A Literature Survey on Multimodal and Multilingual Sexism Detection

Xuan Luo, Bin Liang, Qianlong Wang, Jing Li, Erik Cambria, Xiaojun Zhang, Yulan He, Min Yang, and Ruifeng Xu, *Member, IEEE*

***Abstract*—Sexism has become a pressing issue, driven by the rapid-spreading influence of societal norms, media portrayals, and online platforms that perpetuate and amplify gender biases. Curbing sexism has emerged as a critical challenge globally. Being capable of recognizing sexist statements and behaviors is of particular importance since it is the first step in mind change. This survey provides an extensive overview of recent advancements in sexism detection. We present details of the various resources used in this field and methodologies applied to the task, covering different languages, modalities, models, and approaches. Moreover, we examine the specific challenges these models encounter in accurately identifying and classifying sexism. Additionally, we highlight areas that require further research and propose potential new directions for future exploration in the domain of sexism detection. Through this comprehensive exploration, we strive to contribute to the advancement of interdisciplinary research, fostering a collective effort to combat sexism in its multifaceted manifestations.**

***Index Terms*—Large language models (LLMs), multilingual, multimodal, sexism detection, survey.**

Received 27 July 2024; revised 26 October 2024, 6 January 2025, and 27 February 2025; accepted 8 April 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62176076; in part by the Natural Science Foundation of Guangdong under Grant 2023A1515012922; in part by the Shenzhen Foundational Research Funding under Grant JCYJ20220818102415032; in part by the Major Key Project under Grant PCL2023A09; in part by Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies under Grant 2022B1212010005; and in part by CIPSC-SMP-ZHIPU Large Model Cross-Disciplinary Fund under Grant ZPCG20241119405. Xuan Luo and Jing Li were supported in part by the Research Grants Council of the Hong Kong Special Administrative Region under Grant PolyU/25200821; in part by the Innovation and Technology Fund under Grant PRP/047/22FX; in part by PolyU Research Centre on Data Science and Artificial Intelligence under Grant 1-CE1E; and a gift fund from Huawei Noah's Ark Lab. *(Corresponding author: Ruifeng Xu.)*

Xuan Luo is with Harbin Institute of Technology, Shenzhen 518055, China, and also with The Hong Kong Polytechnic University, Hong Kong 999077, China.

Bin Liang is with The Chinese University of Hong Kong, Hong Kong 999077, China.

Qianlong Wang is with Harbin Institute of Technology, Shenzhen 518055, China.

Jing Li is with The Hong Kong Polytechnic University, Hong Kong 999077, China.

Erik Cambria is with Nanyang Technological University, Singapore 639798.

Xiaojun Zhang is with Xi'an Jiaotong-Liverpool University, Suzhou 215123, China.

Yulan He is with King's College London, WC2R 2LS London, U.K.

Min Yang is with Shenzhen Institute of Advanced Technology, Shenzhen 518055, China.

Ruifeng Xu is with Harbin Institute of Technology, Shenzhen 518055, China, and also with Peng Cheng Laboratory, Shenzhen, China (e-mail: xuruifeng@hit.edu.cn).

Digital Object Identifier 10.1109/TCSS.2025.3561921

I. Introduction

SEXISM, characterized by discrimination or prejudice based on gender, has a long history dating back to ancient civilizations[1] and continues to be a pervasive issue in contemporary society. The advent of digital content platforms has not only facilitated the expression of discriminatory attitudes or behaviors but has also become an arena for the propagation of sexist ideologies, extending its reach into virtual spaces. In particular, gender discrimination manifests in diverse and context-specific ways across critical domains, such as media and advertising, social media moderation, workplace equity, healthcare disparities, legal systems, and educational resources. In light of this, the imperative for effective and efficient methods to detect and address sexism has become increasingly important, encompassing both the digital realm and real-life scenarios.

A. Why This Survey?

While there are a bunch of surveys on hate speech detection [1], [2], [3], [4], [5], [6], [7] where sexism is a subcategory within, there is a lack of literature surveys focusing specifically on sexism detection [8], [9], [10]. This survey aims to fill the gap by providing an overview of the field of sexism detection.

Given the multifaceted feature of sexism, this literature survey begins by categorizing the various tasks associated with sexism detection, followed by a compilation of relevant resources. This survey evaluates the strengths and limitations of different sexism detection models and techniques by examining the adapted models and proposed methodologies for identifying sexist language, stereotypes, and discriminatory patterns in diverse contexts. It provides a clearer presentation of multimodal and multilingual sexism detection by offering well-organized comparisons, outlining challenges, and showcasing the latest evaluation techniques. Moreover, this survey goes further into the works of other disciplines than existing surveys of sexism detection, which mainly focus on social media, text modality, and strictly within the computer science discipline.

B. Scope of the Survey

Sexism, a deeply entrenched social issue, extends its tendrils into diverse spheres of human interaction, manifesting in nuanced ways across different scenarios such as social media,

[1]According to https://en.wikipedia.org/wiki/Sexism#History

workplaces, educational settings, and within the realm of entertainment and advertisements. The natural language processing (NLP) techniques necessitate adaptive and comprehensive detection methods for the identification and mitigation of sexism in these varied contexts. This survey embarks on a multidimensional exploration of sexism detection, encompassing a spectrum of languages and modalities (Fig. 5). Advancements in technology, particularly the rise of large language models (LLMs), have injected new possibilities into the field of sexism detection. This survey examines traditional sexism detection methods to recent LLM methods.

This survey not only engages with the latest research in artificial intelligence and NLP but also reaches across disciplinary boundaries. Particularly, we include the research of computing and society, which provides scenarios of gender discrimination and insights for sexism detection. By drawing connections to sociology and psychology studies, we strive to provide a multifaceted perspective that illuminates the complex social and psychological dynamics underlying gender discrimination.

C. Structure of the Survey

This survey is structured into 12 sections, organized into three parts. 1) Introduction part: I. Introduction, II. Sexism: Definition, Categories, and Scope, III. Survey Methodology; 2) Result part: IV. Tasks, V. Approaches, VI. Techniques, VII. Evaluation; and 3) Discussion part: VIII. Summary and Challenges, IX. Limitations, X. Research in Computing and Society, XI. Future Research and Potential Application, and XII. Conclusion.[2]

II. Sexism: Definition, Categories, and Scope

A. What is Sexism?

According to [11], sexism is defined as "individuals' attitudes, beliefs, and behaviors, and organizational, institutional, and cultural practices that either reflect negative evaluations of individuals based on their gender or support unequal status of women and men". Sexism can manifest in various contexts, including in language and culture, as observed in advertising, pornography, prostitution, media portrayals, and sexist jokes. In its most extreme forms, sexism can lead to sexual violence, including sexual harassment and rape.

Sexism is categorized in several ways, as shown in Fig. 1. Sexism can be.

Misogyny/Misandry: Misogyny is hatred, contempt, or prejudice against women or girls. On the contrary, misandry is against men or boys. Misogyny can perpetuate women's lower social status compared with men, thereby upholding patriarchal social roles. It often manifests through sexual harassment, coercion, psychological techniques aimed at controlling women, and the legal or social exclusion of women from full citizenship. Misandry, the inverse of misogyny, is commonly used as an accusation by men in the manosphere to counter feminist accusations of misogyny.

Hostile/Benevolent/Ambivalent: Sexist beliefs and behaviors that are overtly antagonistic are regarded as hostile sexism. Compared with hostile sexism, benevolent sexism is less obvious since it holds subjective and seemingly positive attitudes. For example, hostile sexism views women as *manipulative* and *deceitful*, while benevolent sexism frames women as *innocent* and *fragile*. Ambivalent sexism [12] is a compound of benevolent and hostile sexism.[3]

Institutional/Interpersonal/Internalized: Sexism operates on different levels in society. Institutional sexism is embedded within institutions and organizations, such as the education system or other workplaces. Interpersonal sexism manifests during interactions with others. Internalized sexism involves an individual's acceptance of sexist beliefs about themselves, such as self-deprecating "blonde jokes".

B. Related Concepts

To ensure a well-defined scope for this survey, this section explains terms that often overlap or co-occur with sexism, clarifying the boundaries between related issues.

1) Gender Bias: Gender bias refers to the systematic unequal treatment based on one's gender, such as wage discrimination and the gap in hiring. It also exists in languages.[4] According to [14], bias in computer systems has three categories: pre-existing, technical, and emergent bias. The preexisting bias (before the creation of the system) is the gender-biased input data originating from individuals, society, or historical context; the technical bias (at the time of creation or implementation) is the gender-biased inference due to the limitations of technical design; the emergent bias (when the system context has changed) is due to the changes in cultural values. In gender bias research, the research focus is commonly directed towards different systems, including language systems [13], [15], journals' peer review system [16], search engines and models [17], spanning from machine translation [18], pretrained models [19], [20], [21], [22], [23], LLMs [24], [25], [26], [27] to word presentation [28]. Conversely, in sexism detection, the focus shifts to individuals. This survey concentrates on the review of individual sexism.

2) Hate Speech: Hate Speech has varied meanings, and no single, consistent definition exists. It could be "intentionally promotes, justifies, or spreads exclusion, contempt, and devaluation of certain groups of the population through which these are humiliated or violated in their dignity in a discriminatory way" as translated by [29], or, a legal term in some countries, "communication that disparages a person or a group based on some characteristic such as race, color, ethnicity, gender, sexual

[2]Specifically, the RESULT part covers the tasks and corresponding resources used in sexism detection research (IV), the approaches adopted to tackle these tasks (V), the techniques applied to improve the performance (VI), and the evaluation research (VII). The DISCUSSION part covers a summary of the results and challenges posed by existing data and models (VIII), limitations of this survey (IX), related research in other disciplines (X), emerging trends and future directions (XI), and conclusion (XII).

[3]For instance, ambivalent sexists would hire someone for their attractive appearance but also fire them if they reject sexual advances.

[4]For example, the concept that the "prototypical human being is male" is ingrained in the structure of many languages. Specifically, syntactical rules are often structured in such a manner that feminine terms typically stem from their corresponding masculine forms [13].

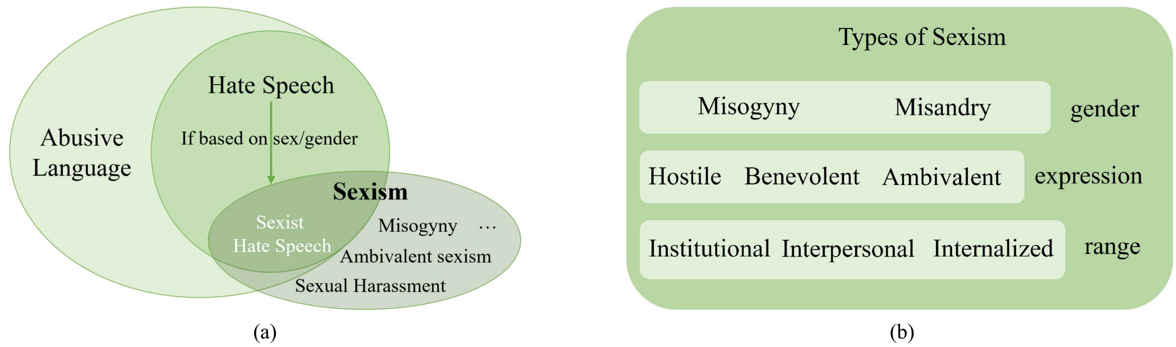


Fig. 1. Relationship of related concepts and divisions of sexism. (a) Related concepts. (b) Examples of different divisions of sexism.

orientation, nationality, religion, or other characteristics" as defined by [30] in the *Encyclopedia of the American Constitution*, or "public incitement to violence or hatred directed against a group of persons or a member of such a group defined based on race, color, descent, religion or belief, or national or ethnic origin" as defined by the *EU Council Framework Decision*. Extreme sexist speech is generally considered a subset of hate speech.

3) Toxic Speech/Abusive Language: This term encompasses a broader range of content than hate speech. It refers to any form of offensive or human rights-violating content, including but not limited to sexism, profanity, obscenity, hate speech, and more [31], [32].

III. Survey Methodology

To review the studies for sexism detection, this survey process started from gathering an initial set of papers from leading academic and research digital libraries, including ScienceDirect, Springer, IEEE Xplore, ACM Digital Library, and Google Scholar[5], focusing on publications between 2016 and January 2025. This involved a comprehensive keyword search, screening of abstracts, and full-text reviews to ensure all highly relevant studies were included.

The search queries for digital libraries: *Sexism* or *Misogyny* for title, abstract, and keyword (if applicable).

The search queries for Google Scholar: 1) *Hate speech detection*, *Sexism detection* for the title; and 2) *Gender bias sexism*, *audio sexism*, *speech sexism*, *image sexism*, *video sexism*, *multimodal sexism*, *multilingual sexism*, *social media sexism*, and *workplace sexism* are used for the whole article.[6]

We employed a snowball sampling approach to collect relevant papers. Starting with an initial set of papers, we iteratively searched for additional papers by examining the references of the previously collected papers. This process continued until no new eligible papers were identified, resulting in 527 unduplicated papers.

[5]Useful for searching research in other disciplines, such as social sciences and humanities, other than computer science and engineering.

[6]Google Scholar is used primarily for the Introduction and Terminology section. We added domain-specific qualifiers to the search string for the whole article to prioritize technically oriented papers within Google Scholar's interdisciplinary results. For example, Generic keywords (e.g., "sexism") were combined with technical terms (e.g., "multimodal") to narrow the results. This adjustment ensured alignment with the paper's focus on computational methods.

A. Eligibility Criteria

1) *Language:* Only papers written in English were considered. 2) *Publication Type:* Papers were primarily sourced from conference proceedings and journals, while degree theses were excluded. 3) *Accessibility:* Papers that were not accessible were excluded. 4) *Relevance and Impact:* Papers were selected based on their relevance, citation frequency, and contributions to the field of sexism detection. 5) *Substantive Content:* Papers that briefly mentioned sexism, such as in the introduction, related work, or future work sections, were excluded. After applying these criteria, a total of 135 eligible papers were selected for this survey.

