IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 January 2023, accepted 14 January 2023, date of publication 25 January 2023, date of current version 22 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3239375

SURVEY

Twitter Hate Speech Detection: A Systematic Review of Methods, Taxonomy Analysis, Challenges, and Opportunities

ZAINAB MANSUR[1,2], NAZLIA OMAR[1], AND SABRINA TIUN[1]

[1]Faculty of Information Science and Technology, Centre for Artificial Intelligence Technology, Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia
[2]Department of Computer Science, Faculty of Sciences, Omar Al-Mukhtar University, Al Bayda, Libya

Corresponding authors: Nazlia Omar (nazlia@ukm.edu.my) and Zainab Mansur (zainabmes04@gmail.com)

This work was supported by Universiti Kebangsaan Malaysia under Grant FRGS/1/2020/ICT02/UKM/02/6 and Grant TAP-K007009.

ABSTRACT Hate speech detection has substantially increased interest among researchers in the domain of natural language processing (NLP) and text mining. The number of studies on this topic has been growing dramatically. Thus, the purpose of this analysis is to develop a resource that consists of an outline of the approaches, methods, and techniques employed to address the issue of Twitter hate speech. This study can be used to aid researchers in the development of a more effective model for future studies. This review focused on studies published over the past eight years, i.e., from 2015 to 2022. This systematic search was carried out in December 2020 and updated in July 2022. Ninety-one articles published within the mentioned period met the set criteria and were selected for this review. From the evaluation of these works, it is clear that a perfect solution has yet to be found. To conclude, this paper focused on presenting an in-depth understanding of current perspectives and highlighted research opportunities to boost the quality of hate speech detection systems. In turn, this helps social networking services that seek to detect hate messages generated by users before they are posted, thus reducing the risk of targeted harassment.

INDEX TERMS Hate speech, classification, automatic detection, twitter, systematic review, natural language processing, social media.

I. INTRODUCTION

Twitter and other social sites have grown exponentially in the past decade. These media promote user anonymity and freedom of speech, thereby driving the growth and transmission of hate speech, as mentioned by [1]. Further, [2] indicated that Twitter is among the most utilized social media site, with 300 million active members monthly. Even though it is popular and relevant, hate speech is frequently spread on Twitter. It is now one of the most widely used social networks for the automatic recognition of in-text hate speech [3], [4], [5] as well as a data source for research into abusive language. Social media is currently witnessing the growing phenomenon of hate speech. This, in turn, creates hostility among users, triggering severe real-life conflicts, and influencing businesses. Social media companies often delete hateful content, preventing them from being published. This study focuses on Twitter social media texts, especially those written in English, as it is a widely known language and the most readily available data source [6]. Whereas manual filtering is inflexible, there is a demand to automate the online hate speech detection process. Non-automated tasks directly affect the reply time, whereas a solution based on computers can perform faster at this task than humans. Consequently, it is imperative to contribute to automated hate speech detection solutions in texts. These facts have motivated research in the natural language processing (NLP) field. There is growing literature concerning hate speech. According to [7], research communities have assigned this challenge as supervised document classification based on NLP and machine learning. Twitter was considered one of the largest social media companies in 2017. It eventually changed the rules of its privacy policy regarding abusive acts. These rules involve all tweets that encourage abuse, harassment, suicide, self-harm,

The associate editor coordinating the review of this manuscript and approving it for publication was Chi-Yuan Chen.

violence, hate, and so forth [8]. Accordingly, researchers have recently increased efforts in identifying hate speech in online content on Twitter. However, only a small dataset is available in languages other than English. It is worth mentioning that English is the most prevalent spoken language globally. It is also the main focus in the detection of hate content. However, it is difficult to recognize hate speech as there are variations to its concept. The most generally utilized word for this phenomenon is hate speech, which is a legal phrase in many countries [7]. In the literature, hate speech is defined in many different ways. Reference [9], defined hate speech based on an analysis of various descriptions available in the literature on this topic: ''Hate speech is a language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics, such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or others, and it can occur with different linguistic styles, even in subtle forms or when humor is used.''. Examples of hate speech on Twitter are as follows: ''Twitter user Pu**y a** ni**a'', and ''You hate football you are a fa**ot.'' [10]. In the last few years, many methods have been developed to address hate speech detection, which is well ahead of their strategies. However, the evaluations are mostly targeted at detecting non-hate content in contrast to identifying and classifying hateful ones [1]. Most of these efforts are still facing some challenges in reaching a feasible solution as the language in social media is developing rapidly [9]. Thus, an in-depth understanding of the current literature on the subject matter is necessary. Although the identification of hate speech has been evolving for a number of years, this field suffers from the lack of a systematic literature review (SLR). SLR papers are essential to facilitate the attainment of the latest updates, such as open issues and research gaps on a precise theme.

Governments and social media have highlighted the need to provide a people-friendly environment through the improvement of hate speech detection methods. An advantage of SLR is that it focuses on enhancing hate speech detection methods, which could help the government and social media companies prevent this phenomenon, as shown in Figure 1. There are various aspects to studying hate speech detection by Twitter data. This review focuses on multiple studies using Twitter in English as the data source. This study, therefore, offers a rich understanding of state-of-the-art methods and recommendations for future work from an extensive analysis of the topic. A systematic review of the computational methods to hate speech detection is provided. The existing challenges are discussed, and the remaining challenges are highlighted for more research opportunities. This paper makes the following noteworthy contributions:

A comprehensive study of the identification of hate speech within social media posts on Twitter;

- Detailed analysis and synthesis of existing studies in this research field;
- Outline of frequently reported problems;
- Creation of taxonomies from the analysis of the literature;
- Outline some significant obstacles to the research and the recognition of potential future trends;

The rest of this paper is arranged into six sections. Section II presents the related works. The methods and procedures for the SLR are explained in Section III, including the study questions, search strategy, and selection criteria. Section VI presents the analysis concerning the research questions, whilst Section V presents the discussion and future work for open issues. Finally, Section VI concludes the work.

II. RELATED WORK

This section highlights the current survey and review articles, focusing on the importance of the contributions of this work. Some reviews and surveys were found about hate speech detection issues, such as by [8], [9], [11], [12], [13], and [14]. In the research conducted by [9], a systematic mechanism for reviewing existing works on hate speech detection from an informatics perspective was applied. It is considered the second survey on this topic after that of [7], which provided a short overview of hate speech detection within NLP. According to [7], the survey is relatively brief and primarily focuses on feature extraction. The survey by [9] provided a comparison of hate speech to other similar forms, a summary of statistics on detection methods, a discussion of the terminologies needed to study hate speech, and the features involved in this domain. Later on, they concentrated on bullying research. They described several English datasets and existing challenges, involving different social media platforms, with less than 20 papers focusing on hate speech. In another study by [8], a more reliable, accurate, and comprehensive classification of anger-linked social media messages for detecting hate speech was established. This will help ensure proper classification because anger eventually leads to extensive participation in hate crimes. In another study by [12], the researchers attempted to review six various hate speech detection models on a variety of social media sites. The methods used were based on the NLP, data mining, machine learning domains, and the variations between these methods were discussed. In a further work by [11], a brief review was conducted on the use of state-of-the-art NLP techniques such as dictionaries, bag-of-words, and n-gram to automatically detect hate speech on online social media sites. Moreover, the study by [14] offered a review of methods for recognizing misogyny in social media, particularly on Twitter. The approaches included standard machine learning and deep learning methods. Furthermore, the findings considered different languages, including English. In a recent review article by [13], the authors employed machine learning techniques to categorize hate speech on Twitter, involving generic metadata de-signs, threshold configurations, and divergences. They also discussed the benefits and weaknesses of individual and integrated machine learning algorithms for the

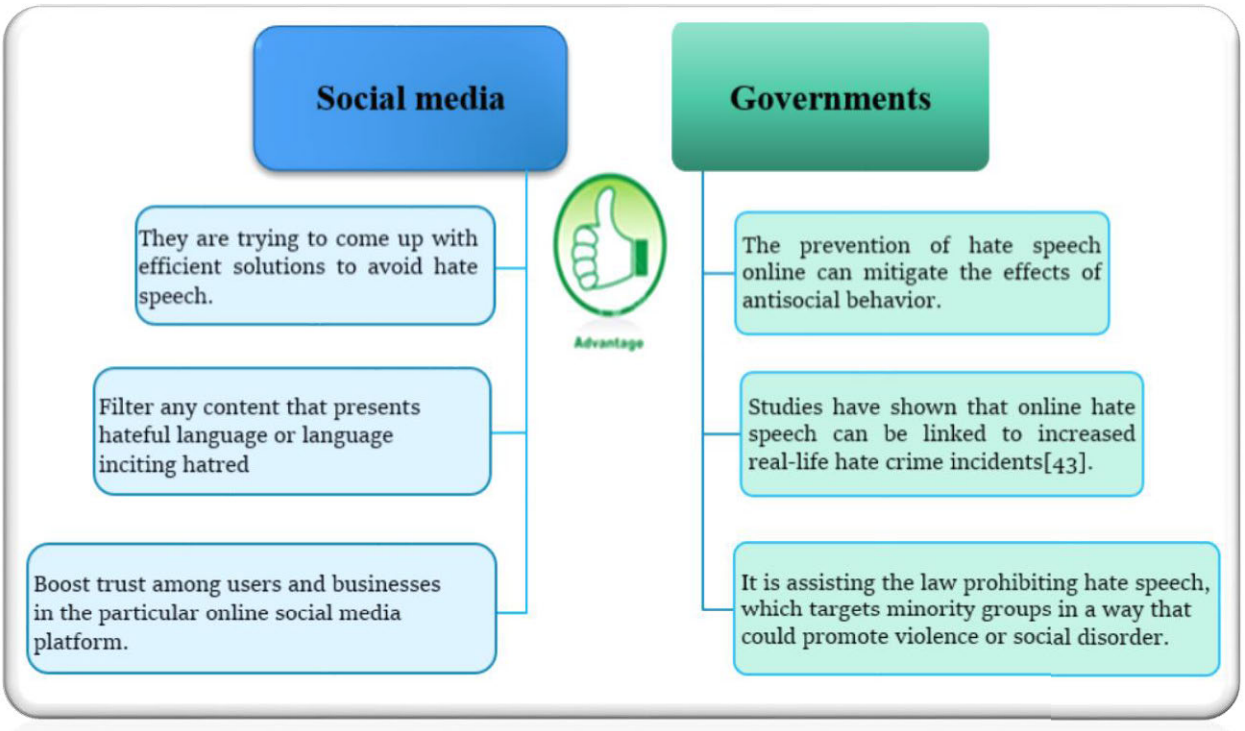

FIGURE 1. Benefits of the study and analysis of hate speech detection methods.

classification process. In addition, they displayed the hate speech benchmark dataset for testing the implementation of the classification paradigm. Even though some surveys and reviews are available on this topic, significant limitations exist. These works partly lack SLR guidelines, up-to-date reviews, and survey studies. Furthermore, these studies are limited in that they did not focus completely on Twitter and, more specifically, on the English language, unlike the survey by [15], which examined the available benchmark datasets used for abusive language and hate speech detection on different social media sites. Their analysis involved the dataset development process, the themes of interest, language coverage, and annotation framework. Although many existing works on hate speech are based on Twitter, previous surveys or reviews lack comprehensive coverage of this particular social site. Twitter ranks among the most frequently used social networks for the automated identification of hate speech in texts [3], [5]. Hence, Twitter has improved connectivity among people worldwide and is a convenient public forum for users. Compared to earlier studies, this paper reviewed a substantially larger number of papers. Additionally, [16] presented a short review of English and non-English literature with some challenges and future research directions. Reference [17] conducted a survey to illustrate the generalizability of current hate speech detection models and explain how hate speech algorithms have an issue in generalizing. Research directions for improving generalization in hate speech detection are discussed. Reference [18] offered an overview of machine learning techniques and techniques for detecting hate speech in online social networks. They explored the primary constituents of hate speech classification using ML algorithms. The failure and capability of each approach are assessed to identify the study gaps and specify the open challenges. Reference [19] discussed different definitions of hate speech, and several challenges were presented concerning data collection and annotation. The authors briefly discussed the differences between nine datasets that used different text languages and platforms. The sources of metadata and the feature selection are also described briefly based on five previous works using machine learning methods. In their paper, a multiple-view SVM model was developed to classify hate speech using three datasets from three different platforms, an interpretation of the model, and an analysis of errors reported. Accordingly, they raise some general challenges. Reference [20] surveyed the racist and sexist class of hate speech methods, focusing on a few factors: data sources, features used, and algorithms of ML. They offered brief descriptions of the text corpus, presented some of the most frequently used approaches for representing features, and made a short comparison between ML models. A short systematic review of the literature was provided by [21], which included articles released earlier than January 2020. Only studies published in English and Indonesian for conferences and journals were considered in their SLR study. A variety of data sources were considered, namely comments from Twitter, Facebook, Wikipedia, Instagram, Online Today, YouTube, and Yahoo. There is only a small finding and a small suggestion by their SLR. More recently,

[22] performed a systematic review of text-based hate speech detection methods and mainly focused on the essential datasets with text-based features and machine learning algorithms. Their collected articles were reviewed according to different themes. They provided three challenge groups and three direction points. Even though their review focused on an English hate speech dataset, our SLR differs from their review in that it provides a more detailed analysis and some taxonomies for our selected studies from a different standpoint.

III. METHODS

This section outlines the procedures followed in this SLR study to provide fair coverage of the reviewed literature. The systematic review process involved several procedures: formulation of the study questions, development of the search string, selection of the study criteria, data extraction, and data synthesis.

A. RESEARCH QUESTION FORMULATION

The goal of the study was to explore recent advances within the topic of hate speech detection and to find, evaluate, analyze, and synthesize the works conducted on the detection of hate speech on Twitter to provide a summary of all the efforts that have been achieved in the study of this subject. The strategies in Kitchenham and Charters (2007) were adopted in conducting the SLR. The current study attempted to answer the following research questions (Qs):

Q1: What were the ratios of the journals and the conferences linked to the databases used by the selected studies?

Q2: What methods and techniques were applied to detect hate speech on Twitter in the selected studies?

Q3: What types of validations were used in this study domain?

Q4: What were the performance metrics commonly used in this study domain?

Q5: What data was available or used for detecting hate speech on Twitter?

Q2: What methods and techniques were applied to detect hate speech on Twitter in the selected studies?

Q3: What types of validations were used in this study domain?

Q4: What were the performance metrics commonly used in this study domain?

Q5: What data was available or used for detecting hate speech on Twitter?

Q6: What were the challenges in this research domain?

Q7: What are the possible future trends in detecting hate speech in English texts on Twitter?

