A Systematic Literature Review on Automatic Sexism Detection in Social Media

Wang Lei
Faculty of Information Engineering & Computer Science, Hebei Finance University, 071000, Baoding, Hebei, China | College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
wanglei648@gmail.com

Nur Atiqah Sia Abdullah
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia | Knowledge and Software Engineering Research Group (KASERG), Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
atiqah684@uitm.edu.my (corresponding author)

Syaripah Ruzaini Syed Aris
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia | Crowd Sourcing Business and Innovation (CBIG), Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
ruzaini@uitm.edu.my

Received: 1 September 2024 | Revised: 1 October 2024 and 10 October 2024 | Accepted: 12 October 2024

ABSTRACT

Sexist content has become increasingly prevalent on social media platforms, underscoring the critical need for the development of efficient Automatic Sexism Detection methods. Previous literature reviews have not encompassed the new advancements in Automatic Sexism Detection observed over the past three years. Hence, the present study conducted a Systematic Literature Review (SLR) that examined 48 primary studies published between 2014 and 17th Sept. 2024, retrieved from six bibliographic databases. This paper aims to present a comprehensive literature review on Automatic Sexism Detection, encompassing the datasets, preprocessing techniques, feature extraction methods, text representations, classification approaches, and evaluation models employed in Automatic Sexism Detection research. The paper includes a discussion of the findings, limitations, and future research directions of the chosen articles. Additionally, it provides an overview of the conclusions drawn from the conducted research. The performed analysis reveals a lack of corpus beyond the English and Spanish language encountered in datasets, with most of the latter being annotated for either misogyny or non-misogyny. Common preprocessing techniques analyzed in the current study include lowercase conversion, text removal, tokenization, stemming, and rewriting. Discrete representations, such as TF-IDF, N-grams, and BoW, are frequently utilized, while distributed representations, like Bert and GloVe, are prominent. Bert is the predominant classification model utilized while combining lexical features can enhance the results in the majority of the discussed scenarios. Accuracy (A) and F1 score (F1) are the most widely deployed evaluation metrics in this field.

Keywords-automatic sexism detection; systematic literature review; social media; features; word representation; deep learning; machine learning; misogyny

I. INTRODUCTION

Social media have become fertile ground for enflamed debates that typically pit 'us' against 'them,' resulting in numerous instances of disrespectful and derogatory language usage [1]. It is the primary venues for social protest, activism, and other activities, where campaigns, like #MeToo, #8M, and #TimesUp, have quickly risen in popularity [2]. The Gamergate controversy began in 2014 and first gained traction on 4chan before spreading to other social media platforms [3]. The organized movement known as "Gamergate", which started online and eventually moved offline, posed a major threat to the lives of women employed in the video game business. During the 2020 worldwide coronavirus pandemic, 35% of the respondents in [4] said that they had experienced online harassment on the basis of identity-based traits. The previous evidence supports the existence of cyber sexism, which has a detrimental effect on society. Hence efforts must be made to obstruct similar occurrences [5]. Furthermore, it is critical to impede the widespread propagation of gender stereotypes, particularly towards young people, considering that a significant portion of Internet users—especially those who utilize social networks—are teens [2, 6].

It takes a lot of time for comment moderators to weed out inappropriate comments. Since it is difficult to completely rely on manual detection of the huge amount of sexist content, more and more researchers are focusing on the automatic identification of sexism. Through a trained model, Automatic Sexism Detection can determine whether a given text contains sexism, as well as the sexism category. The analysis of social data to identify and uncover communication patterns among users and understand their behavior has gained a lot of interest [7]. Many researchers have been engaged in Automatic Sexism Detection in social media, like Twitter and Facebook, in recent years.

Sexism is defined by the Oxford English Dictionary as prejudice, stereotyping, or discrimination, typically against women, based on sex. Sexism can be hostile or benevolent, and involves stereotyping, ideological concerns, sexual violence, and other types of behavior. It can be conveyed in variety of ways, such as direct, indirect, descriptive, or reported [8]. Subtle or covert sexism is less easy to spot than hostile or overt sexism, but it is more pervasive in social media and more harmful to society [8, 9]. The current research concentrates on sexism in online media, and particularly on identifying misogyny or hostility towards women [10]. Misogyny is defined as hatred, or dislike of, or prejudice against women in the Oxford English Dictionary.

Automatic Sexism Detection is a text classification task. There is lexicon-based sexism detection, traditional Machine Learning (ML)-based sexism detection, Deep Learning (DL)-based sexism detection, and the hybrid approach, which constitutes a mixture of the previous three methods. Among these sexism detection approaches, the DL usually achieves better performance. Over the past few years, more people have used the DL method to conduct Automatic Sexism Detection. The former allows individuals to capture subtle, hidden similarities, and distinctions among various abusive behaviors while preventing overfitting [11]. The transformers-based DL, like Bidirectional Encoder Representations from Transformers (BERT), A Robustly Optimized Bert Pretraining Approach (RoBERTa), usually outperforms Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) [6, 8, 12].

Social media platforms, such as Twitter, Facebook, YouTube, and Weibo, connect people all over the world by allowing them to share content, photographs, and videos, as well as express their first-hand thoughts, leave comments, and follow their friends. Due to the convenience provided by many usable features, social media platforms have been rapidly increasing, attracting consumer involvement with content, and providing an inexpensive communication medium that allows anyone to instantly reach millions of users [1, 13]. However, while these systems give an open forum for individuals to express themselves, they also have a negative side.

In [14], a survey on automatic misogyny detection in social media was conducted. It investigated shared tasks, with the misogyny language being only English, Spanish, and Italian. It was found that the swear word count, swear word presence, sexist slurs presence, hashtag presence, and word length were useful features for automatic misogyny detection. The approaches for automatic misogyny detection were identified, including traditional ML models, namely Support Vector Machine (SVM), Naive Bayes (NB), Logistic Regression (LR), as well as Ensembles and deep neural networks, such as LSTM and CNN.

In [15], a literature review on racist and sexist hate speech detection was conducted, focusing on surveying ML models, exploitable features, and accessible datasets. Five datasets were described. It was discovered that, in this task, feature representation was the most important factor, with LSTM being a good classifier. The most popular methods at the feature representation stage included DL, word embedding, n-gram, word-gram, and TF-IDF. The best results were obtained with DL feature extraction algorithms. However, the present review is not only intended for sexism detection.

While Automatic Sexism Detection can target various media formats, like text, images, audio, video, and multi-modal inputs, the present study's focus is primarily on Automatic Text Sexism Detection, predominantly observed in social media platforms. The task of Automatic Text Sexism Detection is particularly challenging. After 2020, though, there have been new advances in the research carried out on the aforementioned task, with the emergence of many datasets, new feature extraction techniques, text representations, and classification methods. Therefore, it is necessary to conduct a new literature review on Automatic Text Sexism Detection to showcase the latest developments.

The purpose of the present research is to perform a review on the Automatic Sexism Detection, referring to sexism detection within a text, in social media. It examines the datasets and techniques employed during this task for greater A/accuracy and effectiveness to be achieved. The limitations and future directions of Automatic Sexism Detection will be also discussed.

II. METHODOLOGY

An SLR, following the guidelines provided for SLRs in Software Engineering [16] and PRISMA 2020 [17], was conducted. It started by identifying research questions. Then articles were searched and selected using keywords in the online database websites and applying the inclusion/exclusion criteria. Data were extracted, analyzed, and synthesized. Finally, a thorough discussion was provided and conclusions were drawn. The SLR flow is illustrated in Figure 1.