IV. Tasks

In this section, we succinctly categorize the primary classification tasks in the sexism detection domain, provide lists of the associated datasets and open challenges dedicated to these tasks, and mention the explicitly proposed codebooks for data annotation.[7]

The mainstream tasks involving sexism detection are binary or fine-grained multi-label classification tasks, depending on how the relevant datasets are annotated.

1) Multilabel hate speech categorization, where sexism is labeled as a subcategory [33], [34], [35], [36], [37], [38].
2) Binary classification, where the data is labeled as either sexism/non-sexism [39], [40], [41], [42] or misogyny/nonmisogyny [43], [44], [45].
3) Multilabel categorization, where the data is classified into subcategories of sexism [46], [47], [48], [49], [50] or misogyny [51], [52], [53] based on specific divisions.

A. Resources

The related datasets and open challenges for sexism detection tasks, including those with other label aspects and hierarchical classification tasks, are listed in Tables I and II. An analysis of

[7]While most of the resource papers do not detail their annotation codebooks, a few papers specifically outline them.

TABLE I
LIST OF DATASETS. FOR RESOURCES THAT DO NOT HAVE SPECIFIC NAMES, WE PROVIDE A BRIEF DESCRIPTION OF THE TASK THEY PERFORM

Dataset	Ref.	Category	Language	Source	Mode	Size(K)	Year
H.S.D.	[34]	Sexism, Racism, NOT	en	T	T	16.9	2016
H.S.D.	[33]	Sexism, Racism, Neither, Both	en	T	T	6.9	2016
M.D.	[54]	Misogyny or NOT	en	T	T	4.3	2016
OFFCOMBR	[35]	Offensive or NOT	pt	BW	T	1.3	2017
		Sexism, Racism, Cursing. . .					
S.D	[46]	Benevolent, Hostile, Others	en	T	T	10.1	2017
S.D.	[39]	Sexism or NOT	en	F	T, I	-	2018
M.D.	[55]	Misogyny(5) or NOT	en	T	T	4.5	2018
M.D.	[52]	Misogyny(5) or NOT	es, en	T	T	8.1	2018
M.D.	[51]	Misogyny(5) or NOT	en, it	T	T	10.0	2018
H.C.	[56]	Harassment Type(3) or NOT	en	SafeCity	T	9.9	2018
H.C.	[57]	Harassment Type(5) or NOT	en	T	T	25.0	2018
MEME	[58]	Sexism or NOT, Aggressive, Ironic	en	F, T, I, R	T, I	0.8	2019
S.D.	[47]	Sexism(4) or NOT	en	T	T	3.1	2019
S.D.	[59]	Sexism(23) or NOT	en	ESP	T	13.0	2019
H.C.	[60]	Harassment Theme(3)	en	ESP	T	2.4	2019
H.S.D.	[61]	Hate (81) or NOT	pt	T	T	5.7	2019
MMHS150K	[36]	Sexism, Racism, . . ., NOT	en	T	T, I	150.0	2020
MeTwo	[62]	Sexism, Doubtful, NOT	es	T	T	3.6	2020
RUHSOLD	[63]	Sexism, Religious Hate, . . .	ur (Roman Urdu)	T	T	10	2020
S.D	[40]	Sexism or NOT	en	T	T	1.1	2020
H.C.	[64]	Harassment Theme(4), Retaliation	en	survey	T	2.4	2020
M.D.	[43]	Misogyny or NOT	en, hi, bn	F, T, Y	T	12.1	2020
Sent.C.	[65]	Sentiment(3) or NOT	ar	T, Y,	T	1.7	2020
S.D.	[66]	Sexist content(3) or NOT	fr	T	T	11.8	2020
S.D.	[37]	Sexism, Homophobia, . . .	ru	Y	T	-	2020
CallMeSexist	[41]	Sexism or NOT	en	T	T	3.8	2021
RP-Mod & RP-Crowd	[67]	Sexism, Racism, . . .	de	RP	T	85.0	2021
Let-Mi	[68]	Misogyny(7) or NOT	ar	T	T	6.6	2021
		Target(2)					
M.D.	[53]	Misogyny(4) or NOT(3)	en	R	T	6.6	2021
M.D.	[69]	Abusive, Misogyny(6), None	da	T, F, R	T	27.9	2021
Stereotype Classification	[70]	Gender stereotype(3)	fr	T	T	9.2	2021
Misogynistic-MEME	[45]	Misogyny or NOT	en	F, T, I, R	T, I	0.8	2022
		Aggressive or NOT					
		Ironic or NOT					
ArMIS	[44]	Misogyny or NOT	ar	T	T	1.0	2022
CoRoSeOf	[50]	Sexist content(3) or NOT(2)	ro	T	T	39.2	2022
SWSR	[48]	Sexism(3) or NOT	zh	Weibo	T	9.0	2022
		Target(2)					
S.D., M.D.	[71]	Sexism/Misogyny(10) or NOT	en	GitHub	T	10.0	2022
Challenge & Suggestion	[72]	Challenge(8) and Suggestion(6)	en	survey	T	0.1	2022
H.S.D.	[38]	Sexism, Racism, General Hate, . . .	ar	T	T	11.0	2022
LAHM	[73]	Sexism, Racism, . . .	en, hi, ar, fr, de, es	T	T	228.0	2023
EDOS	[49]	Sexism(4,11) or NOT	en	R, Gab	T	20.0	2023
SMSC	[74]	Sexism (3)	en	-	T, I	0.6	2023
		Emotional Reaction(3)					
S.D.	[42]	Sexism or NOT	en	Y	T	200.0	2023
GalMisoCorpus2023	[75]	Misogyny or NOT	gl	T, M	T	12.0	2024
MultiHate	[76]	Sexism or NOT	en	-	T	1,760.8	2024
S.D.	[77]	Sexism(5) or NOT	tr	T, Y	T	6.9	2024
S.D.	[78]	Sexism or NOT	en, es	TikTok	V	3.7	2024
		Source Intention(2)					
		Sexism Categorization(5)					
S.D., M.D.	[79]	Level of Sexism/Misogyny	de-at	news fora	T	8.0	2024
M.D.	[80]	Misogyny or NOT	en	movie	T	10.0	2024
		Misogyny Categorization(12)					
		Severity					
M.D.	[81]	Optimistic, Pessimistic, or Neutral	hi-en(code-mixed)	Y	T	12.7	2025
		Appreciation, Criticism, . . .					
BeyondGender	[82]	Sexism or NOT	en, zh	Y, Weibo	T	21.1	2025
		Gender (man or woman)					
		Phrasing (hostile or mild)					
		Misogyny or NOT					
		Misandry or NOT					

Note: H.S.D., hate speech detection; S.D., sexism detection; M.D., misogyny detection; H.C., harassment classification, and Sent.C., sentiment classification. Language is presented by two-letter lowercase abbreviations (ISO 639). Roman Urdu refers to the Urdu language written with the Latin script. Source is where the data are collected from Twitter, Facebook, YouTube, Instagram, Reddit, Mastodon, Rheinische post (RP), Brazilian Web (BW), and everyday sexism project (ESP). Some are collected by survey. Mode are text, image, and video.

TABLE II
LIST OF OPEN CHALLENGES

Open Challenges	Tasks	Language	Mode	Ref	#Data
SemEval-2019 Task 5	1 - Hate Speech or NOT 2 - Aggressive behavior and target classification	es, en	T	[83]	20K
SemEval-2022 Task 5	1 - Misogynous or NOT 2 - Misogyny categorization(4)	en	T	[84]	11K
SemEval-2023 Task 10	1 - Sexist or NOT 2 - Sexism categorization(4) 3 - Fine-grained sexism vectors(11)	en	T	[49]	20K
GermEval-2024	1 - Binarized and multiclass categorization 2 - Label distribution prediction	de	T	[85]	8K
Tamil-ACL 2022	1 - Abusive comment detection(7)	ta, ta-en	T	[86]	13K
AMI-IberEval 2018	1 - Misogyny identification(2) 2 - Misogynistic behavior categorization(5) 3 - Target classification(2)	es, en	T	[52]	8K
AMI-Evalita 2018	1 - Misogyny Identification (2) 2 - Misogynistic behavior categorization (5) 3 - Target classification (2)	it, en	T	[51]	10K
AMI-Evalita 2020	1 - Misogyny and aggressive behavior identification 2 - Unbiased misogyny identification	it	T	[87]	8K
EXIST-IberLEF 2021	1 - Sexist or NOT 2 - Sexism categorization (5)	es, en	T	[88]	11K
EXIST-IberLEF 2022	1 - Sexist or NOT 2 - Sexism categorization (5)	es, en	T	[89]	11K
EXIST-CLEF 2023	1 - Sexist or NOT 2 - Source intention (3) 3 - Sexism categorization (5)	es, en	T	[90]	10K
EXIST-CLEF 2024	1.1 - Sexist or NOT 1.2 - Source intention (3) 1.3 - Sexism categorization (5)	es, en	T	[91]	10K
	2.1 - Sexist or NOT in memes 2.2 - Source intention in memes (2) 2.3 - Sexism categorization in memes (5)	es, en	I	[91]	5K
EXIST-CLEF 2025	3.1 - Sexist or NOT in memes 3.2 - Source intention in memes (2) 3.3 - Sexism categorization in memes (5)	es, en	T, V	[92]	3K

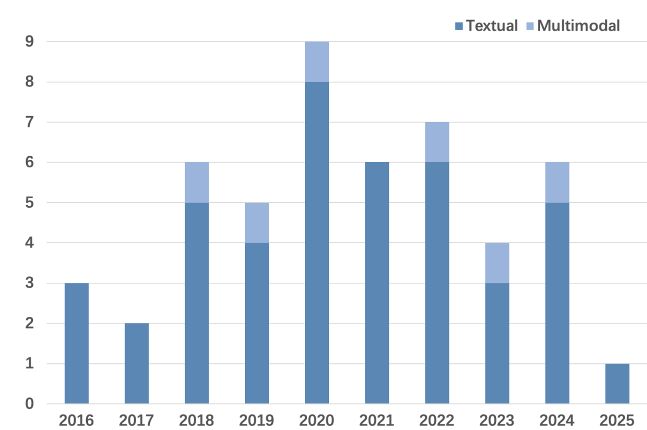


Fig. 2. Number of resource paper publications per year from 2016 to 2025 related to sexism detection in NLP.

these tables reveals two trends: 1) the modalities used in sexism detection have expanded from textual data to include images and, more recently, videos; and 2) hierarchical classification tasks are becoming increasingly common, with a growing number of languages being explored in recent years.

Figs. 2–4 provide visualizations of the publication statistics of modalities used in sexism detection resources, sources and task categories, and language distribution. Fig. 3 shows that Twitter is the primary data source, followed by Facebook, Reddit, and YouTube. Additionally, misogyny detection is frequently tackled as a separate task rather than a subcategory of sexism detection. Fig. 4 indicates that English is the predominant language used in sexism detection research, followed by Spanish.

While there are a few resources for multimodal tasks, significant deficiencies still exist, particularly concerning auditory elements. More platforms, such as live streaming and podcasting, should be explored for automatic sexism detection.

B. Codebooks

Samory et al. [41] aligned various dimensions of sexism with psychological scales measuring sexism and related constructs. Drawing from these scales, they formulated a codebook for detecting sexism on social media. This codebook was then applied to annotate both existing and newly created datasets, revealing their limitations in terms of breadth and validity concerning the concept of sexism. Having systematically reviewed the literature of 10 primary studies that characterized misogynistic and sexist texts in various domains, Sultana et al. [93] developed a rubric specifically designed to identify misogynistic remarks and sexist jokes within the software engineering domain. Similarly, Sultana [71] built a labeling rubric based on prior studies on sexist, misogynistic, and discriminatory detection.

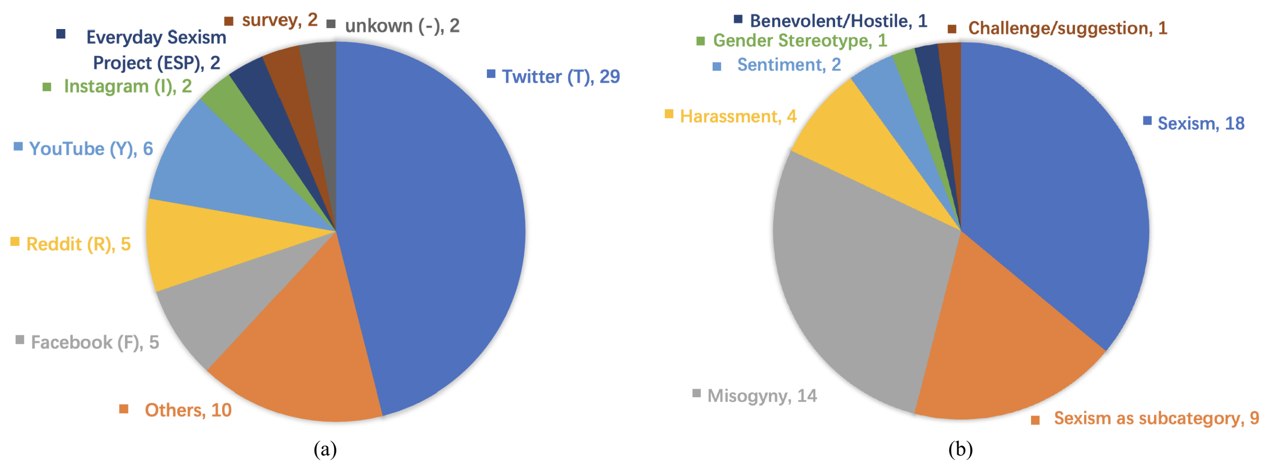


Fig. 3. Sources and label types of previous sexism detection datasets, from 2016 to 2025. (a) Data source. (b) Classification categories.

