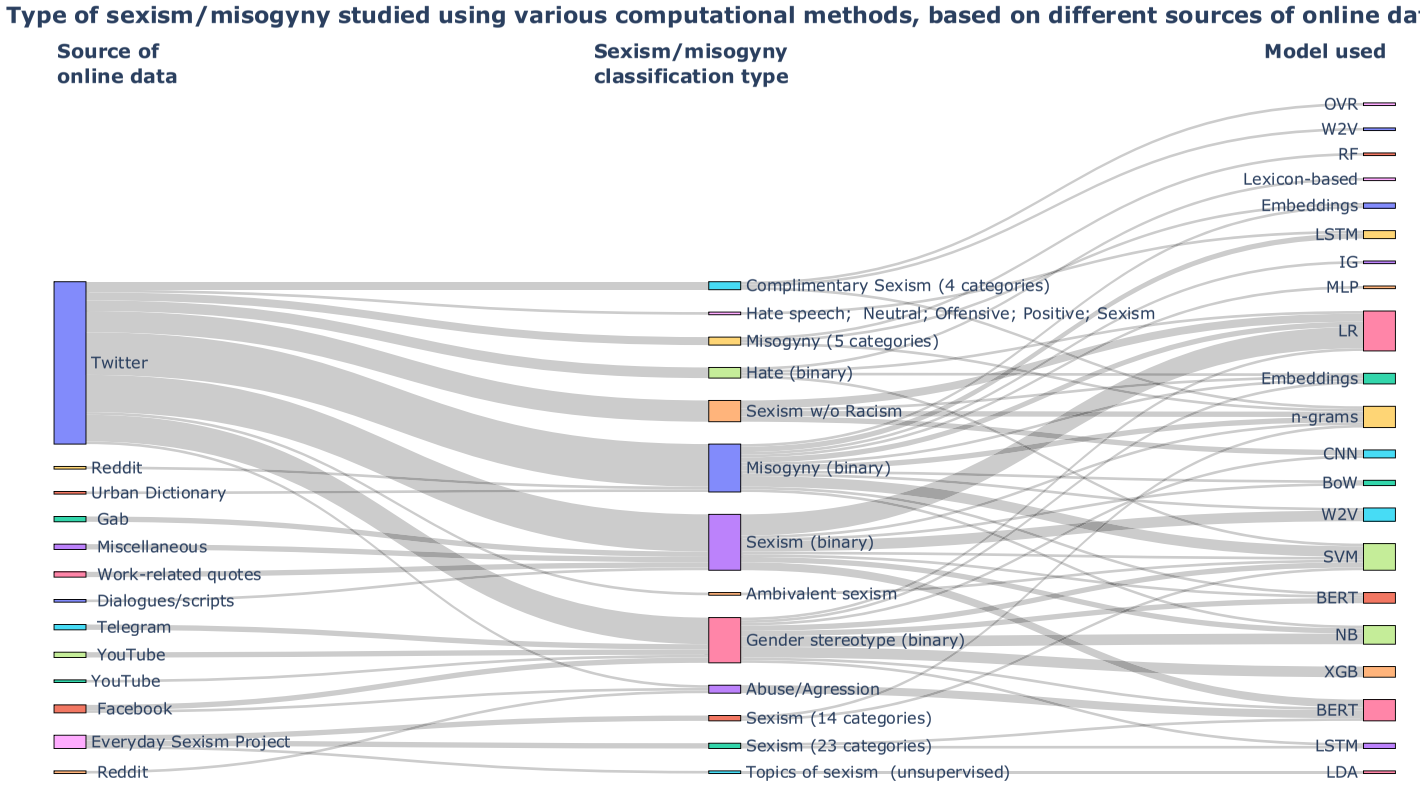

Fig. 7 Sankey diagram of the link between each categories of online data, classification type and computational models. The Sankey Diagram allows to visualize flow between various points in a system. In our system, we show the flow (i.e., count of association) between sources of online data, sexism/misogyny classification type, and NLP model used.

they served to be effective in some instances, they also presented with limitations of their own. For instance, Grosz and Conde-Cespedes (2020) define sexism as the prejudicial and discriminatory nature of sexist behavior pervading in the social context, especially for women. Using theoretical concepts, along with the typology of abuse presented in earlier research, Guest et al. (2021) define misogyny as content "directed abuse at women or a closely-related gendered group (e.g., feminists)." Whereas Lynn et al. (2019) define misogyny as a hate crime, a result of a "cultural attitude of hatred for females because they are female", and presented with two bidirectional-DL models (Bi-LSTM and Bi-GRU) with dropout layers, which performed well in sensitivity and accuracy even with a slightly imbalanced dataset.

Beyond binary classification, combining techniques can enhance model performance. For example, Attanasio and Pastor (2020) demonstrated improved results using a multi-agent classifier built upon sentence embeddings, TF–IDF weighting, and lexicons specific to misogynistic language.Plaza-del Arco et al. (2021) proposed Multi-Task Learning (MTL) system with hard parameter sharing approach (sharing the hidden layers between all tasks, while keeping several task-specific output layers) using BERT-based models for utilizing the transferred knowledge from multiple other (but related to sexism identification) tasks like polarity and emotion classification and offensive language detection classification helped in the identification, both in binary and multiple categories. Though emotion was not helpful in categorizing, MTL shows promising generalization to the original task.

Additionally, Frenda et al. (2019) exploit stylistic, semantic and topic information about misogynistic speech to identity misogyny and classify it to different categories. For gathering linguistic features, they propose an approach based on stylistic features captured by means of character *n*-grams, on sentiment information and on a set of lexicons built by examining the misogynistic tweets from training data provided by the organizers. Each text was represented by a vector composed of all specific topic features (set of lexicons), pondered with Information Gain, and character *n*-grams, weighted with TF–IDF measure. This set of features were experimented employing a Support Vector Machine (SVM) algorithm and an ensemble technique, reaching promising results. Canós (2018) worked on the same data and task, experimenting SVM alongside TF–IDF with a one-vs.-one and one-vs.-rest classifier approach, where the later proved better for English, presumably because of larger vocabulary. On the other hand, Nozza et al. (2019) defined several templates to create a balanced synthetic dataset for their proposed DL model—Universal Sentence Encoder (USE), which further debiased their model features to be less-sensitive to identity terms and yet obtain a better categorization.

Overview of performance evaluation

When it comes to evaluating performance, classification models with traditional computational approaches seem to fair relatively similar or better (in some cases) at automated sexism/misogyny identification tasks. Indurthi et al. (2019)'s work shows how different set of pretrained embeddings trained from different state-of-the-art architectures and methods when used with simple machine learning (ML) classifiers like SVM and XGBoost perform very well in binary classification tasks. Kohli et al. (2021) used two kinds of methods: first using an ensemble approach comprising of XGBoost, LightGBM and Naïve Bayes; and second employing BERT-based architecture. Both the models performed well on binary identification task, but differently on different languages and aggression label analysis, one of which was gendered, due to the overlapping context

in all. Overall, SVM is seen to be the best-performing conventional classifier (hence taken as a baseline for some of the works), and a lot of papers have used it in their work as a standalone classifier, or as an ensemble voting classifier (Frenda et al., 2018; Nascimento et al., 2022)—alongside other classifiers like Gradient Boosting and Random Forest (RF). Regression models like Logistic Regression have also been used by a lot of studies, especially for binary tasks. While Decision Tree (Plaza-Del-Arco et al., 2020), and RF (Singh et al., 2021) has also been used, they do not show much success among the conventional ones whereas most DL models use Fully Connected (FC) layers for classification (Bashar et al., 2019). Yet Convolutional Neural Network (CNN) with various semantic features is also seen to work well in categorizing hate-speech—particularly sexism, even outperforming Logistic Regression for some of the performance metrics (Gambäck & Sikdar, 2017).

However, there is no definitive indication of whether traditional models or neural networks (DL models)—particularly transformer-based architectures—are generally preferred for detecting or identifying sexism or misogyny. While some studies did not perform comparative analyses across different model architectures (e.g., (Waseem, 2016; Shah et al., 2020; Frenda et al., 2018; Canós, 2018)), others explicitly compared selected models to highlight performance differences. For instance, some studies such as Plaza-Del-Arco et al. (2020), Schutz et al. (2022), and Butt et al. (2021) found that traditional models (e.g., SVM, Random Forest) outperformed neural networks, whereas others like Sharifirad et al. (2018), Lynn et al. (2019), and Guest et al. (2021) reported the opposite. Additionally, some studies showed comparable results across different model types (e.g., (Grosz & Conde-Cespedes, 2020; Singh et al., 2021; Bashar et al., 2019)). Transformer models have been shown to offer clear advantages over conventional neural network models in several studies (e.g., (Chiril et al., 2021, 2020; Parikh et al., 2019)). Learning sequence and contextual information through embeddings yielded competitive performance in certain cases, particularly using LSTM and BERT-based models (e.g., (Badjatiya et al., 2017; Rodriguez-Sanchez et al., 2020; Chiril et al., 2021)). More broadly, the integration of embeddings was generally found to enhance model performance (e.g., (Fersini et al., 2021; Anzovino et al., 2018; Jha & Mamidi, 2017)). Furthermore, incorporating text features such as *n*-grams (e.g., (Bhattacharya et al., 2020; Waseem & Hovy, 2016; Butt et al., 2021)) or combining both embeddings and *n*-gram features (e.g., (Abburi et al., 2021; Sharifirad et al., 2018)) also contributed to improved outcomes.

Yet, studies such as Gröndahl et al. (2018) and Arango et al. (2022) suggest that most research emphasizes model development, with limited attention to evaluating whether these models generalize to other contexts or understanding the factors that influence their success or failure. In fact, Arango et al. (2022) found a *significant drop* in performance while replicating some state-of-the-art methods across multiple datasets. Herd and Burton (2024) explicitly state that "*relying on such single point estimates to evaluate safety requirements is problematic since they only provide a partial and indirect evaluation of the true safety risk associated with the model and its potential errors*". Given the probabilistic result the metrics provide over a binary domain, it may overestimate the model performance since their trustworthiness depends on other secondary factors such as sample size, the model calibration, the quality of the dataset, among others (Herd & Burton, 2024). This raises the question of whether strong performance metrics alone are truly reliable indicators of a robust or effective model.

Is classification the only way?

Though almost all the computational methods employ classification techniques, it is not the only way. But it is favorable, for good reasons. Clustering techniques are mostly useful for content analysis and to study discourse, to help identify implicit themes/topics from the data which may be (unintentionally) omitted during manual inspection, and reassignment into its overarching categories for better interpretation, even though may sometimes provide superficial results (Siddiqi et al., 2018). Karami et al. (2019) employs unsupervised text-mining approaches like LDA topic modeling. Utilizing the themes they found, they performed qualitative thematic analysis before finally moving to a theoretical thematic analysis to group the previously identified topics into four categories of sexism. Melville et al. (2019) also uses LDA for grouping 7 topics, and alongside clustering based on Louvain algorithm (Blondel et al., 2008) for grouping 20 topics. They define sexism based on themes and sites associated with the experience of sexism from Everyday Sexism (Bates, 2015) and journalism. From these studies, it is evident that clustering is more useful for content analysis, rather than for the detection/identification tasks.

Challenges

In this section, we outline the challenges for the interdisciplinary approaches that are likely due to the disciplinary divide, and ways of addressing them. We identify that these challenges could be because of two broad reasons—(i) Use of different computational strategies; and (ii) Linking social science theories to the tested computational strategies. The first part essentially talks about the different strategies used, compares them based on different parameters in each subsection, and weighs the advantages and limitations of each approach. The second part focuses more on how existing literature has sexism and misogyny in their work. By analyzing how the same terms are defined in social science theories, we form an argument about how the existing computational research could benefit from a more fine-grained categorization of sexism (or misogyny) to improve their automated identification task.

Use of different computational strategies

This subsection is intended to shed light on the different computational strategies that have been used to quantify sexism and misogyny, while also segregating the strategies based on some differentiators like the small dataset size and the dataset languages used.

Beyond Binary Classification

Most of the studies have used binary classification of sexism (e.g., (Singh et al., 2021; Schutz et al., 2022; Grosz & Conde-Cespedes, 2020)) and misogyny (e.g., (Bashar et al., 2019; Lynn et al., 2019; Bashar et al., 2019)) for their detection tasks. While practical, such models often overlook theoretical nuance, providing surface-level predictions that may fail to capture the complex manifestations of sexism (Parikh et al., 2019; Abburi et al., 2021). Previous studies have highlighted the difficulty in capturing the nuanced meanings of terms like sexism and misogyny in computational settings (Samory et al., 2021), emphasizing the importance of evaluating both the intensity and specific type of misogynistic behavior within a given context (Lynn et al., 2019).

Tasks such as hate speech or harassment detection inherently involve subjectivity, with no universally accepted definitions or absolute truths (Rottger et al., 2022). While certain beliefs may be broadly agreed upon, they do not fully encapsulate the complexity of these concepts. Moreover, annotation processes are often conducted without direct oversight from dataset creators, resulting in partially subjective datasets that may not align clearly with specific downstream applications (Rottger et al., 2022). Even with detailed label descriptions, human annotators often struggle with intuitively unclear (Vidgen & Derczynski, 2020) or closely related (Parikh et al., 2019) categories. Some improvements have been observed when annotation guidelines are adjusted to address edge cases (Zeinert et al., 2021),though further distinction—such as between active and passive language—may offer additional clarity (Anzovino et al., 2018). Moreover, perceptions of toxicity are highly subjective and often shaped by the receiver's interpretation of the speaker's expression (Sap et al., 2019).

