Multimedia Systems (2023) 29:1203–1230
https://doi.org/10.1007/s00530-023-01051-8

REGULAR PAPER

A literature survey on multimodal and multilingual automatic hate speech identification

Anusha Chhabra[1] · **Dinesh Kumar Vishwakarma**[1]

Received: 19 September 2022 / Accepted: 9 January 2023 / Published online: 20 January 2023

Abstract
Social media is a more common and powerful platform for communication to share views about any topic or article, which consequently leads to unstructured toxic, and hateful conversations. Curbing hate speeches has emerged as a critical challenge globally. In this regard, Social media platforms are using modern statistical tools of AI technologies to process and eliminate toxic data to minimize hate crimes globally. Demanding the dire need, machine and deep learning-based techniques are getting more attention in analyzing these kinds of data. This survey presents a comprehensive analysis of hate speech definitions along with the motivation for detection and standard textual analysis methods that play a crucial role in identifying hate speech. State-of-the-art hate speech identification methods are also discussed, highlighting handcrafted feature-based and deep learning-based algorithms by considering multimodal and multilingual inputs and stating the pros and cons of each. Survey also presents popular benchmark datasets of hate speech/offensive language detection specifying their challenges, the methods for achieving top classification scores, and dataset characteristics such as the number of samples, modalities, language(s), number of classes, etc. Additionally, performance metrics are described, and classification scores of popular hate speech methods are mentioned. The conclusion and future research directions are presented at the end of the survey. Compared with earlier surveys, this paper gives a better presentation of multimodal and multilingual hate speech detection through well-organized comparisons, challenges, and the latest evaluation techniques, along with their best performances.

Keywords Hate speech · Multilingual · Multimodal · Machine learning · Deep learning · Online social media

1 Introduction

With the growing internet and technology, a large amount of information content is present on online community networks as multimodal data (Text, Pictures, and videos). If we look at the statistics, we can visualize that a large number of people using social media is escalating at a very great speed, and people can easily present their views to each other via various social media platforms. The type of content on the online social media stages contributes to the propagation of hate speech and misleading people. Right now, controlling this kind of media information is very important. Therefore, hate speeches harm individuals and impact society by raising hostility, terrorist attacks, child pornography, etc. Figure 1 shows a portion of hate speech and offensive expressions posted on social media or the web. Figure 1(a) shows a clear example of encouraging violence during huge fights against CAA, NRC, and NPR across India in Jan 2020 [1]. Figure 1(b) shows the tweet released under #putsouthafricansfirst, a person openly tweeting to attack the foreigners working in South Africa. Figure 1(c) shows a tweet posted in 2014 advocating killing Jewish people for fun after the synagogue shooting in Pittsburg [2]. Figure 1(d) shows a post posted in Jan 2018 that a supreme leader is giving a genuine threat statement to the US for war [3].

The recent instances of high-profile politicians making speeches were an apparent attempt at inciting violence, which led to large-scale violence. These instances are yet to be dealt with by law enforcement agencies. Hence, the integrity of identifying hate instances is one of the most significant challenges in social media stages, and research-based analysis of this type of content is necessary. The following

✉ Dinesh Kumar Vishwakarma
dinesh@dtu.ac.in

Anusha Chhabra
anusha.chhabra@gmail.com

[1] Biometric Research Laboratory, Department of Information Technology, Delhi Technological University, Delhi 110042, India

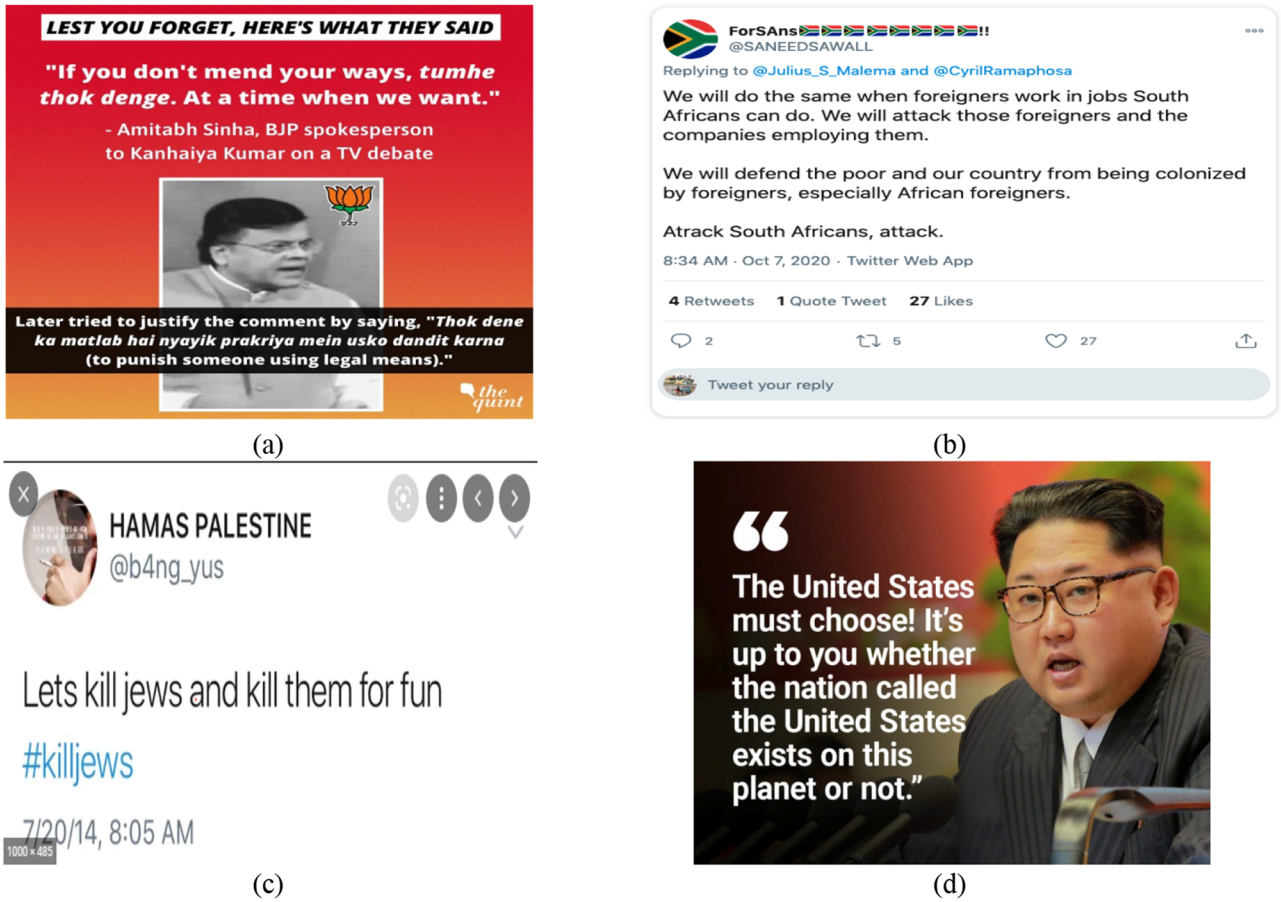

Fig. 1 Examples of hate speech and offensive expressions present over social media

section describes the definition analysis of hate speech from various sources.

1.1 Hate speech: definition perspective and analysis

There is a general agreement among researchers to define hate speech, and researchers have described it as a language that attacks an individual or a society dependent on characteristics like race, shading, nationality, sex, or religion [4]. This section provides some state-of-the-art definitions of hate speech (Table 1). Although many authors and social media platforms have given their purposes for hate speech, researchers are following them to understand the forms and classifications of hate speech. The definitions from the various sources are as follows:

- Some of the scientific definitions include the community's perspective.
- Major social networking sites like Facebook, YouTube, and Twitter are the most used platforms where hate speech occurs regularly.

The definition analysis (Table 2) mainly relies on various sources like multiple definitions from scientific papers and powerful social media platforms. The dimensions used for analysis are "violence," "attack," "specific targets," and "status."

After a thorough definition analysis, we have also portrayed the definition of Hate Speech as follows:

> "Hate Speech is a toxic speech attack on a person's individuality and likely to result in violence when targeted against groups based on specific grounds like religion, race, place of birth, language, residence, caste, community, etc."

1.2 Hate speech: forms and related words

Figure 2 shows significant hate forms of speech like Cyberbullying, Toxicity, Flaming, Abusive Language, Profanity, Discrimination, etc., and Table 3 presents the definitions of the above forms of hate speech found in the literature with their distinction from hate speech.