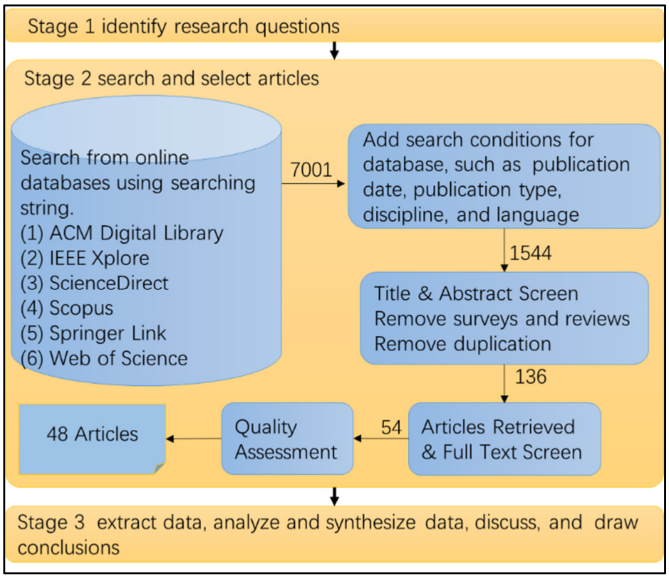
Stage 1 identify research questions
Stage 2 search and select articles
Search from online databases using searching string.
(1) ACM Digital Library
(2) IEEE Xplore
(3) ScienceDirect
(4) Scopus
(5) Springer Link
(6) Web of Science
7001
Add search conditions for database, such as publication date, publication type, discipline, and language
1544
Title & Abstract Screen
Remove surveys and reviews
Remove duplication
136
Articles Retrieved & Full Text Screen
54
Quality Assessment
48 Articles
Stage 3 extract data, analyze and synthesize data, discuss, and draw conclusions

Fig. 1. Systematic literature review flow.

A. *Research Questions*

Automatic Text Sexism Detection is influenced by various factors involving the utilized data resources, preprocessing steps, selection of features representing input text, developed models or algorithms, and the evaluation metrics used for assessing the proposed solutions. To ensure the coverage of these factors, the Research Questions (RQ) are:

RQ1: What datasets did the researchers use to develop and test their Automatic Sexism Detection models or algorithms?

RQ2: What kind of pre-processing was utilized to prepare the data for Automatic Sexism Detection research?

RQ3: What feature and word representation techniques were employed to generate the text representations for the Automatic Sexism Detection model?

RQ4: What algorithms and predictive models were employed for Automatic Sexism Detection?

RQ5: What evaluation criteria were used to evaluate the proposed techniques?

B. *Search Strategy*

To obtain the literature resources related to the research questions, six online databases were selected, namely, ACM Digital Library, IEEE Xplore, ScienceDirect, Scopus, Springer Link, and Web of Science. These databases cover a range of Computer Science academic journals and conferences [18-19]. The keywords were divided in four parts: (1) sexism and its synonyms, (2) detection and its synonyms, (3) social media and its synonyms, (4) automatic, with the same term being used for automated technology. All four conditions must be simultaneously met.

The search terms are: ("sexism" OR "gender discrimination" OR "sexist" OR "sexual discrimination" OR "sexual harassment" OR "misogyny") AND ("detection" OR "classification" OR "classify" OR "categorizing" OR "category" OR "detect") AND ("social media" OR "cyber" OR "digital" OR "internet" OR "online" OR "social network" OR "web") AND ("machine learning" OR "automatic" OR "automatically" OR "data mining" OR "deep learning" OR "lexicon").

The inclusion criteria cover the studies published from 2014 to 17th Sept. 2024, which are written in English, entailing research papers presented at conferences, journals, workshops, etc. The disciplines investigated were Computer Science, Mathematics, Engineering, Social Sciences, Decision Sciences, Multidisciplinary, Neuroscience, etc. The studies must relate to Automatic Text Sexism Detection or to Sexism datasets in social media.

The exclusion criteria include non-English publications, studies handling mainly audio, visual, or multimodal Sexism Detection, informal studies, such as unknown journals or conferences, and irrelevant articles to the research questions. The articles were eliminated from 7001 to 1544 by inclusion/exclusion criteria. After screening the title and abstract, removing surveys, reviews and duplication, 136 articles were left. Subsequently, full text screening was conducted and lastly 54 articles were obtained.

C. *Quality Assessment*

After a selection of 54 papers, the quality of their research articles was evaluated based on Quality Assessment (QA). In Table I, the Six QA questions found in [20] and three out of the 11 QA questions encountered in [21] were combined [21]. The following factors could be used to evaluate the quality scoring: if the condition is fully satisfied, the score is 1; if it is partially met, the score is 0.5; if the criterion is not or nearly not met, the score is 0. Each article receives a quality rating score between 0 and 9. Six articles were removed because they did not meet the evaluation criteria. As a result, there are 48 articles with a score of 5.5 or higher that were reserved.

TABLE I. QUALITY ASSESSMENT CRITERIA

Item	Assessment Criteria
1	Is there a clear statement in research aims?
2	Are the data collection method(s)/datasets adequately described?
3	Are the pre-processing/ features used in the study clearly described?
4	Does the study present a detailed description of the approach (classifier/ techniques)?
5	Does the study present a detailed evaluation of the approach?
6	Is there a comparison with any other approach?
7	Are the results compared with those of previous research?
8	Are the findings clearly stated and supported by the results?
9	Are the research limitations presented?

III. RESULTS

This section covers the SLR results, beginning with a summary of the selected original study. Next, the data collected from the included papers are analyzed to answer the pre-formulated research questions.

A. *Overview of the Selected Studies*

After the quality assessment process, 48 studies remained, which are closely related to this paper's research area. Figure 2 displays the distribution of the primary studies by year. Even though articles from 2014 to 17th Sept. 2024 were gathered, the chosen studies were published between 2018 and 17th Sept. 2024, suggesting that Automatic Sexism Detection began to receive more attention in 2018.

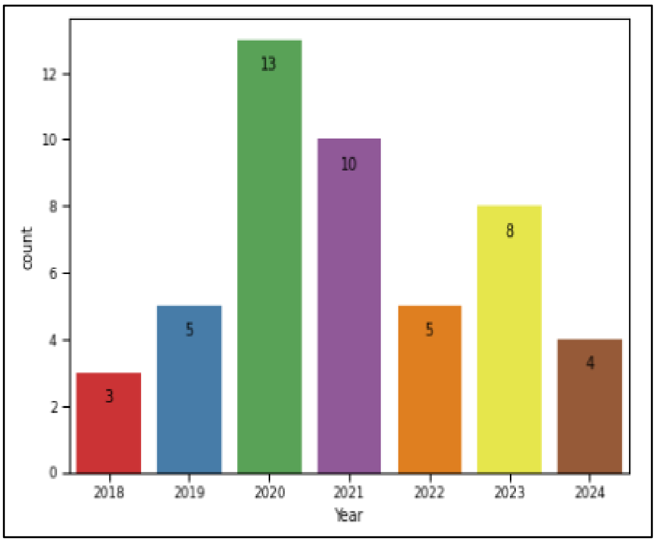

Fig. 2. Number of studies over the years.

B. RQ1

The dataset quality is very important to the training result, as well as the performance of the model. As different languages may have different kinds of pre-processing, word embedding, and modeling techniques, dataset language was investigated.

According to Table II, 45 datasets were used for Automatic Text Sexism Detection. The most utilized languages in the datasets are English (22), Spanish (10), Arabic (3), Bangla (2), and Hindi (2). Chinese, French, Italian, Mexican, Romanian, and Turkish respond to only one dataset, respectively. It was found that there is a lack of corpus in most languages except English and Spanish.