Despite increasing interest, research addressing the nuanced forms of sexism and the intersectional ways women experience it—online and offline—remains limited (Melville et al., 2019). Barak (2005) identifies four types of gender-based harassment in cyberspace: active verbal, passive verbal, active graphic, and passive graphic. These categories are influenced by both objective features (e.g., explicitness, repetition) and subjective factors (e.g., recipient's attitudes, sensitivities). As a result, perceived severity varies widely across individuals—a principle equally applicable to sexism and misogyny. Similarly, Sharifirad et al. (2018) propose a four-type classification of sexism: information threat, and indirect, sexual, and physical harassment. To overcome data scarcity, they use ConceptNet to semantically enrich text via augmentation techniques—leveraging relations like "IsA" and "RelatedTo"—with three replacement strategies: all words (most effective), nouns, and verbs. They find that such semantic enrichment, especially through text generation alone, considerably boosts classification performance.

Swim et al. (2004) point out that using only sexist language—as done in the above studies—limits the generalizability of findings to other, less overt forms of sexist behavior. Still, multi-label classification has shown promise. Echoing this approach, Talavera et al. (2021) also employ a multi-label model with fewer (five) categories. They reinforce earlier findings that pretrained language models, when fine-tuned to the specific task domain, are particularly effective in low-resource settings. Most of the existing research on multi-class classification of sexism or misogyny consider at most five categories of sexism (such as (Anzovino et al., 2018; Sharifirad et al., 2018; Jha & Mamidi, 2017)). Anzovino et al. (2018) classify misogyny into five subcategories: Discredit, Harassment & Threats of Violence, Derailing, Stereotype & Objectification, and Dominance. Guest et al. (2021) extend from their work to develop a hierarchical taxonomy with three levels to define misogynistic content, with four overarching categories of misogyny: (i) Misogynistic Pejoratives, (ii) descriptions of Misogynistic Treatment, (iii) acts of Misogynistic Derogation and (iv) Gendered Personal attacks against women. Studies like Frenda et al. (2018), Canós (2018), and Shah et al. (2020) use the IberEval 2018 shared task dataset (Fersini et al., 2018) where the data was categorized to five categories, namely: *Stereotyping and Objectification*, which involve oversimplified portrayals of women or emphasis on their physical appearance; *Dominance*, which asserts male superiority to reinforce gender inequality; *Derailing*, which shifts blame away from men or disrupts conversations to refocus them around male comfort; *Sexual Harassment and Threats of Violence*, which encompass unwanted advances, coercion, or threats intended to exert power over women; and *Discredit*, which targets women with slurs or insults without any broader argumentative context or intent.

Likewise, the EXIST shared tasks in 2021 and 2022 (Rodríguez-Sánchez et al., 2021; Rodriguez-Sanchez et al., 2022) begin with a binary definition of sexism based on the Oxford English Dictionary—"prejudice, stereotyping, or discrimination, typically against women, based on sex"—and extend it into a multi-class taxonomy comprising five expert-defined categories: *Ideological and inequality*, which includes anti-feminist rhetoric or denial of structural gender inequality; *Stereotyping and dominance*, referring to assertions of traditional gender roles or male superiority; *Objectification*, where women are reduced to physical appearance or sexualized attributes; *Sexual violence*, encompassing harassment, coercion, or threats of a sexual nature; and *Misogyny and non-sexual violence*, involving expressions of hatred, hostility, or aggression directed at women. Both datasets support the detection of a broad spectrum of sexist content, ranging from explicit misogyny to more subtle, implicit expressions of sexist behavior, as opposed to most other research which only focus on the detection of explicit contents (such as (Waseem & Hovy, 2016; Katsarou et al., 2021; Anzovino et al., 2018)).

The utility of multi-label classification is highlighted in Parikh et al. (2019), who illustrate the co-occurrence of various sexism categories in user-reported first-person accounts. Their dataset comprises 23 categories (defined by a social scientist) derived from real-world reports of sexism. The annotation process followed a rigorous three-phase protocol involving pre-training, pilot testing, and final quality checks, ultimately reducing the labels to 14 for classification. To address domain-specific challenges, they fine-tuned a BERT model using masked language modeling and next sentence prediction, further enhancing it with distributional word embeddings and a linguistic feature vector. These linguistic features drew from prior work on biased language detection and incorporated affective and sentiment lexicons, including PERMA (Positive Emotion, Engagement, Relationships, Meaning, Accomplishment) features and emotion scores. Their proposed multi-label, multi-class LSTM-based framework significantly outperformed various traditional machine learning and deep learning baselines.

Building on this work, Abburi et al. (2021) take a more fine-grained approach by reintroducing the full 23-category schema. Using self-trained semi-supervised learning, they augment the labeled dataset to allow co-occurrence of categories and improve class distribution. Their approach enhances textual diversity and focuses on hard-to-classify samples using confidence scores and selective intersection. In addition to the domain-tuned BERT with a BiLSTM and attention mechanism, they incorporate a custom loss function that leverages label confidence scores. Their domain-tuned BERT–BiLSTM–Attention model achieved notable gains, though the data focused exclusively on sexist examples and did not generalize to broader detection tasks.

Data Limitations and Model Performance

Even when models perform well, limited dataset size can undermine reliability and restrict exploration of more sophisticated deep learning (DL) architectures, such as those employing advanced attention mechanisms requiring larger amounts of data (e.g., Grosz & Conde-Cespedes, 2020). For example, in their binary classification of misogyny, Guest et al. (2021) developed a hierarchical taxonomy across three levels: the second level featured four non-mutually exclusive categories—misogynistic pejoratives, treatment, derogation, and gendered personal attacks—with further sub-categorization at the third level. Although the classification is sound, and logistic regression and both weighted and unweighted BERT achieved strong performance, the dataset's limited size and low proportion of misogynistic content (8.1%) posed challenges, especially since classification was performed only on the instances labeled True for misogyny.

To address small dataset issues, Schutz et al. (2022) explored transfer learning approaches, including the use of multilingual transformers pre-trained on external datasets and data augmentation by integrating similar external content. Their experiments demonstrated that fine-tuning the entire model with domain-specific data improved performance. However, pre-training generally proved more effective, as fine-tuning showed tendencies toward overfitting and did not yield consistent improvements on external datasets. As DL models typically require large datasets to achieve optimal performance, traditional models may give a comparative performance (such as (Samory et al., 2021)) or even outperform them in data-constrained settings—particularly when augmented with rich linguistic features such as *n*-grams (Plaza-Del-Arco et al., 2020; Butt et al., 2021).

Dependence on external benchmark datasets

Several previous studies highlight that the main challenge in quantifying sexism stems from the lack of high-quality datasets required to train robust, scalable automated detection systems (Guest et al., 2021). Most computational studies rely heavily on external benchmark datasets, which can affect the quality, reliability, and representativeness of the data. As Zeinert et al. (2021) point out, "When abusive language is annotated, classes are often created based on each unique dataset (a purely inductive approach), rather than leveraging established terminology from social science or psychology (a deductive approach, building on existing research)."

While benchmark datasets are frequently used in shared tasks.[14] for automated misogyny detection and related challenges, they are often found to be misrepresented or imbalanced. For example, the IberEval2018 dataset contained uneven category representation—certain misogynistic behaviors, like 'derailing,' were underrepresented (< 2%), while categories like 'active' tweets were significantly overrepresented (> 85%). Additionally, notable discrepancies appeared between language groups in the same dataset (Canós, 2018). Similarly, in the TRAC2020 shared task dataset (Bhattacharya et al., 2020), texts containing multiple languages were categorized under a single language, complicating analysis for non-speakers unfamiliar with the other languages' socio-cultural contexts. Furthermore, the proportion of texts exhibiting hate speech varied significantly across languages (Gordeev & Lykova, 2020).

These inconsistencies underscore the necessity for dedicated, theoretically-informed, 'reliable' data collection and annotation practices grounded in social science frameworks, before performing any experimentation with NLP tools.

Datasets are mostly in Western Languages

The majority of studies on online hate speech detection have been conducted in English (e.g., (Katsarou et al., 2021; Waseem & Hovy, 2016; Parikh et al., 2019), inter alia). Spanish and Italian follow closely, particularly through shared tasks—for example, studies in Spanish (Schutz et al., 2022; Plaza-Del-Arco et al., 2020; Canós, 2018) and in Italian (Attanasio & Pastor, 2020; Ou & Li, 2020; Nozza et al., 2019), inter alia. More recently, a limited number of works have begun to explore other languages such as Hindi, Bangla (e.g., (Bhattacharya et al., 2020)) and Chinese (e.g., (Jiang et al., 2021)). But even within same languages, there lies substantive differences in the peculiar lexical choice and morphological structures rising from the regional colloquial usages, leading to linguistic and cultural heterogeneity (Bhattacharya et al., 2020). However, this fundamental issue remains

[14] "Shared tasks are collaborative efforts in which researchers and practitioners come together to solve a common problem using shared data and evaluation measures. They promote competition, collaboration, and progress in research, and have become an important part of many academic and industrial communities" (SIGEDU, https://sig-edu.org/sharedtasks).

largely unexamined in the current literature. Though models like fine-tuned cross-lingual multitask BERT shows promising performance even with non-English languages, with it performing better on Bangla when experimented alongside English, presumably owing to the dataset peculiarity or specific features of the language itself (Gordeev & Lykova, 2020). But when it comes to representation, there is a huge gap between only the use of Indo-European languages (especially English) and other languages. When working with multiple languages, "back-translation" has been used in the said languages to augment the data and translate all of them to a single uniform language, which could be one of the source languages or different.

Butt et al. (2021) performed the same technique on Spanish and English (source) languages to convert it into English, with German being the second language, using the deep-translator python library. And their results on all of their tried ML algorithms show an improvement with the augmentation; even indicating that with proper pre-processing, it could give competitive results in comparison to deep learning models. Zeinert et al. (2021) too had experimented translating misogynistic posts provided by Anzovino et al. (2018) to Danish using translation services in an attempt to augment the minority class data. But it did not prove as useful in providing a sampling alternative, hence inferring that language-specific investigation is important for cultural discovery, for the sake of automatic detection systems. Rodriguez-Sanchez et al. (2020) found improved cross-domain generalization when models were trained on their Spanish dataset and tested on an Italian dataset. This suggests that some datasets may offer broader coverage of sexist content—including explicit, implicit, and context-dependent forms—and are therefore more transferable for studying the same task across different languages. However, we observe a lack of experiments involving languages that differ significantly in cultural–linguistic context and core linguistic dimensions.

In fact, Waseem (2016) suggests against boosting the minority class in the interest of mimicking reality in the datasets, even if it causes larger misclassification for the class. Rahali et al. (2021) uses gender swap data augmentation and data consolidation with feature ablation, which is seen to improve the learning of the model, especially when used with the same language. But using multi-language datasets does not help much, since English does not consolidate well with other languages (e.g., Arabic and French) with limited samples as compared to English, inevitably giving rise to data imbalance, a data bias. So, there is a need to look beyond English.

Singh et al. (2021) converted the whole dataset from multiple languages to a uniform English dataset, by transliterating the sentences belonging to other languages using Indic-Transliterator. But that required transferring a word from the alphabet of one language to another, which could give faulty outcomes. Given the linguistic variety and limitations that could be faced when delving into other languages, improving the existing gaps in sexism identification tasks in English should be of primary focus.

Biases Bias is a broad term that can be defined in various ways, depending on the field and context. In this section, we summarize: (a) how studies in our dataset define or quantify biases, (b) whether these studies acknowledge biases present in their own research, and (c) what measures, if any, they adopt to mitigate these biases. Social science studies tend to address bias in their work. In NLP, however, definitions of bias often depend heavily on domain-specific concerns like biases in word embeddings, annotator labels, or amplified predictions related to demographic characteristics (Hovy & Prabhumoye, 2021). In this work, we adopt the definition by Shah et al. (2020), where bias refers specifically to "the mismatch of ideal and actual distributions of labels and user attributes in the training and application of a system." Additionally, the rapid advancement of NLP may itself

exacerbate bias by outpacing the field's ability to adapt effectively to emerging contexts (Hovy & Prabhumoye, 2021).

Gender bias

Eagly and Mladinic (1989) use attitude theory to explain how stereotypes emerge from attitudes toward men and women. According to them, the cognitive component of an attitude—defined as a person's thoughts about an object—can manifest through attributes assigned to social groups. When attributes assigned to these groups carry evaluative meanings (positive or negative), they give rise to stereotypical perceptions. Their study reveals that even when women are positively evaluated, they tend to be stereotypically viewed as inferior to men in agentic or instrumental (traditionally masculine-positive) qualities, yet superior in communal or expressive (traditionally feminine-positive) qualities.

Further supporting this, Schmid Mast (2004) provide empirical evidence of an implicit hierarchical gender stereotype, showing that men are associated more strongly with hierarchy and women with egalitarianism. The magnitude of this stereotype reflects a deeply rooted societal bias. Social media platforms further amplify these biases through their widespread influence. Consequently, biased social information from such sources, when used to train machine learning models, can unintentionally lead to gender biases. Models may thus propagate negative stereotypes about various social groups.

To address this issue, Dinan et al. (2020) proposed measuring bias through a semantic and pragmatic framework, evaluating three dimensions derived from "knowledge of the conversational and performative aspects of gender." By separately analyzing how author gender influences the dataset, they aimed for a deeper understanding and mitigation of gender bias.