The goal of Q1 was to highlight the ratios for journals and conferences linked to databases used by the selected studies. The purpose of Q2 was to underline the most applied machine learning techniques in the selected papers. The aim of Q3 was to illustrate the common practical validation methods used in the selected articles.

The goal of Q4 was to demonstrate the popular performance measures used in this study area. The aim of Q5 was to discuss the data most used in this study domain, including the types of classes, the sizes of tweets that are privately or publicly available, and the possible links. Q6 aimed at identifying the most critical challenges that the community attempted to tackle in the selected papers. Finally, the goal of Q7 was to highlight potential directions for further investigation in this study domain.

B. REVIEW PROCEDURE

The selected databases exhibited abundant scientific competence in several high-impact research papers, as shown in Figure 2. They were considered to be diverse and reliable for this study topic. Moreover, these databases contained cross-disciplinary studies on explorations by different academic fields, such as the sciences and social science, and were thus deemed to be sufficient for this SLR. The eight digital databases involved in this SLR are presented in Table 1. The search query was structured to find as many papers as possible that were relevant to the subject of interest. The initial stage involved establishing the keywords, where thus far, 'hate speech' is the most frequently used term by the scientific community. This term can be defined as the most popular expression for this type of harmful, user-generated content, and it is even a legal term in many regions [7]. Thus, it is considered to be the more focused term in this study domain. The two other synonymous terms, 'cyber hate' and 'hateful language' have been used in some published studies. The term 'cyber hate' was used in [35], [37], [56], and [69], whilst the term 'hateful language' was used in [12] and [75]. These keywords have been used in connection with the Boolean (AND) and (OR) to form the search query. The effective collection of the search string keywords was as follows: ('cyber hate' OR 'hate speech' OR 'hateful language') along with the main terms of detection ('detection' OR 'recognition' OR 'classification'). The SLR search query is as indicated below:

("cyber-hate" OR "hate speech" OR "hateful language") AND ("detection" OR "recognition" OR "classification").

TABLE 1. Digital databases utilized.

Digital Databases	Category	URL
Web of Science	Multidisciplinary	http://wokinfo.com
ACM Digital Library	Discovery engine	http://dl.acm.org
IEEE Explore	Engineering, technical.	http://ieeexplore.ieee.org
Wiley Online Library	Authentic artworks.	https://onlinelibrary.wiley.com
SpringerLink	Superior quality.	http://link.springer.com
Taylor and Francis	Reliable research.	https://taylorandfrancis.com
Scopus	Biggest database.	https://www.scopus.com
ScienceDirect	Multidisciplinary	https://www.sciencedirect.com/

FIGURE 2. Selected Digital Databases and Libraries.

C. REVIEW PROCEDURE CRITERIA

Particular criteria were set to ensure whether an article was to be included or omitted in the SLR. The inclusion and exclusion criteria were used in the selected studies to identify the most suitable studies. The papers collected were regarded as the most closely related articles without repetition and duplication. The inclusion and exclusion criteria were specifically crafted to exclude those extracted papers that did not fulfill the target of this study, as shown in Table 2 and Table 3.

1) INCLUSION CRITERIA

The inclusion criteria entail articles printed in English and studies published in the last eight years (2015–2022) about hate speech detection. In other words, only studies on the detection of hate speech on Twitter using English datasets were included. The main focus was on hate speech and its categories, including reviews or survey articles. Articles on proposals for methods and methodologies, and comparative and evaluation studies were also included.

English datasets were included. The main focus was on hate speech and its categories, including reviews or survey articles. Articles on proposals for methods and methodologies and comparative and evaluation studies were also included.

2) EXCLUSION CRITERIA

The exclusion criteria were: non-English papers and studies using data from other social media, such as YouTube, Facebook, Instagram, etc. Studies with only textual data were considered; those that include other multimedia (images, videos, and audio) were excluded. Articles with data on multilingual hate speech on Twitter and from multiple platforms were also not included. Studies that did not report on an approach for addressing hate speech-related issues were also excluded, including articles that did not focus on machine learning techniques, NLP, detection and classification models, or performance evaluations. Papers that did not focus on hate speech at all were also excluded.

D. QUALITY ASSESSMENT

Besides the inclusion and exclusion criteria, each article selected was evaluated for quality. Typically, quality biased research results. Moreover, it provides the reader with confidence that each of them fulfils the SLR requirements.

TABLE 2. Inclusion criteria.

Inclusion criteria
Studies published throughout 2015–2022 related to hate speech detection.
Research written in English.
Studies involving hate speech detection from Twitter only use the English dataset.
Reviews or surveys.
Comparative and evaluation studies.
Proposal of methods and methodology.
The main focus is on hate speech and its categories.

TABLE 3. Exclusion criteria.

Exclusion criteria
Studies using other social media, such as YouTube, Facebook, Instagram, etc.
-Only text-based studies are accounted for, and those with any media (image, video, and audio) are not included.
Studies using languages different from English and multilingual hate speech on Twitter (for example, Dutch, etc.) are not included.
Datasets from multiple platforms are not included.
Does not report an approach for addressing hate speech-related issues.
Does not focus on machine learning techniques, NLP, detection and classification models, or performance evaluation.
Quantitative studies.
Those not focused on hate speech at all.

We adopted a set of criteria from [23] for assessing the quality of each of the 91 articles covered by this SLR. These criteria are not meant to criticize works by scholars [23], but rather to offer comfort to readers regarding the coverage of the SLR questions. The adapted criteria are based on the following questions:

QA1: Have the inclusion and exclusion criteria in the research article been described properly?

QA2: Does the literature search appear to have enclosed all the related studies?

QA3: Have the associated data used in the studies been well described?

QA4: Has the study subject/context been clearly defined?

QA1 was given ‘‘yes’’ if the article clearly states the inclusion criteria, ‘’’no’’ if the inclusion criteria are not provided and cannot be presented for the study, and ‘‘not fully’’ if the

inclusion criteria are not explicitly mentioned in the study. QA2 was specified as "'yes" if the authors of the paper conducted their search in more than five electronic libraries at most with references, given "'no" If the number of electronic libraries checked was between three and four but less than five, and "not fully" if only a few digital libraries were searched. QA3 was given "yes" if the detailed data used in the study was clearly stated, given "no" if the data used in the study were not provided and could not be reported, and "not fully" if the data used were not mentioned in the study. QA4 was specified as "'yes" if the researchers clearly state the information about the study, is given "no" f the information about the study cannot be easily accessed, and "not fully" if the information about the study is partially presented. We concentrated on finding papers that provided enough evidence to address our RQs. We can estimate an individual's level of confidence that the article meets the requirements. According to the findings, all 91 studies are eligible for more analysis.

E. SEARCH PROCESS

The scope of this SLR covered all the study papers published over the last eight years, i.e., from January 2015 to December 2022. Firstly, 1692 articles were collected: 235 articles from Scopus, 448 articles from ACM Digital Library, 84 articles from Taylor and Francis, 397 articles from Springer-Link, 30 articles from Wiley Online Library, 205 articles from WOS, 234 articles from IEEE Explore, and 59 articles from Science Direct. After filtering, 56 out of the 1692 articles were found to be duplicates. After reading the title and abstract, 1500 articles out of 1636 were excluded. Furthermore, 66 articles were excluded from the final full-text reading and data extraction process. A final total of 91 articles that met the inclusion criteria based on the search process were reviewed in this SLR study. Throughout this process, certain notes were recorded, which were later transformed into useful insights to form the final structure of this study. The focus of every paper was on the relevant details concerning the journal or conference sources and names, the dataset used, methods and techniques applied, practical validation and evaluation methods, detection issues, and suggestions for further study. Information was collected to answer each of the SLR questions. The search strategy was based on the PRISMA guidelines by [24] as the key element of this study design, as presented in Figure 3. Table 4 presents the complete statistical information on the included and excluded publications and the total number of papers that were stated, screened, eligible, and included in the current SLR study.

TABLE 4. The number of research studies found.

Digital Library	Based on the Query	Based on the title &abstract	Based on a full reading
Science Direct	59	8	5
IEEE Explore	234	57	27
Web of Science	205	68	10
Wiley Online Library	30	5	0
Taylor and Francis	84	11	3
Springer Link	397	89	13
ACM Digital Library	448	259	19
Scopus	235	25	14
Total	1692	522	91

IV. ANALYSIS

As a consequence of the search strategy described above, 91 out of 136 papers, in general, were identified in terms of the inclusion and exclusion criteria. The scope of the search started in 2015, and no publication papers were found in this year, according to the search query. Figure 4 presents the hate speech detection studies in the last seven years. There was a considerable rise in the number of articles in this area in 2018 and 2019. It was also noted that there was a decrease in 2022, mostly because the year is still ongoing, and many studies would have been published at the end of the year. Figure 5 indicates that the largest number of journals were published in 2018, with 13 papers; on the other hand, the largest number of conferences was in 2019, with 18 papers. Researchers used conferences to present and discuss their ideas on this topic. Out of the selected articles, 53% were conference papers and 47% were journal papers.

Q1. What Are the Ratios of the Journals and Conferences Linked to the Databases Used by the Selected Studies?

Most of the published papers were from the IEEE and ACM Digital Library, i. e, 27 (30%) papers and 19 (21%) papers respectively. The ACM Digital Library had four journal studies and 19 (21%) papers respectively. The ACM Digital Library had four journal studies and 15 conference papers, while the IEEE included 4 journal papers and 23 conference papers. In addition, Scopus had 14 (15%) published articles, of which 7 were journal papers and another 7 conference papers. This was followed by Springer, with 13 (14%) published papers, out of which 10 were journal papers and three conference papers, as illustrated in Figure 6. Web of Science had 10 (11%) journal articles. Elsevier had 5 (6%) published articles, all of which were journal papers. Taylors Francis had 3 (3%) journal studies.

Q2. What Were the Methods and Techniques Applied for the Detection of Hate Speech on Twitter in the Selected Studies?

This section outlines most of the methods and techniques that were developed in the selected studies. Figure 7 shows the taxonomy of the methods adopted in the chosen papers on hate speech detection. Based on these methods, we identified three main categories, as follows:

1) Machine learning techniques (ML)

Most classification tasks are performed by machine learning, which typically entails linking certain output vectors to several input vectors. Essentially, machine learning algorithms could be used as single or hybrid algorithms.

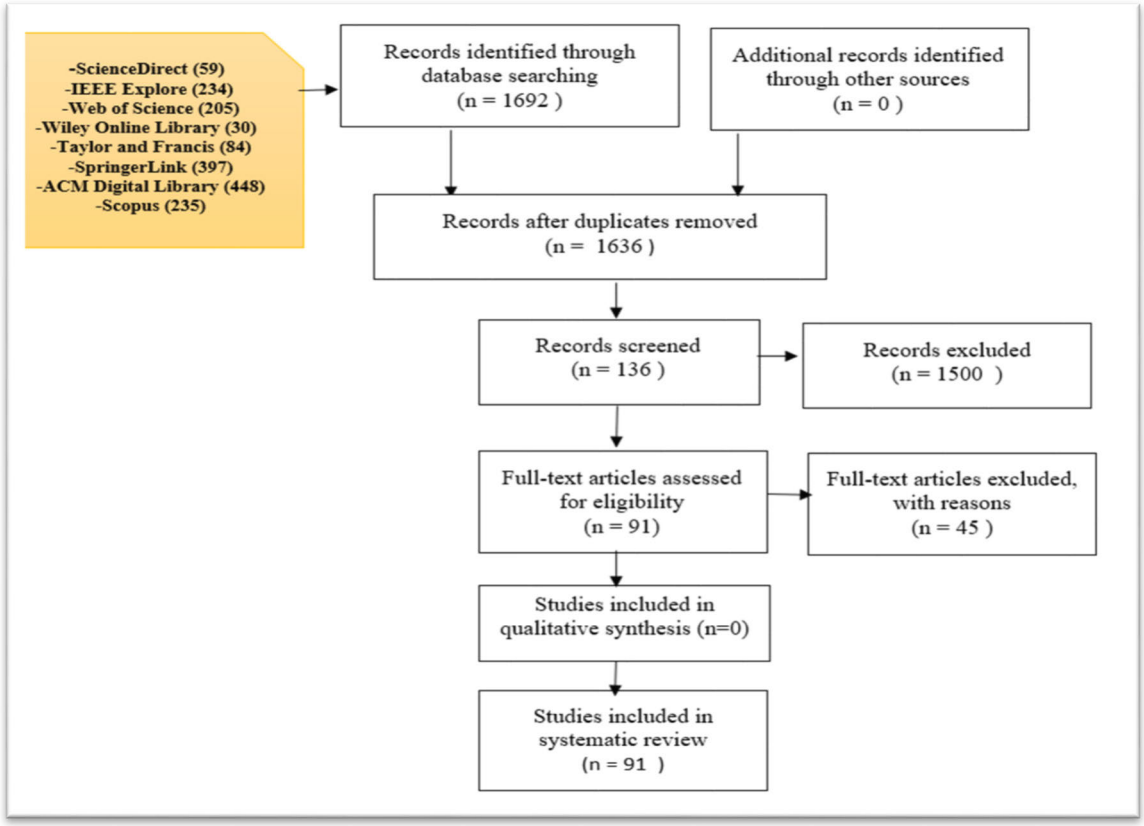

FIGURE 3. PRISMA flow diagram.

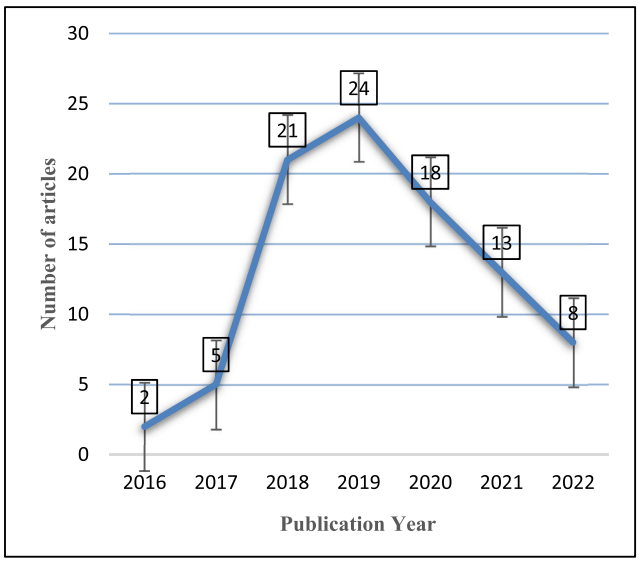

FIGURE 4. The number of publications per year.