Figure 3 portrays the quantity of datasets annually released in each language. Even though there was an English dataset available in 2012, it had not been used until 2021 [22] for Automatic Sexism Detection. Since 2018, datasets in both English and Spanish have been released almost every year. 2018 also saw the release of the first datasets for Italy and Turkey. The first datasets in Hindi, Bengali, and French were available in 2020. The first releases in Arabic and Mexican were in 2021, while in Chinese and Romanian in 2022. More languages of corpus were developed in recent years, indicating that more academics are becoming interested in Automatic Text Sexism Detection in many languages.

The research centers on sexism identification, sexism type categorization, sexism object identification, which is either individual or generic, sexism place categorization, sexism disclosure type categorization, and so on. There were 12 ways to label text as a binary classification in the selected studies. The following were triple classes which had 5 label manners. In D7, the text was divided into up to 23 categories. The most popular annotated style, misogyny or not misogyny, was applied to 19 datasets. Datasets D11, D18, D40, D42, D43, and D44 were annotated as sexist or not sexist. If not strict, two datasets, D19 and D24, can be also included into the sexist or not sexist class. D13 ws found to be distinguishing masculine from feminine stereotypes. It is different from other annotations, where neither sexism towards females nor gender ignorance occurs.

The datasets employed have more than one set of labels, while most datasets have only one set of labels. Most of the utilized datasets are in English, with a different number of labels. Besides, there are 18 datasets annotated with labels, such as misogyny or not misogyny, which are in English, Hindi, Italian, and Spanish. D42 labeled the corpus to three sets: individual, generic, or non- sexist; and five labels: SA, SCB, MA, SO, and non-sexist. Particularly, D7 was a multi-label task and we categorized the text to 14 or 23 classes.

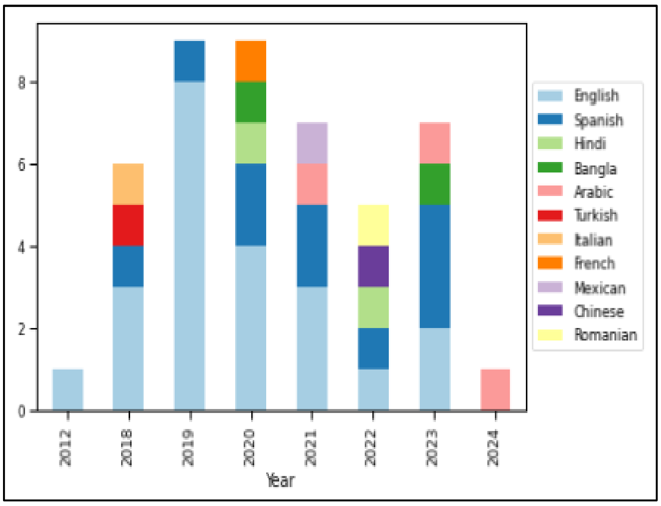

Fig. 3. Dataset in different languages.

English and Spanish were the most widely utilized languages. Misogyny identification and sexist identification are the most used annotating manners. The current corpus can be annotated with different labeling methods for it to be sufficiently utilized. At the same time, more corpus with large datasets and different annotations in different languages needs to be developed. Apart from excluding research on female sexism, the research on masculine sexism should be given more attention while datasets are developed.

C. RQ2

The purpose of pre-processing is to keep features that are related to their labels [64]. The outcomes are significantly impacted by pre-processing [38]. Due to linguistic differences, various languages may require different data pre-processing techniques to produce the best results for Automatic Sexism Detection.

There are more data preprocessing techniques for English and Spanish than for other languages, since there is substantially more Automatic Sexism Detection research in these languages than in any other language. The pre-processing steps, such as lowercase conversion, text removal, tokenization, stemming, and rewriting are frequently employed. Some studies also deployed lemmatization, data augmentation, and breaking large statements into shorter ones. In [10, 56], lemmatization and POS tagger, were, respectively, utilized. During the text removal procedure, stop words, punctuation, URLs, numbers, user mentions, extra spaces, hashtags, emojis, non-alpha-numeric characters, etc. have been removed from the corpus by some researchers. At the same time, some researchers have rewritten URLs, user mentions, hashtags, emojis, number, slang, etc., instead of having removed them. In [29], stop words were removed from the text for the models based on BoW and neural networks, whereas they were maintained for the BERT model.

Overall, the most used pre-processing techniques can be/have been utilized in the present research. When the data are prepared, if a sentence is too long, it may be divided into smaller sentences, suggested in [21, 22, 37, 38]. Data augmentation can be deployed if the data are imbalanced or the corpus must be expanded, as mentioned in [39, 48, 50, 53, 62]. In addition, question marks and exclamation marks may be reserved, as in [41], while punctuation will be eliminated.

D. RQ3

The text is unstructured, and it needs to be converted into structured features for further text classification. Linguistic features, discrete text representation, and distributed text representation are widely used.

In [10, 12, 23, 32, 35, 51, 54, 60], linguistic features were employed. Features, such as the number of characters, words, URLs, punctuation, uppercase letter were prioritized over characteristics, like number of user mentions, sentences, syllables, emojis, hashtags, digits, percentage of Hashtags, misspelled words, length of the text, and average length of the words. In [12, 35, 46, 51, 53], lexicon or sentiment based linguistic features, such as number of swear words, percentage of swear words, sexist slurs, women words, ASF, ASM, and PR, were also utilized.

TF-IDF, N-grams, and Bag of Words are the most used techniques of discrete representation. One-hot was only employed in [43]. Some researchers combined discrete text representation with linguistic features, such as TF-IDF lexical vector [8], bag of hashtags, and bag of emojis [12].

Distributed representation is also called "word representation" or "word embedding", which was first proposed in 1986. In natural language processing, a group of language modeling and feature learning techniques, known as word embedding, are deployed to map words or phrases from the lexicon to vectors of real numbers. BERT and GloVe are the most widely employed distributed representation techniques, with ELMo, Word2Vec, and FastText following. Word-vectors were also trained with LSTM [27] and Bi-directional Long Short-Term Memory (BiLSTM) [22].

Overall, BERT, GloVe, TF-IDF, and N-grams were the most used features and word representation techniques for the selected articles' best performance model to be constructed. Even though other techniques did not appear as much as the previous four techniques, they also contributed to the optimal performance model. In [60], text quality was evaluated using FleschKincaid Grade Level and Flesch Reading Ease scores, having taken text quality as a feature. In [40, 45], feature selection was performed through the utilization of different techniques to realize dimensionality reduction.

E. RQ4

Many methodologies available for Automatic Sexism Detection, include lexicon-based approaches, traditional ML approaches, DL approaches, and hybrid methods.

It was found that the most common approach for Automatic Sexism Detection is DL, 47%, n=27, followed by traditional ML, 31%, n=18, and the Hybrid method, 22%, n=13. Compared to the traditional ML and DL, lexicon-based models were not the highest performing ones in the selected articles. However, they were used as a comparation model in [65]. Lexicon-based methods may come in rather handy in situations when a supervised approach cannot be trained on a big enough dataset [44].

BERT, CNN, RNN, LSTM, BiLSTM, Gated Recurrent Unit (GRU), and their combination constituted the most widely utilized DL method in Automatic Sexism Detection. The most popular DL model deployed to get the best results, was BERT. It should be noted that different languages have their own pre-trained BERT, for example, uncased Bert [26] and Bert-cased [12] for English, Bert-base [6] for French, AraBert [57] for Arabic, AlBerto [61] for Italian, BanglaBertBase and SahajBert [56] for Bangla, BETO [29, 49, 51, 52], BERTIN, RoBERTuito, and MarIA [53] for Spanish, BERT, BERT-wwm, and RoBERTa [8] for Chinese, while there is also the Multilingual case [6] and BERT-multi-cased [12] for multilingual. Some researchers combined BERT with other DL techniques, being also regarded as DL approaches. CNN [33, 34], LSTM [27, 40], and BiLSTM [34] were, respectively, used along with BERT, having reached the best performance, respectively.