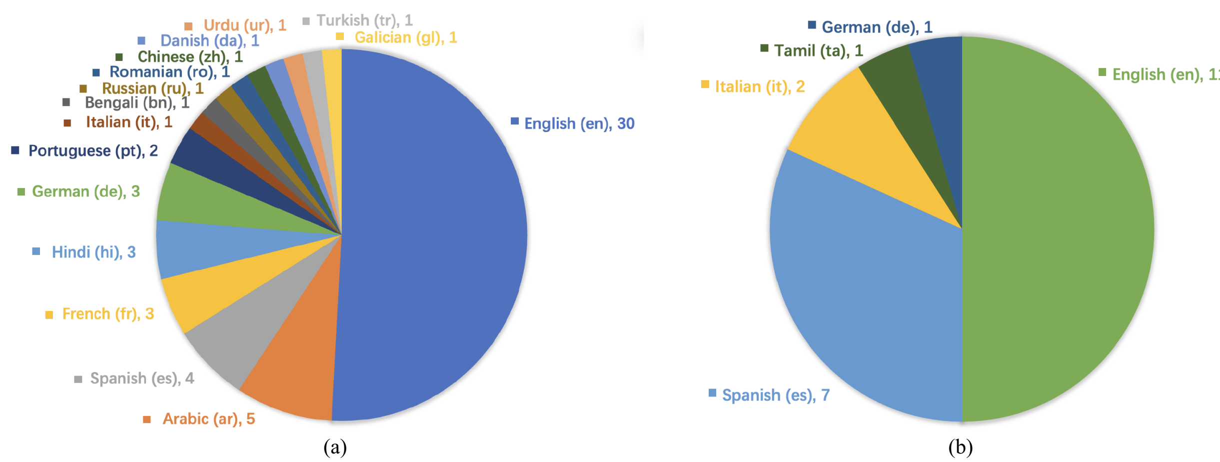


Fig. 4. Amount of previous sexism detection datasets and open challenges in different languages, from 2016 to 2025. (a) Datasets. (b) Open challenges.

V. APPROACHES

In this section, we present a roadmap of the approaches for sexism detection, which is organized into four categories according to the input features and model frameworks [Fig. 6(a)]: 1) statistical learning-based approach; 2) word embedding-based approach; 3) pretrained language models-based approach; and 4) LLMs-based approach. In each part, we start from textual modal to multimodal methods. For a comprehensive overview, we compile Tables III and IV for unimodal and multimodal approaches, respectively.

A. *Statistical Learning-Based Approach*

This approach involves extracting traditional statistical features from text data, such as TF-IDF [46], [94], word N-grams [35], [39], [55], and document-level statistics (e.g., sentence length, punctuation, etc.) [34]. These features are then used to train a classifier using classical machine learning (ML) algorithms.[8] Commonly used algorithms include support vector machines (SVM), logistic regression (LR), or random forests (RF), to train a classifier. Ensembles of these models have demonstrated significant success, as evidenced in the Evalita-2018 [51] and IberEval-2018 [52] shared tasks.

Waseem and Hovy [34] proposed a hate speech classification approach using an LR model with various feature sets. The most indicative features for their best model were character n-grams, while the inclusion of location or length had a negative impact. Anzovino et al. [55] used features such as part of speech, text embeddings, and n-grams with supervised classification models for misogynistic language identification and categorization. Mustapha et al. [94] used SVM, along with the TF-IDF feature, to detect the harassment towards women on Twitter during the Covid-19 pandemic.

1) Multimodal: To detect sexism in the memes, Fersini et al. [58] considered the bag-of-word model for textual feature representation and handcrafted visual features, taking low-level grayscale features, low-level colored features, photographic features, and semantic concepts-related features into account. They found that for unimodal classifiers, textual features are more informative in predicting sexist content. They also noticed that early fusion in multimodal approach is worse than either unimodal approach.

[8] Also denoted as ML in Tables III and IV.

TABLE III
METHODS FOR TEXTUAL DETECTION TASKS

Ref	Category	Models	Language	Task	#Data	Year
[34] (baseline)	ML	LR	en	Hate Speech DET.	16k	2016
[95]	ML, NN	CNN/LSTM + GloVe + GBT	en	Hate Speech DET.	16K	2017
[46] (baseline)	ML	SVM + TF-IDF,	en	Sexism DET.	10K	2017
	RNN	LSTM, FastText				
[35] (baseline)	ML	SVM/NB + n-grams	pt	Offisive Content CLA.	1K	2017
[56]	NN	CNN + LSTM	en	Harassment CLA.	10K	2018
[96]	ML, NN	SVM, NB, CNN, LSTM	en	Sexism DET.	-	2018
	DA					
[56]	NN	CNN + LSTM	en	Harassment CLA.	10K	2018
[55]	ML	RF, NB, MPNN, SVM	en	Misogyny DET.	4K	2018
[59]	PLM, NN	BERT + biLSTM	en	Sexism CLA.	13K	2019
	DA					
[47] (baseline)	NN, ML	CNN + LSTM + NB	en	Harassment CLA.	3K	2019
[61]	RNN	GloVe + LSTM	pt	Hate Speech DET.	6K	2019
[97]	NN	GloVe + CNN + GRU	en	Hate Speech DET.	25K	2019
[98]	TM	LDA	en	Sexism ANA.	79K	2019
[60]	TM	LDA	en	Sexism ANA.	2K	2019
[40] (baseline)	RNN	GloVe + LSTM	en	Sexism DET.	1K	2020
[66]	PLM	BERT	fr	Sexism DET.	12K	2020
[99]	PLM	BERT	en	Hate Speech DET.	16K	2020
	TL					
[63]	PLM, NN	BERT + CNN	RU	Hate Speech DET.	10k	2020
	TL					
[100]	PLM, NN	BERT + BiLSTM	en	Sexism CLA.	13K	2020
	DA, MTL	+ ELMo + GloVe				
[64]	TM	LDA	en	Sexism ANA.	2K	2020
[101]	RNN	GRU + multi-attention	en	Harassment DET.	11K	2020
[43]	ML	SVM	en, hi, bn	Misogyny DET.	12K	2020
[70]	PLM	SentenceBERT	fr	Sexism DET.	8K	2021
	DA					
[102]	NN	BERT + ELMo + GloVe	en	Sexism CLA.	13K	2021
		+CNN/RNN		Misogyny DET. & CLA.	5K	
[103]	PLM, NN	BERT + BiLSTM	en	Sexism CLA.	13K	2021
	DA					
[53]	ML, PLM	LR, BERT	en	Misogyny DET.	6K	2021
[104]	PLM	ByT5, TabNet	es, en	Sexism DET.	11K	2022
	Tab.L					
[38]	NN	LSTM, CNN+LSTM	ar	Hate Speech DET.	11K	2022
		GRU, CNN+GRU				
[105]	PLM	BERT, RoBERTa, DeBERTa	es, en	Sexism DET.	11K	2022
[71] (baseline)	PLM	BERT	en	Sexism DET.	10K	2022
				Misogyny DET.		
[48] (baseline)	PLM	BERT/RoBERTa + TF-IDF	zh	Sexism DET. & CLA.	9K	2022
[106]	LLM	ChatGPT	en	Content Moderation	-	2023
[107]	LLM	GPT-NeoX	es, en	Sexism DET.	11K	2023
		BERTIN-GPT-J-6B				
[108]	PLM	mBERT, XLM-RoBERTa	es, en	Sexism DET. & CLA.	10K	2023
[109]	DA	BERT, Word2Vec	en	Sexism DET.	378K	2023
[110]	LLM, PLM	ChatGPT + RoBERTa	en	Sexism DET.	32K	2023
	DA			Hate Speech DET.	71K	
[111]	PLM	BERT, BETO	es, en	Sexism DET.	11K	2023
	MTL					
[112]	PLM	BERT + Word2Vec + LR	en	Sexism DET.	20K	2023
[113]	LLM	ChatGPT	en	Harmful Content DET.	3K	2024
[114]	NN	BERT + BiLSTM	en	Sexism CLA.	13K	2024
	DA, MTL	+ ELMo + GloVe				
[75] (baseline)	ML	RF, SVM, linear SVM	gl	Misogyny DET.	12K	2024
[94]	ML	SVM + TF-IDF,	en	Harassment DET.	3K	2024
[76]	NN	CNN + LSTM, GPT2	en	Sexism DET.	1,760K	2024
[115]	PLM, LLM	RoBERTa, DeBERTa, Llama2	en	Sexism DET.	0.4K	2024
[116]	MTL	XLM-RoBERTa	en, it, hi, de	Misogyny DET.	10K	2024
[117]	PLM, LLM	RoBERTa, DeBERTa, Mistral	en, es	Sexism DET.	10K	2024
[80] (baseline)	PLM	BERT, RoBERTa, DeBERTa	en	Misogyny DET. & CLA.	10.0	2024
[118]	DA, PLM	RoBERTa, MarIA	en, es	Sexism DET. & CLA.	11.0	2024
[119]	RLHF, LLM	Mistral, Llama3	en	Sexism DET. & CLA.	20K	2025
[120]	ML, TM, PLM	TF-IDF + LDA + PLM	en	Misogyny DET.	6K	2025

Note: In Category, DA, data augmentation; TL, transfer learning; MTL, multitask learning; TM, text mining; Tab.L, tabular learning. In Models, NB, Naive Bayes; LR, logistic regression; RF, random forest; GBT, gradient boosted trees; SVM, support vector machine, MPNN, multilayer perceptron neural network. #Data only take datasets containing sexism-related labels into account.

TABLE IV
METHODS FOR MULTIMODAL DETECTION TASKS

Ref	Category	Models	Language	Task	#Data	Year
[39]	NN, ML	CNN + SVM/DT/NN	en	Sexism DET.	(I) 0.4K	2018
		+ Word2Vec, n-grams			(I,T) 0.2K	
[58]	ML	BOW + SVM/NB/DT/NN	en	Sexism DET.	0.8K	2019
[36]	NN	CNN + RNN	en	Hate Speech DET.	150K	2020
[121]	NN	VGG16	en	Misogyny DET.	0.8K	2021
		LSTM + USE				
		LSTM + Clarifai + USE				
[122]	PLM	Clarifai + USE, Visual-BERT	en	Misogyny DET.	11K	2023
[123]	PLM, MLLM	VisualBERT + CLIP + LSTM + Graph	en	Misogyny DET.	11K	2024
[78]	PLM, NN, ML	RoBERTa, Wav2Vec2, BLIP+TF-IDF, SVM	en	Sexism DET.	1.8K	2024
		BETO, MFCC, ViT+LSTM	es		1.9K	
[124]	PLM, MLLM	ViLT, CLIP	en	Sexism DET.	2.5K	2024
			es		2.5K	
[125]	PLM, MLLM	BERT, CLIP, MASK RCNN	en	Misogyny DET.	11K	2025
				Sexism DET.	13.5K	

Note: DA, Data augmentation; Trans.L, transfer learning; Tab.L, tabular learning. In models, RF, random forest, NB, Naive Bayes; NN, nearest neighbour; LR, logistic regression; GBT, gradient boosted trees; SVM, support vector machine; MPNN, multilayer perceptron neural network. Only take datasets containing sexism-related labels into account.

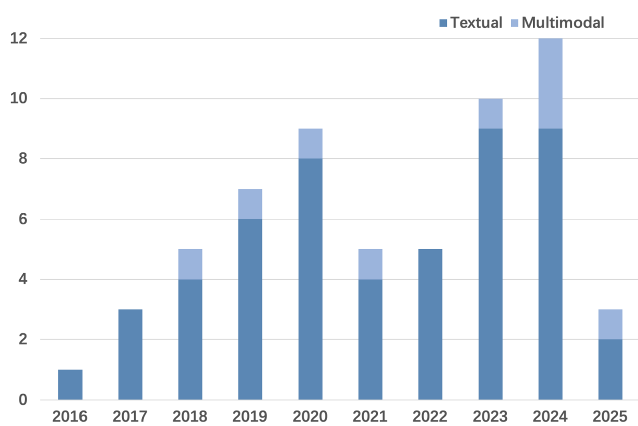


Fig. 5. Number of textual modal approaches and multimodal approaches related to sexism detection in NLP, from 2016 to 2025.

B. Word Embedding-Based Approach

This approach involves utilizing pretrained word embeddings, such as Word2Vec [39], [112], GloVe [40], [61], or FastText [46] to convert words into dense vector representations that capture semantic meaning. These word embeddings can be aggregated to obtain a representation for entire sentences or documents, typically through averaging or weighted summation, resulting in a fixed-size vector representation of the text. These sentence vectors can then be used as input to a neural network (NN).[9] Various neural network architectures can be employed, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs). RNNs include gated recurrent unit (GRU), long short-term memory (LSTM), and bidirectional LSTM (Bi-LSTM).

LSTM is the most popular choice [38], [40], [61]. Usually, ensembles of classical ML models and NN models yield improved results [97]. Karlekar and Bansal [56] applied a CNN-RNN model for single-label classification and employed CNN-based character embeddings along with bidirectional RNNs for multi-label classification, utilizing data from SafeCity's online forum where users share their experiences of harassment and abuse. Karatsalos and Panagiotakis [101] proposed a multiattention approach based on RNN to categorize online harassment, incorporating back-translation to address data imbalance, as sexual harassment was more prevalent than other subcategories. Al-Hassan and Al-Dossari [38] compared the performance of four deep learning models for Arabic hate speech detection, finding that adding a CNN layer to LSTM or GRU improved performance.

1) Multimodal: Gasparini et al. [39] extracted visual features in advertisements using CNNs and applied traditional MLM classifiers. Gomez et al. [36] explored three strategies to integrate textual and visual information, but their multimodal models did not outperform text-only models. Fersini et al. [121] addressed misogynistic content detection in memes [45] using a multimodal approach, combining VGG-16 for the visual component and LSTM with universal sentence encoder (USE) embeddings for the textual component. This approach outperformed both unimodal classifiers and the VisualBERT model, achieving state-of-the-art results.

C. Pretrained Language Models-Based Approach

Pretrained language models (PLMs)[10] are generally referred to as Transformer-based neural network models. Representative encoders include BERT, RoBERTa, and DeBERTa, while representative decoders are BART and T5.