Annotator bias

It is seen that having a lot of categories of misogyny may also impact on the annotators' agreement, both in terms of depth (subcategories) and breadth (different types) of the said categories, owing to the differences in experience and values of the annotators. And their inherent social biases may impact on their choice, especially when working using contexts (Guest et al., 2021). Having different level of understanding of the language in question or personal prejudices, and differing individual world-view are seen as primary issues in inter-annotator disagreements. Bhattacharya et al. (2020) used several rounds of discussions and sensitization towards gender issues among annotators to resolve this issue, by providing with counterexample method and examining annotator votes, alongside using an 'unclear' tag in case of disagreement. Sometimes when adversarial examples (even just 25%) are included while training the dataset, it is seen to help in the robustness of the models and their performance. In fact, providing the models with different aspects of sexism and challenging the models with different examples have shown to be effective for generalizability (Samory et al., 2021).

Some studies use different criteria for selecting the annotators they want in their study, based on both similarities and differences on each. It could be based on region, demographic, education and ethnicity (Guest et al., 2021), native speakers of language (Chiril et al., 2020), feminists (Jha & Mamidi, 2017); but mostly who studied gender (Lynn et al., 2019) and linguistics (Nozza et al., 2019). A comparison on amateur (crowd-sourced workers) and expert (having both a theoretical and applied knowledge of hate speech) annotators (as most studies use either of them) by Waseem (2016) state the contrasts observed in annotation with both, and the consequent model performances which did not substantially improve on their previous model (Waseem & Hovy, 2016). The emphasis on the most significant features changes from extra-linguistic features for majority-voted amateurs to content of the tweets for the experts, and among the features they experimented with, the ones having highest performances (high $F1$ score) were not necessarily the features with

the best performances. Singh et al. (2021) considers misogyny and sexism as a subset of hate-speech, and used data that was manually annotated by multiple annotators using 'Discursive Methods of Annotation' since it was seen as a pragmatic approach to including the socio-pragmatic phenomenon using social studies, and as a function of both the contextual factors and the discursive experience of the speaker. Zeinert et al. (2021) does an iterative process of raising cases for revision in the discussion rounds, formulating the issue, and providing documentation for annotation, inviting in annotators with diversity in age, occupation/background, region (spoken dialects). Annotation biases can lead to other kinds of bias, like racial bias due to lack of knowledge of different dialects—which could potentially amplify the harm against people from the minority community (Sap et al., 2019).

Other biases

Hirsch (1992) documents the oppression of women through language, as she talks about male-specific words that are positively portrayed in English, in turn reflecting the "consensus reality" of the patriarchal society. While theorizing the language and gender connection (with many of the examples drawn from political discourse) from one of its reviewed books, it talks about how language is used as a tool to further perpetuate patriarchy (Cameron, 2020). The same is seen for the computational models based in English. The datasets taken for studies could also add to the bias owing to the different considerations made due to the data source, hence not representing the diversity in real-world. For example, domain sources where misogyny is assumed to be most likely like women fashion blogs, fitness tips videos, etc. (Bhattacharya et al., 2020); or when sexism is taken as one of the sentiment label, with data collected around some specific cases/instances/event networks like #Coronavirus, #ClimateChange, #Immigrants and #MeToo (Katsarou et al., 2021). Another form of such bias could rise because of subjectivity in mislabeled data. Samory et al. (2021) had performed re-annotation on the external datasets they used in their study, following the sexism annotation codebook they devised themselves. Relying on two baselines: GenderWord (Zhao et al., 2018) and Jigsaw's Perspective API (Hosseini et al., 2017), they found a large majority of sexist tweets were non-sexist, only $\sim 60\%$ of the sexist labels adhering to their ground truth. They found that stratifying misclassification rates helped in giving a more accurate result. Both these points could hinder model performance.

Yet, with the listed biases, always a question remains if they were a cause of systematic errors (both conscious or unconscious) or were a result of a narrowed preference in a particular direction in favor of the said bias. In other words, the use of 'bias' to refer to systematic error is problematic. According to Hammersley and Gomm (1997), it depends on 'truth' and 'objectivity', whose justification and role have been questioned. Due to the ambiguous nature of the term itself, we might question if the forms of bias explored are a result of methodological adaptability; conscious limitations due to the scope of the research (such as research designs); or could arise because of the models themselves. Either way, they may not indicate the research as "being biased". Of the five most common sources of bias in NLP tasks as identified by Hovy and Prabhumoye (2021), we have reviewed almost all of them in this section. This indicates that these biases are well-known across the CSS literature, and can be explored more to mitigate them from all sources, using algorithmic and methodological approaches.

Lexical dependency for characterizing sexism and misogyny

To linguistically characterize misogyny and sexism, many studies have used different theoretical concepts to represent both. Farrell et al. (2019) had built a list of key lexicons for categorizing misogyny using Encyclopaedia of Feminist Theories (Code, 2002) and other pre-existing hate-speech lexicons and studies of the specific rhetoric of manosphere, taken from different corpus. In their observatory work, they study the evolution

of communities where users share in-group characteristics. But even though corroborating the theories and existing ideas helped in providing lexicons, they acknowledged the limitations of using it due to its lack of completeness (shortcomings in capturing all the words that might be relevant). Other times, studies use words 'typically associated' with misogynistic content created by domain experts (Lynn et al., 2019) which is used as neologisms for identification of emerging or cloaked misogyny.

Lexical dependency can cause NLP models to overfit because of too much influence of certain identity terms. This eventually results in false positives, severe unintended bias, and lower performance. Rodriguez-Sanchez et al. (2020) acknowledge the biases inherent in keyword-based dataset collection, noting that models tend to exhibit bias toward specific terms and struggle to detect instances featuring shorter length, subtle expressions of sexism, irony, or context-dependent language. Bashar et al. (2019) acknowledge that misogynistic abusive tweets might contain certain keywords, but would not necessarily always contain such slurs. To work around that, they show that classifiers can work with small-labeled datasets, provided that the word vectors used are pre-trained on the context domain of the problem and paired with careful customization and regularization. This proves that a large-labeled dataset is not always required for training purposes. In fact, if the word vectors are pre-trained in the context of the problem domain, alongside careful customization of the model, the classifiers could also be trained on small datasets. On the other hand, Plaza-del Arco et al. (2021) generates linguistic resources using a set of word embeddings, with the initial seed lexicon eventually getting populated with words and *n*-grams more attuned to the domain because of linguistic similarities. Using a voting schema rule with logistic regression and multinomial Naïve Bayes, alongside the lexicon-based system and combinations of unigrams and bigrams gave a good result with the Spanish dataset. Observations show that some expressions of hate when combined with other terms change the sense entirely and hence better-supervised learning begins with larger data.

For larger datasets, the issue is elevated with the imbalanced nature of the datasets and their disproportionate dependence on these determinate terms, having a high correlation to minority class (Nascimento et al., 2022). Using such identity terms, or samples from target domains during the training phase requires a priori knowledge but can often lead to the introduction of further bias. Introducing a regularization approach to the models to add some degrees of contextualization using Entropy-based Attention Regularization (EAR) could mitigate the problem to some extent, as they are seen to show competitive performance, along with an improvement in the bias metrics (Attanasio et al., 2022). Consequently, developing classifiers that can decompose gender bias within full sentences into semantic dimensions can be used, since it can be contextually determined (rather than being explicitly gendered). This has in turn shown to give a better performance in controlling gender differences (Dinan et al., 2020). Ou and Li (2020) find limitations of only using the pooler output of DL multilanguage models like XLM-RoBERTa, and hence obtains deeper and more abundant semantic features by extracting from its hidden layer state which gives better performance. Data correction strategy focused on gender bias, consisting of two-stage modules—bias detection and replacement of the said bias-sensitive words (BSWs), is seen to reduce the differentiation of similar terms related to gender, and in turn, contribute to mitigating the unintended bias. Since the frequency of female identity terms is high (even when representing similar groups/classes or other social identities) in datasets related to sexism and misogyny, they replaced these potential bias terms with ⟨identity⟩ tag without compromising the model accuracy. Their proposed multi-view stacked classifier is seen to outperform other state-of-the-art models and diminish gender bias (Nascimento et al., 2022).

Natural language processing (NLP) applications like sentiment analysis are crucial for analyzing and detecting online sexism/misogyny. Incorporating polarity and emotion information is seen to be useful for the benefit of the task as they portray the usually emotional, expression of negative emotion and polarity towards the recipient (Plaza-Del-Arco et al., 2020; Plaza-del Arco et al., 2021). Using feature representations has further helped in training the model, by adding representations of the text in terms of various lexical, syntactic, and morphological features. While the most common types of features used are the bag-of-words representations of text, and/or the embeddings, adding to the features also helps in the performance. Many papers have used it to enhance their model performance. The idea is to map out the various aspects of sexism as seen in the everyday social constructs and use it to comprehensively map them out for the benefit of the identification tasks (Samory et al., 2021).

Also looking beyond the text, Chiril et al. (2020) performed further characterization of the binary sexist classification by distinguishing cases where the addressee is directly addressed from those where she is not. The three categories being: (i) directed assertions—sexist tweet directly addressed to a woman or a group of women; (ii) descriptive assertions—sexist tweets not directed to an addressee; and (iii) reported assertions—tweets containing report of an experience or a denunciation of sexist behavior. On performing classification based on results per class, they identified the absence of context with the utterance, humor, and satire, and the use of stereotypes or metaphors to be the causes of misclassification through their manual error analysis in their best performing model—BERT. As Frenda et al. (2019) had also stated one of their principle problems is the use of linguistic devices like irony and sarcasm. In general, it has been highlighted in multiple research (e.g., (Guest et al., 2021; Rodriguez-Sanchez et al., 2020)) that the more subtle forms of sexism (mostly depended on the context) are not picked up well by the models. As Singh et al. (2021) note in their error analysis, many of the confounding variables were specific lexical items that either denoted explicitly sexual content or conveyed strongly negative sentiment.

Different psychometric scales can also be used to map out various aspects of sexism/misogyny as a social construct, to comprehensively detect the different categorizations. King and King (1997) reaffirm the previously stated theory on modern sexists, and describe them as "people who while rejecting old-fashioned discrimination and stereotypes, may believe that discrimination against women is a thing of the past, feel agnostic against women who are making political and economic demands, and feel resentment about special favors for women, such as policies designed to help women in academics and work." In other words, the distinction between old-fashioned and modern sexism lies in the fact that the former showcases an obvious unequal treatment of women while questioning their intelligence, while the latter is less sympathetic to women's issues (if at all they perceive them to be issues) since they presume greater equality in the workforce than what exists. The Modern Sexism (MS) scale this study provides aims to be a good indicator to detect modern sexism, which could be both overt and covert in nature. People endorsing MS beliefs are hence less likely to detect the occurrence of a normative sexist behavior (Swim et al., 2004). In the review by Swim and Cohen (1997) on the MS scale, they indicate the same as they observe that it measures the subtle forms of sexism that are built upon cultural and societal norms. They also review another general measure of sexism, namely the Attitude Toward Women Scale (AWS), which measures overt or blatant sexism. And through their analysis, they indicate that even with these distinctive differences, both share related constructs. These social constructs are often perpetrated as discriminatory attitudes towards a feminine gender role, which are traditionally allocated and differentiated by sex. García-Cueto et al. (2015) propose a scale to assess the gender role

attitude, showing how sexist attitudes can be modified using the theoretical perspectives of gender equality.

Linking social science theories to computer science research

Following from the previous subsection where we introduced our argument that sexism/misogyny is not a binary task, in this section, we expand on that point by providing social science theories and scales to explain the need to not computationally limit the classification to the binary output. To support that, alongside including the theories and scales, we also analyze how some studies have aided their work with these theories in any capacities (i.e., the extent of adaptation—using one or more categories of the scaling) and implemented them at any stage of their research. We distinguished each subdivision into two parts: the **concept** and the **applications**, to help us differentiate between the concepts themselves and on how they are implemented in studies.

Sexism is not always hostile

Concept.

Grosz and Conde-Cespedes (2020) state, that models can perform detection tasks easier on datasets containing large amounts of "hostile" sexism, since it hinges on some words, regardless of their context. But that does not provide a real-world scenario. In general, sexism is said to have two components: hostility towards women and endorsement of traditional gender roles, and most of the sexist attitude measures so far have stemmed from there. But it is not always so. Through their anthropological research on sexism, Glick and Fiske (1997) call sexism "fundamentally ambivalent", adding the subjectively benevolent nature of sexism to the previously perceived singularly hostile nature. They argue that the "simultaneous existence of male structural power and female dyadic power" creates an ambivalent ideology. While the hostile ideology seeks justification of their male position through derogatory characterization of women (HS), benevolent ideology relies on kinder and gentler justification, which may inherently look as subjectively positive for the sexist as they encompass feelings of protectiveness and affection towards women (BS). By drawing parallels from paternalism, which also has two ideologies—dominative and protective, they demonstrate that the protectiveness is particularly strong when women (e.g., wives, mothers, daughters) are dyadically dependent on men, as a feeling something akin to the sense of "ownership". The hierarchical stereotype ideology explained before constitutes the belief contributing to the gender differentiation. Like paternalism, it also consists of both hostile and benevolent side. Competitive gender differentiation being the hostile kind, delves on negative stereotypes of women implying men to be the better gender; and the complementary gender differentiation (the benevolent kind) stems from the traditional stereotypes of women through assigned gender roles and men's dyadic dependence on women, albeit in an extremely positive light (Eagly & Mladinic, 1994). Similarly, for heterosexuality, which has a hostile side when viewing women as mere sexual objects who use sexual attraction to gain power over men; and intimate or benevolent side that romanticizes the former belief, viewing women as necessary for men to feel "complete".

Applications.