Traditional ML methods, such as SVM, LR, eXtreme Gradient Boosting (XGBoost), Random Forest (RF), Light Gradient Boosting Machine (lightGBM), and NB+LR Vote were the models which achieved the best performance. The more excellent ones were SVM and LR. However, most of the research did not employ DL approaches during the comparison procedure. Hybrid approaches were the combination of the lexicon method, traditional ML method, and DL method. Especially, lexicon was widely utilized in [8, 10, 12, 23, 35, 46, 51, 53]. Additionally, most research combined lexicon with DL methods, such as BERT, RoBERTa, and BiLSTM, having demonstrated that the hybrid approach combined with lexicon could improve performance. Overall, DL approaches usually outperform lexicon approaches and traditional ML approaches. A hybrid method, especially one combined with lexicon, may improve model performance. The voted method in [33, 44] can be also employed to improve model performance.

TABLE II. SUMMARY OF AUTOMATIC SEXISM DETECTION

Language / Count	Name / Year	Total Size	ID	Cited by	Best performance	Pre-processing	Feature/Word Representation	Evaluation metrics
English (22)	PAN 2012 task 2: CSP (2012)	20000	D1	[22]	GRU+BiLSTM [22]	Lowercase, removing	GloVe, Word-vectors with BiLSTM	F0.5:0.927, A:0.9727
	AMI-EVALITA (2018)	5000	D2	[5, 12, 23-29]	Attention-based BiLSTM [23]	Lowercase, removing	Word2Vec, debiased word embeddings, tf-idf, lexicons	F1:0.824, A:0.872
	AMI-IBEREVAL (2018)	3977	D3	[5, 12, 25, 29]	SVM [12]	Tokenizer, stemming	Count of swear words and links, presence of swear words, hashtags, sexist slurs and women words	A:0.9132, P:0.8716 R:0.9116, F1:0.8912
	MeToo_Places (2018)	1024	D4	[30]	CNN [30]	-	-	A:0.83
	CSH_SafeCity (2019)	9892	D5	[22, 31, 32]	GRU+BiLSTM [22]	Lowercase, removing	GloVe	A:0.9912, F0.5: 0.925
	MeToo_Severity (2019)	278765	D6	[33]	RF+LR+DT+LSTM Voting Classifier [33]	Lowercase, removing, replacing	-	A:0.89
	Multi-Label Sexism (2019)	13023	D7	[24, 34-36]	Hierarchical neural architecture [24]	Lowercase, removing, breaking	ELMo, GloVe, tBERT	F1:0.777, Fmac:0.705
	HatEval (2019)	6600	D8	[12]	-	-	-	-
	#meToo Sexual Harassment Disclosures (2019)	5117	D9	[37]	Weight Dropped AWD-LSTM [37]	Removing	Pretrained on corpus	A:0.96, P:0.95 R:0.98, F1:0.96
	SIMAH (2019)	10622	D10	[38-42]	GCN [42]	Removing, tokenizer, lemmatizer	Word2Vec, Tf-idf, 45 features, Sentence-Bert	A:0.94
	SWR (2019)	5006	D11	[25]	SVM [25]	Removing, NLTK lemmatizer	Characters n-grams	A:0.8932
	Urban Dictionary (2019)	2285	D12	[23, 43]	Bi-GRU [43]	-	One-hot	A:0.9310 , Se:0.9208 Sp:0.9396
	GSOA (2020)	4333	D13	[44]	-	-	-	-
	QMI (2020)	5000	D14	[27]	LSTM [27]	Replacing, Stemming	Pretrained LSTM	A:0.846
	SI (2020)	14139	D15	[45]	LR [45]	Removing	X^2 Feature Selection	A:0.8533
	TRAC (2020)	4263	D16	[23, 46]	RF [46]	Removing, tokenizer, stemming	Tf-idf, N-gram, VADER lexicon	A:0.96, P:0.91 R:0.96, F1:0.92
	CSH (2021)	about 25k	D17	[22]	GRU+BiLSTM [22]	Lowercase, removing	GloVe, Word-vectors with BiLSTM	F0.5:0.925, A:0.9912
	EXIST (2021)	3436	D18		-	-	-	-
	EN-SWS (2021)	1142	D19	[23, 47]	GloVe+BiLSTM+Attention [47]	Removing, Replacing	GloVe	P:0.84, R:0.93 F1:0.88
	Lyrics (2023)	24234	D20	[29]	-	-	-	-
	SgPh (2023)	4240	D21	[29]	-	-	-	-
	EDOS (2022)	11398	D44	[48]	SVM [48]	Removing, Replacing, Over sampling	Tf-idf	A:0.9464, P:0.95 R:0.85, F1:0.95
Spanish (10)	AMI-IBEREVAL (2018)	4138	D22	[5, 10, 12, 28, 29, 49]	BETO(Bert) [49]	-	-	A:0.846, P:0.7964 R:0.867, F1:0.8302
	HatEval (2019)	6600	D23	[23, 44, 50-52]	Bert+LR [50]	Data augmentation	Bert	A:0.86, P:0.87 R:0.90, F1:0.90
	MeTwo (2020)	3600	D24	[10]	mBERT (text features) [10]	Lowercase, removing, replacing, tokenizer, stemming	Word2Vec, length of the tweet, bert	A:0.74, F1:0.64 R:0.66, P:0.63
	MisoCorpus-(2020)	7682	D25	[51]	SVM [51]	Lowercase, removing, replacing	LF, AWE	A:0.85175
	MLAS (2021)	7191	D26	[50]	Bert+LR [50]	Data augmentation	Bert	A:0.84, P:0.83 R:0.89, F1:0.89
	EXIST (2021)	3541	D27	[52, 53]	MTL-TAI [52]	-	BETO(Bert)	A:0.809
	MisoFB-22 (2022)	2468	D28	[49]	BETO (Bert) [49]	-	-	A:0.8785, P:0.867 R:0.8884, F1:0.8775
	Lyrics (2023)	8856	D29	[29]	-	-	-	-
	SgPh (2023)	2822	D30	[29]	-	-	-	-
	VAW (2023)	7100	D31	[54]	XGBoost [54]	Removing, tokenizer, stemming, SMOTE	Chi^2	A:0.9845232 P:0.9845232
Hindi (2)	TRAC (2020)	3984	D32	[23, 46]	RF [46]	Removing, tokenizer, stemming	Tf-idf, N-gram, VADER lexicon	A:0.93, P:0.90 R:0.93, F1:0.89
	Hindi Sexually Harassing (2022)	8446	D33	[55]	CNN-LSTM [55]	Removing, replacing, tokenizer	-	A:0.9353
Bangla (2)	TRAC (2020)	-	D34	-	-	-	-	-
	Bangla Sexist (2023)	3752	D35	[56]	BERT+LSTM [56]	Removing, POS	BanglaBertBase	A:0.8259, F1:0.8117 CK:0.6498
Arabic (3)	Let-Mi (2021)	7866	D36	[57]	AraBERT [57]	Removing, stemming,	AraBERT	A:0.910

Language / Count	Name / Year	Total Size	ID	Cited by	Best performance	Pre-processing	Feature/Word Representation	Evaluation metrics
						tokenization		
	CASH (2023)	56245	D37	[58]	TCN-BiLSTM [58]	Removing, replacing, tokenizer	FastText, Word2Vec	A:0.9665, F0.5:0.969 AUC:0.969
	CSTFS (2024)	7487	D45	[59]	XGBoost [59]	Removing, replacing	Tf-idf	A:0.86, P:0.87 R:0.86, F1:0.86
Turkish (1)	HSTW (2018)	1288	D38	[60]	SVM [60]	Lowercase, stemmer	N-grams, tf-idf, BoW	F1:0.68
Italian (1)	AMI-EVALITA (2018)	5000	D39	[5, 12, 23, 28, 61]	Attention-based BiLSTM [23]	Lowercase, removing	Word2Vec, debiased word embeddings, tf-idf, lexicons	F1:0.893 A:0.894
French (1)	SDFT (2020)	11834	D40	[6]	BERT [6]	Replacing	FastText, Bert	A:0.790, P:0.767 R:0.759, F1:0.762
Mexican (1)	GVMT (2021)	32500	D41	[62]	DNN [62]	Removing, ROS, tokenizer	BoW	AUC:0.8993 Sp:0.9801, Se:0.8488
Chinese (1)	SWSR (2022)	8969	D42	[8]	Lexion+RoBERTa [8]	Removing	Tf-idf	Fmac:0.780, A:0.804
Romanian (1)	CoRoSeOf (2022)	39245	D43	[63]	SVM [63]	-	N-grams, tf-idf	F1:0.8314, P:0.8307 R:0.8324

Note: A: Accuracy, P: Precision, R: Recall, Sp: Specificity, Se: Sensitivity, Fmac: Macro F1, -: not mentioned or used for training or data augmentation.