Parikh et al. [102] developed a framework that integrates BERT sentence representations with ELMo word embeddings and linguistic features from CNN or RNN. Their methods outperformed traditional machine learning baselines in tasks such as sexism [59] and misogyny classification [51]. Younus and Qureshi [104] pointed out that existing deep learning methods

[9] Also denoted as NN in Tables III and IV.

[10] Also denoted as PLM in Tables III and IV.

often overlook platform- or language-specific idiosyncrasies when building classifiers. They proposed a framework combining the token-free ByT5 model[11] and the attention-based TabNet[12], integrating language and platform dependencies. This framework can effectively handle both numeric and categorical data. Das et al. [112] improved classification performance on the SemEval-2023 dataset by integrating user gender information encoded by Word2Vec with textual features from Sentence-BERT. The gender information was predicted by a classifier trained on a gender prediction dataset.

For French sexism detection, Chiril et al. [66] compared several models and found that BERT yielded the best results. In Chinese sexism classification, Jiang et al. [48] presented benchmark results using BERT-based models outperforming other approaches, particularly when leveraging lexicons.s For multilingual sexism detection, Vaca-Serrano [105] reviewed PLMs pre-trained in English and Spanish, determining the best-performing models for low-text volume tasks. An ensembling strategy was adopted to reduce biased predictions and achieve the highest performance in the EXIST 2022 competition. Similarly, de Paula et al. [108] utilized multilingual BERT and RoBERTa for the EXIST 2023 challenges, achieving top positions in multiple tasks.

1) Multimodal: Rizzi et al. [122] explored both unimodal and multimodal approaches for misogynous meme detection. Reformatting as a textual classification, unimodal methods included textual transcription via OCR, image tags from the Clarifai API, and captions generated by the visual vocabulary, all encoded by USE or Text-BERT. For multimodal methods, they used Visual-BERT with the early fusion of transcription-tags and transcription-captions. Their results showed that while textual features were crucial, multimodal approaches were necessary for effective detection. In addition, they introduced multimodal bias estimation to address distortion from biased elements[13] in memes, using Bayesian Optimization to mitigate it. Arcos and Rosso [78] examined three modalities: text (PLMs embeddings of transcriptions, OCR, and titles), audio (MFCCs or Wav2Vec2 embeddings), and video (ResNet or ViT, temporally modeled by LSTM or captioned by BLIP). Their results showed a 4.4%–4.8% improvement in multimodal performance over unimodal models for Task 2, while the textual model outperformed multimodal model in Tasks 1 and 3.

D. LLMs-Based Approach

LLMs[14] have a similar architecture to generative PLMs but with a larger model scale and parameter count. They are typically used for text generation, text understanding, and various text-related tasks. Multimodal LLMs (MLLMs)[15] are designed to process and generate data across multiple modalities, such as text, images, and audio. This capability allows them to understand and create content that combines different types of information, enhancing their versatility in applications such as image captioning, video analysis, and interactive chatbots. Notable examples of multimodal LLMs include OpenAI's CLIP, which connects vision (images) and language (text) for improved understanding.

Li et al. [113] explored generative AI's role in detecting harmful content on social media, showing that ChatGPT can match human accuracy (80%) in annotating toxic, offensive, and hateful content, although performance is prompt-dependent. Franco et al. [106] integrated LLMs into content moderation pipelines to address biases against minorities and vulnerable users. Their model showed promising results in analyzing sex-related and gender stereotypes, benefiting particular minority users. Tian et al. [107] employed two GPT-based LLMs with ensembling and cascading strategies. The first LLM was utilized to predict the sexism label. Subsequently, a confidence checker is employed to differentiate between hard and easy samples. The hard samples are then assigned to the latter LLM. They achieved the highest F1 scores in the EXIST 2023 challenge by fine-tuning models on hate speech datasets. Abercrombie et al. [115] examined the correlation between annotator demographics and gender-based violence annotations, finding that LLMs performed worse than tailored RoBERTa on sexism detection tasks. Khan et al. [117] proposed two fusion approaches for sexism identification in EXIST-2024 [91], using a dual-transformer network (DTFN) and ensembling outputs from PLMs, LLMs, and DTFN, ranking No.1 in English and No.4 in both English and Spanish. Riahi Samani et al. [119] proposed a reinforcement learning from human feedback (RLHF) fine-tuning framework for sexism detection, leveraging LLMs' contextual learning to provide clear insights into why certain content is flagged as problematic. Results with Mistral-7B and LLaMA-3-8B models highlighted the importance of RLHF in building explainable systems for online discourse, enabling more transparent and effective sexism detection.

1) Multimodal: Barua et al. [124] addressed the meme classification (Task4-6) of EXIST-2024 [91] using a five-component model architecture: 1) ViLT (pretrained vision-language model) to generate image-aware text representation and text-aware image representation for image-text pairs; 2) semantics representations of the memes pooled by the ViLT model; 3) attention-enhanced context vectors based on the significance of tokens and patches, respectively; 4) modality fusion achieved by concatenating the vectors of each modality; and 5) logits classification. Their approach outperformed multimodal models (CLIP and ViLT) by at least 7% and 6% in multiclass and multilabel classification tasks, respectively.

In addition, Kumari et al. [123] proposed a CTXSGM-Net framework to mitigate the unintended bias from meme classifiers, including three modules: an unbiased scene graph, VisualBERT, and a memory network using CLIP and LSTM. The contextual information is obtained by a CLIP-LSTM-based memory network. On the other hand, the unbiased semantic relationships between objects in memes are captured by the unbiased scene graph module. They trained with

[11]ByT5 is a pretrained byte-to-byte model.
[12]TabNet is designed for attentive interpretable tabular learning.
[13]Resulting in certain features being strongly and misleadingly associated with the target classes.
[14]Also denoted as LLM in Tables III.
[15]Also denoted as MLLM in Tables IV.

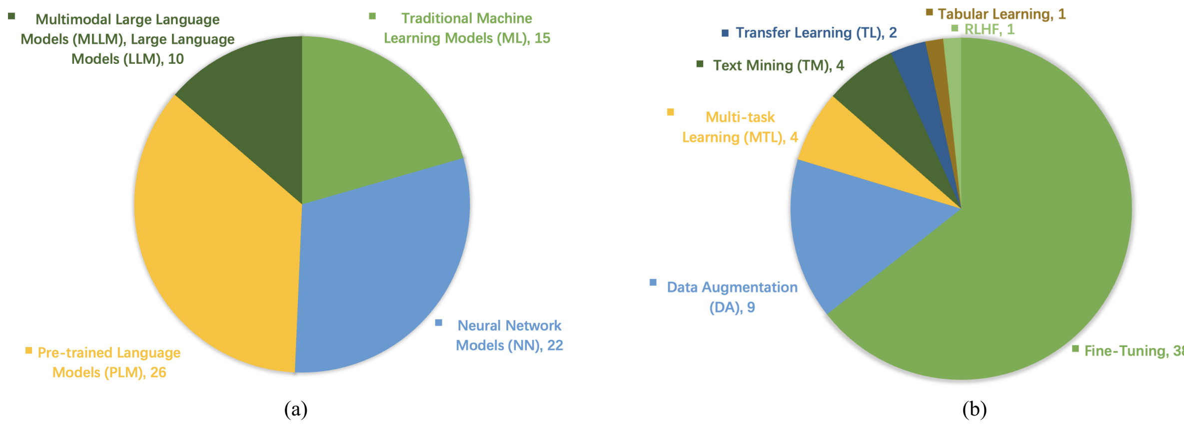


Fig. 6. Models and techniques used in approach papers, from 2016 to 2025. (a) Models. (b) Techniques.

supervised contrastive learning and cross-entropy loss jointly to improve multimodal representations. Their model outperformed the SOTA model in the task of SemEval-2022 Task 5 and showed efficacy across a few benchmark meme datasets. Rehman et al. [125] utilized an adaptive gating-based multimodal context-aware attention mechanism to selectively focus on pertinent visual and textual information, thereby generating contextually relevant features. Additionally, we utilized a graph neural network to reconstruct unimodal features and a context-aware attention module to provide multimodal features. Various feature extraction techniques were incorporated.

From the model perspective, as models evolve, newer architectures generally demonstrate better performance. However, for specific tasks or languages, some older models remain valuable. Some research underscored the value of monolingual and lighter models in the nuanced field of language-specific detection tasks [115], [126], offering insights into their competitive edge against LLMs such as ChatGPT and Llama. Moreover, there has been relatively little focus on multimodal sexism detection compared with the more established work in multimodal hate speech detection [127].

From the modality perspective, current studies suggest that text remains the dominant modality. Visual content, such as memes and advertisement posters, often perpetuates sexism through stereotypical imagery, objectification, or contextual juxtaposition with text. While early work relied on handcrafted features such as grayscale and color features, recent studies have applied CNNs, Visual-BERT, and ViT to extract features from visual data. On the other hand, audio content can convey sexism through interruptions, tone-based microaggressions, and paralinguistic cues such as sarcastic tone or dismissive laughter. However, there is a significant research gap in utilizing acoustic features due to the lack of speech corpora. Combining text, visual, and audio data can potentially disambiguate context and provide a more comprehensive understanding of sexism, but it also introduces complexities in integrating multiple modalities. Despite efforts to fuse multiple modalities, challenges such as modal alignment and noise remain. Addressing these challenges is crucial to developing effective multimodal sexism detection systems.

VI. Techniques

To gain a deeper understanding of the practical applications of the aforementioned approaches in sexism detection, we now explore key techniques. This section provides a detailed discussion of methods such as data augmentation, transfer learning, multi-task learning, and text mining [Fig. 6(b)]. These techniques enhance model performance and improve its adaptability and generalization in complex scenarios.

A. Data Augmentation

Data augmentation techniques (DA) are methods that use existing data to create new data samples that can improve model optimization and generalizability.

Parikh et al. [59] infused domain-specific features by fine-tuning BERT with unlabeled accounts of sexism. Their top-performing method surpasses traditional ML baselines and several deep learning baselines significantly. Although the dataset contains labels for 23 classes, they only consider 14 sexism categories by merging certain categories. Later, [103] introduced a multi-level training approach with a self-training strategy to address the 23-category classification task. The self-training strategy iteratively augmented the original labeled set by incorporating pseudo-labeled accounts, selecting those predictions with high confidence of correctness. The augmented data were finally utilized to train the multilabel classifier. The multi-level method involved sequentially training models at different categorization levels to mitigate class imbalances, beginning with reduced category sets. Sharifirad et al. [96] utilized knowledge sources (ConceptNet and Wikidata) and applied cosine similarity for word mapping between the two knowledge sources to augment data for a small-sized sexism detection dataset. Chiril et al. [70] studied the impact of gender stereotypes on sexism detection. They annotated a dataset for gender stereotype detection and augmented it based on sentence similarity to train a gender stereotype detector. Further, they detect sexism as an auxiliary task and found that multiclass gender stereotypes detection benefits sexism classification.

To mitigate vocabulary differences, which caused the performance gap across subtypes of sexism, Rodríguez-Sánchez et al. [118] increased the amount of data for minority classes or those with high heterogeneity. To mitigate the extent of class imbalance in hate speech detection, Rizos et al. [97] proposed three data augmentation techniques: 1) embedding-based synonym replacement; 2) word tokens shifting and warping; and 3) class-conditional sentence generation. They further demonstrate the generalization properties of the augmentation techniques by applying them to various architectures and testing on different hate speech datasets. Additionally, Sen et al. [110] utilized Counterfactually Augmented Data (CADs)[16] to enhance model robustness for detecting harmful language in out-of-domain contexts. They explored the feasibility of automating this task using generative NLP models, as manually creating CADs is laborious and costly. They used the model Chat-GPT, Polyjuice, and Flan-T5 to generate CADs and assess their effectiveness in enhancing model resilience compared with manually crafted CADs. Results from various out-of-domain test sets indicated that manually crafted CADs remained the most effective, closely followed by CADs generated by Chat-GPT.

To offer insights into sexist language usage within a highly influential aspect of popular culture, Betti et al. [109] analyzed lyrics from over 377K songs in the WASABI database, which contains two million songs [128]. They examined the manifestation of sexism over five decades (1960–2010) and quantified gender biases. The sexism classifier, utilizing the dataset and code from [41], identified sexist lyrics on a larger scale than prior studies, which were limited to small samples. Their findings revealed a rise in sexist content over time, especially in popular songs and from male artists. Moreover, songs exhibit varying language biases depending on the genders of singers, with songs by male solo artists displaying more pronounced biases.

B. Transfer Learning

Transfer learning (TL) is a technique where a model trained on one task is reused for another related task, to boost performance on the related task.

Mozafari et al. [99] analyzed the contextual information extracted from BERT's pretrained layers and then fine-tuned it with four strategies: 1) BERT-based fine-tuning; 2) adding nonlinear layers before the final activation function; 3) adding a Bi-LSTM layer to process all the outputs of the latest transformer encoder before the final activation function; and 4) adding a CNN layer to process the matrix of the output vectors from each transformer encoder. The CNN-based fine-tuning strategy surpassed previous works by capturing syntactical and contextual information embedded across various transformer encoder layers. Additionally, their model can spot certain biases that may arise during the data collection or annotation.

Rizwan et al. [63] introduced a CNN-gram architecture that leveraged n-gram information to learn specific patterns from text efficiently. Additionally, they trained domain-specific embeddings with PLMs on over 4.7 million tweets in Roman Urdu. The results indicated that BERT demonstrated superior performance in domain adaptation and transfer learning.

C. Multitask Learning

Some research adopts multitask learning (MTL) with data in the same domain to develop more robust and effective models by leveraging shared information and enhancing generalization performance.