Table 1 Some of the prominent definitions by some state-of-the-art

References	Hate speech definitions
[5]	An antagonistic, malevolent speech focused on an individual or a social event of people taking into account a part of their genuine or intrinsic qualities. It communicates unfair, scary, objecting, hostile, or potentially biased perspectives toward those attributes, including sex, race, religion, identity, shading, public beginning, incapacity, or sexual direction
[6]	Hate speech is a conscious and hardheaded public assertion expected to slander a gathering of individuals
[7]	Hate speech is a quick attack on individuals subject to race, identity, sex, character, and veritable sickness or impediment. We portray assail as horrible or dehumanizing talk, clarifications of deficiency, or calls for dismissal or seclusion
[8]	Hate speech alludes to content that advances viciousness or scorn against the public dependent on specific ascribes, like ethnic or race beginning, religion, inability, sex, age, veteran reputation, and sexual direction/sex personality
[9]	Content that attacks people based on actual or perceptual race, ethnicity, country of origin, religion, gender, sexual orientation, disability, or illness is not permitted. It is considered a potential threat or attack for the content that many people find offensive (jokes, Stand-up comedy, lyrics of popular songs, etc
[10]	Hate speech attacks an individual or get-together depending on characteristics like religion, race, ethnicity, insufficiency, sexual heading, or sex character
[11]	Hate speech attack others dependent on racism, ethnicity, public start, sexual bearing, sex, character, age, handicap, or genuine illness
[12]	The language used to convey hate speech towards a selected bunch
[13]	Hate Speech is a purposeful attack on a specific social occasion of people motivated by the pieces of the group's character
[14]	Hate Speech assails or prompts malignance against gatherings in light of explicit qualities like actual looks, religion, ethnicity, sexism, and many more. Moreover, individuals with diverse phonetic styles in unobtrusive construction can happen

Table 2 Definition Analysis

Ref	Dimensions			
	Specific targets	Status	Violence	Attack
[5]	Yes	No	Yes	No
[6]	Yes	No	Yes	No
[7]	Yes	Yes	No	Yes
[8]	Yes	No	Yes	No
[9]	Yes	Yes	Yes	No
[10]	Yes	No	No	Yes
[11]	Yes	No	No	Yes
[12]	Yes	No	Yes	No
[13]	Yes	Yes	Yes	Yes
[14]	Yes	Yes	Yes	No

Hence, analyzing hate speech on the web is one of the critical areas to study due to the following reasons:

- Reduce conflicts and disputes created among human beings due to toxic language and offensive expressions.
- The broad availability and notoriety of online web-based media, like Facebook, Twitter, Instagram, web journals, microblogs, assessment sharing sites, and YouTube, boost communication and allow people to freely share information in the form of their thoughts, emotions, and feelings among strangers.
- Moreover, click baiting takes massive attention and encourages visitors to click on the link, harming readers' emotions.

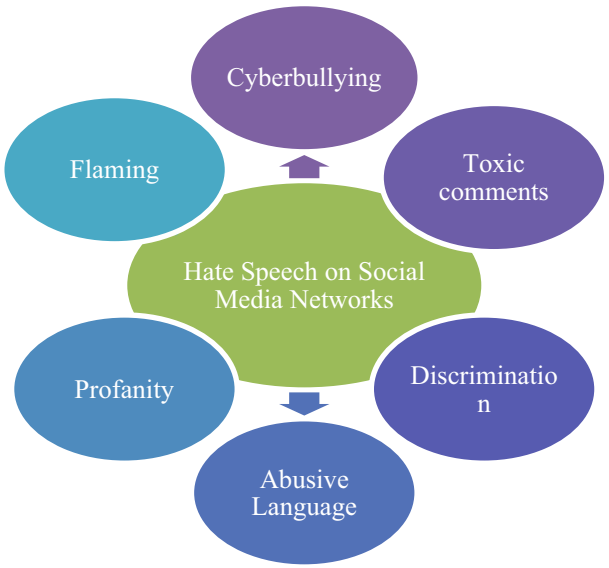

Fig. 2 Forms of Hate speech

- Hate speeches can incite violence and cause irreparable loss of life and money.
- The latest incident was triggered by online hate speech in the Philippines, citing the example of the Christchurch mosque shooting in 2019 [20].
- To forestall bigot and xenophobic viciousness and separation spread among Asians and individuals of Asian drop uniquely in this pandemic. As per the report distributed by US Today in May 2021, more than 6600 hate

Table 3 Comparison between Hate speech and its various forms

Forms	Definitions of forms	Distinction from hate speech
Cyberbullying	Characterized as a deliberate demonstration completed by a social occasion or individual using electronic stages [15]	Hate speech is abusive speech explicitly directed toward a unique, non-controllable attribute of a group of people
Discrimination	Interaction via a distinction and afterward utilized as the premise of unreasonable treatment [16]	Hate speech is a virulent form of discrimination
Flaming	Flaming describes antagonistic, profane, and threatening remarks that can upset and offend other members of the forums, generally called trolls [17]	Unlike flaming, hate speech can occur in any context
Abusive Language	The term abusive language seeks to diminish or humiliate some person or group [18]	Hate Speech is a type of abusive language
Profanity	Hostile or indecent words or expressions	Hate speech can use profane words but not always
Toxic language	Conveying content that is disrespectful, abusive, unpleasant, and harmful [19]	Not all toxic comments contain hate speech

and offensive incidences against Asian- Americans and Asians have been accounted for [6].
- To save our society from being gravely damaged.

From the points mentioned above, it has been observed that detecting and restraining hate speech at an initial stage is very crucial and, indeed, a challenging task. Major online media stages like Facebook, Twitter, and YouTube are trying to eliminate hate speeches and other harmful content at an initial step as part of their ongoing projects, using advanced AI techniques. However, keeping an eye on an individual is vital to have hate off platforms. Social media platforms and an individual can adopt the following suggestions:

- The most significant source of hate speech on the internet is trolls. A person should block, mute, or report these trolls instead of giving recognition.
- A person should do a proper data analysis and facts before forwarding the posts.
- Social media firms should follow strong policy rules against abusive behavior.

The following section describes the general framework of hate speech detection adopted by several researchers.

1.3 Motivation

Recently, it has been observed that the number of users are actively involving on social media in the forms of WhatsApp post, Facebook posts, YouTube shorts, reviews, comments etc. on various topics. People are sharing their views resulting in tremendous amount of data on the web. The data should be analyzed for further research. Giving the various hate form definitions, their analysis and to highlight the motivation behind the hate content detection in every aspect, we briefly discuss the recent works in this area in terms of various methodologies, modalities, performances, benchmarks etc. The further future trends are also highlighted giving the motivation to researchers for detecting hate content.

1.4 General framework of hate speech detection

Figure 3 provides a framework for the process of hate speech identification. The foremost step is to search the powerful source platform where most hate speech/ offensive languages occur. Most state-of-the-art adopted significant social media firms like Facebook and Twitter. The second step is to collect data either in the form of posts or tweets. Gathering a great measure of information from web-based media stages nowadays is one of the significant research challenges for researchers and academia. The platforms provide a simple and quick approach to gathering and storing information through inbuilt APIs [21]. Figure 4 shows the types of data accessible and non-accessible via social media, respectively.

A large amount of hate speech data collection is from two powerful social media platforms: Twitter, Instagram, Facebook, and these two platforms are actively working on combating hate speeches. The next phase includes data normalization and feature extraction for training a model, and the last step performs classification to classify the problem.

Most literature surveys ([14, 25–27, 4, 28–31]) have been published till now. Table 4 compares our survey with related surveys in various aspects like definition analysis, comparison with other hate forms, NLP aspect in terms of modalities, and explanation of models and datasets. This paper also gives an itemized portrayal of hate speech identification in multimodal information by considering major phases like data collection, text mining approaches in automatic hate speech detection, and different machine and deep learning approaches. This paper is a more detailed and systematic survey considering various parameters in terms of datasets, methods, etc. as follows:

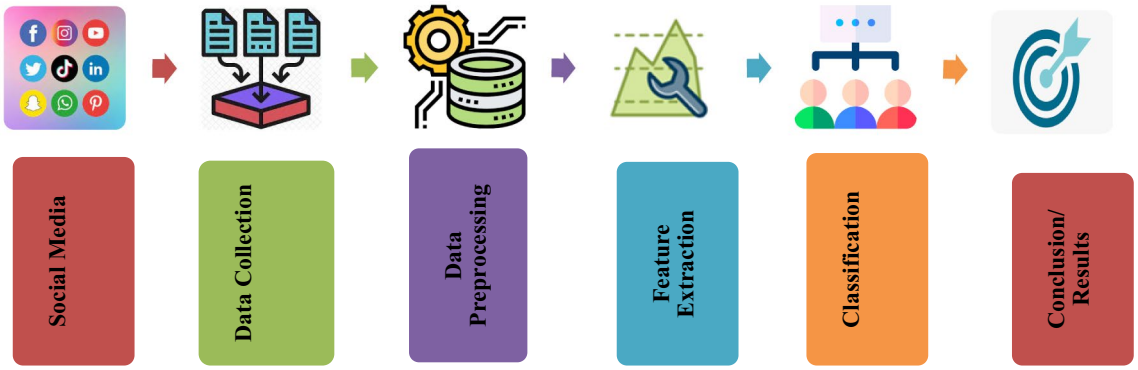

Fig. 3 General framework of hate speech detection

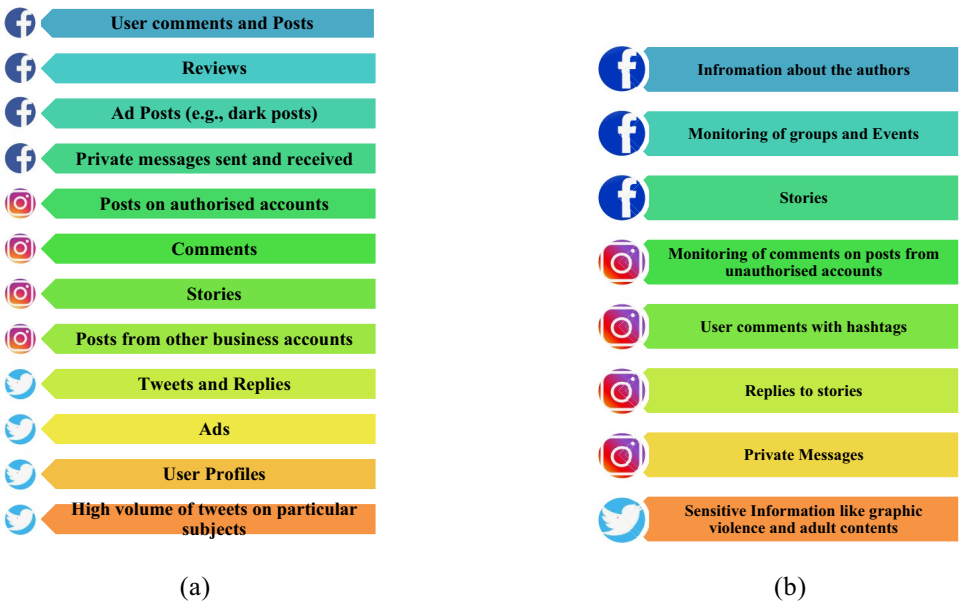

Fig. 4 **a** Type of data accessible via some social media networks and **b** Type of data not accessible via some social media networks [22–24]