F. RQ5

There are 11 evaluation metrics, including A, F1, Precision (P), Recall (R), Macro F1 (Fmac), Micro F1 (Fmic), Confusion Matrix (CM), Area Under Curve (AUC), Cohen Kappa (CK), Weighted F1 (Fw), and F0.5 used in the selected articles to evaluate the performance of the models. For non-multi-label tasks, A, n=31, 76%, is the most used evaluating metric, followed by F1, n=21, 51%, P, n=21, 51%, and R, n=18, 44%. Other metrics are lower than 23%. If Fmac and Fw are regarded as F1, the percentage of F1 used is 68%. For multi-label tasks, A, F1, Fmac, and Fmic were employed to evaluate the models introduced in [21, 22, 37, 38]. In [5], only Fmac was utilized for its multi-label task to be evaluated. As a conclusion, A calculation is the simplest method of evaluation but does not work for imbalanced datasets. Regarding the latter, it is better to take other evaluation methods, like F1 and AUC, into consideration.

IV. DISCUSSION

A. Significant Findings

In terms of datasets, 45 datasets in total emerge from the selected studies. The most common languages detected in the datasets are English and Spanish. Only a few datasets were developed as multi-label tasks, whereas most of them constituted non-multi-label tasks. There are either binary class tasks to identify sexism, sexist/sexism or misogyny, or not. Misogyny categorization is also a prevalent task. There are also researchers having identified sexism spaces, meaning that the sexism target is either individual or general and stereotypes address to the masculine or feminine gender.

Concerning data pre-processing, raw data should be pre-processed to reduce data noise [37]. The most used pre-processing techniques include lowercase, removing stop words, URLs, punctuation and numbers, tokenization, stemming, removing user mentions, extra spaces, hashtags, emojis, non-alpha-numeric characters and special characters, rewriting URLs, user mentions and hashtags, etc.**Error! Reference source not found.** The techniques of breaking into sentences, data augmentation, and lemmatization were also employed. However, the data could be fed into a Bert model without pre-processing, as in [29]. Moreover, question marks and exclamation marks were useful for the classification procedure [41, 66].

As for the features and word representation techniques, the text should be converted to numbers or vectors, which can be used by classification models by extracting linguistics features, discrete representation techniques, and distributed representation techniques. Most widely deployed techniques include Bert, GloVe, N-grams, TF-IDF, FastText, Word2Vec, ELMo, BoW, and several characters. What should be paid more attention is that lexicon and sentiment-based features, which are useful for improving model performance, were extracted [12, 35, 51, 53]. Additionally, some studies combined discrete text representation with linguistic features, such as TF-IDF lexical vector [8], bag of hashtags, and bag of emojis [12]. Furthermore, to achieve the best performance and save time, it is not required to feed all features into the model, so feature selection techniques, such as truncated SVD dimensionality reduction [40] and $\chi 2$ feature selection [45], were adopted to reduce dimension.

To realize Automatic Sexism Detection, the numbers or vectors obtained from feature and word representation should be fed into the classification model. There are lexicon-based, traditional ML, DL, and hybrid approaches. The DL approach, especially Bert, which can acquire information with contextual and grammatical properties [10], usually outperform lexicon-based and traditional ML approaches [6, 10, 37, 43] unless the dataset is too small [44, 65]. The hybrid approaches, combining other approach types, including lexicon and linguistic features, can improve model performance [25, 37, 53]. It should be noted that BERT tokenizer, which is also the most utilized DL-based classification model, can be used for word representation.

There are 11 evaluation metrics employed by the selected studies. For non-multi-label tasks, A, 76%, and F1, 68%, were the most utilized metrics, with F1 being considered more suitable for imbalanced datasets as low R or P are penalized by F1, whereas high A values are maintained when the model performs well in the main classes [38]. However, in addition to A and F1, Fmac and Fmic were utilized [21, 22, 37, 38] for a multi-label task.

B. Limitations of the Existing Studies

Through the analysis of the selected articles, it is observed that there is a lack of benchmark datasets and lexicons for various languages. The annotated dataset size is not big enough to achieve better performance for the supervised methods. Some datasets, such as D24 and D40, are imbalanced affecting the result. It was also found that some dataset links, such as D7 and D37, are not available now. Moreover, some datasets, such as D11 and D24, used a tweet id to retrieve data, but some tweet content has been deleted by tweeter. So, researchers did not use the very same corpus to conduct research.

Besides, there was not a unified paradigm for utilizing linguistic features, suggesting that the best features for Automatic Sexism Detection have not been formed. The split of the trainset and test set was different for the same dataset utilized by different researchers, leading to an unreasonable direct comparison between two research works. Furthermore, most classification models were solely trained on one dataset, with their generalization ability being, thus, questionable. For imbalanced datasets, the evaluation metrics F1 and AUC are more suitable than A. Some proposed models, such as those presented in[10, 37, 43, 49], etc., did not totally outperform other models in all evaluation metrics selected by researchers,

Finally, according to [6, 10, 37, 44, 60], implicit sexism slurs, such as irony, are difficult to be detected. Some sentences without explicit slurs related to sexism are unable to be detected by models.

C. Trends of Automatic Sexism Detection Studies

There are several remarkable trends in the Automatic Sexism Detection that can be potential future works. To address the lack of corpus, more annotation datasets may be developed [61]. Although the traditional data collection and annotation method is the most accurate, it is time-consuming and inefficient. There are some techniques for addressing the lack of corpus. At the pre-processing stage, data augmentation methods, such as back-translation [39], RNN Generate [50], ROS [62], SMOTE [54], over sampling [48], etc., can be employed to tackle this problem. Additionally, data augmentation can also tackle imbalanced datasets [42, 58], while combining data augmentation with monolingual models can enhance sexism detection performance [53].

At the word representation stage, a pre-trained word embedding, like the BERT [10], GloVe [22], or domain-adaptive pretraining method [49] can be applied, as they have been trained on a large scale corpus. At the classification stage, the semi-supervise method [34] and transfer learning [27] can be deployed to alleviate the influence of corpus shortage, as they can utilize the unlabeled data sufficiently. Considering the trends of feature selection and word representation, feature selection techniques are usually employed by traditional ML approaches, while word representation is usually utilized by DL methods. In a DL method, the corpus can be fed to the model without feature selection. However, in the hybrid approach, feature selection and word representation will be combined to get better results. So, an in-depth research is needed to ascertain the impact of features and the best way to combine them for DL [10]. It is also important to investigate how to apply various text representation techniques, such as Bert and GloVe, to get better categorization accuracy [22].