To perform 23-category sexism classification such that the categories can co-occur, Abburi et al. [100] proposed an MTL approach involving topic proportion distribution estimation, cluster label prediction, and sexism detection tasks. They utilized unlabeled data from the same domain for estimation and clustering through unsupervised learning and employed weakly-labeled negative data from another corpus. Additionally, they explicitly leveraged the cooccurrences of multilabels in the training data [59]. Later, Abburi et al. [114] introduced a knowledge-based cascaded multitask framework involving several tasks. For homogeneous tasks, they utilized intradomain data and designed the same tasks as [100]. For heterogeneous tasks, they leveraged cross-domain data for emotion classification and sarcasm detection, considering that accounts of sexism may exhibit sarcasm and emotion. A knowledge module was employed to generate external representations for domain-specific keywords. They achieved SOTA performance by training the model with all auxiliary tasks.

To cope with the constantly evolving form and targets of abusive content, Hangya and Fraser [116] proposed a two-step approach to build models economically for new target/language, leveraging existing datasets related to the target domain. Their model was first trained in a multitasking fashion and then performed the target task with few-shot adaptation. The model acquired a general understanding of abusive language and achieved better performance in both monolingual and cross-lingual setups.

MTL is also a solution for training robust models when data are scarce or costly to obtain, as it enables information sharing between tasks to improve performance across multiple related tasks simultaneously. However, negative transfer remains a challenge in MTL, where the sharing of noisy information can degrade performance. De Paula et al. [111] introduced a novel method to alleviate the negative transfer problem by leveraging the task awareness concept. It was implemented in two unified architectures where task-aware input and task embedding are added before and after the encoder. For detecting toxic language, hate speech, and sexism, the proposed method effectively reduced negative transfer compared with traditional MTL methods, achieving SOTA performance on the EXIST-2021 benchmark [88].

D. Text Mining

Text mining (TM) is the practice of analyzing vast collections of textual materials to capture key concepts, trends, and hidden relationships.

[16]CADs make slight alterations to existing training data points and invert their labels, potentially reducing the model's reliance on spurious features when trained on them.

Melville et al. [98] employed topic modeling to reveal the most prominent manifestations of sexism by analyzing 79K posts from the ESP. In the low-resolution picture (with seven topics), they observed a significant link between public space/street harassment and domestic abuse/sexism in personal relationships. In the high-resolution picture (with 20 topics), for instance, they observed a layering of experiences of sexism in public spheres such as work and education, atop the sexism experienced at home. Moreover, they noted the evident occurrence of sexism in learning environments for young women.

Similarly, Karami et al. [60] adopted topic modeling to disclose the hidden topic in their collected data. Specifically, they applied latent Dirichlet allocation (LDA) to mine the topics and themes related to workplace sexism and sexual harassment reported on the ESP's website. For further topic analysis, they used thematic analysis to interpret the themes' conceptual meanings. The subsequent study [64] applied LDA to mine the topic and manually coded the themes related to sexual harassment in academia using web survey data. The themes identified in the data align with existing literature on sexual harassment, including sexual coercion, gender harassment, sex discrimination, and unwanted sexual attention. Additionally, the theme of retaliation emerged in instances where individuals experienced bullying or threats for reporting harassment or resisting the harassers.

From the technical perspective, one training approach involves designing effective features using existing data. Another training approach is to increase the training data size, which can be achieved through data augmentation (generating new data) or transfer learning (leveraging the knowledge learned from related tasks); data augmentation and multitask learning often complement each other to enhance performance. For text mining applications, LDA is the primary method employed. In addition, current methods for multimodal sexism detection are primarily text-centered, which may overlook the potential contributions of visual and auditory elements.

VII. Evaluation

A. Metrics for Evaluation

Most datasets and models evaluate performance using metrics such as accuracy, precision, recall, F1-score, and ROC.

B. Model Evaluation

1) Generalizability: Samory et al. [41] leveraged their annotated dataset, *CallMeSexist*, to generate adversarial examples with the help of crowdworkers. They employed these examples to assess the reliability of sexism detection methods. The findings revealed that existing MLMs identify only a narrow range of linguistic indicators for sexism, displaying poor generalization to out-of-domain data. However, by incorporating adversarial and varied samples during the training phase, models exhibited improved generalization and increased robustness. Although CAD is constructed to enhance out-of-domain generalizability, Sen et al. [129] found that models trained on CAD exhibit higher false positive rates compared with those trained on the original dataset. They tested BERT and LR models for sexism and hate speech detection with CAD that contained gendered and identity terms in nonsexist and nonhateful contexts. They also found that using a diverse set of CAD helps mitigate unintended bias (Table V).

Compared with LLMs, MLMs generally struggle with out-of-domain cases and often require additional training techniques, such as data augmentation and adversarial training, to improve generalization. While more versatile across domains, LLMs tend to exhibit more severe biases inherited from their pretraining data. It amplifies the need for fine-tuning and bias mitigation to handle nuanced sexism detection tasks.

2) Interpretability and Bias: Mohammadi et al. [132] introduced a novel approach, combining BERT architectures with a CNN framework, to enhance model interpretability in sexism detection at a granular level. By leveraging Shapley additive explanations (SHAP) values, they identified the most important terms contributing to sexist content and assigned Sexism Scores to specific parts of a sentence. This approach provided a deeper understanding of the model's decision-making process, enabling decision-makers and researchers to understand how the model arrives at its predictions.

Muntasir and Noor [131] employed the local interpretable model-agnostic explanations (LIME, explainable AI) technique to identify the most relied-on word features that contributed to the transformer-based models' predictions. The results revealed that the model exhibits a significant bias in its predictions, highlighting its inability to recognize sexism in gender-swapped sexist sentences.

This issue arises from imbalanced datasets, which are often skewed towards women and lack sufficient examples of men, resulting in biased models that haven't seen enough examples from underrepresented groups. Similar biases have been revealed in other studies, including exacerbated gender bias due to larger model sizes or greater alignment in LLMs [133] and sexual objectification in language-vision AI models [134]. The consequences of such biases can be severe, particularly in social media content moderation, where biased models can perpetuate gender biases and unfair treatment. It emphasizes the need for explainable AI approaches to ensure fair and transparent decision-making.

3) Cross-Lingual: Yadav et al. [73] employed LAHM to assess SOTA multilingual and MTL methodologies in different classification settings: monolingual, cross-lingual, and machine translation tasks. For monolingual experiments, BERT-based language-specific hate speech models were utilized. For cross-lingual, mBERT was utilized to perform few-shot binary classification experiments. Results showed that mBERT performed much better in English than in other languages. They adopted machine translation to convert English data into multiple languages, fine-tuning mBERT to improve the overall performance in several languages. Following the multilingual HateCheck [135] framework, Das et al. [130] evaluated the effectiveness of ChatGPT across eleven languages. They observed that while ChatGPT excels in detecting hateful posts, it misclassified nonhateful counter-speeches as hate speech. Moreover, its

TABLE V
METHODS FOR MODEL EVALUATION

Ref	Method	Models	Language	Task	#Data
[73]	Machine Translation	BERT, APIs	en, hi, ar, fr, de, es	Hate Speech DET.	300K
[129]	Data Augmentation	LR, BERT	en	Sexism DET.	6K
				Hate Speech DET.	28K
[41]	Data Augmentation	(manual)	en	Sexism DET.	4K
[130]	Functionality Tests	ChatGPT	11 langs	Hate Speech DET.	40K
[131]	LIME	BERT, RoBERTa, DistilBERT, SqueezeBERT	en	Sexism DET. & CLA.	20K
[132]	SHAP	BERT + CNN	en	Sexism DET. & CLA.	20K

Note: In models, LR, logistic regression. #Data only takes datasets containing sexism-related labels into account.

proficiency in distinguishing between nonprotected and protected target groups was more effective for English than for other languages. Regarding emoji-based hate speech, it performed inadequately, particularly when positive emojis are employed in hateful posts.

Under multilingual settings, the dominant approach indeed involves pretraining a PLM on multiple languages and then fine-tuning it on a specific task using a training set or augmented data. Moreover, analysis by [118] indicated that monolingual models may achieve comparable or even superior performance to multilingual models when trained on a similar scale of data. An alternative approach may be incorporating more linguistic insights, such as transfer learning tailored to language characteristics, leveraging unique properties and structures.

VIII. Summary and Challenges

A. Tasks and Resources

The majority of datasets and open challenges have been introduced for social media moderation, as evidenced by the data source and aims of ongoing open challenges. Since 2020, there has been an increasing exploration of languages beyond English, although English remains the dominant language due to its global prevalence. However, resources for multimodal detection were scarce before 2024, with the size of datasets being fewer than a thousand samples.

B. Model Evolution

As detection architectures advance, newer models generally outperform predecessors, though task- and language-specific contexts reveal exceptions. Monolingual or lightweight models (e.g., fine-tuned BERT variants) demonstrate competitive efficacy over LLMs such as ChatGPT in nuanced, language-specific tasks.

C. Modality Utility

Multimodal sexism detection methods rely on fusion strategies (early, late, attention-based). Text is the dominant modality compared with the visual modality, with NLP techniques such as transformer models excelling at detecting overt and covert linguistic patterns. Audio is understudied due to scarce speech corpora.

D. Training Paradigms

Data Augmentation with synthetic data is the most common technique to improve models' performance. Multilingual detection has seen the dominance of multilingual PLMs such as mBERT, although monolingual models can rival their performance with sufficient data, highlighting the potential for tailored linguistic insights, such as morphology-aware transfer learning, to enhance detection capabilities.

E. Limitations of Available Data

1) *Language Disparities*: Main languages (such as English and Spanish) have been collected, but data for minor languages is scarce, impacting the model's generalization across diverse linguistic contexts. 2) *Biased Data Source*: The collection of offline sexist comments in daily life poses challenges, limiting dataset diversity and potentially leading to biases in trained models.

F. Annotation Challenges

1) *Culture-Dependent*: Annotating sexist content can be culture-dependent, limiting the datasets' applicability to specific cultural backgrounds. 2) *Exposure to Toxic Content*: The annotation process may expose annotators to toxic content, raising ethical concerns regarding their well-being and mental health. 3) *Rely on Well-Trained Annotators*: Besides hostile sexism, identifying benevolent sexism is challenging. Annotators need to be well-trained and familiar with various forms of sexism to build a high-quality dataset.

G. Model's Generalizability

Existing models face challenges in effectively adapting to new, unseen scenarios. Furthermore, LLMs exhibit sensitivity to prompts, leading to inconsistencies in predictions. Addressing these issues is crucial for enhancing the generalizability of models for practical utility. Techniques such as cross-validation during training, using diverse datasets, and employing robust evaluation metrics contribute to a model's ability to generalize across various conditions.

H. Model's Interpretability and Reliability

Current models lack intrinsic transparency and often rely on superficial patterns, making it difficult to audit why a statement is flagged as sexist. This reduces trust in model outputs.

On the other hand, explanations from post-hoc interpretability tools (e.g., LIME, SHAP) often prioritize keyword-based rationales, overlooking nuanced context. Moreover, models exhibit inconsistent performance across linguistic and social groups. Addressing biases in the training data and regularly updating models with new and relevant data are key practices for enhancing reliability over time.

IX. Limitations

First, our literature search was limited by the maximum displayed results on Google Scholar, which may have resulted in some relevant studies being missed. To mitigate this, we included other prominent digital libraries in computer science to enrich our paper collection. Although the highly relevant research is top-ranked, some good studies might have been overlooked. This limitation may affect the comprehensiveness of our survey and potentially introduce bias into our results.

Furthermore, this survey primarily focuses on datasets with sexism-related labels and methods. It lacks a comprehensive evaluation of sexism detection approaches across various languages and universal datasets, as the authors have not released their code, posing challenges for reimplementing and reproducing the results. Additionally, it does not evaluate approaches for similar tasks [136]. This limitation may restrict the applicability of our findings to other related research areas.

Finally, the rapid evolution of LLMs and the time lag between submission and publication of conference and journal papers mean that our survey may not reflect the most latest developments in the field. Some of the challenges and limitations identified in our survey have likely been addressed in recent studies, which may not be included in our analysis.

X. Research in Computing and Society

To inspire interdisciplinary research aimed at developing more robust and inclusive solutions for effectively combating sexism, this section highlights key advances in detecting sexism from Computing and Society. Many studies have dedicated to contributing valuable insights into fostering a more inclusive and supportive environment, including in male-dominated realms such as computer game culture [137], the music industry [138], [139], [140], and areas such as military conscription and intimate partner violence [141], mother-blaming of prisoners' [142], political election [143], and children's education [144].

A. Online Sexism and Solution

Sexist content spreading on media platforms negatively impacts users' psychological well-being. Nakandala et al. [145] analyzed over one billion chat messages of 200 female and 200 male streamers from Twitch, revealing the prevalence of gendered conversation and objectification. Similarly, research [146] case studied the harassment experiences of 25 women and LGBTQ Twitch live streamers; [147] interviewed 13 women live streamers facing gender stereotypes and misogyny; [148] examined the emotional labor of women live streamers. Sasse and Grossklags [149] suggested that making sexist content invisible or visible counterspeech can contribute to a sense of safety for both men and women users. Although there is limited research on automatic sexism detection in live streaming, several studies have focused on moderation in this context [150], [151].

1) Sexism in Workplace: Grosz and Conde-Cespedes [40] pointed out that the anonymity of social media leads to a more aggressive and "hostile" version of sexism, which is easier to detect with clue words. To solve real-life cases, they presented sexist statements that are likely to appear in the workplace. Jaijee et al. [152] examined the frequency of sexism experienced by male and female cardiologists and explored the different types of sexism encountered in the field of cardiology. Trinkenreich et al. [72] surveyed 94 women working in a global technology company to investigate the challenges that women encountered in software development teams. Specifically, the study figured out eight factors that encouraged women to leave their jobs and proposed six strategies to mitigate the identified obstacles. Dray and Sabat [153] investigated the common differences in confrontations of workplace sexism and implications.

2) Sexism in Education: Tang et al. [154] investigated gender inequalities in researchers' knowledge status and the division of female labor in science and scientific research. The findings indicated that women tend to be more engaged in topics characterized by lower levels of knowledge, and they are of less assistance. Therefore, the authors emphasized the importance of addressing the knowledge gap within the scientific community and advocated for initiatives that encourage women to contribute to unexplored topics and areas. Biurrun-Garrido et al. [155] filled the gap in clinical nursing settings using online questionnaires. They found that everyday sexism was perceived within the nursing school, and since it did not occur in practicums, care settings, or during classroom teaching, nursing students found it challenging to consciously be aware of these behaviors.