Table 4 Comparison table of related surveys

Ref	Definition analysis	Comparison with other hate forms	Feature Extraction (NLP Aspect)	Modality (text-T, images- I, videos – V)	Linguistic aspect	Models aspect	Datasets
[4]				T	Monolingual		
[14]				T	Monolingual		
[25]				T	Multilingual		
[26]				T	Multilingual		
[27]				T	Multilingual		
[28]				T	Monolingual		
[29]				T	Multilingual		
[30]				T	Multilingual		
[31]				T	Monolingual		
Our survey				T + I + V	Multilingual		

- The study of detecting online hate content has been growing only in the last few years, and machine learning is more prominent. This survey covers the job done in deep and hybrid architectures to determine the issue of recog-

nizing hate speech. This survey also covers the feature extraction methods used in automatic hate speech detection.

- The beauty of this literature covers possible identified merits and demerits of the recent state-of-the-art works, their fundamental aspects, and techniques used in tabular form.
- Another beauty of this survey is that it covers publicly available datasets, dataset challenges, and benchmark models.
- Most previous surveys are on textual data, and limited literature is particularly for detecting hate speech in multimodal information. Therefore, this review paper also considers multimedia data (such as text, images, and videos) to highlight the detection process and the previous works on multilingual hate speech detection. However, this survey incorporates some famous works done for multilingual and multimodal data in this field.
- The survey also considers the current challenges and possible future directions; which researchers can view as further work in this area.

1.5 Review technique

However, various investigations have been distributed in earlier years identifying hate speech, yet this survey contains some noticeable works in this field. We considered influential journals, conferences, and workshops from various online databases such as IEEE Xplore, Science Direct, Springer, ACM Digital Library, MDPI, CEUR Proceedings, etc. The comprehensive survey contains a review of more than 120 articles based on keywords like "hate speech detection," "offensive language detection," "multilingual," "images," "videos," etc. It has been observed from Fig. 5 that the number of articles published on hate speech detection seemed to increase yearly in the last five years.

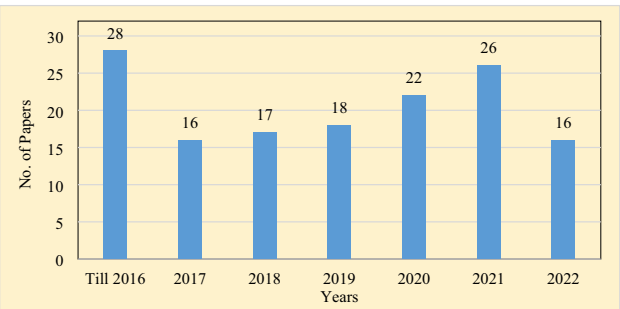

Fig. 5 Year-wise contribution of a research article on Hate Speech Detection over the last five years

This survey presents a comprehensive analysis of hate speech detection research arena as shown in Fig. 6. It breaks down the hate speech detection into several meaningful categorizations such as types of hate speeches, approaches, datasets, feature extraction methodologies etc. Special attention has been paid in the exploration of multimodal and multilingual approaches of better classification capabilities. Specifically, feature extraction methods such as Bag-of-Words, N-grams, Lexicon & Sentiment based features, TF-IDF, part of speech, word references and rule based are analyzed and also presented visually in Fig. 7. Next, hate speech detection methods are discussed in great detail categorizing them into traditional machine learning based and deep learning based approaches. A separate section elaborates the merits and demerits of the same. Special analysis of publicly available multimodal hate speech datasets is also presented that dives into the challenges posed by each of the multimodal hate speech datasets. Evaluation metrics and performance benchmarks are presented to highlight the effectiveness of current state-of-the-art approaches for hate speech detection.

This survey is organized as follows. The introduction is discussed in Sect. 1, whereas Sect. 2 describes possible feature extraction techniques in context with NLP Aspect for automatic hate speech detection. Section 3 covers the most vital work using different methodologies like machine and deep learning and the conversation on multilingual work. Section 4 highlights the challenges related to hate speech datasets and their benchmark models. In contrast, Sect. 5 depicts the various evaluation metrics and performance measures. Finally, Sects. 6 and 7 portrays the conclusion and further future directions respectively.

2 Feature extraction techniques in automatic hate speech detection

A feature is the closed characteristics of an entity or a phenomenon. [32] Focus on natural language processing (NLP) to explore the automation of understanding human emotions from texts. This section provides various text features used to extract hate content (Fig. 7). Word references and lexicons are the most straightforward and basic approaches for feature extraction in text analysis. Identifying the appropriate features for classification is more tedious when using machine learning. The fundamental step in traditional and deep learning models is tokenization, in which the primary and straightforward approach is dictionaries/ lexicons. Dictionary is a method that generates a set of words to be looked at and included in the text. Frequencies of terms are used directly as features. Features play an essential role in machine learning models. Machine learning approaches cannot work on raw data, so feature extraction techniques are needed to convert text into vectors of features. Many

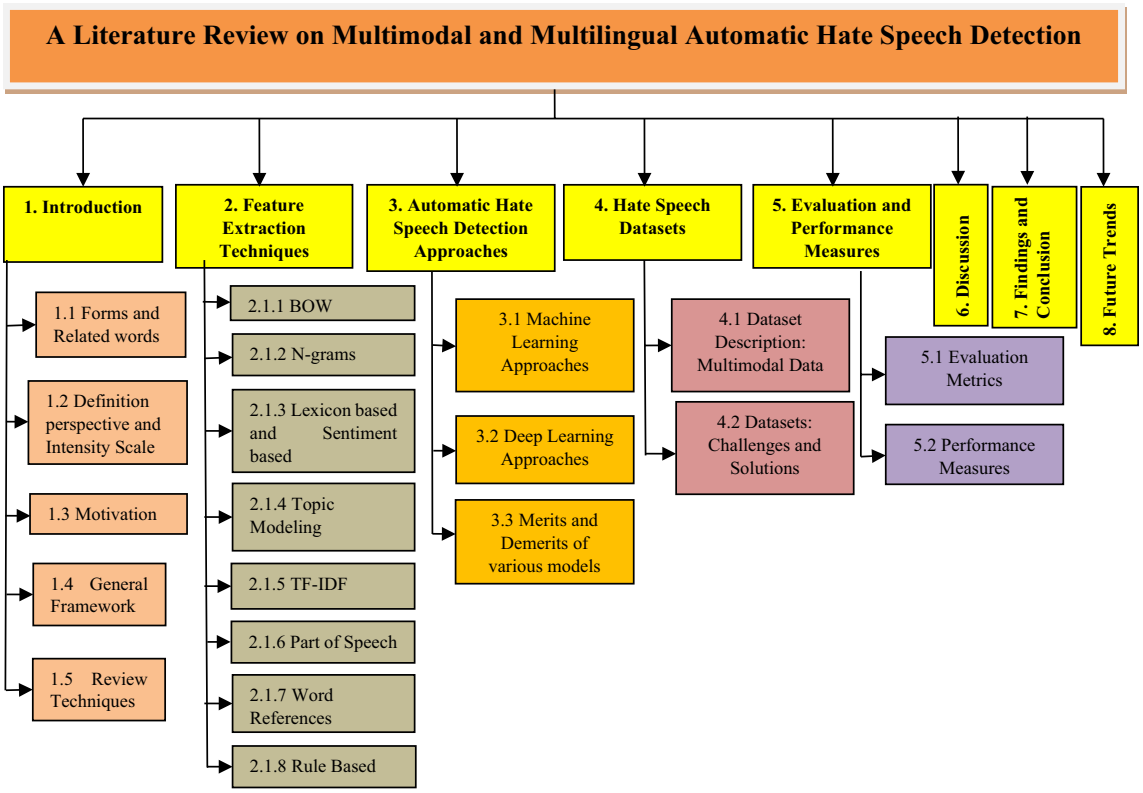

Fig. 6 Organization of the survey

basic features like BOW, Term Frequency- inverted Term Frequency, Word references, etc., are used.

2.1 Bag-of-words (BOW)

BOW is an approach like word references extensively used for document classification ([14, 33–35]). The frequency of each word is used as a characteristic for training a classifier after gathering all the words. The burden of this technique is that the sequencing of words is disregarded, whether it is syntactic or semantic information. Both pieces of information are crucial in detecting hate content. [36] Used BOW to represent Arabic hate features as text pre-processing before applying various machine learning classifiers. [37] Derived a method for detecting Arabic religious hate speech using different features with the machine and deep learning models. Consequently, it can prompt misclassification of whether the terms are utilized in multiple contexts. N-grams were executed to overcome the issue.