Concerning the sexism detection models, DL-based models, such as BERT, BiLSTM, and GRU, can be applied to Automatic Sexism Detection. They may be combined with linguistic features [8, 10, 23, 35] to improve the results. Novel, intricate, and more feasible DL approaches can be followed for Automatic Sexism Detection, enhancing its performance by incorporating features, boosting strategies, alternative pretrained word embeddings, and more advanced attention mechanisms [6, 47]. A robust model is the goal in [50], while contextual information may be useful for the dependability of sexism detection [53].

In cross-domain and cross-lingual contexts, transfer learning is a viable remedy for the problem of domain adaptation [12, 44]. Another option for classifying sexism is to employ few-shot learning [36]. Researchers can focus on multi-label tasks, such as the level of aggressiveness [5] and discrimination against not only women, but also men in the same sexist corpus [25]. The multi-class tasks, for instance, the intensity of the misogynistic speech [43], should be given greater attention, whilst further research on multilingual and cross-lingual sexism detection and classification should be conducted [8, 34].

Given that implicit sexism without explicit slurs is very difficult to be detected [5, 50, 60], there is an urgent need to address this issue. An Explainable Sexism Detection is required, especially for DL methods [8, 43]. Bayesian [27], SHAP, or LIME analysis [32], and attention mechanisms [47] may be used to increase the explainability of models. Sexual cyberbullying [67] is another increasing concern in the digital environment. Therefore, a browser add-on feature for sexism detection should be developed. This could be a Google Chrome add-on to be integrated into the misogynistic text detection system [56].

V. CONCLUSION

Sexist content has become increasingly prevalent on social media platforms, highlighting the critical need for the development of efficient Automatic Sexism Detection methods. A Systematic Literature Review (SLR) on Automatic Sexism Detection in social media was conducted. Six distinct bibliographic databases yielded a total of forty-four source studies. A thorough study of the included papers was carried out, looking at several variables that affected how well the suggested approaches performed. These included the datasets, preprocessing techniques, chosen features, word representation, prediction model, and evaluation metrics. Furthermore, the significant findings and limitations of the existing studies were discussed, and new directions of research were proposed. It was found that English and Spanish were the most used languages. Most datasets were labeled as either misogynist or not. Lowercase conversion and stop word removal were the most used pre-processing techniques. Bert and GloVe are usually deployed for word representation. A Deep Learning (DL) approach usually provides better performance than traditional Machine Learning (ML) methods. Linguistic features are useful for Automatic Sexism Detection. However, lack of corpus, implicit sexism, etc., limit the development of Automatic Sexism Detection. It is, thus, useful to combine DL with linguistic features for higher performance. Even though Automatic Sexism Detection has gained attention from researchers since 2018, there are still a few issues that need to be resolved by the scientific community. Therefore, it is anticipated that this study will offer a wealth of information about current techniques and resources for Automatic Sexism Detection, inspiring more academics to work in this area.

ACKNOWLEDGMENT

The authors express their gratitude to the College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia for all the provided support. The authors would also like to thank the Journal Support Fund under the Institute of Postgraduate Studies, Universiti Teknologi MARA and Science Research Project of Hebei Education Department (QN2024149).

REFERENCES

[1] M. Mondal, L. A. Silva, and F. Benevenuto, "A Measurement Study of Hate Speech in Social Media," in *28th Conference on Hypertext and Social Media*, Prague, Czech Republic, Jul. 2017, pp. 85–94, https://doi.org/10.1145/3078714.3078723.

[2] L. Plaza *et al.*, "Overview of EXIST 2023: sEXism Identification in Social NeTworks," in *European Conference on Information Retrieval*, Dublin, Ireland, Apr. 2023, pp. 593–599, https://doi.org/10.1007/978-3-031-28241-6_68.

[3] M. Bailey, "Haters: Harassment, Abuse, and Violence Online . By Bailey Poland. Lincoln, NE: Potomac Books, 2016.," *Signs: Journal of Women in Culture and Society*, vol. 43, pp. 495–497, Jan. 2018, https://doi.org/10.1086/693771.

[4] F. Husain and O. Uzuner, "Transfer Learning Approach for Arabic Offensive Language Detection System -- BERT-Based Model." arXiv, Feb. 09, 2021, https://doi.org/10.48550/arXiv.2102.05708.

[5] M. Anzovino, E. Fersini, and P. Rosso, "Automatic Identification and Classification of Misogynistic Language on Twitter," in *International Conference on Applications of Natural Language to Information Systems*, Paris, France, Jun. 2018, pp. 57–64, https://doi.org/10.1007/978-3-319-91947-8_6.

[6] P. Chiril, V. Moriceau, F. Benamara, A. Mari, G. Origgi, and M. Coulomb-Gully, "An Annotated Corpus for Sexism Detection in French Tweets," in *12th Conference on Language Resources and Evaluation*, Marseille, France, Dec. 2020, pp. 1397–1403.

[7] S. Bazzaz Abkenar, M. Haghi Kashani, E. Mahdipour, and S. M. Jameii, "Big data analytics meets social media: A systematic review of techniques, open issues, and future directions," *Telematics and Informatics*, vol. 57, Mar. 2021, Art. no. 101517, https://doi.org/10.1016/j.tele.2020.101517.

[8] A. Jiang, X. Yang, Y. Liu, and A. Zubiaga, "SWSR: A Chinese dataset and lexicon for online sexism detection," *Online Social Networks and Media*, vol. 27, Jan. 2022, Art. no. 100182, https://doi.org/10.1016/j.osnem.2021.100182.

[9] L. Richardson-Self, "Woman-Hating: On Misogyny, Sexism, and Hate Speech," *Hypatia*, vol. 33, no. 2, pp. 256–272, Apr. 2018, https://doi.org/10.1111/hypa.12398.

[10] F. Rodriguez-Sanchez, 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, Jan. 2020, https://doi.org/10.1109/ACCESS.2020.3042604.

[11] A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, "A Unified Deep Learning Architecture for Abuse Detection," in *11th ACM Conference on Web Science*, Boston, MA, USA, Jul. 2019, pp. 105–114, https://doi.org/10.1145/3292522.3326028.

[12] E. W. Pamungkas, V. Basile, and V. Patti, "Misogyny Detection in Twitter: a Multilingual and Cross-Domain Study," *Information Processing & Management*, vol. 57, no. 6, Nov. 2020, Art. no. 102360, https://doi.org/10.1016/j.ipm.2020.102360.

[13] S. Alshamrani, "Detecting and Measuring the Exposure of Children and Adolescents to Inappropriate Comments in YouTube," in *29th ACM International Conference on Information and Knowledge Management*, Oct. 2020, pp. 3213–3216, https://doi.org/10.1145/3340531.3418511.

[14] E. Shushkevich and J. Cardiff, "Automatic Misogyny Detection in Social Media: A Survey," *Computacion y Sistemas*, vol. 23, no. 4, pp. 1159–1164, Dec. 2019, https://doi.org/10.13053/cys-23-4-3299.

[15] O. Istaiteh, R. Al-Omoush, and S. Tedmori, "Racist and Sexist Hate Speech Detection: Literature Review," in *International Conference on Intelligent Data Science Technologies and Applications*, Valencia, Spain, Oct. 2020, pp. 95–99, https://doi.org/10.1109/IDSTA50958.2020.9264052.

[16] [16] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," Tech. Rep. EBSE 2007-001, Keele Univ. Durham Univ. Jt. Rep., 2007.

[17] M. J. Page *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *Systematic Reviews*, vol. 10, no. 1, Mar. 2021, Art. no. 89, https://doi.org/10.1186/s13643-021-01626-4.

[18] A. Cavacini, "What is the best database for computer science journal articles?," *Scientometrics*, vol. 102, no. 3, pp. 2059–2071, Mar. 2015, https://doi.org/10.1007/s11192-014-1506-1.