Sexism in ambivalent theory (Glick & Fiske, 1996) is thus hypnotized to encompass these three sources of male ambivalence, which has been used by Jha and Mamidi (2017) to computationally identify benevolent sexism, and classify sexist content based on the two components. They confirm the hypothesis that HS is evidently negative and easily identifiable, while BS is retweeted much more and is camouflaged, seemingly harmless or noble and hence, harder to

detect. It was seen that while SVM showed high precision for both, recall was quite low for HS; their Seq2Seq model (LSTM-based bi-directional RNN) showed a higher recall for both, even though its precision was not as high, presumably because it takes in the structure of the tweet. But owing to the bag-of-*n*-grams feature of FastText (and lesser parameters to tune), it outperformed both the former classifiers. On the other hand, Singh et al. (2021) used the hostile side of the three sources of male ambivalence to define sexism binarily and annotate dialogues in popular sitcoms. Using these concepts, they manually annotated the external datasets (source domain) and used a semi-supervised domain-adaptive learning approach to generate classes in the model for the unannotated data (target domain), thus further augmenting the training data and improving the final classification performance. However, error analysis showed certain false positives like incorrectly classifying aggressive negative statements to a particular woman, or contents with explicit sexual terms and mentions of marriages or weddings as sexist. This could be the underlying drawback of not using a diverse dataset since the authors had included dialogues that included derogatory terms and dialogues justifying stereotypes against women or gender roles. But Mishra et al. (2019) use the concepts from previous research rather differently, by taking inspiration from studies that use randomly initialized user embeddings for improving performances, and inter and intra-user representations based on tweets. Instead of the former semi-supervised approach, they use graph convolutional networks (GCN) based approach, applied to the heterogenous graph representation of two types of nodes—authors and their tweets, to generate richer author profiles. The intention was to use such heterogenous representation to enable the model to learn both community structure and linguistic behavior of authors in such communities. Even with this improvement, several abusive tweets were misclassified, primarily due to the presence of abusive content in the URL (not in the tweet itself), and the deliberate obfuscation of words and phrases by the authors to evade detection.

Subtle forms of sexism/misogyny

Concept.

Since most of the sexism measurement scales are focused on hostile sentiments, it fail to capture the contemporary forms of subtle sexism, which are often cloaked in the guise of egalitarian views and harbor (more) traditional beliefs. Only some of the previous works (such as (Rodríguez-Sánchez et al., 2021; Rodriguez-Sanchez et al., 2022; Samory et al., 2021)) have addressed the more subtle and covert forms of sexism. Yet, due to the increase in social awareness of sex discrimination, the more blatant form of sexism is reduced, replaced with the subtle forms of indirect indices. And the lack of conceptual framework of understanding, coupled with methodological problems were indicated in the simulation study conducted by Beattie and Diehl (1979), where they observe the use of indirect means to interpret the gender and hence influence the evaluation criteria. This gave suggestive evidence to a new form of sexism called "neosexism", which was first introduced by Tougas et al. (1995), and defined as "a manifestation of a conflict between egalitarian values and residual negative feelings towards women". They used a predictive model of 'attitude to affirmative actions' to test the discriminatory bias and evaluated the practical implications of neosexism through their Neosexism Scale (NS). The study indicated that "neosexist beliefs were linked with opposition to programs designed to facilitate integration of both women and minorities", which leads to further proves the importance of understanding the existing prejudicial beliefs of women to understand the different forms of sexism.

Applications.

An analysis of the cross-sectional data during the 2016 US presidential election and the #MeToo movement by Archer and Kam (2020) shows its significant correlation to

neosexism, and the various degrees of dismissal of the respondents to the existing gender discrimination, hence indicating its existence in online platforms. Zeinert et al. (2021) had used NS in their work on Danish tweets to add neosexism to their taxonomy along with the previously categorized forms of sexism. Interestingly, while annotating, they found that neosexism formed the most common form of misogyny and accounted for most of the annotation challenges based on disagreements, primarily due to the challenge of understanding the author's intentions, the degree of abuse (since misrepresentation could harm the subject or the fact) and lack of world knowledge. This further added to the class imbalance in the last stage of sexism labeling in their dataset which affected the reliability of the performance, even though they started with a 1:1 class balance at the initial stage (labeling abusive or not) of their iterative labeling scheme based on the MALER framework proposed by Finlayson and Erjavec (2017). To prevent such bias caused by an imbalanced dataset, Indurthi et al. (2019) process the training dataset using SMOTE (Chawla et al., 2002) which synthetically oversamples data and ensure all classes have an equal number of instances. While the existence of subtle forms of sexism and misogyny is undeniable, having unbiased data representative of the same is essential to gain a better computational outcome.

Summary of general strategies used and existing challenges

Our summary of key research findings identified through the literature review reflects the current drawback in the study of sexism and misogyny identification tasks. Irrespective of the different measures taken by the literary works, some limitations remain consistent, which further hinder obtaining a robust model capable of quantifying sexism or misogyny. As Vidgen and Derczynski (2020) suggest, "More standardization is an important aspiration as research continues to mature, although it must be balanced with enabling research innovation and freedom." Therefore, we summarise the research findings in the following points: (1) We identify the various forms of online sexism—from direct abuse to implicit biases, indicating a need for nuanced detection and classification mechanisms; (2) Most studies focus on Western contexts and often overlook intersectional identities such as race, age, and sexual orientation, limiting the generalizability of findings; (3) There is inconsistency in how online sexism is defined and studied, with diverse methodologies and terminology making comparative analysis difficult; (4) The complex nature of online sexism requires insights from multiple disciplines—such as Computer Science and Social Science to form holistic solutions; (5) While some automated solutions for such a huge amount of online content show promise, there is a need for theoretically grounded exploration of online sexism beyond the scope of current research; (6) A high performance score does not necessarily indicate a robust model, which consequently highlight the potential need to incorporate additional metrics for a more comprehensive evaluation. However, we acknowledge that most of the recent research articles included in this study are from 2022. As illustrated in Fig. 2, the field experienced rapid growth up to that point and has continued to expand since. This temporal limitation implies that some recent advancements in sexism and misogyny detection may not be captured in our review. As the field has continued to evolve rapidly since 2022, certain challenges identified in this study may have already been addressed or reframed in more recent works. Consequently, our findings may not fully reflect the most recent developments, but it indicates the need for ongoing scholarly attention in this fast-growing area of research.

Conclusion

In this systematic literature review, we examined the multifaceted and evolving phenomenon of online sexism and misogyny, with particular attention to its categorization, detection, and the methodologies employed by existing literature across diverse digital platforms. Our analysis revealed that online sexist discourse manifests in a wide spectrum, ranging from explicit abuse and harassment, to implicit and context-dependent expressions. While advancements in NLP and machine learning have enabled meaningful progress in detecting overt forms of online sexism, subtler and more nuanced forms remain underexplored to maintain good consistency and accuracy using various methodologies. We also found that existing taxonomies and datasets are often limited in scope, predominantly rooted in Western contexts (both culturally and linguistically), and frequently lack intersectional perspectives. Furthermore, we observed that the research in this domain remains largely siloed, with insufficient integration between computational, sociological, and feminist theoretical frameworks. The findings of this review emphasizes the necessity for more inclusive, interdisciplinary, and context-aware approaches to the study and mitigation of online sexism. Future research should prioritize the creation of culturally diverse datasets, the development of unified conceptual taxonomies taking into account the range of expressions for sexism and misogyny, and possible incorporation of marginalized voices in both the design and evaluation of detection systems for better perceptivity. By fostering collaborations across disciplines and centering equity in technological development, we can advance more effective and socially responsive strategies to combat sexism and misogyny in online digital spaces. By highlighting these challenges, we aim to guide future research toward deeper theoretical integration, improved computational methods, and more representative datasets, ultimately fostering more nuanced understandings of sexism and misogyny in digital spaces. Through this work, we hope to contribute to further development on this topic ensuring updated resources on the same, and encouraging investigation on the change in dynamics of online sexism and misogyny.

Appendix 1: Systematic literature review strategy

Draft search string

Draft string length: 256 character limit

(1) (misogyny OR sexism)
(2) (hate OR toxic OR abusive OR offensive)
(3) (detection OR identification OR prediction OR classification)
(4) ("natural language processing" OR NLP OR "deep learning" OR "machine learning" OR ML OR "artificial intelligence" OR AI)
(5) 1/ AND 2/ AND 3/ AND 4/
(6) Limit 5 to (english language and yr="2012 -Current")

Inclusion and exclusion criteria

(1) Remove posts from online publishing platforms, online research platforms or similar (e.g. blogs).
(2) Remove papers outside the year range (2012–2022).
(3) Remove papers not written in English.
(4) Remove dissertations, theses, books, and whole conference proceedings; but include pre-prints within the period.
(5) Remove symposium submissions.
(6) Limit by date of external events (2000–current).
(7) Limit by the platform used for study- comparative study across platforms maybe included.
(8) Remove studies not looking at text data (so images, video, etc.).
(9) Remove studies that look into offline instances of sexism and misogyny.
(10) Remove studies that do not look into online social platforms (like Meta, Twitter, Reddit, etc.).
(11) Remove studies that focus on the mental and physical impact of online hate speech from the aforementioned platforms.
(12) Only keep papers that measure misogyny and/or sexism.

(a) This means removing studies with no quantitative methods, papers proposing guidelines, policy recommendations, discussions, tutorials, dataset descriptions, research briefs, working papers, purely theoretical approaches, opinion pieces, position papers, case studies, etc.
(b) Removing studies where frameworks are only stated without any measurements/ results proceeding it.
(c) It will include papers that measure misogyny/sexism with other forms of online hate, such as toxicity, hate-speech, aggression, etc.
(d) It can include gender-bias classification studies that fall close to the definition of sexism/misogyny as generic terms, depending on the context it is being used.

Appendix 2: Citation database queries

Table 2 Citation databases and their respective queries

Citations and their search queries	
Google Scholar	((misogyny OR sexism) AND (hate OR toxic OR abusive OR offensive) AND (detection OR identification OR prediction OR classification) AND ("natural language processing" OR NLP OR "deep learning" OR "machine learning" OR ML OR "artificial intelligence" OR AI) AND (language="English" AND yr="2012 -2022"))
ArXiv	(*all* : *sexism* + *OR* + *all* : *sexist* + *OR* + *all* : *misogyny* + *OR* + *all* : *misogynist* + *OR*+ *all* : %22*gender* + *discrimination*%22 + *OR* + *all* : %22*gender* + *violence*%22+ *OR* + *all* : %22*gender* + *stereotype*%22)
Elsevier	('misogyny detection OR misogyny identification OR misogyny prediction OR misogyny classification OR sexism detection OR sexism identification OR sexism prediction OR sexism classification')
Scopus	TITLE-ABS-KEY ((misogyny OR sexism OR gender AND violence OR gender AND discrimination) AND (detection OR identification OR prediction OR classification) AND PUBYEAR > 2011 AND PUBYEAR < 2023 AND PUBYEAR > 2011 AND PUBYEAR < 2023 AND (LIMIT-TO (SUBJAREA, "SOCI") OR LIMIT-TO (SUBJAREA, "COMP") OR LIMIT-TO (SUBJAREA, "PSYC")) AND (LIMIT-TO (LANGUAGE, "English"))
Semantic Scholar	('online sexism misogyny')
Web of Science (Social Science)	TS=((misogyn* OR sexis* OR (gender NEAR/10 discrim*) OR (gender NEAR/10 stereoty*) OR (gender NEAR/10 violence) OR (gender NEAR/10 based)) NEAR/200 (detect* OR identif* OR predict* OR classif*)) AND WC=((("History" OR "Political Science" OR "Womenś Studies" OR "Social Sciences" OR "International Relations" OR "History %26 Philosophy Of Science" OR "Linguistics" OR "Anthropology" OR "Sociology" OR "Social Work" OR "Language %26 Linguistics" OR "Information Science" OR "Psychology" OR "Social" OR "Ethnic Studies" OR "Philosophy" OR "Psychiatry") NOT ("Computer Science" OR "Artificial Intelligence" OR "Theory %26 Methods" OR "Engineering" OR "Software Engineering" OR "Scientific Disciplines" OR "Automation %26 Control Systems" OR "Mathematical" OR "Mathematics" OR "Mathematical Methods")) AND PY=2012-2022
Web of Science (Computer Science)	TS=((misogyn* OR sexis* OR (gender NEAR/10 discrim*) OR (gender NEAR/10 stereoty*) OR (gender NEAR/10 violence) OR (gender NEAR/10 based)) NEAR/200 (detect* OR identif* OR predict* OR classif*)) AND WC=((("Computer Science" OR "Artificial Intelligence" OR "Theory %26 Methods" OR "Engineering" OR "Software Engineering" OR "Scientific Disciplines" OR "Automation %26 Control Systems" OR "Mathematical" OR "Mathematics" OR "Mathematical Methods") NOT ("History" OR "Political Science" OR "Womenś Studies" OR "Social Sciences" OR "International Relations" OR "History %26 Philosophy Of Science" OR "Linguistics" OR "Anthropology" OR "Sociology" OR "Social Work" OR "Language %26 Linguistics" OR "Information Science" OR "Psychology" OR "Social" OR "Ethnic Studies" OR "Philosophy" OR "Psychiatry")) AND PY=2012-2022