XI. Future Research and Potential Application

The complex nature of sexism necessitates a multifaceted approach to detection and understanding. In this section, we explore several promising avenues for research.

A Multitask Learning

MTL is promising in sexism detection, as the content related to sexism encompasses emotion and implicit expressions. Previous research has shown the effectiveness of MTL on emotion classification and sarcasm detection. In the context of multimodal detection, tasks related to tone and facial expressions can be learned concurrently, enhancing the overall understanding of sexist content.

B Multimodal Research

Investigate methods to enhance the integration of multiple modalities (text, images, videos) for a more comprehensive understanding of sexist content. Currently, there is a notable lack of auditory resources, such as podcasts, as well as limited video resources, particularly from live streaming platforms. Moreover, several challenges remain to be addressed. For instance, it is still unclear why multimodal features do not always lead to improved performance, how different modalities contribute to the models, and how to fully leverage the heterogeneous information present in multimodal data. Addressing these challenges is crucial to unlocking the full potential of multimodal sexism detection approaches.

C Code-mixed Language Research

Code-mixed language, characterized by the usage of multiple languages within the same discourse, enriches our expression and is prevalent in global social media. The situation is common, particularly in post-colonial regions such as India and Hong Kong. The rapid pace of globalization has further accelerated this linguistic fusion. To address the sexism detection in code-mixed context, new methods or models are expected.

D. Other Genders

Extend research to delve deeper into gender issues affecting men and other minority genders. While existing resources predominantly address sexism toward women, reflecting prevalent societal trends, there is a growing recognition of the need to examine how sexism and gender biases impact individuals of all genders. This includes investigating the unique challenges faced by men[17], nonbinary individuals, and those in the LGBTQ+ community.

E. Internalized and Institutional Sexism

The gap in Institutional and internalized sexism detection is significant. Current sexism datasets primarily capture external expressions from individuals directed towards others, omitting instances of internalized sexism where individuals describe themselves. Detecting internalized sexism is beneficial for monitoring the psychological well-being of netizens, especially adolescents. Institutional sexism is suitable for longitudinal studies to track changes in institutions over time, investigating how societal transformations, policy shifts, and cultural trends influence the occurrence and expression of sexism within institutions.

F. Foster Interdisciplinary Research and Practices

Encourage interdisciplinary collaborations to enhance the study of sexism detection. Incorporate perspectives from psychology, sociology, and other fields to create comprehensive models that grasp the diverse facets of sexist content and explore its impact on self-esteem, identity formation, psychological health, and coping mechanisms.

G. Global and Cross-Cultural Perspectives

Compare sexism across different cultural, socio-economic, and geographical contexts to identify commonalities and differences in its prevalence, mechanisms, and impacts. Consider how cultural norms, traditions, and power structures shape attitudes and behaviors related to gender.

[17]We noticed a new resource paper for men gender [82] published in recent months, therefore we include it in Table I.

H. Potential Applications

Automated sexism detection tools are widely deployed on platforms such as Twitter and Facebook to identify misogynistic language, harassment, and hate speech. Potential applications are workplace equity audits and educational content screening. Particularly, organizations employ text analysis to audit internal communications, job postings, and performance reviews for gendered language; educational institutions use detection systems to review learning materials and student online interactions for gender stereotypes.

XII. Conclusion

This survey fills a gap in the existing literature by focusing specifically on sexism detection. By systematically analyzing multimodal and multilingual sexism detection tasks and approaches, this survey provides a comprehensive overview of existing methodologies and identifies critical challenges and future trends in this field. This survey serves as a foundational resource for researchers and practitioners in the field of sexism detection but also encourages collaborative efforts to develop more nuanced and culturally sensitive strategies for combating sexism.

References

[1] P. Fortuna and S. Nunes, “A survey on automatic detection of hate speech in text,” *ACM Comput. Surv.*, vol. 51, no. 4, pp. 1–30, Jul. 2018.

[2] G. Kovács, P. Alonso, and R. Saini, “Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources,” *SN Comput. Sci.*, vol. 2, pp. 1–15, Feb. 2021.

[3] N. S. Mullah and W. M. N. W. Zainon, “Advances in machine learning algorithms for hate speech detection in social media: a review,” *IEEE Access*, vol. 9, pp. 88364–88376, 2021.

[4] F. Alkomah and X. Ma, “A literature review of textual hate speech detection methods and datasets,” *Information*, vol. 13, no. 6, p. 273, 2022.

[5] M. Subramanian et al., “A survey on hate speech detection and sentiment analysis using machine learning and deep learning models,” *Alexandria Eng. J.*, vol. 80, pp. 110–121, Oct. 2023.

[6] A. Chhabra and D. K. Vishwakarma, “A literature survey on multimodal and multilingual automatic hate speech identification,” *Multimedia Syst.*, vol. 29, pp. 1203–1230, Jan. 2023.

[7] A. Gandhi et al., “Hate speech detection: A comprehensive review of recent works,” *Expert Syst.*, vol. 41, no. 8, 2024, Art. no. e13562.

[8] K. Stanczak and I. Augenstein, “A survey on gender bias in natural language processing,” 2021, *arXiv:2112.14168*.

[9] O. Istaiteh, R. Al-Omoush, and S. Tedmori, “Racist and sexist hate speech detection: Literature review,” in *Proc. Int. Conf. Intell. Data Sci. Technol. IEEE Int. Symp. Spread Spectr. Tech. Appl. (IDSTA),* Piscataway, NJ, USA: IEEE Press, 2020, pp. 95–99.

[10] E. Shushkevich and J. Cardiff, “Automatic misogyny detection in social media: A survey,” *Computación y Sist.*, vol. 23, no. 4, pp. 1159–1164, 2019.

[11] J. K. Swim and L. L. Hyers, “Sexism,” in Handbook of Prejudice, Stereotyping, and Discrimination, 2009, pp. 407–430.

[12] P. Glick and S. T. Fiske, “The ambivalent sexism inventory: Differentiating hostile and benevolent sexism,” *J. Pers. Social Psychol.*, vol. 70, no. 3, pp. 491–512, 1996.

[13] M. Menegatti and M. Rubini, “Gender bias and sexism in language,” in Oxford Res. Encyclopedia of Communication, 2017.

[14] B. Friedman and H. Nissenbaum, “Bias in computer systems,” *ACM Trans. Inf. Syst. (TOIS)*, vol. 14, no. 3, pp. 330–347, 1996.

[15] T. Spinde et al., “Automated identification of bias inducing words in news articles using linguistic and context-oriented features,” *Inf. Process. Manage.*, vol. 58, no. 3, 2021, Art. no. 102505.

[16] F. Squazzoni et al., “Peer review and gender bias: A study on 145 scholarly journals,” *Sci. Adv.*, vol. 7, no. 2, 2021, Art. no. eabd0299.

[17] T. Garg, S. Masud, T. Suresh, and T. Chakraborty, “Handling bias in toxic speech detection: A survey,” *ACM Comput. Surv.*, vol. 55, no. 13s, pp. 1–32, 2023.

[18] A. Piergentili, B. Savoldi, D. Fucci, M. Negri, and L. Bentivogli, “Hi guys or hi folks? benchmarking gender-neutral machine translation with the gente corpus,” 2023, *arXiv:2310.05294*.

[19] Z. Xie and T. Lukasiewicz, “An empirical analysis of parameter-efficient methods for debiasing pre-trained language models,” 2023, *arXiv:2306.04067*.

[20] S. Honnavalli et al., “Towards understanding gender-seniority compound bias in natural language generation,” 2022, *arXiv:2205.09830*.

[21] W. Ye et al., “Adversarial examples generation for reducing implicit gender bias in pre-trained models,” 2021, *arXiv:2110.01094*.

[22] H. R. Kirk et al., “Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models,” *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 2611–2624, Dec. 2021.

[23] M. Nadeem, A. Bethke, and S. Reddy, “Stereoset: Measuring stereotypical bias in pretrained language models,” 2020, *arXiv:2004.09456*.

[24] D. Oba, M. Kaneko, and D. Bollegala, “In-contextual bias suppression for large language models,” 2023, *arXiv:2309.07251*.

[25] R. Hada, A. Seth, H. Diddee, and K. Bali, “fifty shades of bias”: Normative ratings of gender bias in GPT generated English text,” 2023, *arXiv:2310.17428*.

[26] V. Thakur, “Unveiling gender bias in terms of profession across LLMS: Analyzing and addressing sociological implications,” 2023, *arXiv:2307.09162*.

[27] T. Wambsganss, X. Su, V. Swamy, S. P. Neshaei, R. Rietsche, and T. Käser, “Unraveling downstream gender bias from large language models: A study on AI educational writing assistance,” 2023, *arXiv:2311.03311*.

[28] Y. C. Tan and L. E. Celis, “Assessing social and intersectional biases in contextualized word representations,” *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 1–12, 2019.

[29] M. R. Mohseni, “Motives of online hate speech: results from a quota sample online survey,” *CyberPsychol., Behav., Social Netw.*, vol. 26, no. 7, pp. 499–506, 2023.

[30] J. T. Nockleby, “Hate speech,” in Encyclopedia of the Amer. constitution, vol. 3, no. 2, 2000, pp. 1277–1279.

[31] A. Kumar et al., “Comment toxicity detection via a multichannel convolutional bidirectional gated recurrent unit,” *NeuroComputing*, vol. 441, pp. 272–278, Jun. 2021.

[32] A. Khatua et al., “Tweeting in support of LGBT? a deep learning approach,” in *Proc. ACM India Joint Int. Conf. Data Sci. Manage. Data*, 2019, pp. 342–345.

[33] Z. Waseem, “Are you a racist or am I seeing things? annotator influence on hate speech detection on twitter,” in *Proc. 1st Workshop NLP Comput. Social Sci.*, 2016, pp. 138–142.

[34] Z. Waseem and D. Hovy, “Hateful symbols or hateful people? predictive features for hate speech detection on Twitter,” in *Proc. NAACL Student Research Workshop*, San Diego, California: Assoc. for Comput. Linguistics, Jun. 2016, pp. 88–93.

[35] R. P. De Pelle and V. P. Moreira, “Offensive comments in the Brazilian web: a dataset and baseline results,” in Proc. Anais Do VI Brazilian Workshop Social Netw. Anal. Mining, São Paulo, 2017.

[36] R. Gomez et al., “Exploring hate speech detection in multimodal publications,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2020, pp. 1470–1478.

[37] N. Zueva, M. Kabirova, and P. Kalaidin, “Reducing unintended identity bias in Russian hate speech detection,” in *Proc. 4th Workshop Online Abuse Harms*, 2020, pp. 65–69.

[38] A. Al-Hassan and H. Al-Dossari, “Detection of hate speech in Arabic tweets using deep learning,” *Multimedia Syst.*, vol. 28, no. 6, pp. 1963–1974, 2022.

[39] F. Gasparini et al., “Multimodal classification of sexist advertisements,” in *Proc. 15th Int. Joint Conf. e-Business Telecommun. (ICETE)*, vol. 1. Porto, Portugal: SciTePress, 2018, pp. 399–406.

[40] D. Grosz and P. Conde-Cespedes, “Automatic detection of sexist statements commonly used at the workplace,” in *Proc. Trends Appl. Knowl. Discovery Data Mining (PAKDD) Workshops, DSFN, GII, BDM, LDRC and LBD*, Singapore: Springer, May 2020, pp. 104–115.

[41] M. Samory, I. Sen, J. Kohne, et al., “call me sexist, but. . .” : Revisiting sexism detection using psychological scales and adversarial samples,” in *Proc. Int. AAAI Conf. Web Social Media*, vol. 15, no. 1, May 2021, pp. 573–584.

[42] T. Bertaglia et al., “Sexism in focus: An annotated dataset of Youtube comments for gender bias research,” in *Proc. 3rd Int. Workshop Open Challenges Online Social Netw., Ser. (OASIS)*, New York, NY, USA: ACM, 2023, pp. 22–28.

[43] S. Bhattacharya et al., “Developing a multilingual annotated corpus of misogyny and aggression,” in *Proc. Second Workshop Trolling, Aggression Cyberbullying, Marseille, France: Eur. Lang. Resour. Assoc. (ELRA)*, May 2020, pp. 158–168.

[44] D. Almanea and M. Poesio, “ARMIS - the Arabic misogyny and sexism corpus with annotator subjective disagreements,” in *Proc. Lang. Resour. Eval. Conf.*, Marseille, France: Eur. Lang. Resour. Assoc., Jun. 2022, pp. 2282–2291.

[45] F. Gasparini, G. Rizzi, A. Saibene, and E. Fersini, “Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content,” *Data Brief*, vol. 44, 2022, Art. no. 108526.

[46] A. Jha and R. Mamidi, “When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data,” in *Proc. Second Workshop NLP Computat. Social Sci*., Vancouver, Canada: Assoc. for Comput. Linguistics, Aug. 2017, pp. 7–16.

[47] S. Sharifirad and S. Matwin, “When a tweet is actually sexist. A more comprehensive classification of different online harassment categories and the challenges in NL,” 2019, *a*rXiv:*1902.10584*.

[48] A. Jiang, X. Yang, Y. Liu, and A. Zubiaga, “SWSR: A Chinese dataset and lexicon for online sexism detection,” *Online Social Netw. Media*, vol. 27, 2022, Art. no. 100182.

[49] H. Kirk, W. Yin, B. Vidgen, and P. Röttger, “SemEval-2023 task 10: Explainable detection of online sexism,” in *Proc. 17th Int. Workshop Semantic Eval*., Toronto, Canada: Assoc. for Comput. Linguistics, Jul. 2023, pp. 2193–2210.

[50] D. C. Höfels, Ç. Çöltekin, and I. D. Mădroane, “CoRoSeOf-an annotated corpus of Romanian sexist and offensive tweets,” in *Proc. 13th Lang. Resour. Eval. Conf.*, 2022, pp. 2269–2281.