2.2 N-grams

The N-grams approach is the most utilized procedure in identifying hate speech and offensive language ([12, 18, 33, 38–41]). The most widely recognized N-grams approach combines the words in sequence into size N records. The objective is to enumerate all size N expressions and check their events. It further increases the performance of all classifiers since it incorporates each word context [42]. Rather than utilizing words, it is additionally conceivable to use the N-grams approach along with characters. [43] Proved character N-gram features are more predictive in detecting hate speech than token N-gram features, whereas it is not valid in the case of identifying offensive language. Although N-grams also have limitations, like all the related words have maximum distance in a sentence [33], an answer for this issue lies in incrementing the N value. However, it lowers the processing speed [15]. [39] Proved that greater *N* values perform better than lower *N*-values (unigrams and trigrams). The authors [4, 38] observed that character N-gram features

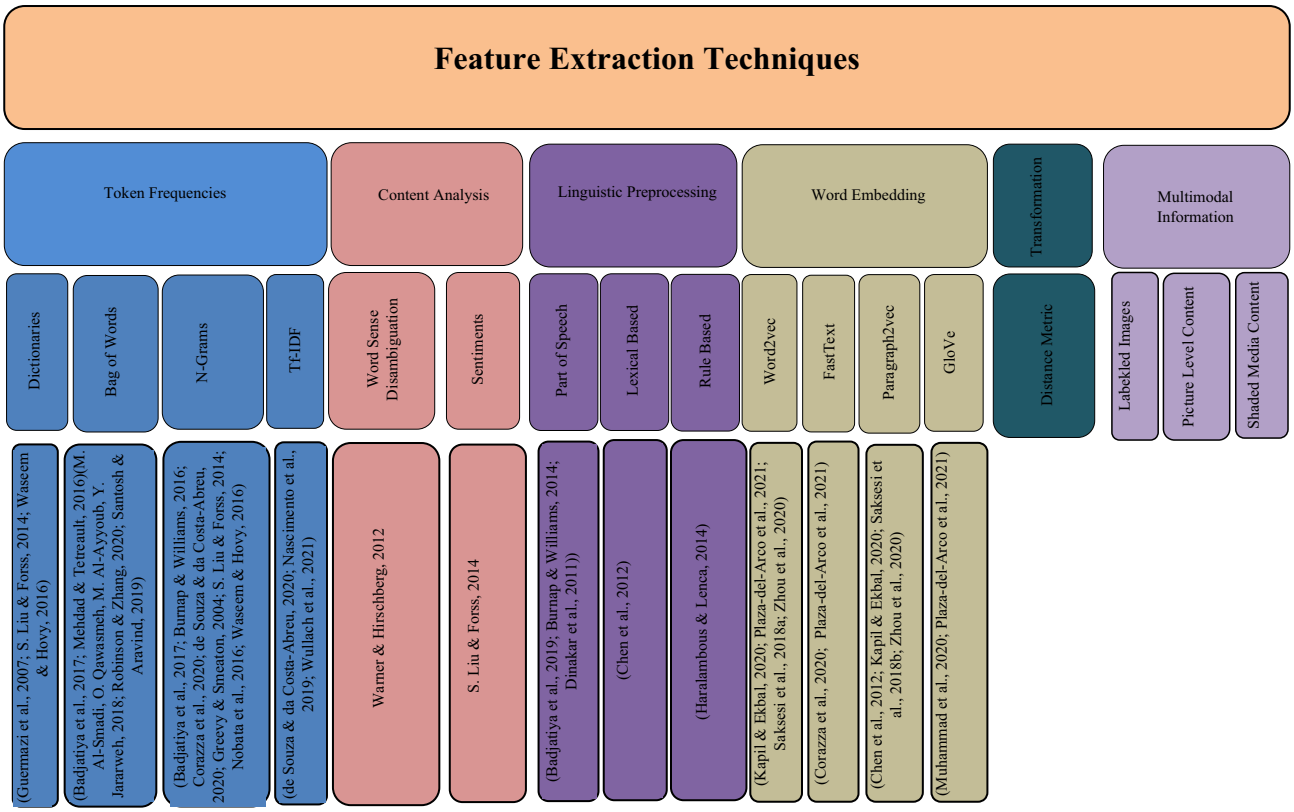

Fig. 7 Common feature extraction techniques

perform better when combined with extra-linguistic features. The authors generated one hot N-gram and N-gram embedding feature to train the model and analyzed better performance by N-gram embedding [44].

2.3 Lexicon-based and sentiment based

Lexical features use unigrams, and bigrams of the target word, whereas syntactic features include POS tags and various components from a parse tree. The parser used in NLP, proposed by the Stanford NLP Group [45], was used to catch the linguistic conditions inside a sentence [15]. Lexicon-based methods are crucial in identifying the sentiments of speech. For example, nigga is an offensive word and must be prohibited in ordinary language [46]. Hateful speech on a social stage cannot be a positive polarity because awful grammar provides a negative inclination by the speaker to the listeners and readers. Authors in ([12, 39, 47–50]) [51]) consider sentiments as a characteristic for identifying hate speech. Some authors [39] used the sentiment features in combination with others, which proved in result enhancement. [52] Presents metaheuristic approach for sentiment analysis and proved that the optimization methods can be alternatively used against machine leaning models with promising results.

2.4 Topic modeling

This method is also famous for topic classification, which focuses on extracting topics that occur in a corpus. Topic modeling is also used for detecting hateful comments from central social media platforms like Youtube [53]. [54]) used the Latent Dirichlet Allocation model ([55]) to discover abstract topics and use them in classifying multimodal data. [56] Derived text clusters from LDA for multilingual hate speech detection and proved that topic modeling is not giving any major incite for classification.

2.5 TF-IDF

TF-IDF is a scoring measure broadly used in information retrieval and is planned to reflect how important a term is in a given record. TF-IDF is the most common feature extraction technique used by traditional classification methods for hate speech identification ([35, 57, 58]). TF-IDF differs from a bag of words technique or N-gram technique because the word recurrence offsets the frequency of each term in the corpus, which clarifies that a few words show up more often than expected (for example,

stop words). [59] Used N-grams and TF-IDF values to perform a comparative analysis of the machine learning models to detect hate speech and offensive language and claimed that the L2 normalization of TF-IDF outperforms the baseline results.

2.6 Part-of-speech

POS tagging is a well-known task in NLP. This approach refers to the technique of classifying words into their parts of speech. Moreover, it improves the value of the context and identifies the word's role in the context of a sentence [60]. Some authors [40] used this approach to classify racist text. PoS tagging with TF-IDF gives a better result in Indonesian Hate Speech Detection [61].

2.7 Word embedding

The most widely recognized technique in text analysis of hate content is the utilization of word references. This methodology comprises all words (the word reference) that are looked at and included in the message. The frequencies are utilized straightforwardly as features and for calculating scores. In NLP, Word embedding is used for representing of words while performing text analysis [62]. Uses word2vec embedding for extracting hate content features for grouping the semantically related words. [63] Applies attention based neural networks and word embedding feature extraction methods for classification. Hate speech detection in Spanish language [64] uses word embedding methods like Word2Vec, Glove, FasText for feature extraction.

Another procedure used in text analysis of hate content is the distance metric, which can be used to supplement word reference-based methodologies. A few investigations have called attention when the negative words are obscured with a purposeful incorrect spelling [65]. Instances of these terms are @ss, sh1t [18], nagger, or homophones, for example, joo [65].

2.8 Rule-based approach

Text analysis uses a rule-based feature selection technique for finding the regularities in data, for example, IF–THEN clauses. [66] Proves that rule-based methods do not include learning but depends on word reference of subjectivity pieces of information. This particular approach is used to extract subjective sentences to generate hate content classifiers for unlabeled corpus [48]. [67] works on the combination of dictionary-based classifiers along with rule-based classifiers to generate the semantic features for hate speech classification.

3 Automatic hate speech identification approaches

This segment describes the research on hate speech identification using various models by establishing a thorough subjective and quantitative examination of what specifies multilingual hate speech. This section compares different machine and deep learning models for detecting hate speech in multiple languages, along with the labels/classification and datasets used**.** Authors also compares the deep learning based models with shallow based [68] learning models. Figure 8 shows various traditional and deep learning models used to identify hate speeches. It has also been observed that in the past few years, most work has been done on the general English language using various machine-learning models. It has also been seen that the results of deep learning and hybrid learning models outperformed using precision and recall. Following two Sects. 3.1 and 3.2, describes the sub-domain AI approaches to multilingual data.