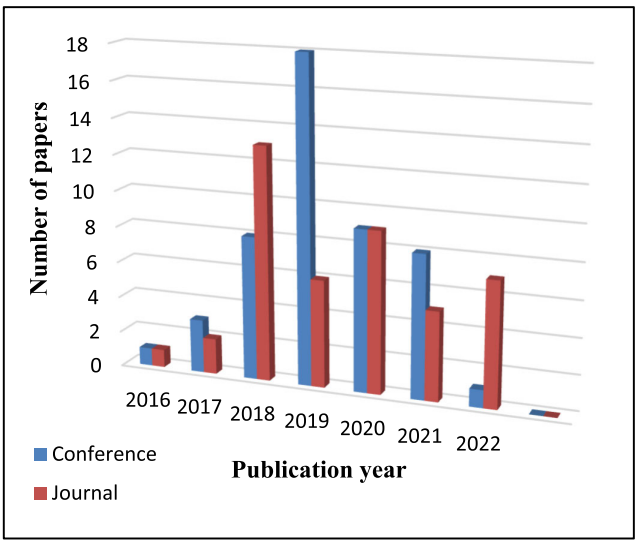

FIGURE 5. The number of publications per year by publication type.

2) Natural Language Processing (NLP)

Adapted NLP techniques enable computer systems to understand and perform natural-language commands. Through these techniques, the text can be read and understood Language processing functionality relies on linguistic knowledge. A number of features can be determined using NLP techniques, including semantic and syntax features. The features used in the hate speech detection methods are the main distinction among the approaches [7].

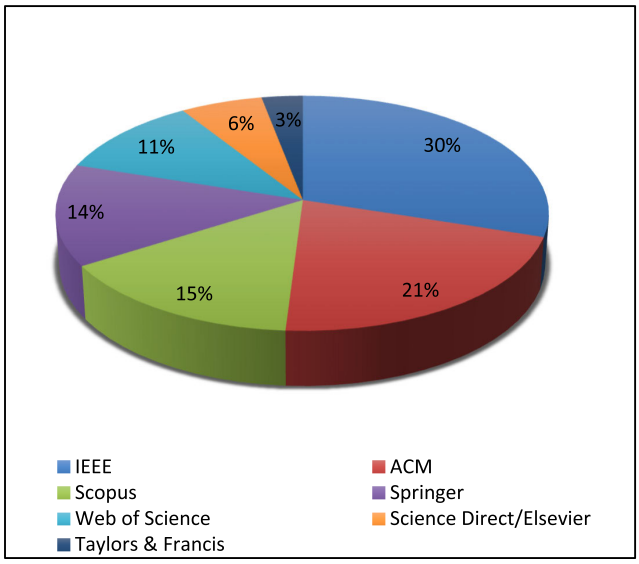

FIGURE 6. Percentage of papers based on the publication databases.

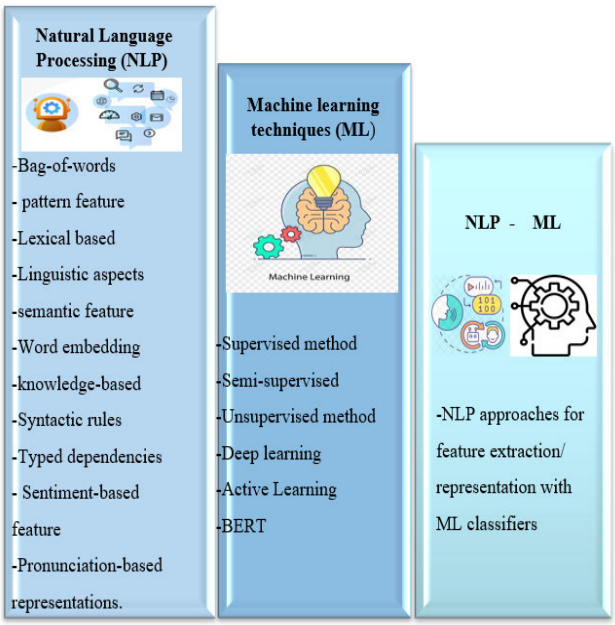

FIGURE 7. Taxonomy of the hate speech detection methods used in the selected studies.

3) ML-NLP

An NLP approach combined with machine learning has several advantages, such as automatic attribute detection [25]. In automatic classification, feature extraction is a crucial step. The machine learning classifier analyses the extracted features from NLP techniques to learn a pattern. In Table 6, we describe the methods and techniques used in the selected studies. In addition, we outline the aim, the data used, the hate speech categories, and the performance measures of each study. Various techniques are involved in hate speech detection, based on the problem under consideration. Additionally, some papers used more than one hybrid, ensemble, or comparative approach. A hybrid machine learning method is a way to combine different algorithms from existing ones or incorporate methods of other fields into a machine learning workflow. Multiple learning algorithms are used in ensemble learning, which provides better predictive outcomes than can be obtained via any one of the fundamental learning algorithms alone. The main difference between the hybrid and ensemble methods is that the hybrid methods predict one single outcome that does not take into account voting, while the ensemble methods operate independently to vote the result. As shown in Table 6, the taxonomy of approaches used in the selected studies is based on different models. It was observed that the most frequently used techniques were the traditional classifiers, such as logistic regression (LR), random forests (RF), support vector machines (SVM), and Naive Bayes (NB). These classifiers, used recently in many comparative studies, have been demonstrated to be useful in the issue of text classification [18]. However, applying suitable features is critical to the success of these traditional classifiers. Aside from choosing the optimal method for feature extraction, it is also essential to consider the architecture of the data [26].

It is possible to improve the quality of classification by combining different feature selection methods [27]. However, deep learning techniques do not rely on handcrafted features. The adoption of deep learning for hate speech detection has been growing significantly since 2017 [28]. Their strength links to the ability to explore data representations suitable for classification [29]. long short-term memory (LSTM) and Convolutional neural networks (CNNs) are well-known deep learning techniques. CNNs are sufficient for extracting context features that offer state-of-the-art outcomes in audio, video, image, and text classification processes. Long short-term memory (LSTM) networks are a unique type of recurrent neural network (RNN) structure that can acquire long-term dependencies by their central memory. The gated recurrent unit (GRU) is another type of RNN, where the gating mechanism of the GRU enables the learning of long-distance connections between words. The bi-directional long short-term memory (BiLSTM) is used to process information in forward and reverse directions. However, different structures of deep learning models gave a different performance in each study. Some studies applied single deep learning techniques, while others combined two forms of deep learning methods, such as CNN and GRU, in a single model. Table 7 shows the deep learning techniques that were used in the selected studies.

In a recent advance, the pre-trained BERT model was usedfor hate speech detection models and obtained superior results [30], [31]. A genetic programming (GP) mode was introduced for identifying hate speech [32]. The model depicts each chromosome as a classifier. Their suggested GP model outscored all advanced systems. However, their model was tested in binary class classification, in which the performance of the GP model in multiclass classification is unknown. Conversely, a few selected studies applied unsupervised learning techniques, which learned patterns in

TABLE 5. A summary table of the methods used in the selected studies.

Ref	Aim of the study	Method used	Method category	Data used	Hate speech categories	F1 score	Accuracy	AUC
[1]	Identify non-hate and hate speech in the textual content.	Deep learning for feature extractors with different classifiers.	ML	Provided by [93][35][77][78]	Various groups	Vary		
[2,4, 74]	Conduct a measurement study about the core targets of hate speech on social media.	Using sentence structure to capture hate	NLP	gathered data	9 categories	-	-	-
[3]	Detect occurrences of coded hate content.	SVM	ML	Collected tweets	Benign, Hateful		79.44%	
[5]	Distinguish a different abusive behavior.	Deep learning with metadata, Textual patterns.	NLP-ML	Provided by [93][77][78]	Various groups	Vary		92% to 98%
[6,43]	Reimplement two states of the art of two prior works.	Language-independent features and supervised machine learning.	NLP-ML	Provided by [93][77][78]	Various groups	Vary		
[10]	Present offensive language and hate speech defense system.	Deep Long Short-term Memory (LSTM).	ML	Provided by [93]	Hate, offensive, neither		90.82%, 89.10% on hate speech and offensive	
[25]	Recognize hate speech in audio and video formats.	Hybrid of NLP and machine learning techniques.	NLP-ML	Provided by [93]	Hate, offensive, neither		98.71%	
[33]	Eliminate harmful content in social media.	Topic modeling technique and unsupervised machine learning.	ML	Provided by [93]	Hate, offensive, neither			
[34]	Reproduce the results of five state-of-the-art models.	Defense schemes and attack replication process.	NLP-ML	Provided by [93][35]	Various groups	0.98 %	0.98%	
[35]	Design hate speech dataset.	Deep neural network combining convolutional and gated recurrent networks.	ML	7 twitter data	Various groups	Vary	-	
[36]	Perform a feature selection investigation.	Feature selection with SVM.	ML	Provided by [93][35][77][87]	Various groups	Vary		
[37]	Given an overview of the salient features.	Deep learning with multi-faceted text representations.	ML	Provided by [93][77][82][87]	Various groups	92.4%		
[38]	Realize various types of abusive language datasets.	Contextual attention for deep learning.	ML	Provided by [93] [77]	Various groups	Vary		
[39]	Determine hateful social media content.	Ensemble of Recurrent Neural Network (RNN).	ML	Provided by [77]	Racism, Sexism, none	0.93%		
[40]	Tackle the issue of costly and inaccurate human annotation.	Emotion analysis and RF.	ML	Provided by [93]	Hate, offensive, neither	0.71%		
[41]	Boost the classification of a variety of hateful and hostile expressions.	Linguistic features.	NLP-ML	Provided by [76]	variety	0.99%		
[42]	Identify hate speech in the context of multitasking.	Fuzzy ensemble approach.	ML	Provided by [76]	Race, Disability, Sexual Orientation			
[43]	Determine the level of individual and interjectional religious hate.	Typed dependencies	NLP-ML	Provided by [76]	Race, sexual orientation, disability	0.68%		
[44]	Differentiate common profanity from hate speech.	Classifier ensembles	ML	Provided by [93]	Hate, offensive, neither	0.79%	79.8%	
		Term						

TABLE 5. *(Continued.)* **A summary table of the methods used in the selected studies.**

[45]	Distinct hate speech from profanity.	representations are based on clustering, skip-grams, and n-grams.	NLP-ML	Provided by [93]	Hate, offensive, neither		80%	
[46]	Selection of misogynous language.	NLP features and ML models.	NLP-ML	Collected data	Misogynous vs no-misogynous	0.38%		
[47]	Detecting misogyny and sexism.	Linguistic features with SVM.	NLP-ML	Provided by [93] [79][80]	Variety of classes		Vary	
[48]	Detecting misogynistic tweets effectually.	Pre-trained on a task-specific to training a CNN.	ML	Collected data	Misogynistic and non- misogynistic	0.76%		0.76%
[50]	Use entirely text-based input.	CNN-LSTM combination, word embeddings, and character n-grams.	ML	Provided by [77]	Sexism, racism, neither	79.24%		
[51]	Detect and distinguish between different strengths of Islamophobia.	Feature selection with six different algorithms.	ML	Collected data	Non- Islamophobic, weak Islamophobic, strong Islamophobic		77.6%	
[52]	Categorize hate speech in social media.	Lexicons and machine learning.	NLP-ML	Provided by [93]	Hate, offensive and neither		80.56%	
[53]	Use DL combined with simple ML for categorizing tweets.	Integrate features extracted from (CNN) trained on word embedding with syntactic and word n-gram features.	ML	Provided by [93]	Racism, Sexism, none.	0.97%		
[54]	To recognize hateful expressions.	Unigrams and patterns. toolkit Weka.	NLP-ML	Provided by [93]	Hate, offensive, neither		87.4% on binary, and 78.4% on ternary classification	
[55]	Analyze how a lack of attention to dialect can lead to bias in labeling.	GloVe vectors and BiLSTM with attention.	ML	Provided by [82]	Abusive, Hateful, Normal, Spam. hate speech, offensive, neither			
[56]	Study the ability of BERT to capture hateful context.	Transfer learning approach.	ML	Provided by [93]	Racism, Sexism, none.	88%		
[57]	Address the class imbalance problem.	Text-based data augmentation.	NLP-ML	Provided by [87]	Sexism, racism, both (racism & sexism), neither	74.1%		
[58]	auto-classification of toxic speech.	Deep learning with embedding representations.	ML	Provided by [93]	Hate, offensive, neither	84.0%		
[59]	Detecting hate content.	Bag of words and TFIDF approach.	NLP-ML	Collected dataset	0 non-hateful, 1 hateful			94.11%
[60]	Require a clean and safe social media network environment.	BIRNN, SVM, LR, LSTM, GRU.	ML	Two Kaggle dataset	Sexist or racist, neither sexist nor racist. sexist or racist, neither	Vary	vary	
[61]	Analyze how hate speech copes with the responses that are used to combat it.	Lexical, linguistic, and psycholinguistic analysis.	NLP-ML	Collected data	Hate tweets counter speech replies tweets.	0.77%		
[62]	Determining how much a particular instance fits each class.	The modified fuzzy method includes two training stages.	ML	Collected data set	Religion, race, disability, and sexual orientation	Vary		

TABLE 5. *(Continued.)* **A summary table of the methods used in the selected studies.**