[19] R. Obiedat, D. Al-Darras, E. Alzaghoul, and O. Harfoushi, "Arabic Aspect-Based Sentiment Analysis: A Systematic Literature Review," *IEEE Access*, vol. 9, pp. 152628–152645, Jan. 2021, https://doi.org/10.1109/ACCESS.2021.3127140.

[20] N. A. S. Abdullah and N. I. A. Rusli, "Multilingual Sentiment Analysis: A Systematic Literature Review," *Pertanika Journal of Science and Technology*, vol. 29, no. 1, pp. 445–470, 2021, https://doi.org/10.47836/pjst.29.1.25.

[21] A. H. Alamoodi *et al.*, "Multi-perspectives systematic review on the applications of sentiment analysis for vaccine hesitancy," *Computers in Biology and Medicine*, vol. 139, Dec. 2021, Art. no. 104957, https://doi.org/10.1016/j.compbiomed.2021.104957.

[22] N. A. Hamzah and B. N. Dhannoon, "The Detection of Sexual Harassment and Chat Predators Using Artificial Neural Network," *Karbala International Journal of Modern Science*, vol. 7, no. 4, pp. 301–312, Dec. 2021, https://doi.org/10.33640/2405-609X.3157.

[23] A. Rahali, M. A. Akhloufi, A.-M. Therien-Daniel, and E. Brassard-Gourdeau, "Automatic Misogyny Detection in Social Media Platforms using Attention-based Bidirectional-LSTM," in *IEEE International Conference on Systems, Man, and Cybernetics*, Melbourne, Australia, Oct. 2021, pp. 2706–2711, https://doi.org/10.1109/SMC52423.2021.9659158.

[24] P. Parikh, H. Abburi, N. Chhaya, M. Gupta, and V. Varma, "Categorizing Sexism and Misogyny through Neural Approaches," *ACM Transactions on the Web*, vol. 15, no. 4, Mar. 2021, Art. no. 17, https://doi.org/10.1145/3457189.

[25] S. Frenda, B. Ghanem, M. Montes-y-Gómez, and P. Rosso, "Online Hate Speech against Women: Automatic Identification of Misogyny and Sexism on Twitter," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 5, pp. 4743–4752, Jan. 2019, https://doi.org/10.3233/JIFS-179023.

[26] J. M. Coria, S. Ghannay, S. Rosset, and H. Bredin, "A Metric Learning Approach to Misogyny Categorization," in *5th Workshop on Representation Learning for NLP*, Online, Jul. 2020, pp. 89–94, https://doi.org/10.18653/v1/2020.repl4nlp-1.12.

[27] M. A. Bashar, R. Nayak, and N. Suzor, "Regularising LSTM classifier by transfer learning for detecting misogynistic tweets with small training set," *Knowledge and Information Systems*, vol. 62, no. 10, pp. 4029–4054, Oct. 2020, https://doi.org/10.1007/s10115-020-01481-0.

[28] S. Lazzardi, V. Patti, and P. Rosso, "Categorizing Misogynistic Behaviours in Italian, English and Spanish Tweets," *Procesamiento del Lenguaje Natural*, vol. 66, pp. 65–76, Mar. 2021, https://doi.org/10.26342/2021-66-5.

[29] R. Calderon-Suarez, R. M. Ortega-Mendoza, M. Montes-Y-Gomez, C. Toxqui-Quitl, and M. A. Marquez-Vera, "Enhancing the Detection of Misogynistic Content in Social Media by Transferring Knowledge From Song Phrases," *IEEE Access*, vol. 11, pp. 13179–13190, 2023, https://doi.org/10.1109/ACCESS.2023.3242965.

[30] A. Khatua, E. Cambria, and A. Khatua, "Sounds of Silence Breakers: Exploring Sexual Violence on Twitter," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Barcelona, Spain, Aug. 2018, pp. 397–400, https://doi.org/10.1109/ASONAM.2018.8508576.

[31] P. Yan, L. Li, W. Chen, and D. Zeng, "Quantum-Inspired Density Matrix Encoder for Sexual Harassment Personal Stories Classification," in *IEEE International Conference on Intelligence and Security Informatics*, Shenzhen, China, Jul. 2019, pp. 218–220, https://doi.org/10.1109/ISI.2019.8823281.

[32] V. Madaan, S. K. Das, P. Agrawal, C. Gupta, and D. Goel, "Fusion of ML models to Identify Sexual Harassment Cases," in *International Conference on Computing Sciences*, Phagwara, India, Dec. 2021, pp. 260–264, https://doi.org/10.1109/ICCS54944.2021.00058.

[33] F. H. A. Shibly, U. Sharma, and H. M. M. Naleer, "Automatic Detection of Online Hate Speech Against Women Using Voting Classifier," in *6th International Conference on Innovative Computing and Communication*, New Delhi, India, Feb. 2023, pp. 735–745, https://doi.org/10.1007/978-981-19-2821-5_62.

[34] H. Abburi, P. Parikh, N. Chhaya, and V. Varma, "Fine-Grained Multi-label Sexism Classification Using a Semi-Supervised Multi-level Neural Approach," *Data Science and Engineering*, vol. 6, no. 4, pp. 359–379, Dec. 2021, https://doi.org/10.1007/s41019-021-00168-y.

[35] P. Parikh *et al.*, "Multi-label Categorization of Accounts of Sexism using a Neural Framework," in *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, Nov. 2019, pp. 1642–1652, https://doi.org/10.18653/v1/D19-1174.

[36] H. Abburi, P. Parikh, N. Chhaya, and V. Varma, "Multi-task learning neural framework for categorizing sexism," *Computer Speech & Language*, vol. 83, Jan. 2024, Art. no. 101535, https://doi.org/10.1016/j.csl.2023.101535.

[37] A. Ghosh Chowdhury, R. Sawhney, P. Mathur, D. Mahata, and R. Ratn Shah, "Speak up, Fight Back! Detection of Social Media Disclosures of Sexual Harassment," in *Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, Minneapolis, MN, USA, Jun. 2019, pp. 136–146, https://doi.org/10.18653/v1/N19-3018.

[38] M. Saeidi, S. B. da S. Sousa, E. Milios, N. Zeh, and L. Berton, "Categorizing Online Harassment on Twitter," in *Machine Learning and Knowledge Discovery in Databases*, Wurzburg, Germany, Sep. 2019, pp. 283–297, https://doi.org/10.1007/978-3-030-43887-6_22.

[39] C. Karatsalos and Y. Panagiotakis, "Attention-Based Method for Categorizing Different Types of Online Harassment Language," in *Machine Learning and Knowledge Discovery in Databases*, Wurzburg, Germany, Sep. 2019, pp. 321–330, https://doi.org/10.1007/978-3-030-43887-6_26.

[40] F. S. F. Pereira, T. Andrade, and A. C. P. L. F. de Carvalho, "Gradient Boosting Machine and LSTM Network for Online Harassment Detection and Categorization in Social Media," in *Machine Learning and Knowledge Discovery in Databases*, Wurzburg, Germany, Sep. 2019, pp. 314–320, https://doi.org/10.1007/978-3-030-43887-6_25.

[41] M. Bugueno and M. Mendoza, "Learning to Detect Online Harassment on Twitter with the Transformer," in *Machine Learning and Knowledge Discovery in Databases*, Wurzburg, Germany, Sep. 2019, pp. 298–306, https://doi.org/10.1007/978-3-030-43887-6_23.

[42] M. Saeidi, E. Milios, and N. Zeh, "Graph Convolutional Networks for Categorizing Online Harassment on Twitter," in *20th IEEE International Conference on Machine Learning and Applications*, Pasadena, CA, USA, Dec. 2021, pp. 946–951, https://doi.org/10.1109/ICMLA52953.2021.00156.