3.1 Machine learning approaches to hate speech detection

Several machine learning models are being created to perform tasks like classification, prediction, clustering, etc. Machine learning models are also able to take advantage of data availability. Labeled data, which is utilized for training the model to achieve reliable accuracy, comes under the classification task. The machine learning algorithms performance directly depends on how accurately the features are identified or extracted. Classification algorithms perform detection tasks after normalizing the text. The efficacy of a model on a combination of several datasets is always better than training on a specific dataset [69]. Machine-learning algorithms are categorized as supervised, semi-supervised, and unsupervised methods. Researchers used these methods to detect online hate data in various languages**.** Out of different machine learning models, researchers primarily use SVM to classify social media data as hate or non-hate. Random Forest holds the second position, and so forth [70]. From Table 5, it has been clearly seen that most research is being conducted on the general English language using supervised machine learning methods. Some authors investigated the impact of pre-processing techniques [36] to improve text quality and mainly to retain the features without losing information [71] for better performance. A piece of research on multi-class classification on some datasets [18] carried out a machine learning-based approach for classifying online user comments into four classes (Clean, Hate, Derogatory, and Profanity) on the Amazon dataset. As Supervised, the

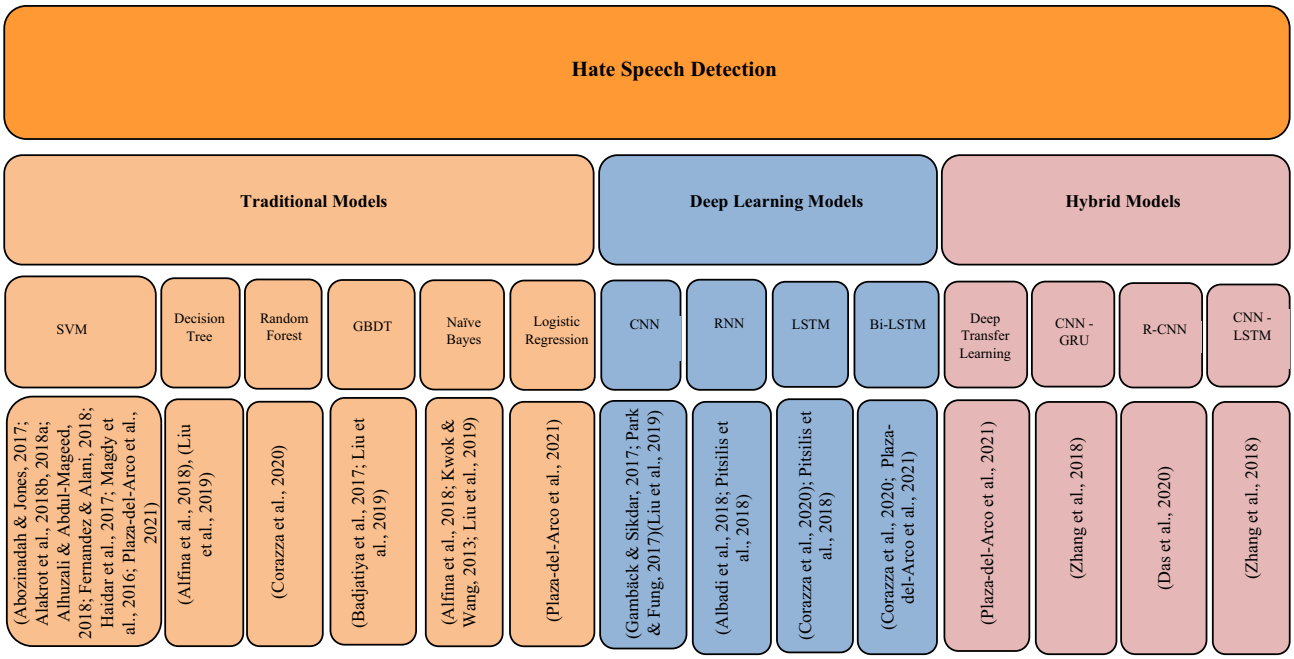

Fig. 8 Taxonomy of Hate Speech Detection considering various models

learning approach is area subordinate since it depends on manually marking a massive volume of information.

The advantage of manual labeling is its efficiency for domain-dependent tasks, while limitation occurs in execution time. The authors trained a supervised machine learning text classifier and used human-annotated data from the Twitter dataset to train and test the classifiers [72]. [72] The Bayesian logistic regression model is used to classify twitter data into hateful and antagonistic labels. Authors [73] focus on South Asian Languages for evaluating and comparing the effectiveness of various supervised techniques for hate speech detection. Semi-supervised learning algorithms are prepared to utilize both labeled and unlabeled information. Labeling information related to unlabeled information can viably upgrade efficiency. [74] Analyzed that unsupervised learning has a limited capacity to deal with limited-scale events, whereas supervised learning can adequately catch small-scale events; however, manually labeling the informational collection lowers the model scalability. [75] Utilized several choices of machine learning classifiers with various vector representations like TF-IDF, Count Vectorizer, and Word2vec as baselines with their own created Urdu dataset. KNN is the most widely used choice when a classification task is considered in a supervised learning approach [76], [77]. [78] Build an ensemble system utilizing various traditional machine learning (LR, SVM, RF, MNB, and XGB) for detecting sentiments.

When working with a particular language, the task can be considered an area-dependent task. The first racial-oriented research was carried out by [34], who carried out a supervised model to distinguish bigoted tweets [72]. Trained a supervised machine learning text classifier and used human-annotated data from Twitter to train and test the classifiers. [50] Proposed an approach for distinguishing hate speech for the Italian language on Facebook.

Authors [79] explored the capacity to recognize hate in the Indonesian language. The best outcomes from [48] were acquired when they consolidated semantic hate and theme-based components. SVM outperformed CNN and the ensemble approach in all subtasks of HaSpeeDe [80], using hate-rich embeddings [81]. Due to the many users available worldwide, multilingual hate speeches are spreading across the continent in different forms.

3.2 Deep learning approach to hate speech detection

Deep learning architectures (Table 6) represent a promising future in text analysis tasks. It relies totally upon artificial neural networks to investigate the patterns in the text with extra depth. In a couple of years, deep learning methods have outperformed machine learning methods in terms of performance due to the availability of large datasets. From the previous works, RNN and CNN are the most generally utilized deep learning models for NLP tasks. The execution of these two profound neural networks is a bit troublesome because of their intricate architectures. RNN has two sorts: LSTM and GRU, which upholds sequential architectures

Table 5 Recent state-of-the-artwork for detecting hate speech in various languages using machine learning models

Machine learning				
Methods	Ref	Models/Algorithms	Classification	Dataset
Supervised ***Methods***	General (English)			
	[82]	Fasttext	__	Youtube
				Myspace
				Slashdot
	[83]	SVM	__	Twitter
				Wikipedia
				Usenet
	[58]	SVM	(Sexuality, Race, Intelligence)	YouTube
	[84]	Multinomial- Naïve Bayes	__	Formspring
		Stochastic gradient descent		
	[72]	Bayesian logistic regression	(Hateful, Antagonistic)	Twitter
	[85]	Logistic regression	(Hateful, Clean)	Yahoo
	[86]	Logistic regression	(Hate, Not Hate)	Twitter
	[87]	SVM	(Hate, Offensive, OK)	Twitter
	[18]	NLP	(clean, hate, derogatory, profanity)	Amazon
	[12]	SVM	(Hate, Offensive, and Neither)	Twitter
		Naïve Bayes		
		Decision tree		
		Random forest		
	[88]	SVM	(Benevolent, Hostile, and others)	Twitter
		Fasttext		
	General (Roman-Urdu)			
	[75]	Logistic regression	(Hate, Offensive, and Neutral)	Twitter
		SVM		
		Random forest		
	General (Hindi-EnglisH Coded Mixed)			
	[89]	BoW	(Hate and Non-Hate)	Twitter
		Character n-grams		
		Word n-grams		
		Negation words		
		Punctuation Marks		
	Cyberbullying (Turkish)			
	[90]	Multinomial- Naïve Bayes	__	Twitter
				Instagram
	Radicalisation (English)			
	[91]	SVM	(Pro-ISIs, Non-Pro-ISIs, both, None)	Twitter
	Racial (English)			
	[34]	Naïve Bayes	(Racist, Non-Racist)	Twitter

Table 5 (continued)

Machine learning				
Methods	Ref	Models/Algorithms	Classification	Dataset
	Religious (English)			
	[65]	SVM	(Anti-Semitic, Not Anti-Semitic)	Yahoo news group
	General (Italian)			
	[50]	SVM	(Hate, Non-Hate)	Face Book
	General (Indonesian)			
	[79]	Random forest	(Hate, Non-Hate)	Twitter
		Decision tree		
	Abusive (Arabic)			
	[92]	Naïve Bayes	(Abusive, Normal)	Twitter
	[93]	SVM	(Abusive, Non- Abusive)	Twitter
	[94, 95]	SVM	(Offensive, In-Offensive)	Youtube
	Adult (Arabic)			
	[96]	SVM	(Adult, Normal)	Twitter
	Cyberbullying (Arabic)			
	[97]	SVM	(Hate, Non-Hate)	Facebook
				Twitter
	Terrorism (Arabic)			
	[98]	SVM	(Pro-ISIs, Anti-ISIs)	Twitter
Semi-Supervised	General (Spanish)			
	[64]	Logistic regression	(Hate and non-Hate)	HaterNet and HatEval
		SVM		
		LSTM		
		Bi- LSTM		
		Transfer learning		
	General (English)			
	[99]	Logistic regression	(Profane, Non-Profane)	Twitter
	[48]	Rule-based	(No Hate, Weak Hate, Strong Hate)	Blog
	Cyberbullying (English)			
	[100]	Fuzzy SVM	—	MySpace
				Slashdot
	Radicalisation (EnglisH)			
	[101]	LibSVM	(Hate, Extremism Promoting)	Twitter
	Terrorism (Arabic)			
	[102]	AdaBoost	(Jihadism, Anti-Jihadism)	Twitter
Unsupervised	General (English)			
	[15]	Match Rules	(Pejorative/ Profanities, Obscenities)	YouTube
	Cyberbullying (English)			
	[103]	K-Means	—	YouTube
				Formspring
				Twitter
	ViolenT (Arabic)			
	[104]	K-Means	(Violence, Non-Violence)	Twitter