[51] E. Fersini et al., “Overview of the Evalita 2018 task on automatic misogyny identification (AMI),” in *Proc. CEUR Workshop Proc.*, vol. 2263, Turin, Italy: CEUR-WS, 2018, pp. 1–9.

[52] E. Fersini et al., “Overview of the task on automatic misogyny identification at Ibereval 2018,” in *Proc. CEUR Workshop Proc.*, vol. 2150. Sevilla, Spain: CEUR-WS, 2018, pp. 214–228.

[53] E. Guest et al., “An expert annotated dataset for the detection of online misogyny,” in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics: Main Volume*, 2021, pp. 1336–1350.

[54] S. Hewitt, T. Tiropanis, and C. Bokhove, “The problem of identifying misogynist language on twitter (and other online social spaces),” in *Proc. 8th ACM Conf. Web Sci. Ser. (WebSci)*, New York, NY, USA: ACM, 2016, pp. 333–335.

[55] M. Anzovino, E. Fersini, and P. Rosso, “Automatic identification and classification of misogynistic language on Twitter,” *Nat. Lang. Process. Inf. Syst.*, vol. 10859, pp. 57–64, May 2018.

[56] S. Karlekar and M. Bansal, “SafeCity: Understanding diverse forms of sexual harassment personal stories,” in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, Brussels, Belgium: Assoc. for Comput. Linguistics, Oct./Nov. 2018, pp. 2805–2811.

[57] M. Rezvan et al., “A quality type-aware annotated corpus and lexicon for harassment research,” in *Proc. 10th Acm Conf. Web Sci.*, 2018, pp. 33–36.

[58] E. Fersini, F. Gasparini, and S. Corchs, “Detecting sexist meme on the web: A study on textual and visual cues,” in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interaction Workshops Demos (ACIIW),* Piscataway, NJ, USA: IEEE Press, 2019, pp. 226–231.

[59] P. Parikh et al., “Multi-label categorization of accounts of sexism using a neural framework,” in *Proc. Conf. Empirical Methods Nat. Lang. Process. 9th Int. Joint Conf. Nat. Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, China: Assoc. for Comput. Linguistics, Nov. 2019, pp. 1642–1652.

[60] A. Karami et al., “Hidden in plain sight for too long: Using text mining techniques to shine a light on workplace sexism and sexual harassment,” *Psychol. Violence*, vol. 14, no. 1, pp. 1–13, 2019.

[61] P. Fortuna, J. R. da Silva, L. Wanner, et al., “A hierarchically-labeled Portuguese hate speech dataset,” in *Proc. 3rd Workshop Abusive Lang. Online*, 2019, pp. 94–104.

[62] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, and L. Plaza, “Automatic classification of sexism in social networks: An empirical study on twitter data,” *IEEE Access*, vol. 8, pp. 219563–219576, 2020.

[63] H. Rizwan, M. H. Shakeel, and A. Karim, “Hate-speech and offensive language detection in roman Urdu,” in *Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP),* 2020, pp. 2512–2522.

[64] A. Karami et al., "Unwanted advances in higher education: Uncovering sexual harassment experiences in academia with text mining," *Inf. Process. Manage.*, vol. 57, no. 2, 2020, Art. no. 102167.
[65] O. El Ansari, Z. Jihad, and M. Hajar, "A dataset to support sexist content detection in Arabic text," in *Proc. Image Signal Process.: 9th Int. Conf. (ICISP)*, Marrakesh, Morocco: Springer, Jun. 2020, pp. 130–137.
[66] P. Chiril et al., "An annotated corpus for sexism detection in French tweets," in *Proc. 12th Lang. Resources Eval. Conf.*, Marseille, France: Eur. Lang. Resour. Assoc., May 2020, pp. 1397–1403.
[67] D. Assenmacher, M. Niemann, K. Müller, M. Seiler, D. M. Riehle, and H. Trautmann, "RP-Mod & RP-crowd: Moderator- and crowd-annotated German news comment datasets," in *Proc. 35th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track (Round 2)*, 2021.
[68] H. Mulki and B. Ghanem, "Let-MI: An Arabic Levantine twitter dataset for misogynistic language," in *Proc. 6th Arabic Nat. Lang. Process. Workshop*, 2021, pp. 154–163.
[69] P. Zeinert, N. Inie, and L. Derczynski, "Annotating online misogyny," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Nat. Lang. Process. (Volume 1: Long Papers)*, 2021, pp. 3181–3197.
[70] P. Chiril, F. Benamara, and V. Moriceau, "Be nice to your wife! the restaurants are closed": Can gender stereotype detection improve sexism classification?" in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*, 2021, pp. 2833–2844.
[71] S. Sultana, "Identifying sexism and misogyny in pull request comments," in *Proc. 37th IEEE/ACM Int. Conf. Automated Softw. Eng.*, 2022, pp. 1–3.
[72] B. Trinkenreich et al., "An empirical investigation on the challenges faced by women in the software industry: A case study," in *Proc. ACM/IEEE 44th Int. Conf. Softw. Eng.: Softw. Eng. Soc.*, 2022, pp. 24–35.
[73] A. Yadav, S. Chandel, S. Chatufale, and A. Bandhakavi, "Lahm: Large annotated dataset for multi-domain and multilingual hate speech identification," 2023, *arXiv:2304.00913*.
[74] H. Buie and A. Croft, "The social media sexist content (SMSC) database: A database of content and comments for research use," *Collabra: Psychol.*, vol. 9, no. 1, 2023, Art. no. 71341.
[75] L. M. Álvarez-Crespo and L. M. Castro, "A Galician corpus for misogyny detection online," in *Proc. 16th Int. Conf. Comput. Process. Portuguese*, P. Gamallo, D. Claro, A. Teixeira, L. Real, M. Garcia, H. G. Oliveira, and R. Amaro, Eds. Santiago de Compostela, Galicia/Spain: Assoc. for Comput. Linguistics, Mar. 2024, pp. 22–31.
[76] A. Vetagiri, P. Pakray, and A. Das, "A deep dive into automated sexism detection using fine-tuned deep learning and large language models," *Eng. Appl. Artif. Intell.*, vol. 145, 2024, Art. no. 110167.
[77] L. S. Mut Altin and H. Saggion, "A novel corpus for automated sexism identification on social media," in *Proc. 3rd Annu. Meet. Special Interest Group Under-Resourced Lang. (LREC-COLING)*, M. Melero, S. Sakti, and C. Soria, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 10–15.
[78] I. Arcos and P. Rosso, "Sexism identification on Tiktok: a multimodal AI approach with text, audio, and video," in *Proc. Int. Conf. Cross-Lang. Eval. Forum Eur. Lang.*, Grenoble, France: Springer, 2024, pp. 61–73.
[79] B. Krenn, J. Petrak, M. Kubina, and C. Burger, "GERMS-AT: A sexism/misogyny dataset of forum comments from an Austrian online newspaper," in *Proc. Joint Int. Conf. Comput. Linguistics, Lang. Resources Eval. (LREC-COLING)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., Torino, Italia: ELRA and ICCL, May 2024, pp. 7728–7739.
[80] B. Sheppard et al., "BIASLY: An expert-annotated dataset for subtle misogyny detection and mitigation," in *Proc. Findings Assoc. Comput. Linguistics (ACL)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Assoc. for Comput. Linguistics, Aug. 2024, pp. 427–452.
[81] A. Singh, D. Sharma, and V. K. Singh, "Misogynistic attitude detection in Youtube comments and replies: A high-quality dataset and algorithmic models," *Comput. Speech Lang.*, vol. 89, 2025, Art. no. 101682.
[82] X. Luo et al., "Beyondgender: A multifaceted bilingual dataset for practical sexism detection," *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 23, pp. 24750–24758, Apr. 2025.
[83] V. Basile et al., "SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter," in *Proc. 13th Int. Workshop Semantic Eval.*, 2019, pp. 54–63.
[84] E. Fersini et al., "Semeval-2022 task 5: Multimedia automatic misogyny identification," in *Proc. 16th Int. Workshop Semantic Eval.*, 2022, pp. 533–549.
[85] S. Gross, J. Petrak, L. Venhoff, and B. Krenn, "GermEval2024 shared task: GerMS-detect – sexism detection in German online news fora," in *Proc. GermEval Task 1 GerMS-Detect Workshop Sexism Detection German Online News Fora (GerMS-Detect)*, B. Krenn, J. Petrak, and S. Gross, Eds., Vienna, Austria: Assoc. for Comput. Lingustics, Sep. 2024, pp. 1–9.
[86] R. Priyadharshini et al., "Overview of abusive comment detection in Tamil-ACL 2022," in *Proc. Second Workshop Speech Lang. Technol. Dravidian Lang.*, Dublin, Ireland: Assoc. for Comput. Linguistics, May 2022, pp. 292–298.
[87] E. Fersini et al., "Ami@evalita2020: Automatic misogyny identification," in *Proc. 7th Eval. Campaign Nat. Lang. Process. Speech Tools Italian (EVALITA)*, 2020.
[88] F. J. Rodríguez-Sanchez et al., "Overview of exist 2021: sexism identification in social networks," *Proces. del Leng. Nat.*, vol. 67, pp. 195–207, Sep. 2021.
[89] F. Rodrıguez-Sánchez et al., "Overview of exist 2022: sexism identification in social networks," *Procesamiento Del Lenguaje Nat.*, vol. 69, pp. 229–240, Sep. 2022.
[90] L. Plaza, et al., "Overview of exist 2023: sexism identification in social networks," in *Proc. Eur. Conf. Inf. Retrieval*, Dublin, Ireland: Springer, 2023, pp. 593–599.
[91] L. Plaza et al., "Exist 2024: sexism identification in social networks and memes," in *Proc. Advances in Information Retrieval*, N. Goharian, N. Tonellotto, Y. He, et al., Eds. Cham, Switzerland: Springer Nature, 2024, pp. 498–504.
[92] L. Plaza et al., "Exist 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos." Accessed: Apr 13, 2025. [Online]. Available: https://nlp.uned.es/exist2025/
[93] S. Sultana, J. Sarker, and A. Bosu, "A rubric to identify misogynistic and sexist texts from software developer communications," in *Proc. 15th ACM/IEEE Int. Symp. Empirical Softw. Eng. Meas. (ESEM)*, 2021, pp. 1–6.
[94] W. N. A. W. Mustapha et al., "Detection of harassment toward women in Twitter during pandemic based on machine learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 3, pp. 1035–1043, 2024.
[95] P. Badjatiya et al., "Deep learning for hate speech detection in tweets," in *Proc. 26th Int. Conf. World Wide Web Companion*, 2017, pp. 759–760.
[96] S. Sharifirad, B. Jafarpour, and S. Matwin, "Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs," in *Proc. 2nd Workshop Abusive Lang. Online (ALW2)*, 2018, pp. 107–114.
[97] G. Rizos, K. Hemker, and B. Schuller, "Augment to prevent: short-text data augmentation in deep learning for hate-speech classification," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, 2019, pp. 991–1000.
[98] S. Melville, K. Eccles, and T. Yasseri, "Topic modeling of everyday sexism project entries," *Front. Digit. Hum.*, vol. 5, p. 28, Jan. 2019.
[99] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A Bert-based transfer learning approach for hate speech detection in online social media," in *Proc. 8th Int. Conf. Complex Netw. Their Appl.*, Lisbon, Portugal: Springer, 2020, pp. 928–940.
[100] H. Abburi et al., "Semi-supervised multi-task learning for multi-label fine-grained sexism classification," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 5810–5820.
[101] C. Karatsalos and Y. Panagiotakis, "Attention-based method for categorizing different types of online harassment language," in *Proc. Mach. Learn. Knowledge Discovery Databases: Int. Workshops (ECML PKDD)*, Würzburg, Germany: Springer, Sep. 2020, pp. 321–330.
[102] P. Parikh et al., "Categorizing sexism and misogyny through neural approaches," *ACM Trans. on the Web TWEB*, vol. 15, no. 4, pp. 1–31, 2021.
[103] H. Abburi et al., "Fine-grained multi-label sexism classification using a semi-supervised multi-level neural approach," *Data Sci. Eng.*, vol. 6, no. 4, pp. 359–379, 2021.
[104] A. Younus and M. A. Qureshi, "A framework for sexism detection on social media via ByT5 and TabNet," 2022.
[105] A. Vaca-Serrano, "Detecting and classifying sexism by ensembling transformers models," in *Proc. CEUR Workshop*, 2022.
[106] M. Franco, O. Gaggi, and C. E. Palazzi, "Analyzing the use of large language models for content moderation with ChatGPT examples," in

Proc. 3rd Int. Workshop Open Challenges Online Social Netw., 2023, pp. 1–8.

[107] L. Tian, N. Huang, and X. Zhang, "Efficient multilingual sexism detection via large language models cascades," in *Proc. Work. Notes CLEF*, 2023.

[108] A. F. M. de Paula, G. Rizzi, E. Fersini, and D. Spina, "AI-UPV at exist 2023–sexism characterization using large language models under the learning with disagreements regime," 2023, *arXiv:2307.03385*.

[109] L. Betti, C. Abrate, and A. Kaltenbrunner, "Large scale analysis of gender bias and sexism in song lyrics," *EPJ Data Sci.*, vol. 12, no. 1, p. 10, 2023.

[110] I. Sen, D. Assenmacher, M. Samory, I. Augenstein, W. Aalst, and C. Wagner, "People make better edits: Measuring the efficacy of LLM-generated counterfactually augmented data for harmful language detection," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Assoc. for Comput. Linguistics, Dec. 2023, pp. 10480–10504.

[111] A. F. M. de Paula, P. Rosso, and D. Spina, "Mitigating negative transfer with task awareness for sexism, hate speech, and toxic language detection," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN),* Piscataway, NJ, USA: IEEE Press, 2023., pp. 1–8.

[112] A. Das, M. Rahgouy, Z. Zhang, et al., "Online sexism detection and classification by injecting user gender information," in *Proc. IEEE InterNat. Conf. Artif. Intell., Blockchain, and Internet Things (AIBThings),* Piscataway, NJ, USA: IEEE Press, 2023, pp. 1–5.