Appendix 3: Terminologies and their meaning

Table 3 Terminologies

Construct	Abbreviation	"A construct is an abstract concept that is specifically chosen (or 'created') to explain a given phenomenon. Constructs used for scientific research must have precise and clear definitions that others can use to understand exactly what it means and what it does not mean" ((Bhattacherjee, 2019))
Computational Social Science	CSS	Computational social science is an interdisciplinary academic sub-field concerned with computational approaches to the social sciences. It leverages the capacity to collect and analyze data with an unprecedented breadth and depth and scale ((Lazer et al., 2020))
Hostile Sexism	HS	Hostile sexism refers to negative views toward individuals who violate traditional gender roles. For example, some people disparage girls who enter traditionally masculine domains such as science or sports ((Daniels & Leaper, 2011)). Part of ambivalent sexism ((Glick & Fiske, 1996))
Neo-sexism Scale	NS	A scale designed to tap into a new type of gender prejudice, called neo-sexist beliefs ((Tougas et al., 1995))
Benevolent Sexism	BS	Benevolent sexism includes valuing feminine-stereotyped attributes in females (e.g., nurturance) and a belief that traditional gender roles are necessary to complement one another. Benevolent sexism also includes the view known as paternalism that females need to be protected by males. Benevolent sexism contributes to gender inequality by limiting women's roles ((Daniels & Leaper, 2011)). Part of ambivalent sexism ((Glick & Fiske, 1996))
Bidirectional Encoder Representations from Transformers	BERT	BERT is a language representation model, which is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers ((Devlin et al., 2018))
Language Models (or Large Language models)	LM (or LLM)	A large language model is a computational model capable of language generation or other natural language processing tasks. As language models, LLMs acquire these abilities by learning statistical relationships from vast amounts of text during a self-supervised and semi-supervised training process
Bag-of-Words	BoW	"A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things: 1. A vocabulary of known words. 2. A measure of the presence of known words" http://tinyurl.com/5n6d9knt
Descriptive Paradigm	–	"The descriptive paradigm encourages annotator subjectivity to create datasets as granular surveys of individual beliefs. Descriptive data annotation thus allows for the capturing and modeling of different beliefs" ((Rottger et al., 2022))
Perspective Paradigm	–	"The prescriptive paradigm, on the other hand, discourages annotator subjectivity and instead tasks annotators with encoding one specific belief, formulated in the annotation guidelines. Prescriptive data annotation thus enables the training of models that seek to consistently apply one belief" ((Rottger et al., 2022))

Appendix 4: Experimentation results from other citation databases

For **Google Scholar**, we used both external APIs like SerpAPI for scraping the data, as well as a software named 'Publish or Perish' (Harzing, 2007) to collect the search results. Both of the methods were rejected because of their disadvantages. Such as, Publish or Perish could only extract 1000 results at a time for each search query. While this drawback was overcome by searching for documents with a shorter range of years to stay within the limit, it lacked some of the fields that were needed for this study—abstract and discipline. Alternatively, SerpAPI[15] worked similar to a web scrapping tool and could only scrape the results as the search engine demonstrates, i.e., it only scrapes what Google shows on their Google Scholar pages, nothing more. Even though the fields we got through this API were relevant, they did not contain the full information we needed for the analysis. For example, the full text in the title and abstract was missing and was instead indicated with dotted extensions in the beginning and end of the text.

Table 4 Categories for each area of research

Computer Science	Social Science
• Computer Science • Artificial Intelligence • Theory and Methods • Engineering • Software Engineering • Scientific Disciplines • Automation and Control Systems • Mathematical • Mathematics • Mathematical methods	• History • Political Science • Womenś Studies • Social Sciences • International Relations • History and Philosophy of Science • Linguistics • Anthropology • Sociology • Social Work • Language and Linguistics • Information Science • Psychology • Social • Ethnic Studies • Philosophy • Psychiatry

Appendix 5: Web of Science strategy

We performed automated elimination (or pre-processing) techniques based on the following criteria to narrow down our search results for both areas of study[16]:

- Remove studies that are not published in English.
- Remove studies that do not contain any abstracts.
- Keep only the first abstract in studies that contain more than one abstract.
- Remove certain publication types, such as review articles and editorials.

With the Web of Science API, separate search queries were used for the two broad disciplines (or research areas)—CS and Social Science. The categories of the research areas taken for each of them are as follows:

These disciplines were taken from the Web of Science category list, which branches from five major research areas—out of which we took the two categories **Social Sciences** and **Technology**. The published works present in the Web of Science Core Collection are assigned to at least one Web of Science category. Each of the said Web of Science

[15] https://serpapi.com/.

[16] More details can be found here: https://images.webofknowledge.com/images/help/WOS/hp_advanced_search.html.

categories (as listed in Table 4) is mapped to one research area found in the classification of research areas.[17]

Appendix 6: ArXiv strategy

The ArXiv API was used following the query search strategy.[18]

We performed automated elimination (or pre-processing) techniques based on the following criteria to narrow down our search results for both areas of study:

- Remove studies that are not between 2012 and 2022.
- Remove studies that do not contain any abstracts.

While combining search results of E and F, care was taken to remove the duplicate studies based on the title and abstract, where we kept the study from the former database. This is to ensure consistency along the data since the published and updated (i.e., when the pre-prints were submitted to ArXiv) years could differ, hence ensuring the published works are not mislabeled as pre-prints.

Appendix 7: Further analysis of the initial search results

Documents by disciplines

Figure 8 shows the frequency of publications per year in the range of 2012–2022, as per each discipline and publication type. Like we had discussed previously in Sect. 4.3, we see a huge disparity in the number of publications between the disciplines which focus on sexism and/or misogyny. This inherently appears to impact on the diversity of the concept

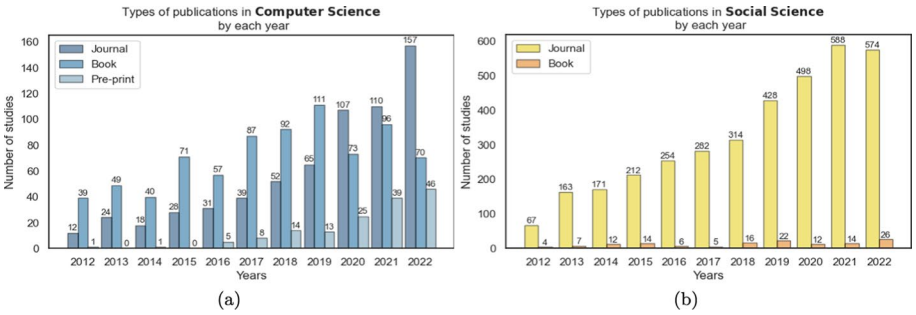

Fig. 8 **a** Type of publications in Computer Science. **b** Type of publications in Social Science

[17] Source: https://images.webofknowledge.com/images/help/WOS/hp_research_areas_easca.html.

[18] More details of the search strategy can be found here: https://info.arxiv.org/help/api/user-manual.html#query_details.

explored by the disciplines, with SS exploring a broader range of themes than CS. Furthermore, we also see that the type of publication too differs quite a bit as CS tend to produce a handful of research as pre-prints on this topic.

Documents focused on social media platforms

The share of documents focusing on different social media platforms, as observed in Fig. 9, reveal that X (formerly Twitter) was the dominant platform for most research in CS, while Facebook (or Meta) was more dominant in SS till 2022. The ease of access to Twitter data during the period could have been a contributing factor to allow application of automated approaches in CS. Whereas, Facebook having more number of active users could have contributed to more research in SS, than any other platforms (including Twitter).

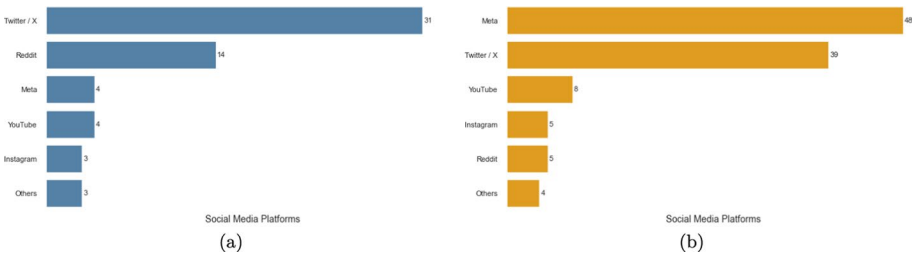

Fig. 9 Publications mentioning social media platforms in titles or (/and) abstracts in: **a** Computer Science, and **b** Social Science

General topics centering around sexism or misogyny

Figure 10 show the different thematic (or topic) representations across the disciplines over the period of 2012–2022. Not only do we see a wider range of themes in SS expanding over more number of research (like we observe in the previous subsection as well), but also a steady rise in most of the topics along the time. Especially the theme of 'Feminism with misogyny/sexism' and the 'Hostile sexism' seems to be of particular interest for SS

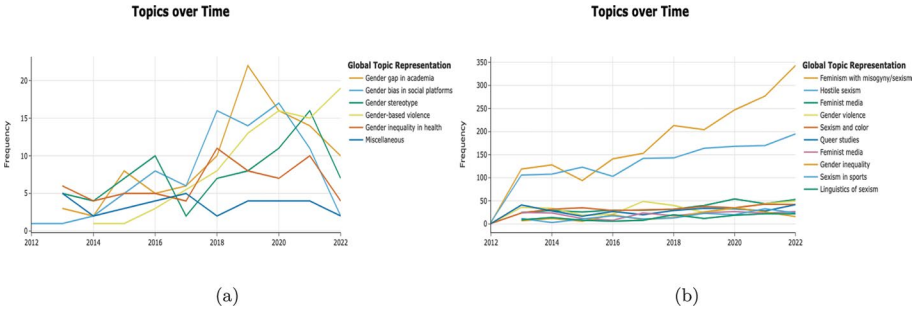

Fig. 10 General topics centered around sexism/misogyny over the years in: **a** Computer Science, and **b** Social Science

research, given the proliferation of sexism and misogyny beyond offline spaces. The theme of 'Linguistics in sexism' are instrumental in capturing the subtle forms of sexism, and is therefore seen to gain traction over the years. The themes of CS research on sexism and misogyny seem to fluctuate in the given period with no consistent rise, except for the 'Gender-based violence'.

Appendix 8: Abbreviation of models

The abbreviations used in Fig. 7 are a collection of the following models as shown in Table 5.

Table 5 Model names and their abbreviation

Abbreviation	Full name of the model(s)
LR	Logistic Regression
RF	Random Forest
SVM	Support Vector Machine
BERT	BERT, RoBERTa, mtBERT, FlauBERT, XLMRoBERTa, BERT-Base, Among Other BERT Based Models
CNN	Convolutional Neural Network
NB	Naïve-Bayes, Multinomial NB
LSTM	LSTM, Bi-LSTM
W2V	Word2Vec, GloVe
LDA	Latent Dirichlet Allocation
GB	Gradient Boosting, CatBoost
DT	Other Decision Tree models
GCN	Graph Convolutional Network
RNN	Recurrent Neural Network
DNN	Deep Neural Network (Unspecified)
XGB	XGBoost
*k*NN	*k*-Nearest Neighbours
BoW	Bag-of-Words
RC	Ridge Classifier
n-grams	Unigrams, Bi-grams and Other Types of *n*-grams
IG	Information Gain
MLP	Multi-layer Perceptron
Embeddings	FastText, InferSent, Universal Sentence Encoder, and Other Types of Embeddings
OVR	One-vs.-Rest
GRU	Gated Recurrent Units

Appendix 9: Most frequent keywords

In this section, we demonstrate the top 100 keywords in the co-occurrence network, like in Sect. 4.3 but based on all the manuscripts of each individual field.

Most frequent keywords in Computer Science

In the Fig. 11, we have the network diagram visualizing the most frequent and relevant keywords in Computer Science literature related to sexism and gender bias. The top 5 most common keywords are labelled in red boxes. Among the top 100 identified keywords, six emerged as particularly dominant: **gender stereotypes, ambivalent sexism, hostile sexism, sexism, benevolent sexism, and sexist attitudes**. These keywords are centrally positioned and form dense clusters, indicating their strong interconnections within the

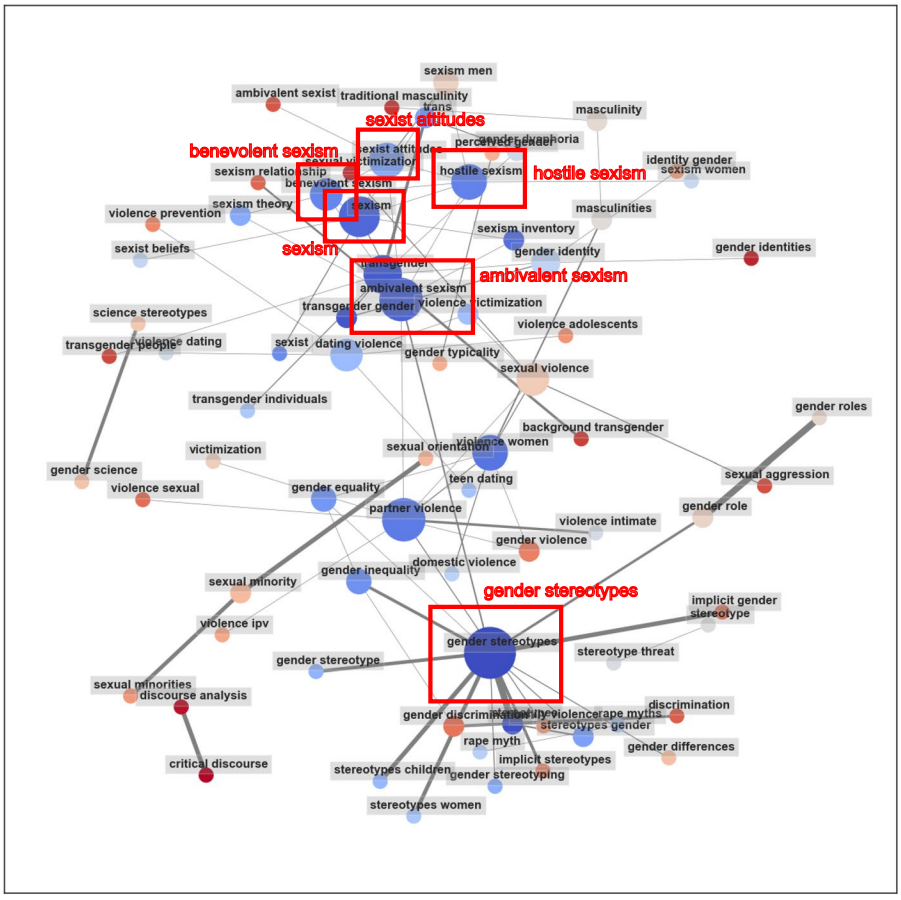

Fig. 11 Network diagram of most frequent keywords in **Computer Science**. Among all the top 100 frequent and relevant keywords, the 6 most common ones (in descending order) are highlighted in the figure: (1) gender stereotypes, (2) ambivalent sexism, (3) hostile sexism, (4) sexism, (5) benevolent sexism, (6) sexist attitudes

literature. The prominence of both psychological constructs (e.g., ambivalent sexism) and sociocultural phenomena (e.g., gender stereotypes) suggests a multidisciplinary engagement with sexism-related topics. The diagram highlights how research in Computer Science often intersects with gender theory, emphasizing the need for nuanced understandings of sexism in digital and computational contexts.