Table 6 Generic deep learning architectures

Architectures	Key features	Merits	Demerits
Multi- layer Perceptron (MLP)	It consists of more than 2- hidden layers Widely applicable for classification and regression	High success rate	The learning process is deficient
Recurrent neural networks (RNN)	Useful where output depends on previous results Share the same weight for each step Mainly used for sequence learning problems The capability of analyzing the data stream	Good response Can memorize sequential events	Learning issue due to gradient problem
Convolution Neural Networks (CNN)	Highly applicable to visual data Every hidden layer filter transforms the inputs to the 3D output volume	Few neuron connections required	Requires large labeled datasets
Auto-encoder	The exact number of Input/output nodes Mainly used for feature extraction and space dimensionality reduction	No need for labeled data	Error-prone training
Deep belief network (DBN)	The hidden layer of each network is visible to the next layer Allows supervised and unsupervised learning	Use a greedy learning approach to initialize the stack Maximizing the likelihood inferences directly	Expensive training due to the sampling process
Deep boltzmann machine (DBM)	Uses Stochastic MAX likelihood method to maximize the lower bound of likelihood	Robust interface for indefinite output	Interfacing demands high-time complexity Parameter optimization is tedious for large datasets

though CNN has a hierarchical architecture. The efficacy of deep learning methods is directly based on the right choice of algorithms, the number of hidden layers, feature representation techniques, and learning high-level features from the data. Due to the exclusive performance factors, deep learning approaches are not better in every case than conventional methods. For hate speech identification,

[105] Utilized RNN model with word frequency and their outcomes beat the present state-of-the-art deep learning methods for hate speech identification. Deep learning techniques like automatic prediction, sentiment analysis, and classification are now being used to process hate images. [106] is a collection of memes from various social media platforms like Reddit, Facebook, Twitter and Instagram. The dataset is prepared from the 2016 U.S. Presidential Election Event, a collection of manually annotated image URLs and text embedded in the images, resulted in 743 memes. With respect to the classification of hateful memes, [107–109] presents various deep learning models to classify on memes dataset. Out of the researches done so far, [109] presents a visio-linguistic model (VILIO) for hateful memes detection and yields benchmark results. Deep learning strategies are recently being utilized in message characterization and sentiment analysis with maximum exactness [110]. Authors [78] used deep learning and transfer-based models (DNN, DNN with Embedding, CNN, LSTM, Bi-LSTM, m-BERT, distil-BERT, XML-RoBERTa, MuRIL) to reduce misclassification rate and to improve prediction rate for understanding code-mixed Dravidian languages. Table 7 shows the recent state-of-the-art for identifying hate speech using deep and hybrid learning methods while considering multiple languages like English, Italian, Arabic, Spanish, etc. As seen in Table 7, deep and hybrid learning models are evolving for classification tasks. Most works have been done on the Twitter dataset in the general English language using supervised approaches ([41, 105, 111–113]). Authors [114] show that LSTM is the most effective machine learning method for hate speech identification. [115] uses rule-based clustering methods which outperform the other baseline and state-of-the-art methods like Naive Bayes, BERT, Logistic Regression, RNN, LSTM, CNN-Glove, GRU-3-CNN in terms of AUC, Accuracy, Precision, Recall and F1-Score. [54] Performs semi-supervised multi-task learning utilizing a fuzzy ensemble approach in which they generated sequential and constructive rules to be added to the rule set and Latent Dirichlet Allocation [55] for implementing topic extraction and identifying hate speech forms

Table 7 Recent state-of-the-art for detecting hate speech in various languages via deep and hybrid learning models

Deep and Hybrid Learning				
Methods	Ref	Models/Algorithms	Classification	Dataset
Supervised methods	General (English)			
	[111]	CNN	(Hate, Non-Hate, Racism, Sexism)	Twitter
	[41]	LSTM	(Sexism, Racism, None)	Twitter
		GBDT		
	[105]	LSTM	(Sexism, Racism, None)	Twitter
		RNN		
	[113]	CNN + GRU	(Sexism, Racism, Both, Non-Hate, and Hate)	Twitter
	Religious (Arabic)			
	[119]	GRU based RNN	(Hate and Non-Hate)	Twitter
	Roman Urdu			
	[73]	Naive Bayes	(Neutral-Hostile, Simple-Complex, Offensive- Hate Speech)	Twitter
		Logistic regression		
		Random forest		
		SVM		
		CNN		
Semi-Supervised ***Methods***	[54]	Fuzzy ensemble approach	(Religious, Race, Disability, Sexual Orientation)	Twitter
		CNN		
		LSTM		
		GBT		
		SVM		
		Naive Bayes		
		Decision Tree		

for four classes from the Twitter dataset. The authors also proved that the fuzzy-based approach [54], metaheuristic approaches ([116, 117]) and Interpretable approach [118] had outperformed other techniques with high detection rate. [119] performs a supervised hybrid learning approach for classifying hate speech into two labels, specifically in the Arabic dialect. Moreover, Baysian attention networks, which follow the architecture of transformer models, are implemented for multilingual (English, Croatian and Slovene) contexts [120].

3.3 Merits and demerits of various models

Hate speech identification is a very much prevalent research field now a day. Researchers worldwide are experimenting with various models for specific field detection with numerous advantages and disadvantages. [121] implemented CNN and BERT models and proved efficient accuracy with intra-domain and cross-domain datasets. ([122, 123]) used FCM, SCM, and TKM for concatenating/combining features extracted from CNN and R-CNN, respectively, on textual and visual Twitter data, giving an advantage resulting in good accuracy compared to other baseline models. [124] used ELMO, BERT, and CNN to improve classification results but with higher time complexity. [125] also have a limitation of higher computational complexity, yet they created their detection system and implemented a deep belief network on labeled and unlabeled data. [116] presents two metaheuristic optimization algorithms (Ant Lion Optimization and Moth Flame Optimization) for the first time to solve Hate Speech Detection Problem with an efficient accuracy of above 90%. [117] implemented enhanced seagull optimization algorithm on CrowdFlower and StormFront datasets claiming the outperforming scores of above 98%. The pros and cons of the latest state-of-the-art works on hate speech detection are shown in Table 8.

4 Hate speech datasets

Social media platforms are prevalent nowadays, and users are increasing tremendously. Due to this, hate speech contents in various forms are at its peak. The presence of a massive amount of data on the web and collecting a good and relevant amount of data is challenging for researchers. Social media stages provide simple and easy approaches to gathering data using their APIs [21]. However, data assortment is not confined to APIs only. Figure 9 shows various ways of accessing data from social media.

Table 8 Merits and demerits depicted from the latest state-of-the-art works in hate speech detection

Refs	Objective	Features	Models	Merits	Demerits
[126]	To include the diacritics in detecting Vietnamese hate content	One hot embedding	● RNN ● LSTM ● GRU ● BERT	Results from models with diacritics generation outperformed others without diacritics generation	Only character-level feature is applied for implementation
[63]	Sentiment knowledge-based hate speech detection	● Word Embedding ● N-grams	Attention-based neural network	Including sentiments increases the overall performance	Small-scale datasets ate chosen for system generation can lead to uncovering bias
[115]	To propose a method for dealing with the problem of hate speech classification	●Tri-gram ●Wordngram	Rule-based clustering	Robustness to data imbalance	Complexity is directly proportional to the size of the dataset
[127]	To extract multimodal information	Character embedding	● CNN ● Attention-based Bi-GRU	High attention score for social context feature	Misclassification due to a few instances of hate words in the dataset
[128]	Hate content detection on textual data, along with image processing	TF-IDF	● CNN ● VGG ● RNN ● LSTM	Good performance on the combination of image, text, and captions	_
[129]	Cross-Lingual hate speech detection using capsule network	FastText	● CNN ● LSTM ● Bi-LSTM	Capsule network learning outperforms the other state-of-the-art models for multiple language pairs	Non-translation of words used in hashtags
[116]	Metaheuristic Optimization based Automatic detection of hate content	● BoW ● TF-IDF ● Word2vec	● Ant-Lion Optimization ● Moth- Flame Optimization	Metaheuristic Optimization-based approach has shown the overall increased performance	Higher time complexity
[130]	Abusive context detection in Greek contents on Refugees and Migrants	● Unigram ● Bigram ● Trigram	● BERT ● Resnet	Good accuracy achieved for specific domain content	Social-based information is missing
[131]	To identify Arabic hate content using multi-task learning	Word Embedding	● BERT	The proposed model outperformed existing models already applied to Arabic datasets	Contextual representation is not good while generating outcomes
[121]	To analyze a large dataset of about 1 million realistic hate and non-hate sequences and highlight the problems that may occur when training a hate speech classifier on small datasets	● POS ● Sentiment features ● TF-IDF	● CNN ● BERT	Good accuracy with intra-domain as well as cross-domain datasets	Implemented on textual tweets only
[64]	Hate speech detection for the Spanish language	● Word2vec ● GloVe ● FastText ● TF-IDF	● Logistic Regression ● SVM ● LSTM ● CNN ● Bi-LSTM ● Transfer Learning	Transfer Learning achieves the best results	Limited to training the model on monolingual language only

Table 8 (continued)