[57]	Examine current detection models according to pronunciation representation.	Pronunciation-based representation.	NLP-ML	Provided by [93]	Hate, offensive, neither		91% hate 90.8% offensive	
[63]	Examine intra-user and inter-user representation for hate speech detection.	Bidirectional LSTM baseline	ML	Provided by [77]	Racism, Sexism, none		0.77%	
[65]	To cope with the problem of the baseline models.	Word embeddings with LSTM, BiLSTM.	ML	Provided by [93][77]	Various groups		86%	
[66]	Find instances of aggression and hate speech against women.	Bag of Words and Ensemble Classification Model.	NLP-ML	Provided by [79]	Non-misogynous misogynous (Discredit, Sexual harassment Stereotypes, Dominance, Derailing).		79.1% in binary classification	
[67]	Measuring and mitigating unintended bias in ML for misogyny detection.	Synthetic template, e Universal Sentence Encoder.	NLP-ML	Provided by [80][82]	Various groups			Up to 0.72% increase
[68]	Rebuild seven current hate speech methods from the previous studies	LR-char, MLP char, CNN+GRU, LSTM.	ML	Provided by [93][35][76][77]	Various groups	Vary		
[69]	Propose three attacks for the hate speech classifier.	Word splitting, word merging, and character dropping, RF.	NLP-ML	Kaggle dataset	Homophobia, white supremacy, racism, and misogyny.	decreases between 19, 2%		
[70]	A new structure for augmenting hate speech datasets.	Deep generative model.	ML	Provided by [77][78][82]	Various groups			
[71]	Extracting and analyzing comments that convey any curse or meaning. curse or swear.	Latent Semantic Analysis (LSA) and cosine similarity, ensemble LR, SVM, RF.	NLP-ML	Collected data	Hate, offensive and neither		93%	
[72]	Boost the performance of hateful content detection.	Improved N-Gram technique (IN-Gram).	NLP-ML	Provided by [77]	Racism, Sexism none		10– 12% increase	
[74]	explore several transformer-based methods.	Transformer-based method.	ML	Provided by [93]	Hate, offensive, neither	0.75%		
[82]	Annotated dataset of 80 thousand tweets.	Crowdsourcing.	Crowdsource	Collected data	Offensive, abusive, hateful, cyberbullying, aggressive, normal, spam.		79%	
[83]	Evaluate eight classifiers over three feature engineering techniques.	TFIDF, word2Vec, and Doc2Vec feature engineering.	NLP-ML	Provided by [93]	Hate speech, offensive, neither			
[84]	Identify hate and offensive speech.	Conventional ML algorithms and deep learning architectures.	ML	Provided by [93]	Hate, offensive, neither	Vary	Vary	
[85]	Distinguish hate speech from offensive language more clearly.	Typed dependency, SVM.	NLP-ML	Provided by [93]	Hate, offensive, neither	0.48% in hate class		
[86]	Grasp of online hate: aimed at individuals or targeting groups of a common protected characteristic.	Lexical Analysis, Hashtag-based dataset, Key phrase-based dataset.	NLP	Collected data and data provided by [93][77]	Various groups	-		
[89]	Capture multiple hate speech categories.	CNNFastText, LSTMFastText, BERT	NLP-ML	6 public Twitter data	Various groups	Vary	Vary	

TABLE 5. *(Continued.)* **A summary table of the methods used in the selected studies.**

[90]	How can a dimensionality reduction strategy boost the model detection performance	Dimensionality reduction approach with LR.	ML	Provided by [93]	Hate, offensive, neither		83%	
[91]	Identify misogyny: hate speech against women	BOW, Lexical Features, SVM	ML	Provided by [79]	Non-misogynous misogynous (Discredit, Sexual harassment Stereotypes, Dominance, Derailing)	0.30%	0.61%	
[92]	Categorize the tweets as hate speech or not.	Doc2vec and word2vec algorithms using R, a coding language.	ML	Collected data	0 for neutral, 1 for offensive, 2 for hate		74.1%.	0.77%
[94]	Analyze and detect hate speech at a user level.	Semi-supervised node embedding approach.	ML	collect and annotate hateful users	Hateful user, Normal user,	0.67%	90.0%	
[95]	Efficient hateful behavior detection.	Hybrid feature representation approach.	ML	Provided by [93][82]	Various groups	F1 score up to 3.0%		
[96]	Use resampling techniques to tackle unbalanced datasets.	Four resampling methods with SVM, NB, and LR.	ML	Provided by [93]	Hate, offensive, neither	0.95%	91%	
[97]	Expose an LM model while fine-tuning to contexts that capture certain semantic features.	Active Learning, an ensemble classifier	NLP-ML	Collected data,[80], [98],	Various groups	Vary	Vary	
[102]	Data augmentation is the primary emphasis of this research.	Bert, BiLSTM	NLP-ML	[93]	Hate, offensive, neither.	Vary		
[103]	Identify hate speech.	Several feature extractions with various classification algorithms.	NLP-ML	[93],[77],[78]	Non-hateful, hateful	Vary		
[104]	Use syntax heavily in Hate speech detection.	Syntax hateful parse trees, BERT, neural network.	NLP-ML	[93],[77],[82]	Various groups	Vary		
[108]	Identify offensive language and hate speech.	Knowledge Graphs, LSTM model.	NLP-ML	Kaggle dataset	Normal Tweets hate tweets.		97%	
[110]	Identify hate and offensive speech	Preprocessing, grid-search SVM with Tf-IDF.	NLP-ML	[93]	Hate, offensive, neither		93%	
[111]	Identify Twitter hate speech.	Linear SVM Model.	NLP-ML	Kaggle dataset	Hateful, not hateful.	0.75%		
[112]	To boost hate speech identification.	Genetic algorithm.	NLP-ML	HatEval dataset [78]	Hateful, not hateful.	62%		

unlabelled data. For example, [33] used the unsupervised self-organizing map (SOM) algorithm. A SOM is a form of an artificial neural network capable of converting nonlinear and complex statistical relationships within data items into pure linear relationships. However, more details on the performance of unsupervised learning were unavailable in this study. Accordingly, further investigation into the performance of different unsupervised learning techniques in hate speech detection is needed.

Q3. What Types of Validation Methods Were in This Study Domain?

The term validation refers to the test set that is never seen during the training of the model. A model can be more effectively evaluated using data that has not been seen before. This is commonly referred to as a train-test split algorithm to evaluate the approach used. The training data is a set of data used in training (for a machine learning model) to find model parameters. To tune model parameters, the validation set of data is used to provide an unbiased evaluation of a model fitting the training dataset. The test dataset allows a fair assessment of the final model fitted to the training dataset. It was essential to highlight the current validation methods used in this search domain to provide an understanding of the model and to assess the reliability of the generalization. Three validation methods were applied in the selected studies, as presented in Table 8. The first one is the k-Fold

TABLE 6. The taxonomy of approaches used in the selected studies based on different models.

ML model used	Reference.
Single- model	[3][10][28][33][30][32][36][38][43][45][47][48][53][55][56][59][62][63][64][65][67][70][72][82][85][95][96]
Hybrid- model	[1][5][6][33][34][35][37][50][53][57][58][60]
Ensemble-model	[25][39][42][44][45][53][66][90][110]
Comparative-model	[5][6][28][33][34][37][40][41][46][49][51][52][53][54][57][60][68][69][71][74][83][86][87][93][94][96][105][106][107][108]

TABLE 7. Most used deep learning techniques in the selected studies.

Techniques	Studies
LSTM	[5][10][19][34][37][39][48][50][53][60][63][65][68]
LSTM with Attention	[84][89]
CNN-GRU	[1][34][35][37][57][68]
BiLSTM	[31][38][50][64][65][102][110]
CNN	[74]
LSTM-GRU	[1][22][37][48][53][84][86][89]
GRU	[60]
CNN-BiLSTM	[5][60]
CNN-LSTM	[58]
BiLSTM-attention	[50][57]
	[55]

cross-validation method. In this method, the main training set is split into k subsamples. One of the k subsets is used as the test set in each fold, and the remaining k−1 subsets are used as the training set. Sixteen selected papers applied fivefold cross-validation, and 20 studies applied a tenfold cross-validation method. The second method entails splitting the data into training, testing, and development sets. The development set is a sample of the dataset used, and it is used to optimize and evaluate the model during the training process while the model hyper-parameters are being tuned. Five studies applied training-testing-development sets. The third and last method entails splitting the data into training and testing sets, where the training model is validated against the test data. A total of 29 studies used the training-testing sets.

Q4. What Were the Performance Metrics Commonly Used in This Study Domain? The measurement of performance is an essential step in machine learning. Multiple evaluation metrics were used by different models for hate speech detection, depending on the technical aim. It was observed that the most commonly used performance metric was the F1 score and accuracy, which involved 29 and 27 selected studies, respectively, as shown in Figure 8. Accuracy measures cases that have been correctly predicted and are used mainly if all the classes are equally significant. Accuracy is an effective measure of unbalanced classes.

Conversely, the F1 score is appropriate once a balance is required between precision and recall as misclassified instances are better measured by the F1 score. The macro-averages essentially compute the metric independently for every class and then take the average (thereby treating every class the same). In contrast, a micro-average will take into account the contributions of each class to determine the average. The micro-average is preferable if there are potential class imbalances. The area under the curve (AUC) measures how well a classifier can identify different classes. When the AUC is higher, the model is more effective at detecting the positive or negative classes. The macro-averaged F1 scores and the AUC were used in 8 studies. The macro-average F-measure is suitable when the class is extremely unbalanced, while the AUC is used to verify or represent the performance of the multi-classification issue.

TABLE 8. Validation methods in the domain of study.

Methods	Studies
5-fold cross-validation.	[1, 34, 35, 36, 37, 38,90, 92, 92, 93, 94, 95, 96]
10-fold-cross validation	[3, 5, 6, 28, 38, 39, 41, 42,43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53]
Training/devel./ testing.	[54, 55, 56, 57, 58,101,102,110,112]
Training/testing	[10, 25, 35, 39, 48, 59, 60, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 74, 91, 93,97, 104, 105, 106, 107, 108, 109, 111]

Q5. What Data Was Available or Used for the Detection of Hate Speech on Twitter?

The general process of collecting Twitter data involved a few steps, as illustrated in Figure 9. Twitter provides a web-based application programming interface (API) for valuable data scraping and services. The Python module Tweepy is used to link the Twitter API and the precise objects and methods offered by the API. The Tweepy module enables

developers to reach Twitter data, including personal information, timelines, tweets, and retweets. However, the number of tweets that a user can create at any one time is limited [75]. The Twitter API only allows a user to access the last 3,200 tweets on a timeline and restricts the use to a period of one hour. To manage the application of the Twitter API, the user should be aware of the risks associated with the rate limitation.

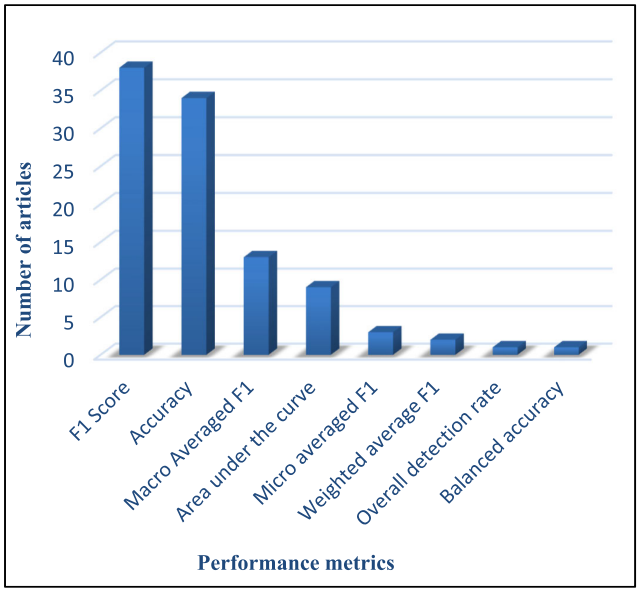

FIGURE 8. Performance metrics in the domain of study.

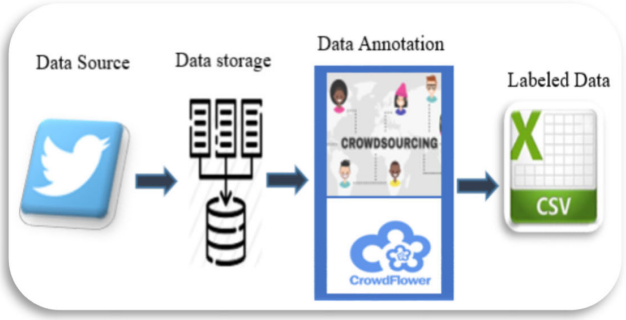

FIGURE 9. General twitter data collection phases.

The amount of time a user must wait after receiving an error grows dramatically with each failed attempt. The tweets collected by the API tool will be stored for the next step of the annotation service. CrowdFlower (CF) is a platform that provides a crowdsourcing service for dataset labeling. The agreement score for CrowdFlower is based on a majority vote from the trusted staff of each annotated class. The final step involves storing the labeled data in CSV files. A comma-separated values (CSV) file is a simple text record containing a sequence of data. Complicated data from an application can be captured to a CSV file, and then exported from the CSV file to another application. The datasets used in the selected studies can be classified as shown in Figure 10 namely private data, publicly available data, Kaggle, and benchmark datasets. These data have different tweet sizes and various types of hate speech categories. In private data, the authors compiled and labeled new posts to evaluate a model for specific issues. The drawback of private data is that it is difficult to compare it with other search findings [9]. However, some private data can be reached upon request from the dataset owner, such as the data compiled by [76]. However, while this study was not in the listed papers, the data collected in this study were used in 4 selected papers; thus, it had to be referred to in the SLR paper. There were some available data, such as the data compiled by [77] and [93], that were more frequently used in this study domain. Reference [77] was also not included in the selected studies. Therefore, it was necessary to mention the data in this SLR. Some benchmark datasets, which can be useful for comparing two different methods, were used in the selected articles. One of the benchmark datasets used in some of the selected papers was SemEval 2019 Task 5 compiled by [78], which is a benchmark English Twitter dataset that holds tweets about hate speech against women and immigrants. Moreover, (AMI) shared tasks, i.e., the IberEval 2018 [79] and EvalIta 2018 [80], which included hate speech tweets against women, were the other benchmark datasets that were used in a few selected papers. Likewise, the TweetEval benchmark dataset by [99] consists of a set of seven NLP tasks, generic criteria for testing, and reference models for measuring the performance of new models. The seven NLP missions include hate detection, emotion detection, emoji detection, sentiment analysis, offensive language detection, stance detection, and a distinct tagged dataset. Another recent Twitter benchmark dataset established by [98], the EAH dataset, includes East-Asian tweets throughout COVID-19. A recent benchmark dataset is made by [113], namely the COVID-HATE dataset. Our selected studies did not include many benchmark datasets mentioned in the original papers. However, they are Twitter datasets used in some of our selected documents. As such, the data were necessary to have in this SLR. On the other hand, this study domain is also considered a collection of hate speech tweets from Kaggle accessible on the Kaggle platform. It holds tweets that represent various topics, including homophobia, white skin, misogyny, and racism. Table 9 shows the data used in this study domain, where some aspects of the data, such as the amount of data (number of tweet instances), type of data classes, data source (the owner of the data), and the sources that used the mentioned dataset, have been highlighted. Figure 11 provides the taxonomy of the data used in the selected study based on the target classes.