[113] L. Li, L. Fan, S. Atreja, and L. Hemphill, "Hot ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media," *ACM Trans. Web*, vol. 18, no. 2, pp. 1–36, 2024.

[114] H. Abburi et al., "Multi-task learning neural framework for categorizing sexism," *Comput. Speech Lang.*, vol. 83, 2024, Art. no. 101535.

[115] G. Abercrombie, N. Vitsakis, A. Jiang, and I. Konstas, "Revisiting annotation of online gender-based violence," in *Proc. 3rd Workshop Perspectivist Approaches NLP (NLPerspectives) @ LREC-COLING*, G. Abercrombie, V. Basile, D. Bernadi, S. Dudy, S. Frenda, L. Havens, and S. Tonelli, Eds., Torino, Italia: ELRA and ICCL, May 2024, pp. 31–41.

[116] V. Hangya and A. Fraser, "How to solve few-shot abusive content detection using the data we actually have," in *Proc. Joint Int. Conf. Comput. Linguistics, Lang. Resources Eval. (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 8307–8322.

[117] S. Khan, G. Pergola, and A. Jhumka, "Multilingual sexism identification via fusion of large language models," in *Proc. Work. Notes CLEF*, 2024.

[118] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, and L. Plaza, "Detecting sexism in social media: an empirical analysis of linguistic patterns and strategies," *Appl. Intell.*, vol. 54, pp. 10995–11019, Sep. 2024.

[119] A. Riahi Samani, T. Wang, K. Li, and F. Chen, "Large language models with reinforcement learning from human feedback approach for enhancing explainable sexism detection," in *Proc. 31st Int. Conf. Comput. Linguistics*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, Eds., Abu Dhabi, UAE: Assoc. for Comput. Linguistics, Jan. 2025, pp. 6230–6243.

[120] A. Mohasseb, E. Amer, F. Chiroma, and A. Tranchese, "Leveraging advanced NLP techniques and data augmentation to enhance online misogyny detection," *Appl. Sci.*, vol. 15, no. 2, pp. 1–24, 2025.

[121] E. Fersini et al., "Misogynous meme recognition: A preliminary study," in *Proc. Int. Conf. Italian Assoc. Artif. Intell.*, Springer, 2021, pp. 279–293.

[122] G. Rizzi et al., "Recognizing misogynous memes: Biased models and tricky archetypes," *Inf. Process. Manage.*, vol. 60, no. 5, 2023, Art. no. 103474.

[123] G. Kumari, A. Sinha, and A. Ekbal, "Unintended bias detection and mitigation in misogynous memes," in *Proc. 18th Conf. Eur. Chapter Assoc. Comput. Linguistics (Volume 1: Long Papers)*, Y. Graham and M. Purver, Eds., St. Julian's, Malta: Assoc. for Comput. Linguistics, Mar. 2024, pp. 2719–2733.

[124] D. D. Barua et al., "Penta ML at exist 2024: tagging sexism in online multimodal content with attention-enhanced modal context," in *Proc. Work. Notes CLEF*, 2024.

[125] M. Z. U. Rehman, S. Zahoor, A. Manzoor, M. Maqbool, and N. Kumar, "A context-aware attention and graph neural network-based multimodal framework for misogyny detection," *Inf. Process. Manage.*, vol. 62, no. 1, 2025, Art. no. 103895.

[126] A. da Silva Oliveira, T. de Carvalho Cecote, J. P. R. Alvarenga, V. L. de Souza Freitas, and E. J. da Silva Luz, "Toxic speech detection in Portuguese: A comparative study of large language models," in *Proc. 16th Int. Conf. Comput. Process. Portuguese - Vol. 1*, P. Gamallo, D. Claro, A. Teixeira, L. Real, M. Garcia, H. G. Oliveira, and R. Amaro, Eds., Santiago de Compostela, Galicia/Spain: Assoc. for Comput. Lingustics, Mar. 2024, pp. 108–116.

[127] K. Perifanos and D. Goutsos, "Multimodal hate speech detection in Greek social media," *Multimodal Technol. Interaction*, vol. 5, no. 7, pp. 1–10, 2021.

[128] G. Meseguer-Brocal et al., "WASABI: a two million song database project with audio and cultural metadata plus webaudio enhanced client applications," in *Proc. Web Audio Conf. (Collaborative Audio# WAC2017)*, 2017.

[129] I. Sen, M. Samory, C. Wagner, and I. Augenstein, "Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, 2022, pp. 4716–4726.

[130] M. Das, S. K. Pandey, and A. Mukherjee, "Evaluating ChatGPT's performance for multilingual and emoji-based hate speech detection," 2023, *arXiv:2305.13276*.

[131] F. Muntasir and J. Noor, "Explainable AI discloses gender bias in sexism detection algorithm," in *Proc. 11th Int. Conf. Netw. Syst. Security Ser. (NSysS)*, New York, NY, USA: ACM, 2025, pp. 120–127.

[132] H. Mohammadi, A. Giachanou, and A. Bagheri, "A transparent pipeline for identifying sexism in social media: Combining explainability with model prediction," *Appl. Sci.*, vol. 14, no. 19, 2024.

[133] X. Dong, Y. Wang, P. S. Yu, and J. Caverlee, "Disclosure and mitigation of gender bias in LLMS," 2024, *arXiv:2402.11190*.

[134] R. Wolfe, Y. Yang, B. Howe, and A. Caliskan, "Contrastive language-vision AI models pretrained on web-scraped multimodal data exhibit sexual objectification bias," in *Proc. ACM Conf. Fairness, Accountability, and Transparency*, 2023, pp. 1174–1185.

[135] P. Rottger et al., "Multilingual hatecheck: functional tests for multilingual hate speech detection models," in *Proc. 6th Workshop Online Abuse Harms (WOAH),* Seattle, Washington: Assoc. for Comput. Linguistics, 2022, pp. 154–169.

[136] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the type and target of offensive posts in social media," in *Proc. Conf. North Am. Chapt. Assoc. Comput. Linguistics: Human Lang. Technol., Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Assoc. for Comput. Linguistics, Jun. 2019, pp. 1415–1420.

[137] Y-n Seo, P. Oh, and W. Y. Kil, "Into the wolves' den: an investigation of predictors of sexism in online games," *Behav. Inf. Technol.*, vol. 41, no. 8, pp. 1740–1754, 2022.

[138] E. G. Armstrong, "Sexism and misogyny in music land," *J. Criminal Justice Popular Culture*, vol. 8, no. 2, pp. 96–126, 2001.

[139] M. D. Cobb and W. A. Boettcher, III "Ambivalent sexism and misogynistic rap music: Does exposure to eminem increase sexism?" *J. Appl. Social Psychol.*, vol. 37, no. 12, pp. 3025–3042, 2007.

[140] M. A. Flynn, C. M. Craig, C. N. Anderson, and K. J. Holody, "Objectification in popular music lyrics: An examination of gender and genre differences," *Sex Roles*, vol. 75, pp. 164–176, Feb. 2016.

[141] M. Amelia Gibbons and M. A. Rossi, "Military conscription, sexist attitudes and intimate partner violence," *Economica*, vol. 89, no. 355, pp. 540–563, 2022.

[142] I. Testoni, G. Branciforti, A. Zamperini, L. Zuliani, and F. A. Nava, "Prisoners' ambivalent sexism and domestic violence: a narrative study," *Int. J. Prisoner Health*, vol. 15, no. 4, pp. 332–348, 2019.

[143] S. M. Utych, "Sexism predicts favorability of women in the 2020 democratic primary… and men?" *Electoral Stud.*, vol. 71, 2021, Art. no. 102184.

[144] A. Chao-Fernández, R. Chao-Fernández, and C. López-Chao, "Sexism in lyrics of children's songs in school and family environment," *Educ. Sci.*, vol. 10, no. 11, p. 300, 2020.

[145] S. Nakandala, G. Ciampaglia, N. Su, and Y.-Y. Ahn, "Gendered conversation in a social game-streaming platform," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 11, no. 1, pp. 162–171, May 2017.

[146] J. Uttarapong, J. Cai, and D. Y. Wohn, "Harassment experiences of women and LGBTQ live streamers and how they handled negativity," in *Proc. ACM Int. Conf. Interactive Media Experiences, Ser. (IMX),* New York, NY, USA: ACM, 2021, pp. 7–19.

[147] K. P. B. Alvarez and V. H. H. Chen, "Community and capital: Experiences of women game streamers in Southeast Asia," *Trans. Soc. Comput.*, vol. 4, no. 3, pp. 1–22, Oct. 2021.

[148] A. M. Guajardo, "it sucks for me, and it sucks for them": The emotional labor of women twitch streamers," in *Proc. DiGRA Conf.: Bringing Worlds Together*, 2022.

[149] J. Sasse and J. Grossklags, "Breaking the silence: Investigating which types of moderation reduce negative effects of sexist social media content," *Proc. ACM Human-Comput. Interaction*, vol. 7, no. CSCW2, pp. 1–26, 2023.

[150] J. Cai, D. Y. Wohn, and M. Almoqbel, "Moderation visibility: Mapping the strategies of volunteer moderators in live streaming micro communities," in *Proc. ACM Int. Conf. Interactive Media Experiences Ser. (IMX)*, New York, NY, USA: ACM, 2021, pp. 61–72.

[151] J. Cai, S. Chowdhury, H. Zhou, and D. Y. Wohn, "Hate raids on twitch: Understanding real-time human-bot coordinated attacks in live streaming communities," *Proc. ACM Hum.-Comput. Interact.*, vol. 7, no. CSCW2, Oct. 2023.

[152] S. K. Jaijee, C. Kamau-Mitchell, G. W. Mikhail, and C. Hendry, "Sexism experienced by consultant cardiologists in the united kingdom," *Heart*, vol. 107, no. 11, pp. 895–901, 2021.

[153] K. K. Dray and I. E. Sabat, "Confronting sexism: Identifying dimensions and exploring impact," *J. Appl. Social Psychol.*, vol. 52, no. 5, pp. 316–340, 2022.

[154] S. Tang, D. Wang, and J. Hou, "Gender inequalities: Women researchers require more knowledge in specific and experimental topics," 2023, *arXiv:2309.01964*.

[155] A. Biurrun-Garrido et al., "Everyday sexism in nursing degrees: A cross-sectional, multicenter study," *Nurse Educ. Today*, vol. 132, 2024, Art. no. 106009.

Xuan Luo is currently working toward the Ph.D. degree in computer science and technology with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China and with the School of Computing, Hong Kong Polytechnic University, Hong Kong, China. Her research interests include natural language processing, sentiment analysis, and social media analysis.

Bin Liang received the Ph.D. degree in computer science from Harbin Institute of Technology, Shenzhen, China, in 2022.

He is a Postdoctoral Fellow with the Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong, China. His research interests include natural language processing, sentiment analysis, deep learning, and machine learning.

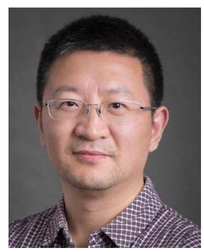
Qianlong Wang is currently working toward the Ph.D. degree in computer science and technology with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, China. His research interests include natural language processing and sentiment analysis.

Jing Li received the B.S. degree in intelligence science and technology from the Department of Machine Intelligence, Peking University, Beijing, China, in 2013, and the Ph.D. degree in systems engineering and engineering management from the Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong, China, in 2017.

From 2017 to 2019, she was a Senior Researcher with the NLP Center, Tencent AI Lab, Shenzhen, China. Since 2019, she has been an Assistant Professor with the Department of Computing (COMP), The Hong Kong Polytechnic University (PolyU), Hong Kong. She established and currently leads the Embodied Artificial Intelligence Lab, Department of Computing (COMP), PolyU, where she is also a member of the Research Centre on Data Sciences and Artificial Intelligence (RC-DSAI).

Erik Cambria received the Ph.D. degree in computing science and mathematics in 2012 through a joint programme between the University of Stirling, Scotland, U.K., and MIT Media Lab, Cambridge, MA, USA.

He is the Founder of SenticNet, a Professor with Nanyang Technological University, Singapore, where he also holds the appointment of Provost Chair in computer science and engineering. Prior to joining NTU, he worked with Microsoft Research Asia, Beijing, China, and HP Labs India, Bangalore, India. His research interests include neurosymbolic AI for explainable natural language processing in domains such as sentiment analysis, dialogue systems, and financial forecasting.



Xiaojun Zhang received the Ph.D. degree in computational linguistics from Nanjing Normal University, Nanjing, China, in 2008.

He is an Associate Professor with Xian Jiaotong-Liverpool University, Suzhou, China and an Honorary Associate with the University of Liverpool, Liverpool, U.K. He is an Adjunct Professor with the Open University of Cyprus, Latsia, Cyprus and Northwestern Polytechnical University, Xian, China. He was an Academic Staff and Researcher at Higher Education Institutes in China, Ireland, and the U.K. His research interest includes translation technology, natural language processing, and practical translation.

Yulan He received the Ph.D. degree in spoken language understanding from the University of Cambridge, Cambridge, U.K., in 2004.

She is a Professor in natural language processing with the Department of Informatics of the King's College London, U.K.

Dr. He is a Turing AI fellow. She has published over 170 papers in the areas of natural language understanding, sentiment analysis and opinion mining, question-answering, topic/event extraction from text, biomedical text mining, and social media analytics.

Min Yang received the Ph.D. degree in computer science from the University of Hong Kong, Hong Kong, China, in 2017.

Currently, she is an Associate Professor with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Beijing, China. Her research interests include natural language processing, deep learning, and machine learning.

Ruifeng Xu (Member, IEEE) received the Ph.D. degree in computer science from The Hong Kong Polytechnic University, Hong Kong, China, in 2006.

Currently, he is a Professor with the College of Artificial Intelligence, Harbin Institute of Technology, Shenzhen, China. He has published more than 200 papers in natural language processing, affective computing, and social media analysis.