Refs	Objective	Features	Models	Merits	Demerits
[132]	To detect hate speech with their proposed system that classifies the text into three classes with 98.71% accuracy	● Sentimental ● Unigram ● Pattern	● Deep classifiers ● Gradient Descent Process ● Ensemble Classifiers ● Logistic Regression	● Proposed method gives good accuracy with minimum loss function values ● Used ensemble approach for the proposed system results in improving the overall hate speech recognition rate	The proposed system is designed and implemented on textual data only
[124]	To improve the results of the classification	Word2vec	● Elmo ● BERT ● CNN	Provides comparable suitability on feature sets	Time complexity is high
[133]	Identification of multilingual (English, Italian, and German) hate speech	● Word Embedding ● N-grams ● Social network- specific feature ● Emoji Embedding ● Emotion Lexica ● FastText	● LSTM ● GRU ● Bi-LSTM	LSTM yields better results on limited length of tweets than Bi-LSTM	Adding unigram and bigram as features with LSTM decreases the performance of all three languages
[125]	To create a system detecting Hate Speech	GloVe	Deep belief network	Deep learning includes supervised and unsupervised algorithms that take both labeled and unlabeled data	The process is expensive due to the sampling technique
[122]	To distinguish hate content in three data sources: the tweet picture, the tweet text, and the text showing up in the image	● Feature Concatenation Model (FCM) ● Spatial Concatenation Model (SCM) ● Textual Kernels Model (TKM)	● CNN ● LSTM	Multi-domain detection with good accuracy	● Multimodal data mainly contains textual processing ● Relations between textual and visual elements are very complex
[123]	To fetch the actual caption and then combine it with multimodal representation to perform binary classification, mainly focused on textual and visual modality	● Feature Concatenation Model (FCM) ● Spatial Concatenation Model (SCM) ● Textual Kernels Model (TKM)	R-CNN	Good accuracy when compared to baseline models	Establishing relations between textual and visual elements is very complex
[134]	To propose a multi-task model for catching the standard features and task-specific features from five labels (Sexism, Racism, Hate, Offensive, Harassment)	● Word Embedding ● CBOW ● Character embedding ● Word2vec	● LSTM ● CNN ● GRU	● Removes the data scarcity problem ● Achieves performance improvement statistically over single task-based systems	Domain-specific embedding is missing

Table 8 (continued)

Refs	Objective	Features	Models	Merits	Demerits
[54]	Setting multiple tasks for classifiers training	_	● Fuzzy Ensemble approach ● CNN ● LSTM ● LDA	● Proposed fuzzy methodology is appropriate for dealing with the diversity as shown by topic detection ● Proposed approach gives an intensity score for each kind of hate speech from a tweet ● Enables the analysis of the correlation between various labels	Approval of fuzzy-based approaches requires comprehensive testing
[135]	To identify hate speech in long archives	● TF-IDF ● Char quad-gram ● Word unigram ● Lexicon Features	● Naive Bayes ● SVM	Accuracy achievement is suitable even for long documents	● Data set is too small ● The features that ascertain the number of words with a positive and negative opinion, ignoring the negative opinions, drive the framework to recognize the sentiment of the words inaccurately
[89]	To automatically detect hateful content on Twitter from code-mixed text	● BoW ● Character n-grams ● word n-grams ● Negation words ● Punctuation marks	● Sub-word level LSTM ● Hierarchical LSTM	Character n-grams feature gives better results	Word and document features do not give better results
[136]	Focus on bias seriousness rather than bias removal from unstructured text data	● PoS ● WordNet ● Centroid Embedding	● Logistic regression ● Multi-layer Perceptron (MLP) ● CNN	Centroid embedding helps in getting bias-free vector representations by calculating the centroid of an adequate number of words, consequently giving improved performance	Bias mitigation ordinarily doesn't prompt any significant change in the ROC-AUC esteems, and it only decreases bias in the learning
[137]	Brazilian Portuguese text classification for detecting hate speech	TF-IDF	● Multinomial Naïve Bayes ● Random Forest ● SVM	SVM is giving the highest performance of about 95%	No relationship between semantics
[138]	To characterize whether a message in the sentence contains components of hate speech or not	Word2vec	Deep learning with RNN (LSTM Based)	● LSTM can conquer the vanishing gradient on standard RNN from long-haul conditions learning ● LSTM is proven to be more viable than standard RNN	Training parameters like learning rate affect the performance of the overall system
[41]	Handling the complexity of natural language constructs with multiple deep learning architectures	● Char n-grams ● TF-IDF ● BOW	● CNN ● LSTM ● Logistic Regression ● Random Forest ● SVM ● GBDT	● comparable words obtained using deep neural network learned embeddings clearly show the hatred towards the objective words ● Minimization of bogus negatives	Concatenations of CNN, LSTM, and FastText embedding as components for GBDTs didn't prompt better outcomes

Table 8 (continued)

Refs	Objective	Features	Models	Merits	Demerits
[139]	Capturing the biases in data annotations and detecting hate content with efficient accuracy	● Word Embedding	● BERT ● Bi-LSTM ● CNN	Enhancing the performance of hate speech detection system	Time complexity may be high

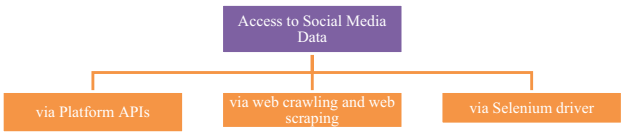

Fig. 9 Prominent ways to access data from social media

4.1 Dataset description

Hate speech identification has become a crucial task in many languages and fields. Recordings play a fundamental role in disseminating content as they can contact a vast crowd, including little youngsters. Appraisals say that 1 billion hours of videos are observed every day on YouTube alone. Detecting hate speech is important to give youngsters a protected climate and a healthy environment for clients in general. Until now, the text has been the most famous configuration utilized by researchers working on it. Subsequently, most current works summarize recognizing hate speech in the text (social platform posts, news remarks, tweets, and so on). While hate speech detection methods primarily use textual inputs, few research contributions exist toward multimodal hate speech detection. Several authors have generated multi-class/ multi-label datasets in various languages for curbing hate content on social media. Hate speech detection (HaSpeeDe) is the prevalent shared task organized within Evalita 2018 [80] and consists of manually annotating Italian messages taken from Twitter and Facebook. This shared task was further categorized into three sub-tasks: HaSpeeDe-FB, HaSpeeDe-TW, and Cross-HaSpeeDe.

4.2 Datasets challenges

- The available and widely used datasets ([38, 140]) have issues in their subjectiveness which introduces bias in the performances. Hate Speech datasets are affected mainly by social, behavioral, racial, temporal, and content production biases [141]. Data imbalance due to bias may lead to misclassification [142].
- One of the significant issues is the unlabeled non-English datasets. Few manually annotated labeled datasets were released for detecting offensive language and hate speech [78]. Moreover, Multilingual hate speech datasets can also share the writing of other languages. For example, a dataset can contain Farsi and Arabic tweets while creating an Urdu hate speech and offensive language dataset [75]
- The problem also arises when the web address of datasets changes [143]. Authors who create a new dataset do not publish those [73].

- Twitter (Lenient data usage policies) is the most prevalent platform. However, the Twitter resources are significant because of the exceptional classification of the Twitter posts, which is limited to short text. Henceforth, contents from other media stages are longer and can be a piece of more extensive conversation in hate speech.
- Datasets differ in their size, degree, and features of the data annotated, which prompts the issue of irregularity in the quantity of hate and non-hate texts within datasets. For example, on a social stage like Twitter, hate speech occurs at a shallow rate contrasted to non-hate. Therefore, researchers can gather data from social media platforms with no character length limit.

Given the above challenges, making data available in a superior arrangement for demand research is essential. Table 9 represents various benchmark models on multiple datasets. Commonly used datasets ([12, 38], Gomez et al., 2020)) benchmarks are also shown in the table.

The overall description of datasets regarding modalities (T-Text, I-Images, V-Videos), classes/ labels, languages, etc., are tabulated in Table 10.

5 Evaluation and performance measures

As datasets play a significant role in testing the performance of hate speech detection. The better-normalized dataset is the best performance an algorithm will give. In this section, metrics for evaluation of machine and deep learning techniques used are F_1-Score, Recall, and Precision, and performance measurement metrics are accuracy and AUC (Area under Curve).

5.1 Evaluation metrics

Most state-of-the-art have utilized accuracy, F_1-Score, Precision, Recall Metrics, and ROC to assess performance metrics. [132] represents several loss functions like mean MSE, cross-entropy, and likelihood loss to anticipate hate speech in the most used dataset, such as Twitter. The loss function is the difference between the predicted value denoted by ŷ and labeled value denoted by y. [143] use four different strong performances indicators (KPIs), which are the percentage of True Positive, the precision, the recall, & F_1- Score defined using Eq. 1:

$$F_1 - Score = 2 \times \frac{P - R}{P + R} \tag{1}$$

[132] uses several loss functions such as Mean Square Error Rate (MSE) [163], given in Eq. 2, Cross-Entropy Loss (CEL) [164] as in Eq. 3, and Likelihood Loss (L) [165] in Eq. 4 to approximate the accuracy of the proposed model in identifying hate speech on the Twitter dataset.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2 \tag{2}$$

Where,

N denoted the quantity of information relative to the predicted value ŷ and labeled y.

$$CEL = \sum_{c=1}^{M} y_{o,c} Log\left(P_{o,c}\right) \tag{3}$$

where,

M represents classes and related features,

O denotes the observed value of the particular class-related feature,

P represents the prediction probability value relevant to O,

Log is the logarithmic function, and.

Y gives the output value as binary values of a specific class.