Most frequent keywords in Social Science

In the Fig. 12, the network diagram for Social Science presents the most frequent keyword connections related to sexism and gender bias within the discipline. Central nodes such as gender stereotypes, gender equality, gender identity, and gender roles indicate the strong thematic focus on societal structures and identity constructs. Compared to Computer Science, the Social Science network is more densely interlinked across themes like violence against women, transgender identity, and critical discourse, reflecting a broader and more intersectional engagement with gendered experiences. While terms like hostile sexism and benevolent sexism still appear, their relative positioning

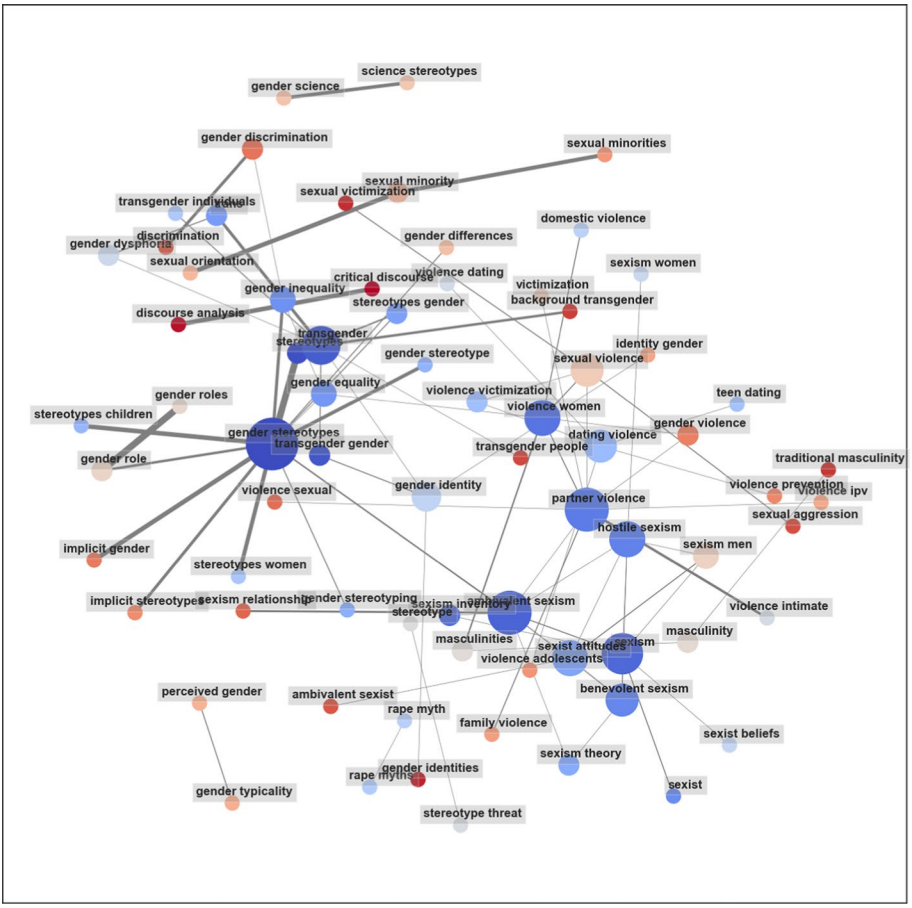

Fig. 12 Network connection of most frequent keywords in Social Science

suggests they are less dominant than structural and sociological concepts. This visualization underscores Social Science's emphasis on the sociocultural and identity-driven dimensions of sexism.

Appendix 10: Expanding on the automated selection techniques used

Topic Modeling approach

The model starts by transforming the input documents (abstracts and the titles) into numerical representations, with the help of embedding, which in these cases is a sentence embedding. Sentence embedding with transformer models maps a text of variable length to a fixed size embedding that should be representative of the meaning of the input text. For our research, we used the sentence transformer 'bge-small-en-v1.5',[19] which maps the each paragraph of our document to a 384 dimensional dense vector space, that was then used to cluster topics of similar semantic structure. In topic modeling, it is key to have a good quality of topic representations to interpret the overall topic and understand patterns in the document, for which we used bag-of-words (BoW) of medium length n-gram value (1–3 n-grams). To further enhance the representative-ness of the topics from BoW, Term Frequency–Inverse Document Frequency (TF–IDF) of our document, which works on a document-level, were adjusted to c-TF–IDF as per their weights, which works on a cluster/categorical/topic level. It considers the differences in documents from different clusters, and can be calculated as:

c-TF–IDF (for a term x within class c)

$$W_{x,c} = ||\mathrm{tf}_{x,c}|| \times \log\left(1 + \frac{A}{f_x}\right), \tag{1}$$

where $\mathrm{tf}_{x,c}$ = frequency of word x in class c,

f_x = frequency of word x across all classes,

A = average number of words per class.

Though both of these approaches did a good job of acquiring the topic representations, we used representation models to fine-tune the topics to refine its representations. For that, we used a combination of three models—a fast keyword extraction model called KeyBERTInspired, PartOfSpeech model, and MaximalMarginalRelevance model. The KeyBERTInspired model increases the coherence and reduces stopwords

Alongside this approach, we tried to further refine our topic representation by fine-tuning using a Large Language model (LLM) named 'Mistral 7B v0.1'—a 7 billion parameter language model, which has shown to outperform other state-of-the-art language models like Llama 13B across all elevated benchmarks (Jiang et al., 2023).

[19] The huggingface page of the model: https://huggingface.co/BAAI/bge-small-en-v1.5.

```
# The main representation of a topic
main_representation = KeyBERTInspired()

# Additional ways of representing a topic
pos_patterns = [
            [{'POS': 'ADJ'}, {'POS': 'NOUN'}],
            [{'POS': 'NOUN'}], [{'POS': 'ADJ'}]
]
aspect_model1 = PartOfSpeech("en_core_web_sm",
            pos_patterns=pos_patterns)
aspect_model2 = [KeyBERTInspired(top_n_words=30, random_state=1234),
            MaximalMarginalRelevance(diversity=.5)]

# LLM model
llm = Llama(model_path="../openhermes-2.5-mistral-7b.Q3_K_M.gguf",
            n_gpu_layers=-1, n_ctx=4096, stop=["Q:", "\n"])
prompt = """ Q:
I have a topic that contains the following documents:
[DOCUMENTS]

The topic is described by the following keywords: '[KEYWORDS]'.

Based on the above information, can you give a short label
of the topic of at most 5 words?
A:
"""
aspect_model3 = LlamaCPP(llm, prompt=prompt)

# Add all models together to be run in a single `fit`
representation_model = {
           "Main": main_representation,
           "Aspect1":  aspect_model1,
           "Aspect2":  aspect_model2,
           "Aspect3": aspect_model3
        }
# The documents to train on are the titles and abstracts of the studies
topic_model = BERTopic(representation_model=representation_model)
                .fit(docs)
```

To assess the model performance, the metrics perplexity and coherence scores were calculated as well. Perplexity is a predictive likelihood that specifically measures the probability that new data occurs given what was already learned by the model. In other words, perplexity characterizes how surprised a model is with new, unseen data. Coherence is typically used to analyze the relationship between two sets of data or the similarity between data sets. In topic modeling, topic coherence measures the quality of the data by comparing the semantic similarity between highly repetitive words in a topic. We used this to maximize intra-topic and minimize inter-topic similarity. We attained a perplexity score of 1.23 and a coherence score of 0.35 from our topic model.

Appendix 11: Analysis of the Computer Science studies: the final selection

In this section, we explore the data statistics for the CS manuscripts which were finally selected before the full-text screening process. The visual analysis is purely based on the text contained in abstracts and titles of the selected studies.

Documents by models

See Tables 6, 7, Figs. 13, 14.

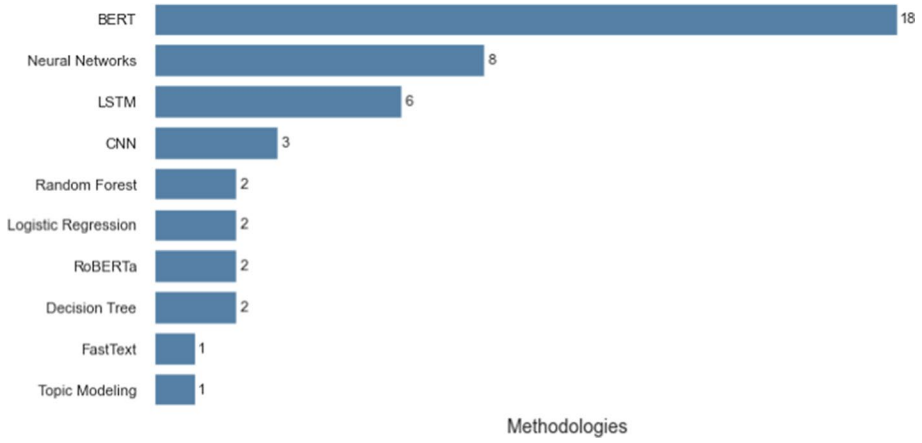

Fig. 13 Models gathered from the abstracts and titles of Computer Science studies

Task types and social platforms it is experimented on

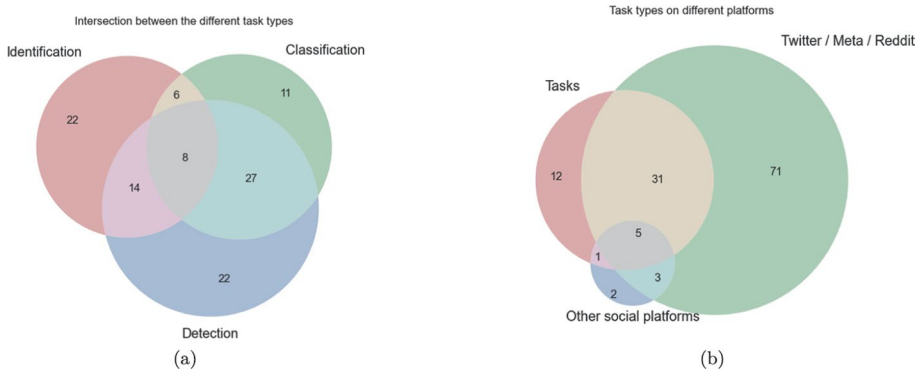

Fig. 14 **a** Types of task present in CS for the quantification of sexism and misogyny. The task types in the figure represent the tasks that we ideally expect a paper to have when quantifying the said terms. Note that, ALL the three task types are relevant for our work, and there are 110 total. **b** The aforementioned tasks and their application on different social media platforms. In this case, Twitter, Reddit and Facebook has shown to be the most research upon, while the other platforms are not. Regardless, a good number of research on those platforms use the specified tasks

Appendix 12: Studies included in the Meta-analysis

Table 6 Documents used for Meta-analysis (Part-1)

Citation	Title
Fersini et al. (2018)	Deep Learning Representations in Automatic Misogyny Identification: What Do We Gain and What Do We Miss?
Rahali et al. (2021)	Automatic Misogyny Detection in Social Media Platforms Using Attention-Based Bidirectional-LSTM
Frenda et al. (2019)	Online Hate Speech Against Women: Automatic Identification of Misogyny and Sexism on Twitter
Bhattacharya et al. (2020)	Developing a Multilingual Annotated Corpus of Misogyny and Aggression
Waseem (2016)	Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter
Gambäck and Sikdar (2017)	Using Convolutional Neural Networks to Classify Hate-Speech
Waseem and Hovy (2016)	Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter
Sen et al. (2022)	Counterfactually Augmented Data and Unintended Bias: The Case of Sexism and Hate Speech Detection
Samory et al. (2021)	"Call me sexist, but...": Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples
Anzovino et al. (2018)	Automatic Identification and Classification of Misogynistic Language on Twitter
Jha and Mamidi (2017)	When Does a Compliment Become Sexist? Analysis and Classification of Ambivalent Sexism Using Twitter Data
Butt et al. (2021)	Sexism Identification Using BERT and Data Augmentation-EXIST2021
Katsarou et al. (2021)	Sentiment Polarization in Online Social Networks: The Flow of Hate Speech
Parikh et al. (2019)	Multi-label Categorization of Accounts of Sexism Using a Neural Framework
Abburi et al. (2021)	Fine-Grained Multi-label Sexism Classification Using a Semi-Supervised Multi-level Neural Approach
Melville et al. (2019)	Topic Modelling of Everyday Sexism Project Entries
Sharifirad et al. (2018)	Boosting Text Classification Performance on Sexist Tweets by Text Augmentation and Text Generation Using a Combination of Knowledge Graphs
Nozza et al. (2019)	Unintended Bias in Misogyny Detection
Chiril et al. (2020)	An Annotated Corpus for Sexism Detection in French Tweets
Chiril et al. (2021)	"Be nice to your wife! The restaurants are closed": Can Gender Stereotype Detection Improve Sexism Classification?
Zeinert et al. (2021)	Annotating Online Misogyny
Guest et al. (2021)	An Expert Annotated Dataset for the Detection of Online Misogyny
Grosz and Conde-Cespedes (2020)	Automatic Detection of Sexist Statements Commonly Used at the Workplace
Shah et al. (2020)	Detecting Hate Speech Against Women
Canós (2018)	Misogyny Identification Through SVM