[43] T. Lynn, P. T. Endo, P. Rosati, I. Silva, G. L. Santos, and D. Ging, "A Comparison of Machine Learning Approaches for Detecting Misogynistic Speech in Urban Dictionary," in *International Conference on Cyber Situational Awareness, Data Analytics And Assessment*, Oxford, UK, Jun. 2019, pp. 1–8, https://doi.org/10.1109/CyberSA.2019.8899669.

[44] F.-M. Plaza-Del-Arco, M. D. Molina-Gonzalez, L. A. Urena-Lopez, and M. T. Martin-Valdivia, "Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies," *ACM Transactions on Internet Technology*, vol. 20, no. 2, Nov. 2020, Art. no. 12, https://doi.org/10.1145/3369869.

[45] A. Karami, S. Swan, and M. F. Moraes, "Space identification of sexual harassment reports with text mining," *Proceedings of the Association for Information Science and Technology*, vol. 57, no. 1, 2020, Art. no. e265, https://doi.org/10.1002/pra2.265.

[46] P. D. Kaware and A. B. Raut, "Automatic Detection of Multilingual Misogynistic Content in Social Media Data Based on Machine Learning Approach," in *International Conference on Integrated Circuits and Communication Systems*, Raichur, India, Feb. 2024, pp. 1–7, https://doi.org/10.1109/ICICACS60521.2024.10499136.

[47] D. Grosz and P. Conde-Cespedes, "Automatic Detection of Sexist Statements Commonly Used at the Workplace," in *Trends and Applications in Knowledge Discovery and Data Mining*, Singapore, Singapore, Dec. 2020, pp. 104–115, https://doi.org/10.1007/978-3-030-60470-7_11.

[48] P. Deb *et al.*, "Evaluating Online Sexism Detection: A Comparative Study of Machine Learning Models using the EDOS Dataset," in *9th International Conference for Convergence in Technology*, Pune, India, Apr. 2024, pp. 1–6, https://doi.org/10.1109/I2CT61223.2024.10543680.

[49] D. A. Rodriguez, J. Diaz-Escobar, A. Diaz-Ramirez, and L. Trujillo, "Domain-adaptive pre-training on a BERT model for the automatic detection of misogynistic tweets in Spanish," *Social Network Analysis and Mining*, vol. 13, no. 1, Sep. 2023, Art. no. 126, https://doi.org/10.1007/s13278-023-01128-2.

[50] E. Aldana-Bobadilla, A. Molina-Villegas, Y. Montelongo-Padilla, I. Lopez-Arevalo, and O. S. Sordia, "A Language Model for Misogyny Detection in Latin American Spanish Driven by Multisource Feature Extraction and Transformers," *Applied Sciences*, vol. 11, no. 21, Jan. 2021, Art. no. 10467, https://doi.org/10.3390/app112110467.

[51] J. A. Garcia-Diaz, M. Canovas-Garcia, R. Colomo-Palacios, and R. Valencia-Garcia, "Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings," *Future Generation Computer Systems*, vol. 114, pp. 506–518, Jan. 2021, https://doi.org/10.1016/j.future.2020.08.032.

[52] 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 *International Joint Conference on Neural Networks*, Gold Coast, Australia, Jun. 2023, pp. 1–8, https://doi.org/10.1109/IJCNN54540.2023.10191347.

[53] F. Rodriguez-Sanchez, J. Carrillo-de-Albornoz, and L. Plaza, "Detecting sexism in social media: an empirical analysis of linguistic patterns and strategies," *Applied Intelligence*, vol. 54, no. 21, pp. 10995–11019, Nov. 2024, https://doi.org/10.1007/s10489-024-05795-2.

[54] E. M. A. Stephanie, L. G. B. Ruiz, M. A. Vila, and M. C. Pegalajar, "Study of violence against women and its characteristics through the application of text mining techniques," *International Journal of Data Science and Analytics*, vol. 18, no. 1, pp. 35–48, Jun. 2024, https://doi.org/10.1007/s41060-023-00448-y.

[55] T. Jain *et al.*, "Detection of Sexually Harassing Tweets in Hindi Using Deep Learning Methods," *International Journal of Software Innovation*, vol. 10, no. 1, pp. 1–15, Jan. 2022, https://doi.org/10.4018/IJSI.309110.

[56] S. S. S. Jahan *et al.*, "Deep Learning Based Misogynistic Bangla Text Identification from Social Media," *Computing and Informatics*, vol. 42, no. 4, pp. 993–1012, Dec. 2023, https://doi.org/10.31577/cai_2023_4_993.

[57] A. Y. Muaad *et al.*, "Artificial Intelligence-Based Approach for Misogyny and Sarcasm Detection from Arabic Texts," *Computational*

Intelligence and Neuroscience, vol. 2022, no. 1, 2022, Art. no. 7937667, https://doi.org/10.1155/2022/7937667.

[58] N. Amer Hamzah and B. N. Dhannoon, "Detecting Arabic sexual harassment using bidirectional long-short-term memory and a temporal convolutional network," *Egyptian Informatics Journal*, vol. 24, no. 2, pp. 365–373, Jul. 2023, https://doi.org/10.1016/j.eij.2023.05.007.

[59] F. Alhayan *et al.*, "Detection of cyberhate speech towards female sport in the Arabic Xsphere," *PeerJ Computer Science*, vol. 10, Jun. 2024, Art. no. e2138, https://doi.org/10.7717/peerj-cs.2138.

[60] H. Sahi, Y. Kilic, and R. B. Saglam, "Automated Detection of Hate Speech towards Woman on Twitter," in *3rd International Conference on Computer Science and Engineering*, Sarajevo, Bosnia and Herzegovina, Sep. 2018, pp. 533–536, https://doi.org/10.1109/UBMK.2018.8566304.

[61] A. Muti, F. Fernicola, and A. Barron-Cedeno, "Misogyny and Aggressiveness Tend to Come Together and Together We Address Them," in *Thirteenth Language Resources and Evaluation Conference*, Marseille, France, Jun. 2022, pp. 4142–4148.

[62] G. Miranda, R. Alejo, C. Castorena, E. Rendon, J. Illescas, and V. Garcia, "Deep Neural Network to Detect Gender Violence on Mexican Tweets," in *7th International Workshop on Artificial Intelligence and Pattern Recognition*, Havana, Cuba, Oct. 2021, pp. 24–32, https://doi.org/10.1007/978-3-030-89691-1_3.

[63] D. C. Hoefels, C. Coltekin, and I. D. Madroane, "CoRoSeOf - An Annotated Corpus of Romanian Sexist and Offensive Tweets," in *Thirteenth Language Resources and Evaluation Conference*, Marseille, France, Jun. 2022, pp. 2269–2281.

[64] S. Islam, A. C. Roy, M. S. Arefin, and S. Afroz, "Multi-label Emotion Classification of Tweets Using Machine Learning," in *International Conference on Big Data, IoT and Machine Learning*, Vienna, Austria, Oct. 2021, pp. 705–722, https://doi.org/10.1007/978-981-16-6636-0_53.

[65] J. Cryan, S. Tang, X. Zhang, M. Metzger, H. Zheng, and B. Y. Zhao, "Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods," in *CHI Conference on Human Factors in Computing Systems*, Honolulu, HI, USA, Apr. 2020, pp. 1–11, https://doi.org/10.1145/3313831.3376488.

[66] Z. Zhang, D. Robinson, and J. Tepper, "Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network," in *European Semantic Web Conference*, Heraklion, Greece, Jun. 2018, pp. 745–760, https://doi.org/10.1007/978-3-319-93417-4_48.

[67] S. Unnava and S. R. Parasana, "A Study of Cyberbullying Detection and Classification Techniques: A Machine Learning Approach," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 15607–15613, Aug. 2024, https://doi.org/10.48084/etasr.7621.