A. CHALLENGES BASED ON THE CONTEXT OF HATE SPEECH

1) HATE SPEECH AS A SERIOUS PROBLEM

Social networking sites are willing to identify user-generated hateful posts before release. An ensemble (RNN) with

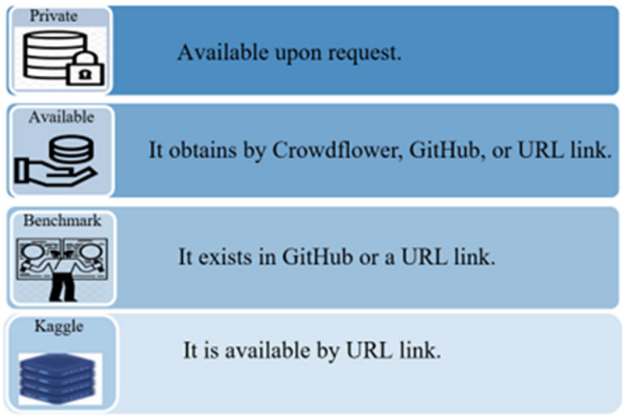

FIGURE 10. Categories of the data sources used in the selected study.

features related to users was implemented for the detection model [39]. Without investigating and identifying hate speech, social networking will not be free of malignant content [33]. In this regard, a common topic modelling technique, such as the Latent Dirichlet Allocation (LDA), was applied with an unsupervised machine learning technique, such as Self-Organizing Maps Additionally, [110] used deep learning to tackle this issue. Likewise, [60] considered hate speech as an unacceptable consequence of free speech and examined different classifiers, namely, BIRNN, SVM, LR, LSTM, and GRU, for hate speech detection. Recently, a genetic algorithm was introduced by [32] as a more reliable and automatic method for identifying and preventing hate speech. Moreover, [112] employed Knowledge Graphs (KGs) to boost hate speech identification. Many methods have been proposed in the literature to tackle this issue. There are still complaints that not enough has been done to address this issue.

2) AN UNDERSTANDING OF ONLINE HATE SPEECH

Hate speech is an overly explicit social networking phenomenon it is crucial to understand it. Reference [2] offered a deeper understanding and an instance of the potential targets of this phenomenon. Furthermore, the study by [81] aimed to understand the proliferation of hate speech on social networking sites, the most popular hated words, a factor of anonymity in hate speech, and the proper categories in all regions. Additionally, [4] outlined how this critical issue emerges online and pointed out that hate speech in the Internet community determining if indicates hate in the offline community. However, determining if a text contains hate speech, even with humans, is not a simple task. There is no universal and individual full agreement on the definition of hate speech.

3) SUBTLE AND UNRELIABLE ANNOTATIONS

Social networking sites provide a rich source of information but are less trustworthy and noisy. Some existing problems associated with hate speech are the subtle instances of Twitter posts and the fact that human annotations can be costly and unreliable. To ensure that an algorithm can critically appraise hate speech features, human annotations must be reliable [77]. It is often difficult to determine if a sentence contains hate or not, mainly if the hate speech is concealed behind sarcasm or if there are no exact words showing hatred, stereotyping, or racism. Moreover, [40] considered suspended account tweets to be a possible source for the retrieval of hateful content. They used emotion analysis for tweets from suspended, active, and natural accounts, and discovered that suspended account users emitted hurtful words that were more descriptive than active account users. Their results revealed that the suspended account tweets could overcome the limitations of using active account tweets. Reference [54] depended on writing patterns, unigrams, and sentiment features to overcome the non-reliability of noisy content. Likewise, as pointed out by [41], by relying on the frequency of actual terms or phrases, the subtle and indirect instances of hate will lead to many false negatives, providing an inaccurate picture of cyber hate patterns. The authors investigated the effectiveness of using linguistic features based on the assertion that using an ''other feature set'' would give a broader meaning to the classifier than words alone. Moreover, due to the challenge of bias in the annotation caused by the dialect and ethnicity, [55] designed tasks that specifically illustrated the inferred dialect in a tweet or the potential ethnicity of the author. Regarding the same concern, [82] tackled the challenge of crowd-sourced workers in differentiating between hate speech and offensive and abusive language. They presented a boosted sampling strategy that preserved an unbiased dataset while providing minority samples with more annotations. Likewise, the lack of sufficiently labelled hate speech data had become an issue in the subject field [56]. The latter focused on transfer learning and examined the ability of BERT to capture hateful contexts within social media posts. Due to considerable ambiguity in existing definitions of hate speech, there was little agreement among the annotators.

4) AMBIGUOUS AND NOISY

Most of the existing methods for categorizing and identifying hate speech concentrate on content shared on online social networks. However, users intend to use typo language to communicate online and distribute their hateful instances to avoid being barred from publishing the post online on social media. Therefore, it would be challenging to gather and annotate hateful speech due to the incompleteness and subjectivity of hate speech. Accordingly, [42] used multi-task learning with a fuzzy ensemble approach, whereby an instance could be given multiple labels, contrary to single-task learning. In this regard as well, [62] used fuzzy approaches in an attempt to overcome the limitation of the previous method, where an instance may not be consistent, i.e., an instance that relates slightly to one class and slightly to another class. Moreover, [10] confirmed that due to the lack of clarity regarding the purpose of using machine learning techniques with a detection model, it was appropriate to concentrate on an explanatory model. Hence, to overcome this limitation,

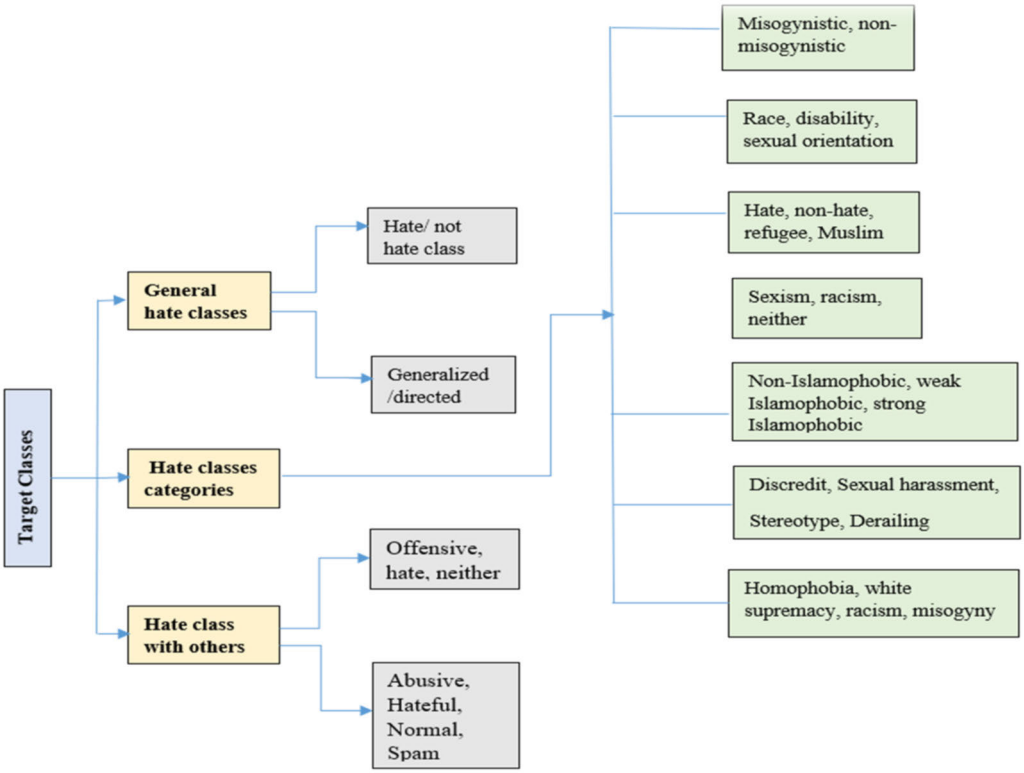

FIGURE 11. Taxonomy of the data used in the selected study based on the target classes.

they implemented the control signals of LSTMs to identify and interfere with hate speech and offensive words. Other issues in prior detection methods were that they made no effort to represent offensive language and hate speech effectively in the training dataset [63]. The bi-directional LSTMs were trained on word representations based on pronunciation. The issue became much more straightforward when only hate-related tweets were taken into account [94]. The authors shifted their focus toward users and developed a model based on a user-embedding node and network connections. Moreover, [64] applied the BiLSTM intra-user and inter-user representation information to constrain the noise in a single tweet. Reference [43] offered dependency as a feature to intersect cyber hate attributes. The noisy and informal nature of tweets could be handled more effectively by advanced pre-processing.

5) SEPARATION OF HATE SPEECH FROM AN OFFENSIVE INSTANCE

For automated hate speech identification on social media, it is crucial to distinguish hate speech from other offensive language instances. The dimensionality reduction approach was implemented in previous studies, such as [90], to address this issue. The offered system relies on feature selection methods, namely, information gain and term frequency-inverse document frequency. Additionally, [95] provided a detailed evaluation of the importance of different semantic feature representations of social media posts. Semantic feature can support and enhance the contextual interpretation of the word senses of a machine learning model. Recently, [108] attempted to distinguish between these two classes using traditional machine learning with superior outcomes. Furthermore, [25] applied an ensemble deep learning classifier using an optimized function and weight-updating process. It is difficult to differentiate between hate speech and offensive speech [92]. Hate speech can also be too violent, unregulated, and aggressive. In [92], the researchers used the text2vec library in R, which is a programming language that is well-established, concise, and powerful. The profane words in the texts often do not imply that the text instance is hateful. Reference [85] used typed dependency as an additional feature in a text instance to consider the connection between long-range words, while providing more details than a word-based feature. However, lexical detection methods suffer from low precision. Such methods can identify all posts containing unique words, such as hate speech, while the existing literature that used supervised learning was unable to differentiate between the two categories. Moreover, [93] used crowdsourcing and LR, NV, DT, SVM, and RF. Based on their findings, tweets without exact hate keywords are even harder to identify. Reference [84] confirmed that there is a good range to boost automated hate speech and offensive identification schemes. The latter provided DT, SVM with Char4-grams, CNN, and LSTM with different word embeddings. The deep learning model offered the best performance among the applied algorithms. Moreover, [65] introduced LSTM for

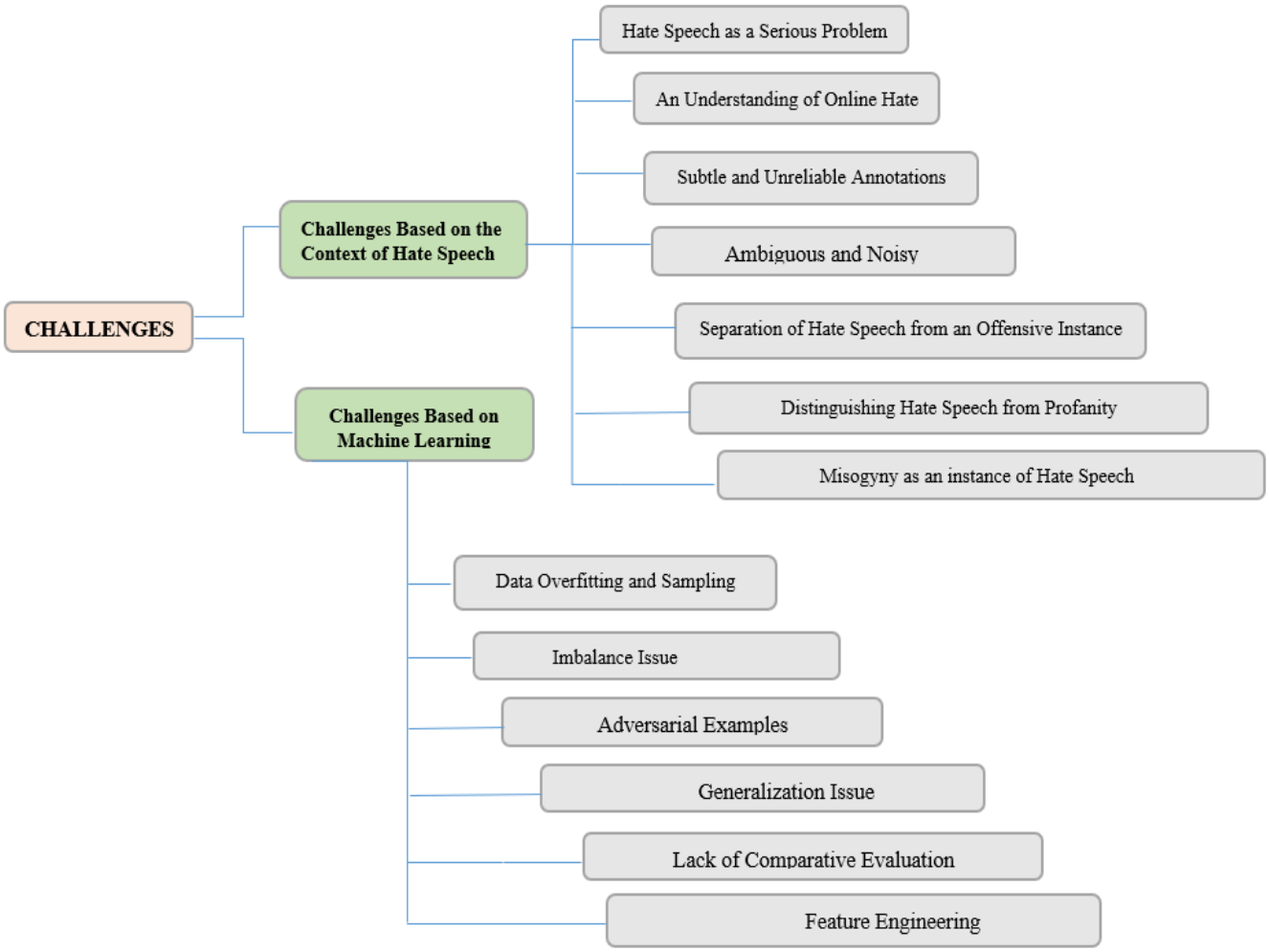

FIGURE 12. The taxonomy of the selected studies is based on the existing challenges.

this issue, and the fastText and BERT embedding have been used as entry features to CNN and BiLSTM classifiers. Hate speech is a tough phenomenon to label. Searching for hateful training data in the absence of specific terms or offensive language could assist in classifying such cases.

6) DISTINGUISHING HATE SPEECH FROM PROFANITY

Another hate speech identification problem is the differentiation between profanity and hate speech. Occasionally, hate speech includes profane language, but the inclusion of those words does not necessarily mean that the text case is hateful [85]. This problem is addressed further by [44] using a linear SVM classifier with standard lexical features. Later, introducing the classifier ensembles to differentiate hate speech and profanity is not easy [45]. Two algorithms have been used by [45] namely a linear SVM kernel and a radial basis function (RBF). Both techniques are effective for smaller data features and can have nonlinear decision-making constraints. There may be challenges here due to subjectivity regarding what is considered offensive. In this challenge, a more in-depth understanding of the semantics of the sentence is encouraged.

7) MISOGYNY AS AN INSTANCE OF HATE SPEECH

Hate against women is a complicated subject that includes traditional and cultural customs. Misogyny is a particular case of hate talk directed at women. Misogyny can take various forms in online social media. Reference [46] provided a corpus of misogynous tweets labeled from multiple viewpoints and used NLP features with ML models. Reference [91] carried out two tasks: task A dealt with the issue of determining whether the binary classification of a tweet was misogynous or not. Task B comprised two parts in dealing with the problem of multiple categories, whereby misogynous tweets were assigned to the correct misogyny group and examined by an LSTM and traditional methods. The tweets mostly consisted of just one or a couple of words, and it was difficult to decide if those very brief tweets were misogynous or not due to the lack of a discourse context [66]. Consequently, misogynous and non-misogynous tweets were automatically labeled, but

TABLE 9. The data used in this study domain.