Table 7 Documents used for Meta-analysis (Part-2)

Citation	Title
Schutz et al. (2022)	Automatic Sexism Detection with Multilingual Transformer Models
Lynn et al. (2019)	A Comparison of Machine Learning Approaches for Detecting Misogynistic Speech in Urban Dictionary
Talavera et al. (2021)	System Description for EXIST Shared Task at IberLEF 2021: Automatic Misogyny Identification Using Pretrained Transformers
Plaza-del Arco et al. (2021)	Sexism Identification in Social Networks using a Multi-task Learning System
Singh et al. (2021)	"Hold on honey, men at work": A Semi-supervised Approach to Detecting Sexism in Sitcoms
Mishra et al. (2019)	Abusive Language Detection with Graph Convolutional Network
Bashar et al. (2019)	Misogynistic Tweet Detection: Modelling CNN with Small Datasets, in: Communications in Computer and Information Science
Frenda et al. (2018)	Exploration of Misogyny in Spanish and English Tweets
Gordeev and Lykova (2020)	BERT of All Trades, Master of Some
Frenda et al. (2018)	Exploration of Misogyny in Spanish and English Tweets
Plaza-Del-Arco et al. (2020)	Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies
Attanasio et al. (2022)	Entropy-Based Attention Regularization Frees Unintended Bias Mitigation from Lists
Rodriguez-Sanchez et al. (2020)	Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data
Attanasio and Pastor (2020)	PoliTeam @ AMI: Improving Sentence Embedding Similarity with Misogyny Lexicons for Automatic Misogyny Identification in Italian Tweets
Indurthi et al. (2019)	FERMI at SemEval-2019 Task 5: Using Sentence embeddings to Identify Hate Speech Against Immigrants and Women in Twitter
Kohli et al. (2021)	ARGUABLY at ComMA@ ICON: Detection of Multilingual Aggressive, Gender Biased, and Communally Charged Tweets using Ensemble and Fine-Tuned IndicBERT
Ou and Li (2020)	YNU_OXZ@ HaSpeeDe 2 and AMI: XLM-RoBERTa with Ordered Neurons LSTM for Classification Task at EVALITA 2020
Dinan et al. (2020)	Multi-dimensional Gender Bias Classification
Nascimento et al. (2022)	Unintended Bias Evaluation: An Analysis of Hate Speech Detection and Gender Bias Mitigation on Social Media Using Ensemble Learning
Badjatiya et al. (2017)	Deep Learning for Hate Speech Detection in Tweets
Fersini et al. (2021)	Deep Learning Representations in Automatic Misogyny Identification: What Do We Gain and What Do We Miss?
Farrell et al. (2019)	Exploring Misogyny Across the Manosphere in Reddit

Funding We thank the University of Exeter for funding the cost to access the Web of Science Expanded API and SerpAPI.

A.D.'s time on the research was funded by the SSIS Global Excellence PhD Studentship from the University of Exeter.

S.B.'s time on the research was funded by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No 101019284).

C.Q.C. thanks the Ewha Frontier 10-10 Project and the DSO National Laboratories Singapore for funding this research.

Data availability No data was generated during this research, but were acquired from online websites or through the API access of the stated citation databases. All the shareable acquired data from the databases collected and used in the research, along with its analysis is made available in the GitHub page: https://github.com/booktrackerGirlsys-lit-review-Sexism. We also include the permissive license to allow users to use, modify and distribute the materials.

Declarations

Conflict of interest The authors have no Conflict of interest to declare that are relevant to the content of this article.

References

Abburi, H., Parikh, P., Chhaya, N., & Varma, V. (2021). Fine-grained multi-label sexism classification using a semi-supervised multi-level neural approach. *Data Science and Engineering, 6*(4), 359–379.

Amnesty International. (2017). *Amnesty reveals alarming impact of online abuse against women*. Amnesty International. Retrieved March 24, 2025.

Anderson, K. J. (2014). *Modern misogyny: Anti-feminism in a post-feminist era*. Oxford University Press.

Anzovino, M. E., Fersini, E., & Rosso, P. (2018). Automatic identification and classification of misogynistic language on Twitter. In *Natural language processing and information systems: 23rd international conference on applications of natural language to information systems, NLDB 2018, Proceedings*, Paris, France, June 13–15, 2018.

Arango, A., Pérez, J., & Poblete, B. (2022). Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems, 105*, 101584.

Archer, A. M., & Kam, C. D. (2020). Modern sexism in modern times public opinion in the # MeToo era. *Public Opinion Quarterly, 84*(4), 813–837.

Attanasio, G., Nozza, D., Hovy, D., & Baralis, E. (2022). Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2022. Association for Computational Linguistics.

Attanasio, G., & Pastor, E. (2020). PoliTeam @ AMI: Improving sentence embedding similarity with misogyny lexicons for automatic misogyny identification in Italian tweets. In *EVALITA evaluation of NLP and speech tools for Italian*, December 17, 2020.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, WWW '17 Companion, Republic and Canton of Geneva (pp. 759–760), CHE. International World Wide Web Conferences Steering Committee.

Barak, A. (2005). Sexual harassment on the Internet. *Social Science Computer Review, 23*(1), 77–92.

Bartow, A. (2009). *Internet defamation as profit center: The monetization of online harassment*. Faculty Publications.

Bashar, M. A., Nayak, R., Suzor, N., & Weir, B. (2019). Misogynistic tweet detection: Modelling CNN with small datasets. In *Data mining: 16th Australasian conference, AusDM 2018*, Revised Selected Papers 16, Bathurst, NSW, Australia, November 28–30, 2018 (pp. 3–16). Springer.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., & Sanguinetti, M. (2019). SemEval-2019 Task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki & S. M. Mohammad (Eds.), *Proceedings of the 13th international workshop on semantic evaluation*, Minneapolis, Minnesota, USA (pp. 54–63). Association for Computational Linguistics.

Basile, V., Croce, D., Maro, M. D., & Passaro, L. C. (2020). EVALITA 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for Italian. In V. Basile, D. Croce, M. D. Maro & L. C. Passaro (Eds.), *Proceedings of the seventh evaluation campaign of natural language processing and speech tools for Italian. Final workshop (EVALITA 2020), CEUR workshop proceedings*, Online event, December 17, 2020 (Vol. 2765). CEUR-WS.org.

Bates, L. (2015). *Everyday sexism*. Schuster UK.

Beattie, M. Y., & Diehl, L. A. (1979). Effects of social conditions on the expression of sex-role stereotypes. *Psychology of Women Quarterly, 4*(2), 241–255.

Bhattacharya, S., Singh, S., Kumar, R., Bansal, A., Bhagat, A., Dawer, Y., Lahiri, B., & Ojha, A. K. (2020). Developing a multilingual annotated corpus of misogyny and aggression. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, 2020, Marseille, France (pp. 158–168). European Language Resources Association (ELRA).

Bhattacherjee, A. (2019). *Social science research: Principles, methods and practices (revised edition)*. University of South Florida.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008.

Bolanos, F., Salatino, A., Osborne, F., & Motta, E. (2024). Artificial intelligence for literature reviews: Opportunities and challenges.

Briggs, J. (2023). Advanced topic modeling with BERTopic.

Butt, S., Ashraf, N., Sidorov, G., & Gelbukh, A. F. (2021). Sexism identification using BERT and data augmentation—EXIST2021. In *IberLEF@SEPLN*, 2021.

Cameron, D. (2020). Language and gender: Mainstreaming and the persistence of patriarchy. *International Journal of the Sociology of Language, 2020*(263), 25–30.

Canós, J. S. (2018). Misogyny identification through SVM at IberEval 2018. In *IberEval@SEPLN*, 2018.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Chiril, P., Benamara, F., & Moriceau, V. (2021). "Be nice to your wife! the restaurants are closed": Can gender stereotype detection improve sexism classification? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, Punta Cana, Dominican Republic (pp. 2833–2844). Association for Computational Linguistics.

Chiril, P., Moriceau, V., Benamara, F., Mari, A., Origgi, G., & Coulomb-Gully, M. (2020). An annotated corpus for sexism detection in French tweets. In *Proceedings of the twelfth language resources and evaluation conference*, 2020 (pp. 1397–1403).

Citron, D. K. (2014). *Hate crimes in cyberspace*. Harvard University Press.

Code, L. (2002). *Encyclopedia of feminist theories*. Routledge.

Daniels, E., & Leaper, C. (2011). Gender issues. In B. B. Brown & M. J. Prinstein (Eds.), *Encyclopedia of adolescence* (pp. 151–159). Academic.

Das, A., Rahgouy, M., Zhang, Z., Bhattacharya, T., Dozier, G., & Seals, C. D. (2023). Online sexism detection and classification by injecting user gender information. In *2023 IEEE international conference on artificial intelligence, Blockchain, and Internet of Things (AIBThings)*, 2023 (pp. 1–5).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint. arXiv:1810.04805

Dinan, E., Fan, A., Wu, L., Weston, J., Kiela, D., & Williams, A. (2020). Multi-dimensional gender bias classification. arXiv e-prints. arXiv:2005.00614

Duggan, M. (2017). *Online harassment 2017*. Pew Research Center.

Eagly, A. H., & Mladinic, A. (1989). Gender stereotypes and attitudes toward women and men. *Personality and Social Psychology Bulletin, 15*(4), 543–558.

Eagly, A. H., & Mladinic, A. (1994). Are people prejudiced against women? Some answers from research on attitudes, gender stereotypes, and judgments of competence. *European Review of Social Psychology, 5*(1), 1–35.

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, WebSci '19*, 2019, New York, NY, USA (pp. 87–96). Association for Computing Machinery.

Fersini, E., Nozza, D., & Rosso, P. (2018). Overview of the EVALITA 2018 task on automatic misogyny identification (AMI). In *EVALITA@CLiC-it*, 2018.

Fersini, E., Rosato, L., Candelieri, A., Archetti, F., Messina, E., et al. (2021). Deep learning representations in automatic misogyny identification: What do we gain and what do we miss? In *CLiC-it*, 2021.

Fiala, D., & Tutoky, G. (2017). Computer science papers in Web of Science: A bibliometric analysis. *Publications, 5*(4), 23.

Finlayson, M. A., & Erjavec, T. (2017). Overview of annotation creation: Processes and tools. In *Handbook of linguistic annotation* (pp. 167–191). Springer.

Fontanella, L., Chulvi, B., Ignazzi, E., Sarra, A., & Tontodimamma, A. (2024). How do we study misogyny in the digital age? A systematic literature review using a computational linguistic approach. *Humanities and Social Sciences Communications, 11*(1), 1–15.

FRA. (2023). *Online content moderation: A fundamental rights-based approach*. FRA.

Frenda, S., Ghanem, B., & Montes-y-Gómez, M. (2018). Exploration of misogyny in Spanish and English tweets. In *IberEval@SEPLN*, 2018.

Frenda, S., Ghanem, B., Montes-y Gómez, M., & Rosso, P. (2019). Online hate speech against women: Automatic identification of misogyny and sexism on Twitter. *Journal of Intelligent and Fuzzy Systems, 36*(5), 4743–4752.

Gambäck, B., & Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. In Z. Waseem, W. H. K. Chung, D. Hovy & J. Tetreault (Eds.), *Proceedings of the first workshop on abusive language online*, 2017, Vancouver, BC, Canada (pp. 85–90). Association for Computational Linguistics.

García-Cueto, E., Rodríguez-Díaz, F. J., Bringas-Molleda, C., López-Cepero, J., Paíno-Quesada, S., & Rodríguez-Franco, L. (2015). Development of the gender role attitudes scale (GRAS) amongst young Spanish people. *International Journal of Clinical and Health Psychology, 15*(1), 61–68.

Ging, D., & Siapera, E. (2018). Special issue on online misogyny. *Feminist Media Studies, 18*(4), 515–524.

Ging, D., & Siapera, E. (2019). *Gender hate online understanding the new anti-feminism: Understanding the new anti-feminism*. Springer.

Glick, P., & Fiske, S. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology, 70*, 491–512.

Glick, P., & Fiske, S. (1997). Hostile and benevolent sexism: Measuring ambivalent sexist attitudes toward women. *Psychology of Women Quarterly, 21*(1), 119–135.

Gordeev, D., & Lykova, O. (2020). BERT of all trades, master of some. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, 2020, Marseille, France (pp. 93–98). European Language Resources Association (ELRA).

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018). All you need is "love": Evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security, AISec '18*, 2018 , New York, NY, USA (pp. 2–12). Association for Computing Machinery.

Grootendorst, M. (2020). KeyBERT: Minimal keyword extraction with BERT.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF–IDF procedure.

Grosz, D., & Conde-Cespedes, P. (2020). Automatic detection of sexist statements commonly used at the workplace. In *Trends and applications in knowledge discovery and data mining: PAKDD 2020 workshops, DSFN, GII, BDM, LDRC and LBD*, Revised Selected Papers, Singapore, May 11–14, 2020.