Data source	Size/Type of classes	Source	Ref.
Private data	5593 tweets:1876Race, 1914 Disability, 1803 Sexual Orientation.	[76]	[41, 42, 43, 68]
Publicly available http://ow.ly/BqCf30jqffN	80k tweets: Abusive, Hateful, Normal, Spam.	[82]	[5, 37,55, 67, 70, 95]
Publicly available (https://github.com/ziqizhang/chase/tree/master/data)	2,435 tweets: hate 414 non-hate 2,021 refugees, Muslim.	[35]	[1, 34, 35, 36, 65, 68]
Publicly available (https://data.world/crowdflower/hate-speech-identification)	25k tweets: 24,802: hate speech, offensive, or neither offensive nor hate speech	[93]	[1, 6, 25, 33, 34, 56, 45, 38, 71, 66, 49, 55, 36, 68, 52, 54, 58, 10, 64, 57, 5, 83, 74, 37, 65,56, 84, 85, 45, 35,86 44, 90, 95, 96]
Publicly available (http://github.com/zeerakw/hatespeech)	16k (16,093) Tweets: 3,149 sexisms, 1,934 racisms, and 11,010 neither.	[77]	[1, 5, 6, 56, 87, 38,49, 50, 39, 36, 68, 54, 53, 70, 57, 37, 65, 56, 72, 64,35,86, 91].
Publicly available (https://competitions.codalab.org/competitions/19935.)	9,000 tweets, 3,783 hateful, and 5,217 as not hateful.	[78]	[6,49,70]
Publicly available https://www.kaggle.com/vkrahul/twitter-hate-speech. https://www.kaggle.com/pandeyakshive97/hate-speech-dataset.	1155 tweets, 8,770 as neither sexist nor racist. 2,242 are sexist or racist, and 29,720 as neither.	Kaggle dataset	[60]
Publicly available https://amiibereval2018.wordpress.com/.	Discredit, Sexual harassment, Stereotype, Derailing.	[79]	[9, 61, 76]
Publicly available https://amiEvalIta2018.wordpress.com	4,000 and 1,000 tweets for the training and test set.	[80]	[61, 67, 91]
Publicly available https://www.kaggle.com/datasets?search=hate+speech	32000 tweets, 2200 tweets homophobia, white supremacy, racism, and misogyny tweets.	Kaggle dataset	[69]
Public data https://amiibereval2018.wordpress.com/ https://amievalita2018.wordpress.com/	4454 tweets, misogynous vs. no- misogynous.	[46]	[46]
Public data https://www.kaggle.com/datasets	0 shows the non-hateful tweet, and 1 shows the hateful tweet.	Kaggle dataset	[59]
Public data *https://github.com/binny- Mathew/*	558 hate speech tweets and 1290 counter speech replies tweets.	[61]	[61]
Public data https://zenodo.org/record/3463560#.XY5LKC2ZOu5	1,364 tweets 470 non- Islamophobic, 484 tweets as weak Islamophobic, 410 tweets as strong Islamophobic.	[51]	[51]
Public data http://github.com/zeerakw/hatespeech	6k Tweets: 946 sexisms, 61 racisms, 18 racism & sexism, 5600 neither.	[87]	[35,36,37,57]
Public data	20,000 classified tweets 0 neutral sentiments, 1 offensive, 2 for hate speech.	CrowdFlower	[66]
Private data	4, 988 users, 544 hateful, 4427 normal.	CrowdFlower	[94]
Private data	1999 tweets, 951 non-hate, 1048 hate.	[3]	[3]
Private data	10k tweets:1800 misogynistic, 3200 no-misogynistic.	[48]	[48]
Private data	554 hateful, 1,125non-hateful.	[31]	[31]
Private data	12,970 tweets: 5,462 hate, 7508 not hate.	[99]	[101]
Public data https://github.com/cardiffnlp/tweeteval	20,000 tweets: very negative, negative, positive.	[98]	[97]
Public data https://zenodo.org/record/3816667	QMI dataset:1800 misogynistic, 3200 non-misogynistic.	[97]	[97]
Public dataASONAM21_COVID_HATE - Dropbox	2319 tweets: Hate, Non-Asian Aggression.	[75]	[32]

as syntactic and grammatical errors were involved, it was difficult to retrieve the text features. Reference [66] proposed an ensemble of five classifiers with bag-of-words for the mentioned issue. Moreover, [47] examined computational linguistic assumptions and variations between sexism and misogyny by using TF-IDF and an SVM with a weighted bag of words. The interpretation of an individual tweet was a key difficulty in the automated identification of misogynous tweets [48]. The CNN model was applied with pre-trained embedding on the undertaken domain to classify misogynistic tweets directed to a person or a community. Additionally, [67] introduced an initial effort to track and reduce unintended bias in machine learning models to detect misogyny. The bias introduced by different identity words in a model is often related to the misogyny class. The appearance of the word 'women' regularly in misogyny posts will lead to unrealistic misogyny scores for most supervised classification models. Close attention is required to classify misogynistic abuse in tweets, which can be difficult to categorize even for humans.

B. CHALLENGES BASED ON MACHINE LEARNING ISSUES

1) DATA OVERFITTING AND SAMPLING

References [6] and [49] re-experimented to provide a clearer understanding of state-of-the-art approaches for identifying hate speech. They focused on two state-of-the-art methods that yielded a productive output for hate speech detection using Twitter data used in the studies by [28] and [88]. Even though the work by [88] was not included among the selected papers, it deserves mention in this study. There were slight problems that needed to be immediately obvious from the explanation of the methods or the corresponding code. The issue in [28] involved the processing features from the data entry. According to [88], there was oversampling of the smaller classes. The correction was made by extracting features based on the training set to prevent data overfitting. Moreover, the oversampling undertaken on the data entry before the train-test split and the generalization error of such methods were revalidated. The model must have the ability to generalize over unobserved data to avoid overfitting issues. In addition, oversampling has to be completed after the train-test data split to avoid overfitting problems.

2) IMBALANCE ISSUE

A significant problem in this field is that the prevalence of hate speech constitutes only a small fraction of the material that can be viewed online. Reference [96] used re-sampling methods, including ROS, SMOTE, and ADASYN, to solve an imbalanced hate speech dataset and evaluated the output of several machine learning classifiers, including SVM, LR, and NB. The efficiency of all the classifiers increased when all the mentioned oversampling techniques were used. To mitigate this issue, [57] applied data augmentation techniques to reduce the degree of class imbalance with a recurrent neural language generation framework. They achieved a noteworthy increase from the baseline in the macro-F1 score and 30% in the hate speech class recall. The minority hate speech class is more important; therefore, the issue is more critical to classification errors than the majority class. Imbalance is a further challenge to the model and may demand distinct mechanisms. To manage the unbalance class dataset, a range of approaches must be used [111].

3) ADVERSARIAL EXAMPLES

There are several ways to invade or deceive text detection systems. A simple attack is carried out by altering the input text so that the individual reader can indeed perceive the intended meaning, but the text is incorrectly identified by detection models [68]. These attacks demonstrate that all the suggested detection techniques are weak against adversaries who can (immediately) introduce typos and adjust word boundaries. Moreover, [34] presented several new defenses whereby significance and readability are preserved, and these systems perform on par with or surpass the results of adversarial retraining. They retraining. Pre-processing barriers alter input data before improving upon the models [68] by pre-processing, and they get at the model and try to reproduce the actual text. In contrast, retraining handles text morphing by training the model using pre-attached data. The same approach was used in [69], where it was verified that a classifier could be deceived into misclassifying a toxic comment as necessary by slightly changing the text. Thus, the false negatives increased by two offered attacks: word splitting and character dropping. An offender could bypass a hate speech detection engine, and harmful content could be labeled benign. Minor changes in a text can alter the meaning or make the content completely worthless. Thus, it is possible to trick a classifier into misclassifying that text.

4) GENERALIZATION ISSUE

The comments collected from a user-generated text can vary between services, and using such data can reduce the generalization of the model. Therefore, detection approaches that do not rely on information from any given platform are essential [50]. Using text-based input can help avoid any information about user metadata that can vary between networks. The absence of generalization can result from training a hate speech algorithm on smaller datasets and testing it in a different data distribution [70]. Deep generative language modeling was used as a data augmentation technique to create massive hate speech sequences. Many factors could affect having a generalized model rather than data variety, either due to the bad hyper-parameter setting or the nature of the machine learning techniques. Several components of hate speech detection are affected by the problem of generalizability, including dataset preparation, model training and verification, and implementation [108].

5) LACK OF COMPARATIVE EVALUATION

As indicated in [7], the lack of a detailed comparison is a significant weakness in this field of work, rendering an evaluation of the impact of existing efforts. In [83], researchers

compared the efficiency between several feature techniques and eight algorithms and tested their performance in three categories using a publicly accessible data set. Reference [35] established other benchmark datasets and introduced convolutional and gated feature techniques and eight algorithms and tested their performance in three categories using a publicly accessible data set. Reference [35] established other benchmark datasets and introduced convolutional and gated recurrent networks with six more datasets. Understanding the nuances of messages from the user with the possible intention of hate led to the introduction of semantic features [95]. Accordingly, semantic features may contribute to an understanding of the contextual text representation in a social media post. A detailed evaluation was performed on the effects of various semantic features in social media messages. Depending on the issue being addressed, various features and techniques were engaged in hate speech detection, which should be evaluated in more comparative research.

6) FEATURE ENGINEERING

Hate speech suffers from a scarcity of typical and special features, and is, therefore, hard to identify in the 'long tail' of a dataset [1]. Deep learning operates as a feature extractor, which is extremely good for catching the semantics of hate speech tweets. CNN with the recurrent neural network (RNN) has been applied to the issue concerned. Moreover, the study by [36] paid attention to feature selection as a very effective potential strategy as it can choose a very limited range of the most predictive features. The study used the SVM with the initial feature set alone and subsequently with both the initial and enhanced feature sets. Dimension reduction was introduced by [71]. LSA and SVD were applied with a generally utilized technique for feature extraction, such as bag-of-words, N-grams, and TF-IDF, in addition to the cosine similarity technique to provide efficient inputs for classifiers. However, neglecting textual information in previous studies led to offering word embeddings, sentiments, and topical information with LSTM-CNN [37]. The feature selection process can improve the performance of the ML model; however, several feature selection processes were analyzed in the hate speech detection task.

C. OTHER CHALLENGES

The literature described several other issues, which could not be grouped, faced by researchers in the hate speech detection process. Islamophobia hate speech messages, for example, are another online social media theme that indirectly communicates hate against Muslims. For this issue, six different algorithms, including deep learning, were implemented to detect Islamophobia hate speech [51]. However, the dataset with a small number of instances containing other classes posed a challenge in the learning process; hence, [72] used TF-IDF, N-gram, and a sentiment lexicon with an SVM for better results. Some recent studies [31], [32] are concerned with detecting hate speech related to COVID-19. Many studies had evaluated the use of various machine learning algorithms for hate speech detection. Nevertheless, they suffered from ineffective sequence representation. As such, the transformer-based method was provided by [74]. In [86], researchers noted that hate speech can be directed at a single person or a group of people (generalized). They adopted a multi-step classification approach for directed and generalized hate speech. The authors in [38] carried out experiments based on BiLSTM with contextual attention toward understanding the overlap between hate speech and other abusive language types. However, [52] used lexical and emotional approaches to overcome the limitations of lexical methods, which tended to have low precision. The authors in [3] built a model based on SVM, which splits the tweets containing hateful code words about racist sentiments into those frequently used words. Reference [61] separated hateful users from counterspeech users by examining the attributes of both accounts. They provided a dataset of tweet-reply pairs, whereby tweets that were hate speech replies represent counter speech. For this later issue, TF-IDF, user profile properties, and lexical properties with LR, RF, SVM, CatBoost (CB), and XGBoost (XGB) were used. Hate speech detection is still an area under consideration to date. Many other issues could appear from more advanced research. There is still a risk of prejudice when using hate speech recognizers, so they should be treated with caution [104].

V. DISCUSSION AND FUTURE WORKS

The identification of hate speech has gained significant attention from researchers, who have introduced various methods. The capture and control of hate speech can prevent hate in social spaces and hate-linked crimes from occurring in the offline world [3]. The NLP hate speech detection approach is still weak, even with the many attempts by researchers [64]. Several attempts have been made to resolve the matter of the proliferation of hate speech, which is well ahead of their strategies. However, most of them still face challenges in reaching a competent solution, as the language in social media is developing rapidly [9]. Moreover, these approaches are inconsistent, and there are still some existing problems. Hence, some possible future directions are suggested to address the remaining challenges. The prevalence of anti-female or misogynous language on social platforms has increased, making it a critical issue that must be investigated [14]. Although many works have addressed this issue, the existing misogyny detection methods in online environments are still in their infancy [46], [67]. Similarly, adopting a multi-label classification approach in misogyny detection could help overcome the growing violence rate [46]. The effect of embedding biases in misogyny detection models remains to be seen. As suggested by [67], it is a process of assessing and comparing the performance output of machine learning models between pre-trained embedding and trained embedding during the training process. Moreover, studies should also include the aspect of aggressive behavior since hate speech is linked to studies on human behavior [8]. In contrast, distinguishing profanity profanity

from hate speech is a challenging task [44]. The presence of hate words in a text does not mean that the text carries the meaning of hate speech. Tweets without obvious hate words are often more challenging to classify [93]. Reference [12] made three recommendations, namely, using unique learning-based features, metadata, and multi-modular data to integrate the context of hate speech. Moreover, a framework has been established that recognizes hateful languages other than English. The performance of the hate speech detection task is significantly impaired when machine-based learning and embedding models are trained on noisy datasets [63]. A deep learning approach is becoming more common for text classification, and it has been suggested to address the growing prevalence of hate speech on a variety of social media platforms [41], [47]. As criticized by [9], this mission demands expertise in social and cultural life. The relatively high disagreement between hate speech and human labeling revealed that this classification might become difficult for machines. Further, [40] indicated that tweets posted on suspended accounts that correlated with a particular event could increase the hate speech dataset. According to [54], a broader hate speech pattern and a unigram dictionary could be useful for detecting hate speech and offensive online messages. The authors in [57] believed that sub-word-level neural embedding is worth exploring in the short-text hate speech detection task. Further examination is needed to seek the impact of imbalanced learning on intensive feature engineering and classification models [96]. Character-level features could offer models that are more immune to attacks instead of word-level features [68], [91], [92]. According to [69], it is yet to be seen whether the attacks are efficient for features extracted using contextual embedding. References [50] and [56] suggested using BERT as a dynamic technique for identifying hate speech. Authors in [112] recommended using a knowledge graph to enhance transformers. In addition, it is necessary to explore the impact of an augmented set of instances of each hate class on the detection performance [33], [42], [66], [83]. Some languages have significantly fewer resources, and the transfer of learning from language to language should be introduced to advance hate speech detection in language-based cases [6], [11]. As [6] also claimed, a practical strategy would be to enhance the generalization of developed English methods. Methods such as unsupervised learning have replaced reliance on labeled data [100]. Future works could consider a careful optimization of hyper-parameters to see how they impact the performance of the method [33], [48], [49]. The focus should be on the classifier ensembles and meta-learning efficiency for this task [44]. More detailed pre-processing is preferred to eliminate unnecessary information from noisy tweets [85]. Studies should focus on the function of humor in hate speech detection tasks [9], [46], [61], [92]. It has been shown that the literature is highly interested in the task of sarcasm detection, called humor tweet detection, in some studies [73]. The focus should be on the dialect to prevent intentional racial biases in hate speech detection, as mentioned by [55]. Moreover, additional hostile keywords should be acquired to monitor potential hateful texts in response to changes in popular hate themes [31]. A further course of work in the future should include exploring different machine-learning techniques and methods to characterize and track social media user-centered content [4], [28], [35], [52], [94], [95]. A creative model for increasing hate speech data should be established using data created through a deep generative strategy trained on limited datasets of hate speech [70]. There are discrepancies between offensive speech and hate speech instances, and further investigation is necessary [58]. The introduction of linguistic features might be a promising path [45], [59]. A further improvement to this task could be to incorporate textual and image data together in the detection models [12], [37], [74]. For possible work corresponding to the hate speech task, the examination of platforms other than Twitter that fertilizes hate speech could be explored, as suggested by [86]. The huge advancements in this field are putting a constraint on the timeliness of this study. Limited data sources exist in languages other than English, which causes limited work in those languages. Since only tweets written in the English language were considered, further research on other languages is encouraged [103]. Additionally, cross-lingual generalization research is merely at the initial stage [17].

VI. CONCLUSION

In recent years, the increasing use of social media has led to highly unacceptable phenomena, such as hate speech language and hate speech-based incidents. Despite ongoing studies aimed at solving the issue of the proliferation of hate speech, there are still challenges in establishing a competent solution for content generated by users. The aim of the current study is to contribute to the existing survey and review papers to advance the investigation in the concerned field. Various aspects can be derived from the selected studies, including the datasets and their categories, the most used machine learning techniques, the performance metrics involved, and the validation methods applied. Moreover, a critical search was carried out on the selected documents that characterized and specified the challenges and recommendations linked to hate speech detection. A potential future study has been recommended to address the issues in previous research. Some of these issues refer to the lack of agreement and bias in data annotations, noisy user-generated posts, small training data, imbalanced data issues, lack of sufficient feature representations, generalization, appropriate user imbalanced data issues, lack of sufficient feature representations, generalization, appropriate user features, and hyper-parameter tuning. Furthermore, it would be interesting to consider the hate speech issue in languages other than English or on other social network sites. The present study analyzed the views in the published papers and provided researchers with a useful reference. This research is essential for additional studies. The research society can further work and concentrate on advanced methods for hate speech detection missions.

REFERENCES

[1] Z. Zhang and L. Luo, "Hate speech detection: A solved problem? The challenging case of long tail on Twitter," *Semantic Web*, vol. 10, no. 5, pp. 925–945, Sep. 2019.

[2] L. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber, "Analyzing the targets of hate in online social media," in *Proc. Int. AAAI Conf. Web Social Media*, 2016, vol. 10, no. 1, pp. 1–4.

[3] R. Magu, K. Joshi, and J. Luo, "Detecting the hate code on social media," in *Proc. Int. AAAI Conf. Web Social Media*, 2017, vol. 11, no. 1, pp. 608–611.

[4] M. Mondal, L. A. Silva, D. Correa, and F. Benevenuto, "Characterizing usage of explicit hate expressions in social media," *New Rev. Hypermedia Multimedia*, vol. 24, no. 2, pp. 110–130, Apr. 2018.

[5] A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, "A unified deep learning architecture for abuse detection," in *Proc. 10th ACM Conf. Web Sci.*, Jun. 2019, pp. 105–114.

[6] A. Arango, J. Pérez, and B. Poblete, "Hate speech detection is not as easy as you may think: A closer look at model validation (extended version)," *Inf. Syst.*, vol. 105, Mar. 2022, Art. no. 101584.

[7] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proc. 5th Int. Workshop Natural Lang. Process. Social Media*, 2017, pp. 1–10.

[8] J. Langham and K. Gosha, "The classification of aggressive dialogue in social media platforms," in *Proc. ACM SIGMIS Conf. Comput. People Res.*, Jun. 2018, pp. 60–63.

[9] 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, 2018.

[10] W. Dorris, R. Hu, N. Vishwamitra, F. Luo, and M. Costello, "Towards automatic detection and explanation of hate speech and offensive language," in *Proc. 6th Int. Workshop Secur. Privacy Anal.*, Mar. 2020, pp. 23–29.

[11] A. Alrehili, "Automatic hate speech detection on social media: A brief survey," in *Proc. IEEE/ACS 16th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Nov. 2019, pp. 1–6.

[12] S. Modi, "AHTDT—Automatic hate text detection techniques in social media," in *Proc. Int. Conf. Circuits Syst. Digit. Enterprise Technol. (ICCSDET)*, Dec. 2018, pp. 1–3.

[13] F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga, "Machine learning techniques for hate speech classification of Twitter data: Stateof-the-art, future challenges and research directions," *Comput. Sci. Rev.*, vol. 38, Nov. 2020, Art. no. 100311.

[14] E. Shushkevich and J. Cardiff, "Automatic misogyny detection in social media: A survey," *Computación Y Sistemas*, vol. 23, no. 4, pp. 1159–1164, Dec. 2019.

[15] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, "Resources and benchmark corpora for hate speech detection: A systematic review," *Lang. Resour. Eval.*, vol. 55, pp. 477–523, Jun. 2020.

[16] T. X. Moy, M. Raheem, and R. Logeswaran, "Hate speech detection in English and non-English languages: A review of techniques and challenges," *Webology*, vol. 18, no. 5, pp. 929–938, Oct. 2021, doi: 10.14704/WEB/V18SI05/WEB18272.

[17] W. Yin and A. Zubiaga, "Towards generalisable hate speech detection: A review on obstacles and solutions," *PeerJ Comput. Sci.*, vol. 7, pp. 1–38, Jun. 2021, doi: 10.7717/PEERJ-CS.598.

[18] 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, doi: 10.1109/ACCESS.2021.3089515.

[19] O. Istaiteh, R. Al-Omoush, and S. Tedmori, "Racist and sexist hate speech detection: Literature review," in *Proc. Int. Conf. Intell. Data Sci. Technol. Appl. (IDSTA)*, Oct. 2020, pp. 95–99, doi: 10.1109/IDSTA50958.2020.9264052.

[20] R. Rini, E. Utami, and A. D. Hartanto, "Systematic literature review of hate speech detection with text mining," in *Proc. 2nd Int. Conf. Cybern. Intell. Syst. (ICORIS)*, Oct. 2020, pp. 1–6, doi: 10.1109/ICORIS50180.2020.9320755.

[21] S. MacAvaney, H. R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, "Hate speech detection: Challenges and solutions," *PLoS ONE*, vol. 14, no. 8, pp. 1–16, 2019, doi: 10.1371/journal.pone.0221152.

[22] F. Alkomah and X. Ma, "A literature review of textual hate speech detection methods and datasets," *Information*, vol. 13, no. 6, p. 122, 2022, doi: 10.3390/info13060273.

[23] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Tech. Rep., 2007.

[24] D. Moher, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *Ann. Internal Med.*, vol. 151, no. 4, p. 264, Aug. 2009.

[25] Z. Al-Makhadmeh and A. Tolba, "Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach," *Computing*, vol. 102, no. 2, pp. 501–522, Feb. 2020.

[26] M. Suhaidi, R. A. Kadir, and S. Tiun, "A review of feature extraction methods on machine learning," *J. Inf. Syst. Technol. Manag.*, vol. 6, no. 22, pp. 51–59, Sep. 2021, doi: 10.35631/jistm.622005.

[27] M. D. Oskouei and S. N. Razavi, "An ensemble feature selection method to detect web spam," *Asia–Pacific J. Inf. Technol. Multimedia*, vol. 7, no. 2, pp. 99–113, Dec. 2018, doi: 10.17576/apjitm-2018-0702-08.

[28] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proc. 26th Int. Conf. World Wide Web Companion*, 2017, pp. 759–760.

[29] A. F. Naswir, L. Q. Zakaria, and S. Saad, "The effectiveness of URL features on phishing emails classification using machine learning approach," *Asia–Pacific J. Inf. Technol. Multimedia J. Teknol. Mklm. Dan Multimedia Asia-Pasifik*, vol. 7, no. 2, pp. 61–69, 2022, doi: 10.17576/apjitm-2022-1102-04.

[30] Y. Yadav, P. Bajaj, R. K. Gupta, and R. Sinha, "A comparative study of deep learning methods for hate speech and offensive language detection in textual data," in *Proc. IEEE 18th India Council Int. Conf. (INDICON)*, Dec. 2021, pp. 1–6, doi: 10.1109/INDICON52576.2021.9691704.

[31] M. Li, S. Liao, E. Okpala, M. Tong, M. Costello, L. Cheng, H. Hu, and F. Luo, "COVID-HateBERT: A pre-trained language model for COVID-19 related hate speech detection," in *Proc. 20th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2021, pp. 233–238, doi: 10.1109/ICMLA52953.2021.00043.

[32] M. K. A. Aljero and N. Dimililer, "Genetic programming approach to detect hate speech in social media," *IEEE Access*, vol. 9, pp. 115115–115125, 2021, doi: 10.1109/ACCESS.2021.3104535.

[33] Y. Saini, V. Bachchas, Y. Kumar, and S. Kumar, "Abusive text examination using latent Dirichlet allocation, self organizing maps and K means clustering," in *Proc. 4th Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, May 2020, pp. 1233–1238.

[34] M. Moh, T.-S. Moh, and B. Khieu, "No 'Love' lost: Defending hate speech detection models against adversaries," in *Proc. 14th Int. Conf. Ubiquitous Inf. Manage. Commun. (IMCOM)*, Jan. 2020, pp. 1–6.

[35] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on Twitter using a convolution-GRU based deep neural network," in *Proc. Eur. Semantic Web Conf.* Cham, Switzerland: Springer, 2018, pp. 745–760.

[36] D. Robinson, Z. Zhang, and J. Tepper, "Hate speech detection on Twitter: Feature engineering vs feature selection," in *Proc. Eur. Semantic Web Conf.* Cham, Switzerland: Springer, 2018, pp. 46–49.

[37] R. Cao, R. K.-W. Lee, and T.-A. Hoang, "DeepHate: Hate speech detection via multi-faceted text representations," in *Proc. 12th ACM Conf. Web Sci.*, Jul. 2020, pp. 11–20.

[38] T. Chakrabarty, K. Gupta, and S. Muresan, "Pay 'attention' to your context when classifying abusive language," in *Proc. 3rd Workshop Abusive Lang. Online*, 2019, pp. 70–79.

[39] G. K. Pitsilis, H. Ramampiaro, and H. Langseth, "Effective hate-speech detection in Twitter data using recurrent neural networks," *Appl. Intell.*, vol. 48, no. 12, pp. 4730–4742, 2018.

[40] W. Alorainy, P. Burnap, H. Liu, A. Javed, and M. L. Williams, "Suspended accounts: A source of tweets with disgust and anger emotions for augmenting hate speech data sample," in *Proc. Int. Conf. Mach. Learn. Cybern. (ICMLC)*, Jul. 2018, pp. 581–586.

[41] W. Alorainy, P. Burnap, H. Liu, and M. L. Williams, "'The enemy among us': Detecting cyber hate speech with threats-based othering language embeddings," *ACM Trans. Web*, vol. 13, no. 3, pp. 1–26, Aug. 2019.

[42] H. Liu, P. Burnap, W. Alorainy, and M. L. Williams, "Fuzzy multi-task learning for hate speech type identification," in *Proc. World Wide Web Conf.*, May 2019, pp. 3006–3012.

[43] P. Burnap and M. L. Williams, "Us and them: Identifying cyber hate on Twitter across multiple protected characteristics," *EPJ Data Sci.*, vol. 5, pp. 1–15, Oct. 2016.

[44] S. Malmasi and M. Zampieri, "Detecting hate speech in social media," 2017, *arXiv:1712.06427*.

[45] S. Malmasi and M. Zampieri, "Challenges in discriminating profanity from hate speech," *J. Experim. Theor. Artif. Intell.*, vol. 30, no. 2, pp. 187–202, Mar. 2018.

[46] M. Anzovino, E. Fersini, and P. Rosso, "Automatic identification and classification of misogynistic language on Twitter," in *Proc. Int. Conf. Appl. Natural Lang. Inf. Syst.* Cham, Switzerland: Springer, 2018, pp. 57–64.

[47] S. Frenda, B. Ghanem, M. Montes-y-Gómez, and P. Rosso, "Online hate speech against women: Automatic identification of misogyny and sexism on Twitter," *J. Intell. Fuzzy Syst.*, vol. 36, no. 5, pp. 4743–4752, May 2019.

[48] M. A. Bashar, R. Nayak, N. Suzor, and B. Weir, "Misogynistic tweet detection: Modelling CNN with small datasets," in *Proc. Australas. Conf. Data Mining*. Cham, Switzerland: Springer, 2018, pp. 3–16.

[49] A. Arango, J. Pérez, and B. Poblete, "Hate speech detection is not as easy as you may think: A closer look at model validation," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2019, pp. 45–54.

[50] J. S. Meyer and B. Gambäck, "A platform agnostic dual-strand hate speech detector," in *Proc. 3rd Workshop Abusive Lang. Online*, 2019, pp. 146–156.

[51] B. Vidgen and T. Yasseri, "Detecting weak and strong islamophobic hate speech on social media," *J. Inf. Technol. Politics*, vol. 17, no. 1, pp. 66–78, Jan. 2020.

[52] R. Martins, M. Gomes, J. J. Almeida, P. Novais, and P. Henriques, "Hate speech classification in social media using emotional analysis," in *Proc. 7th Brazilian Conf. Intell. Syst. (BRACIS)*, Oct. 2018, pp. 61–66.