Guest, E., Vidgen, B., Mittos, A., Sastry, N., Tyson, G., & Margetts, H. (2021). An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online, 2021 (pp. 1336–1350). Association for Computational Linguistics.

Hammersley, M., & Gomm, R. (1997). Bias in social research. *Sociological Research Online, 2*(1), 7–19.

Harzing, A. (2007). Publish or perish. https://harzing.com/resources/publish-or-perish

Herd, B., & Burton, S. (2024). Can you trust your ML metrics? Using subjective logic to determine the true contribution of ML metrics for safety. In *Proceedings of the 39th ACM/SIGAPP symposium on applied computing, SAC '24*, 2024, New York, NY, USA (pp. 1579–1586). Association for Computing Machinery.

Herring, S. C. (1999). The rhetorical dynamics of gender harassment online. *The Information Society, 15*(3), 151–167.

Hewitt, S., Tiropanis, T., & Bokhove, C. (2016). The problem of identifying misogynist language on Twitter (and other online social spaces). In *Proceedings of the 8th ACM conference on Web science, WebSci '16*, 2016, New York, NY, USA (pp. 333–335). Association for Computing Machinery.

Hirsch, S. F. (1992). Julia Penelope, speaking freely: Unlearning the lies of the fathers' tongues. New York: Pergamon, 1990. pp. xxxvii 281. Deborah Cameron (ed.), The feminist critique of language: A reader. London and New York: Routledge, 1990. pp. xi 258. *Language in Society, 21*(1), 136–142.

Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving Google's perspective API built for detecting toxic comments.

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass, 15*(8), e12432.

Indurthi, V., Syed, B., Shrivastava, M., Chakravartula, N., Gupta, M., & Varma, V. (2019). FERMI at SemEval-2019 Task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, 2019, Minneapolis, Minnesota, USA (pp. 70–74). Association for Computational Linguistics.

Jane, E. A. (2016). Online misogyny and feminist digilantism. *Continuum, 30*(3), 284–297.

Jha, A., & Mamidi, R. (2017). When does a compliment become sexist? Analysis and classification of ambivalent sexism using Twitter data. In *Proceedings of the second workshop on NLP and computational social science*, 2017, Vancouver, Canada (pp. 7–16). Association for Computational Linguistics.

Jiang, A., Yang, X., Liu, Y., & Zubiaga, A. (2021). SWSR: A Chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media, 27*, 100182.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b. arXiv preprint. arXiv:2310.06825

Jurasz, O., & Barker, K. (2019). Online misogyny: A challenge for digital feminism? *Journal of International Affairs, 72*(2), 95–114.

Karami, A., Swan, S., White, C. N., & Ford, K. (2019). Hidden in plain sight for too long: Using text mining techniques to shine a light on workplace sexism and sexual harassment. *Psychology of Violence*. https://doi.org/10.1037/vio0000239

Katsarou, K., Sunder, S., Woloszyn, V., & Semertzidis, K. (2021). Sentiment polarization in online social networks: The flow of hate speech. In *2021 Eighth international conference on social network analysis, management and security (SNAMS)*, 2021 (pp. 01–08).

King, L. A., & King, D. W. (1997). Sex-role egalitarian ism scale: Development, psychometric properties, and recommendations for future research. *Psychology of Women Quarterly, 21*(1), 71–87.

Kohli, G., Kaur, P., & Bedi, J. (2021). ARGUABLY at ComMA@ICON: Detection of multilingual aggressive, gender biased, and communally charged tweets using ensemble and fine-tuned IndicBERT. In *Proceedings of the 18th international conference on natural language processing: Shared task on multilingual gender biased and communal language identification*, 2021, NIT Silchar (pp. 46–52). NLP Association of India (NLPAI).

Lazer, D. M., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., et al. (2020). Computational social science: Obstacles and opportunities. *Science, 369*(6507), 1060–1062.

Lynn, T., Endo, P. T., Rosati, P., Silva, I., Santos, G. L., & Ging, D. (2019). A comparison of machine learning approaches for detecting misogynistic speech in urban dictionary. In *2019 International conference on cyber situational awareness, data analytics and assessment (Cyber SA)*, 2019 (pp. 1–8).

Manne, K. (2017). *Down girl: The logic of misogyny*. Oxford University Press.

Mantilla, K. (2013). Gendertrolling: Misogyny adapts to new media. *Feminist Studies, 39*(2), 563–570.

Matsuda, M. J. (2018). Public response to racist speech: Considering the victim's story. In *Words that wound* (pp. 17–51). Routledge.

Megarry, J. (2014). Online incivility or sexual harassment? Conceptualising women's experiences in the digital age. *Women's Studies International Forum, 47*, 46–55.

Melville, S., Eccles, K., & Yasseri, T. (2019). Topic modeling of everyday sexism project entries. *Frontiers in Digital Humanities*. https://doi.org/10.3389/fdigh.2018.00028

Mishra, P., Del Tredici, M., Yannakoudakis, H., & Shutova, E. (2019). Abusive language detection with graph convolutional networks. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long and Short Papers)*, 2019 (Vol. 1).

Mudrak, B. (2018). What are preprints, and how do they benefit authors?

Nascimento, F. R., Cavalcanti, G. D., & Da Costa-Abreu, M. (2022). Unintended bias evaluation: An analysis of hate speech detection and gender bias mitigation on social media using ensemble learning. *Expert Systems with Applications, 201*, 117032.

Nadim, M., & Fladmoe, A. (2021). Silencing women? Gender and online harassment. *Social Science Computer Review, 39*(2), 245–258.

Nozza, D., Volpetti, C., & Fersini, E. (2019). Unintended bias in misogyny detection. In *IEEE/WIC/ACM international conference on Web intelligence, WI '19*, 2019, New York, NY, USA (pp. 149–155). Association for Computing Machinery.

Ou, X., & Li, H. (2020). YNU_OXZ@ HaSpeeDe 2 and AMI: XLM-RoBERTa with ordered neurons LSTM for classification task at EVALITA 2020. In *EVALITA evaluation of NLP and speech tools for Italian* (Vol. *2765*, pp. 102–109). Accademia University Press.

Parikh, P., Abburi, H., Badjatiya, P., Krishnan, R., Chhaya, N., Gupta, M., & Varma, V. (2019). Multi-label categorization of accounts of sexism using a neural framework. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP–IJCNLP)*, 2019.

Parikh, P., Abburi, H., Chhaya, N., Gupta, M., & Varma, V. (2021). Categorizing sexism and misogyny through neural approaches. *ACM Transactions on the Web, 15*(4), 1–31.

Plaza-del Arco, F. M., Molina-González, M. D., López, L., & Martín-Valdivia, M. (2021). Sexism identification in social networks using a multi-task learning system. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII international Conference of the Spanish Society for Natural Language Processing*, 2021, Málaga, Spain (Vol. 2943, pp. 491–499).

Plaza-Del-Arco, F.-M., Molina-González, M. D., Ureña López, L. A., & Martín-Valdivia, M. T. (2020). Detecting misogyny and xenophobia in Spanish tweets using language technologies. *ACM Transactions on Internet Technology, 20*(2), 1–19.

Pykes, K. (2023). What is topic modeling? An introduction with examples.

Rahali, A., Akhloufi, M. A., Therien-Daniel, A.-M., & Brassard-Gourdeau, E. (2021). Automatic misogyny detection in social media platforms using attention-based bidirectional-LSTM. In *2021 IEEE international conference on systems, man, and cybernetics (SMC)*, 2021 (pp. 2706–2711).

Richardson-Self, L. (2018). Woman-hating: On misogyny, sexism, and hate speech. *Hypatia, 33*(2), 256–272.

Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., & Plaza, L. (2020). Automatic classification of sexism in social networks: An empirical study on Twitter data. *IEEE Access, 8*, 219563–219576.

Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., & Donoso, T. (2021). Overview of EXIST 2021: Sexism identification in social networks. *Procesamiento del Lenguaje Natural, 67*, 195–207.

Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L., Mendieta-Aragón, A., Marco-Remón, G., Makeienko, M., Plaza, M., Gonzalo, J., Spina, D., & Rosso, P. (2022). Overview of EXIST 2022: Sexism identification in social networks. *Procesamiento de Lenguaje Natural, 69*, 229–240.

Röttger, P., Vidgen, B., Hovy, D., & Pierrehumbert, J. B. (2022). Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022.

Samory, M., Sen, I., Kohne, J., Flöck, F., & Wagner, C. (2021). "Call me sexist, but..." : Revisiting sexism detection using psychological scales and adversarial samples. In *Proceedings of the international AAAI conference on Web and social media*, 2021 (Vol. 15(1), pp. 573–584).

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, 2019, Florence, Italy (pp. 1668–1678). Association for Computational Linguistics.

Schmid Mast, M. (2004). Men are hierarchical, women are egalitarian: An implicit gender stereotype. *Swiss Journal of Psychology, 63*(2), 107–111.

Schütz, M., Boeck, J., Liakhovets, D., Slijepčević, D., Kirchknopf, A., Hecht, M., Bogensperger, J., Schlarb, S., Schindler, A., & Zeppelzauer, M. (2022). Automatic sexism detection with multilingual transformer models.

Sen, I., Samory, M., Wagner, C., & Augenstein, I. (2022). Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, Seattle, United States (pp. 4716–4726). Association for Computational Linguistics.

Sensales, G., & Areni, A. (2017). Gender biases and linguistic sexism in political communication: A comparison of press news about men and women Italian ministers. *Journal of Social and Political Psychology, 5*(2), 512–536.

Shah, D. S., Schwartz, H. A., & Hovy, D. (2020). Predictive biases in natural language processing models: A conceptual framework and overview. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, Online, 2020 (pp. 5248–5264). Association for Computational Linguistics.

Sharifirad, S., Jafarpour, B., & Matwin, S. (2018). Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, 2018 (pp. 107–114).

Siddiqi, N., Bains, A., Aleem, S., & Aleem, S. (2018). Analysing threads of sexism in new age humour: A content analysis of Internet memes. *Indian Journal of Social Research, 59*, 356.

Singh, S., Anand, T., Ghosh Chowdhury, A., & Waseem, Z. (2021). "Hold on honey, men at work": A semi-supervised approach to detecting sexism in sitcoms. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on Natural Language Processing: Student research workshop*, Online, 2021 (pp. 180–185). Association for Computational Linguistics.

Srivastava, K., Chaudhury, S., Bhat, P., & Sahu, S. (2017). Misogyny, feminism, and sexual harassment. *Industrial Psychiatry Journal, 26*, 111.

Sutton, C., & Gong, L. (2017). Popularity of arxiv. org within computer science. arXiv preprint. arXiv:1710.05225

Swim, J. K., & Cohen, L. L. (1997). Overt, covert, and subtle sexism: A comparison between the attitudes toward women and modern sexism scales. *Psychology of Women Quarterly, 21*(1), 103–118.

Swim, J. K., Mallett, R., & Stangor, C. (2004). Understanding subtle sexism: Detection and use of sexist language. *Sex Roles, 51*, 117–128.

Takkouche, B., & Norman, G. (2011). PRISMA statement. *Epidemiology, 22*(1), 128.

Talavera, I., Fidalgo, D. C., & Vila-Suero, D. (2021). System description for exist shared task at IberLEF 2021: Automatic misogyny identification using pretrained transformers. In *IberLEF@ SEPLN*, 2021 (pp. 484–490).

Tougas, F., Brown, R., Beaton, A. M., & Joly, S. (1995). Neosexism: Plus ça change, plus c'est pareil. *Personality and Social Psychology Bulletin, 21*(8), 842–849.

UN Secretary-General. (2024). *Intensification of efforts to eliminate all forms of violence against women and girls: Technology-facilitated violence against women and girls: Report of the Secretary-General*. UN Secretary-General.

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE, 15*(12), e0243300.

Vitak, J., Chadha, K., Steiner, L., & Ashktorab, Z. (2017). Identifying women's experiences with and strategies for mitigating negative effects of online harassment. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 2017 (pp. 1231–1245).

Waseem, Z. (2016). Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the first workshop on NLP and computational social science*, 2016, Austin, Texas (pp. 138–142). Association for Computational Linguistics.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL student research workshop*, 2016, San Diego, California (pp. 88–93). Association for Computational Linguistics.

WHO. (2013). *Violence against women*. WHO.

Wilson, R. A., & Land, M. K. (2020). Hate speech on social media: Content moderation in context. *Connecticut Law Review, 52*, 1029.

Wrisley, S. P. (2023). Feminist theory and the problem of misogyny. *Feminist Theory, 24*(2), 188–207.

Yasseri, T., Eccles, K., & Melville, S. (2016). Sexism typology: Literature review.

Zeinert, P., Inie, N., & Derczynski, L. (2021). Annotating online misogyny. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing: Long Papers*, Online, 2021 (Vol. 1, pp. 3181–3197). Association for Computational Linguistics.

Zhang, L. (2014). The impact of data source on the ranking of computer scientists based on citation indicators: A comparison of Web of Science and Scopus? *Issues in Science and Technology Librarianship*. https://doi.org/10.5062/F4D798CW

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Short Papers*, 2018, New Orleans, Louisiana (Vol. 2, pp. 15–20). Association for Computational Linguistics.

Authors and Affiliations

Aditi Dutta[1,4] · **Susan Banducci**[1] · **Chico Q. Camargo**[1,2,3,4]

✉ Aditi Dutta
ad882@exeter.ac.uk

Susan Banducci
S.A.Banducci@exeter.ac.uk

Chico Q. Camargo
F.Camargo@exeter.ac.uk

[1] University of Exeter, Exeter, UK

[2] Oxford Internet Institute, University of Oxford, Oxford, UK

[3] Ewha Womans University, Seoul, South Korea

[4] Alan Turing Institute, London, UK