[53] M. Sajjad, F. Zulifqar, M. U. G. Khan, and M. Azeem, "Hate speech detection using fusion approach," in *Proc. Int. Conf. Appl. Eng. Math. (ICAEM)*, Aug. 2019, pp. 251–255.

[54] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate speech on Twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection," *IEEE Access*, vol. 6, pp. 13825–13835, 2018.

[55] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The risk of racial bias in hate speech detection," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1668–1678.

[56] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A bert-based transfer learning approach for hate speech detection in online social media," in *Proc. Int. Conf. Complex Netw. Their Appl.* Cham, Switzerland: Springer, 2019, pp. 928–940.

[57] 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. Manag.*, Nov. 2019, pp. 991–1000.

[58] A. G. D'Sa, I. Illina, and D. Fohr, "BERT and fastText embeddings for automatic detection of toxic speech," in *Proc. Int. Multi-Conf. Org. Knowl. Adv. Technologie (OCTA)*, Feb. 2020, pp. 1–5.

[59] G. Koushik, K. Rajeswari, and S. K. Muthusamy, "Automated hate speech detection on Twitter," in *Proc. 5th Int. Conf. Comput., Commun., Control Autom. (ICCUBEA)*, Sep. 2019, pp. 1–4.

[60] L. Jiang and Y. Suzuki, "Detecting hate speech from tweets for sentiment analysis," in *Proc. 6th Int. Conf. Syst. Informat. (ICSAI)*, Nov. 2019, pp. 671–676.

[61] B. Mathew, N. Kumar, P. Goyal, and A. Mukherjee, "Interaction dynamics between hate and counter users on Twitter," in *Proc. 7th ACM IKDD CoDS 25th COMAD*, Jan. 2020, pp. 116–124.

[62] H. Liu, P. Burnap, W. Alorainy, and M. L. Williams, "A fuzzy approach to text classification with two-stage training for ambiguous instances," *IEEE Trans. Comput. Soc. Syst.*, vol. 6, no. 2, pp. 227–240, Apr. 2019.

[63] R. Hu, W. Dorris, N. Vishwamitra, F. Luo, and M. Costello, "On the impact of word representation in hate speech and offensive language detection and explanation," in *Proc. 10th ACM Conf. Data Appl. Secur. Privacy*, Mar. 2020, pp. 171–173.

[64] J. Qian, M. ElSherief, E. M. Belding, and W. Yang Wang, "Leveraging intra-user and inter-user representation learning for automated hate speech detection," 2018, *arXiv:1804.03124*.

[65] A. Bisht, "Detection of hate speech and offensive language in Twitter data using LSTM model," in *Recent Trends in Image and Signal Processing in Computer Vision*. Berlin, Germany: Springer, 2020, pp. 243–264.

[66] R. Ahluwalia, E. Shcherbinina, E. Callow, A. C. Nascimento, and M. De Cock, "Detecting misogynous tweets," in *Proc. IberEval SEPLN*, 2018, pp. 242–248.

[67] D. Nozza, C. Volpetti, and E. Fersini, "Unintended bias in misogyny detection," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell.*, Oct. 2019, pp. 149–155.

[68] T. Gröndahl, L. Pajola, M. Juuti, M. Conti, and N. Asokan, "All you need is 'Love': Evading hate speech detection," in *Proc. 11th ACM Workshop Artif. Intell. Secur.*, Jan. 2018, pp. 2–12.

[69] R. Oak, "Poster: Adversarial examples for hate speech classifiers," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2019, pp. 2621–2623.

[70] T. Wullach, A. Adler, and E. Minkov, "Towards hate speech detection at large via deep generative modeling," *IEEE Internet Comput.*, vol. 25, no. 2, pp. 48–57, Mar. 2021.

[71] R. I. Rasel, N. Sultana, S. Akhter, and P. Meesad, "Detection of cyber-aggressive comments on social media networks: A machine learning and text mining approach," in *Proc. 2nd Int. Conf. Natural Lang. Process. Inf. Retr.*, Sep. 2018, pp. 37–41.

[72] B. Gupta, N. Goel, D. Jain, and N. Gupta, "A novel IN-Gram technique for improving the hate speech detection for larger datasets," in *Micro-Electronics and Telecommunication Engineering*. Berlin, Germany: Springer, 2020, pp. 611–620.

[73] M. M. Al-Ani, N. Omar, and A. A. Nafea, "A hybrid method of long short-term memory and auto-encoder architectures for sarcasm detection," *J. Comput. Sci.*, vol. 17, no. 11, pp. 1093–1098, Nov. 2021, doi: 10.3844/JCSSP.2021.1093.1098.

[74] R. T. Mutanga and N. Naicker, "Hate speech detection in Twitter using transformer methods," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 9, pp. 1–7, 2020.

[75] R. Joshua. (May 2009). *Tweepy Documentation*. [Online]. Available: http://tweepy.readthedocs.io/en/v3

[76] W. Alorainy, P. Burnap, H. Liu, and M. L. Williams, "The enemy among us': Detecting cyber hate speech with threats-based othering language embeddings," *ACM Trans. Web*, vol. 13, no. 3, pp. 1–26, Aug. 2019.

[77] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in *Proc. NAACL Student Res. Workshop*, 2016, pp. 88–93.

[78] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti, "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.

[79] E. Fersini, P. Rosso, and M. Anzovino, "Overview of the task on automatic misogyny identification at IberEval 2018," in *Proc. IberEval SEPLN*, vol. 2150, Sep. 2018, pp. 214–228.

[80] E. Fersini, D. Nozza, and P. Rosso, "Overview of the evalita 2018 task on automatic misogyny identification (AMI)," in *Proc. EVALITA Eval. NLP Speech Tools Italian*, vol. 12, 2018, p. 59.

[81] M. Mondal, L. A. Silva, and F. Benevenuto, "A measurement study of hate speech in social media," in *Proc. 28th ACM Conf. Hypertext Social Media*, Jul. 2017, pp. 85–94.

[82] A. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of Twitter abusive behavior," in *Proc. Int. AAAI Conf. Web Social Media*, 2018, vol. 12, no. 1, pp. 1–10.

[83] S. Abro, S. Shaikh, Z. Hussain, Z. Ali, S. Khan, and G. Mujtaba, "Automatic hate speech detection using machine learning: A comparative study," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, pp. 1–8, 2020.

[84] A. H. Wani, N. S. Molvi, and S. I. Ashraf, "Detection of hate and offensive speech in text," in *Proc. Int. Conf. Intell. Hum. Comput. Interact.* Cham, Switzerland: Springer, 2019, pp. 87–93.

[85] K. J. Madukwe and X. Gao, "The thin line between hate and profanity," in *Proc. Australas. Joint Conf. Artif. Intell.* Cham, Switzerland: Springer, 2019, pp. 344–356.

[86] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding, "Hate lingo: A target-based linguistic analysis of hate speech in social media," in *Proc. Int. AAAI Conf. Web Social Media*, 2018, vol. 12, no. 1, pp. 1–10.

[87] 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.

[88] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *Proc. Eur. Conf. Inf. Retr.* Cham, Switzerland: Springer, 2018, pp. 141–153.

[89] P. Chiril, E. W. Pamungkas, F. Benamara, V. Moriceau, and V. Patti, "Emotionally informed hate speech detection: A multi-target perspective," *Cognit. Comput.*, vol. 14, no. 1, pp. 322–352, Jan. 2022, doi: 10.1007/s12559-021-09862-5.

[90] N. Rai, P. Meena, and C. Agrawal, "Improving the hate speech analysis through dimensionality reduction approach," in *Proc. 6th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Mar. 2020, pp. 321–325.

[91] R. Ahluwalia, H. Soni, E. Callow, A. Nascimento, and M. De Cock, "Detecting hate speech against women in English tweets," *EVALITA Eval. NLP Speech Tools Italian*, vol. 12, p. 194, Dec. 2018.
[92] J. Dhillon, V. Gupta, R. Govil, B. Varshney, and A. Sinha, "Crowdsourcing of hate speech for detecting abusive behavior on social media," in *Proc. Int. Conf. Signal Process. Commun. (ICSC)*, Mar. 2019, pp. 41–46.
[93] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proc. Int. AAAI Conf. Web Social Media*, 2017, vol. 11, no. 1, pp. 512–515.
[94] M. Ribeiro, P. Calais, Y. Santos, V. Almeida, and W. Meira Jr., "Characterizing and detecting hateful users on Twitter," in *Proc. Int. AAAI Conf. Web Social Media*, 2018, vol. 12, no. 1, pp. 1–10.
[95] Y. Senarath and H. Purohit, "Evaluating semantic feature representations to efficiently detect hate intent on social media," in *Proc. IEEE 14th Int. Conf. Semantic Comput. (ICSC)*, Feb. 2020, pp. 199–202.
[96] H. Rathpisey and T. B. Adji, "Handling imbalance issue in hate speech classification using sampling-based methods," in *Proc. 5th Int. Conf. Sci. Inf. Technol. (ICSITech)*, Oct. 2019, pp. 193–198.
[97] M. A. Bashar and R. Nayak, "Active learning for effectively fine-tuning transfer learning to downstream task," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 2, pp. 1–24, Apr. 2021, doi: 10.1145/3446343.
[98] B. Vidgen, "Detecting East Asian prejudice on social media," in *Proc. Social Inf. Netw.*, 2020, pp. 162–172, doi: 10.18653/v1/2020.alw-1.19.
[99] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke, "TWEETEVAL: Unified benchmark and comparative evaluation for tweet classification," in *Proc. Find. Assoc. Comput. Linguist. Find. ACL (EMNLP)*, 2020, pp. 1644–1650, doi: 10.18653/v1/2020.findings-emnlp.148.
[100] S. M. Al-Ghuribi, S. A. M. Noah, and S. Tiun, "Unsupervised semantic approach of aspect-based sentiment analysis for large-scale user reviews," *IEEE Access*, vol. 8, pp. 218592–218613, 2020, doi: 10.1109/ACCESS.2020.3042312.
[101] D. G. Kyrollos and J. R. Green, "MetaHate: A meta-model for hate speech detection," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2021, pp. 2496–2502, doi: 10.1109/BigData52589.2021.9672023.
[102] K. J. Madukwe, X. Gao, and B. Xue, "Token replacement-based data augmentation methods for hate speech detection," *World Wide Web*, vol. 25, no. 3, pp. 1129–1150, May 2022, doi: 10.1007/s11280-022-01025-2.
[103] R. M. O. Cruz, W. V. De Sousa, and G. D. C. Cavalcanti, "Selecting and combining complementary feature representations and classifiers for hate speech detection," *Online Social Netw. Media*, vol. 28, Mar. 2022, Art. no. 100194, doi: 10.1016/j.osnem.2021.100194.
[104] M. Mastromattei, L. Ranaldi, F. Fallucchi, and F. M. Zanzotto, "Syntax and prejudice: Ethically-charged biases of a syntax-based hate speech recognizer unveiled," *PeerJ Comput. Sci.*, vol. 8, pp. 1–19, Feb. 2022, doi: 10.7717/peerj-cs.859.
[105] V. Mercan, A. Jamil, A. A. Hameed, I. A. Magsi, S. Bazai, and S. A. Shah, "Hate speech and offensive language detection from social media," in *Proc. Int. Conf. Comput., Electron. Electr. Eng. (ICE Cube)*, Oct. 2021, pp. 1–5, doi: 10.1109/ICECube53880.2021.9628255.
[106] A. Razdan and S. Shridev, "Hate speech detection using ML algorithms," in *Proc. Int. Conf. Artif. Intell. Mach. Vis. (AIMV)*, Sep. 2021, pp. 1–6, doi: 10.1109/AIMV53313.2021.9670987.
[107] S. Dascálu and F. Hristea, "Towards a benchmarking system for comparing automatic hate speech detection with an intelligent baseline proposal," *Mathematics*, vol. 10, no. 6, p. 945, Mar. 2022, doi: 10.3390/math10060945.
[108] G. H. Panchala, V. V. S. Sasank, D. R. H. Adidela, P. Yellamma, K. Ashesh, and C. Prasad, "Hate speech & offensive language detection using ML & NLP," in *Proc. 4th Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Jan. 2022, pp. 1262–1268, doi: 10.1109/ICSSIT53264.2022.9716417.
[109] R. T. Mutanga, N. Naicker, and O. O. Olugbara, "Detecting hate speech on Twitter network using ensemble machine learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 3, pp. 331–339, 2022, doi: 10.14569/IJACSA.2022.0130341.
[110] A. Kumar, V. Tyagi, and S. Das, "Deep learning for hate speech detection in social media," in *Proc. IEEE 4th Int. Conf. Comput., Power Commun. Technol. (GUCON)*, Sep. 2021, pp. 1–4, doi: 10.1109/GUCON50781.2021.9573687.
[111] B. Pariyani, K. Shah, M. Shah, T. Vyas, and S. Degadwala, "Hate speech detection in Twitter using natural language processing," in *Proc. 3rd Int. Conf. Intell. Commun. Technol. Virtual Mobile Netw. (ICICV)*, Feb. 2021, pp. 1146–1152, doi: 10.1109/ICICV50876.2021.9388496.
[112] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "ERNIE: Enhanced language representation with informative entities," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1441–1451, doi: 10.18653/v1/p19-1139.
[113] C. Ziems, B. He, S. Soni, and S. Kumar, "Racism is a virus: Anti-Asian hate and counter hate in social media during the COVID-19 crisis," 2020, *arXiv:2005.12423*. [Online]. Available: https://claws.cc.gatech.edu/covid/#dataset

ZAINAB MANSUR received the B.Sc. degree (Hons.) in computer science from Omar Al-Mukhtar University, Libya, in 2003, and the M.S. degree in computer science from Universiti Kebangsaan Malaysia (UKM), Malaysia, in 2011. She is currently a Researcher with UKM. Her research interests include natural language processing, machine learning, text and web mining, and sentiment analysis.

NAZLIA OMAR received the B.Sc. degree (Hons.) from UMIST, U.K., the M.Sc. degree from the University of Liverpool, U.K., and the Ph.D. degree from the University of Ulster, U.K. She is currently an Associate Professor with the Center for AI Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia. Her main research interests include natural language processing and computational linguistics.

SABRINA TIUN received the B.Sc. degree (Hons.) from Bradley University, USA, and the master's and Ph.D. degrees from Universiti Sains Malaysia. She is currently a Senior Lecturer with the Faculty of Information Science and Technology (FTSM), Universiti Kebangsaan Malaysia (UKM), Malaysia. Her research interests include natural language processing, computational linguistics, speech processing, and social science computing.