Table 9 Benchmark models on datasets

Dataset	Refs	Benchmark models	F1-score	Accuracy	Precision	AUC- ROC
Ethos_binary	[144]	Bi-LSTM + Static BE	0.7971	0.8015	0.8037	_
Maha-Hate	[145]	BERT	_	0.909	_	_
HateXplain	[146]	BERT + HateXplain [Attn]	0.687	0.698	_	0.851
APEACH	[147]	BERT	0.8424	_	_	_
Told-Br	[148]	Multilingual BERT	0.75	_	_	_
Hateful memes	[109]	Vilio	_	_	_	0.825
OffVidPT-2/ OffVidPT-3	[149]	M-CNN	0.74	_	_	0.78
MMHS150K	[122]	TKM	0.70	68.2	_	0.731
t-Davidson	[12]	BERTbase + CNN	0.92	_	0.92	_
WaseemHavoy	[38]	BERTbase + CNN	0.88	_	0.89	_

Table 10 Dataset description in terms of size, labels, languages, and modalities

Source	Name of the dataset	Refs	Size	Classes/ labels	Language	Modality
Youtube and Reddit comments	Multi-labEl haTe speecH detectiOn dataSet (ETHOS)	[144]	998	Ethos_Binary ● Hate ● Non-Hate Ethos_Multi-Label ● Violence ● Non-Violence ● Gender ● Race ● National origin ● Disability ● Religion ● Sexual Orientation	English	T
Line today	Hate speech	[46]	~12 K and ~1 M	● Hate ● Offensive ● Normal	Taiwan	T
Twitter	ToxiGen	[150]	~2 M	● Toxic ● Benign	English	T
Twitter	Slovenian twitter Dataset 2018–2020 1.0	[151]	~12 M	● Acceptable ● Inappropriate ● Offensive ● Violent	Slovenian	T
Crowd sourcing platforms	APEACH	[147]	~11 K	● Hate ● Non-Hate	Korean	T
Twitter	Maha-Hate	[145]	25 K	● Hate ● Offensive ● Profane ● Not	Marathi	T
Twitter	Urdu hate speech and offensive language	[75]	~11 K	● Hate ● Offensive ● Neutral	Roman-Urdu	T
Twitter and GAB	HateXplain	[146]	9055 twitter posts, 11,093 GAB posts	● Hate ● Offensive ● Normal	English	T
Handcrafted	HatemojiCheck	[152]	~4 K Test cases	● Hateful ● Not Hateful	English	I
Youtube and Facebook	HS-BAN	[153]	50 K	● Hate ● Non-Hate	Bangla	T
Twitter	ToLD-Br	[148]	21 K	● Toxic ● Non- Toxic	Portuguese	T
Facebook	Hateful Memes	[108]	10 K	● Hateful ● Not Hateful	English	I+T
Youtube	OffVidPT-2/ OffVidPT-3	[149]	400	● Offensive ● Non- offensive	Portuguese	V
Twitter	MMHS150K	[122]	150 K	● No attack ● Racism ● Sexism ● Homophobic ● Religion-based attacks ● Any community attacks to other attacks	English	I+T
Bengali wikipedia dump and Bengali news articles	Bengali_hate_speech v1.0	[154]	3418 Samples	● Personal hate ● Political hate ● Religious hate ● Geopolitical hate ● Gender Abusive hate	Bengali	T

Table 10 (continued)

Source	Name of the dataset	Refs	Size	Classes/ labels	Language	Modality
Youtube	Hate_speech_dataset_videos	[155]	300	● Normal ● Hateful (Racist, Sexist)	English	V
Wikipedia comments + Civil comments	Detoxify	_	~55 K	● Very toxic ● Toxic ● Hard to say ● Non- toxic	English, French, Spanish, Italian, Portuguese, Turkish, Russian	T
Twitter	OffensEval2020	[156]	~9 M	● Offensive ● Non-Offensive	Arabic, Danish, English, Greek, and Turkish	T
Twitter	OffensEval2019	[157]	~14 K	● Offensive ● Non-Offensive	English	T
Facebook and YouTube	T-HSAB	[158]	~6 K	● Normal ● Abusive ● Hate	Tunisian Arabic	T
Twitter	L-HSAB	[159]	~6 K	● Normal ● Abusive ● Hate	Levantine Arabic	T
Twitter	HatEval19	[140]	19 K	● Hate ● Non-Hate	Spanish, English	T
Twitter	Peer-to-Peer Hate	[160]	28 K	● Hate ● Non-Hate	English	T
Twitter and Facebook	Evalita2018	[80]	4 K posts	● No-Hate ● Weak- Hate ● Strong- Hate	Italian	T
Twitter	Twitter Abusive Behaviour	[161]	80 K	● Normal ● Spam	English	T
Twitter	t-Davidson	[12]	25 K	● Hate ● Offensive ● Neither	English	T
News articles	BEEP	[161]	9381 Human Labeled comments & 2,033,893 Unlabeled comments	● Hate ● Offensive but not hate ● None	Korean	T
Twitter	Arabic offensive	[162]	1.1 K and 32 K	● Obscene ● Offensive ● Clean	Modern Standard Arabic	T
Twitter	WaseemHavoy	[38]	~17 K	● Racist ● Sexist ● Neither	English	T

$$L = -\frac{1}{n}\Sigma_{i=1}^{n} \log(\hat{Y}_{(i)}) \tag{4}$$

where,

n gives the number of classes.

y denotes the output.

5.2 Performance of popular hate speech detection methods

Most state-of-the-art on hate speech detection used precision, recall, and F_1-score for evaluation; others used AUC and accuracy for performance measures due to some imbalanced datasets. Table 11 gives evaluation and performance measures from some state-of-the-art works based on accuracy, precision, recall, F_1 score, and AUC. As seen in Table 11, Precision, Recall, and F_1-score are the metrics used by most authors as they provide better insights into the prediction than accuracy and AUC. Deep-learning models have outperformed machine-learning models with high-performance metrics, as presented in Table 11.

Table 11 Performance comparison

Ref	Algorithms	Accuracy	Precision	Recall	F1-Score	AUC
[115]	Rule-based clustering	0.94	0.92	0.91	0.92	0.96
[64]	BERT, XLM, BETO	–	–	–	0.772	–
[149]	BERT, CNN, Random forest, Naive bayes	–	–	–	0.74	0.78
[132]	Deep learning	0.9873	–	–	–	–
[166]	Linear SVM classifier	0.90	0.88	1.0	0.90	–
	Naive bayes classifier	0.92	0.899	0.964	0.924	–
[167]	BERT	–	0.86	0.94	0.77	–
[143]	Random forest, SVM, and J48graft [168]	–	0.88	0.87	0.87	–
[41]	Logistic regression, SVM, DNN, CNN, Random forest, GBDT	–	0.93	0.93	0.93	–
[12]	Logistic regression, SVM	–	0.91	0.90	0.90	–
[50]	SVM, LSTM	-	0.83	0.87	0.85	–
[49]	One-class classifiers, Decision tree, Naïve bayes, Random forest	–	0.73	0.86	–	–
[18]	Skip bigram model	–	0.83	0.83	0.83	–
[110]	Deep learning	0.91	–	–	–	–
[38]	Logistic regression	–	0.72	0.77	0.73	–
[33]	SVM, Random forest, Decision tree	–	0.79	0.59	0.68	–
[169]	SVM	–	0.49	0.43	0.46	–
[170]	Random forest, Decision tree, SVM, Bayesian LR	–	0.89	0.69	0.77	–
[85]	Logistic regression	–	–	–	–	0.80
[39]	Naïve Bayes	–	0.97	0.82	–	–
[72]	Decision tree, Random Forest, SVM	–	0.89	0.69	0.77	–
[34]	Naïve bayes	0.73	–	–	–	–
[65]	SVM	0.94	0.68	0.60	0.63	–

6 Discussion

Hate speech is an emerging issue in social media sites now days. The identification of hate content is one of the major concern and challenge for the researchers. The proposed article shows a systematic order of state-of-the-art works done so far. Feature extraction methods such as distance metric and multimodal information, especially related to hate content detection, are not used to the best of my knowledge. Both directional models such as RNN and LSTM, and non-directional models such as Transformers ad BERT are utilized in identifying hate content. Although machine learning has shown its growth in the last decade, but NLP has also shown the steepest growth by including the evolutionary models such as BERT and Transformers. The variants of BERT like ALBERT, RoBERTa, DistilBERT etc. are used increasingly in solving real life problems because of their self-attention mechanism. The researchers also use LSTM model as it yields subsequently higher results than BERT on small datasets does. The pros and cons of various models are described in detail in the Sect. 3.3. From the last two years, the metaheuristic optimization algorithms such as Ant Lion optimization, Flame Moth optimization, Seagull optimization are also considered in this area with the promising results.

7 Findings and conclusion

Hate speech attempts to marginalize different classes and groups of persons already in the minority due to their race, language and religion. This article reviewed the most outstanding work on automatic hate speech identification. Firstly, we introduced some state-of-the-art hate speech definitions and analysis on the basis of some specific dimensions. This survey also highlights some of the NLP aspects in this area. There is also a good comparison between hate speech definition and definitions of various hate forms. Then, we presented a taxonomy of automatic hate speech detection, including sub-domains of AI approaches. Metaheuristic algorithms which are very new with context of hate speech detection are also mentioned in this manuscript. Paper also covers various works done in multilingual and multimodal hate speech detection along with various datasets description.

8 Future trends

Our studies recommend some future trends from the following angles:

- We have explored some standard hate speech datasets along with their key features, classifications, objectives, and types of data format available. Most datasets are available in textual form. Very few datasets like (MultiOFF, MMHS150K) are found on hateful memes. No video dataset is found publicly as per the best of my knowledge. So, creating a new dataset of images and videos can be further seen as a future task. Moreover, numerous analysts look at the significant challenge of the datasets availability as few publicly available datasets exist. Authors do not use them, and if they create a new dataset, they do not publish them, making it too difficult to compare results and conclusions.
- Choosing informative, independent, and discriminating features are crucial in classification problems. This paper covers commonly used text analysis features for hate classification tasks. Hence, automatic feature engineering for generating specific hate features can be a future aspect.
- For the last few years, authors have been focusing on multilingual hate speech identification by creating their datasets. But very few labeled datasets are found in non-English languages. Various benchmark models can be applied to non-English labeled datasets also.
- We have also covered important work for hate speech identification in various languages. Hence, the models that understood only the English language are not efficient in processing the input from different Indian languages [78]. So, building a system for code-mixed languages can be considered a future aspect.
- Nowadays, emojis are also used to show feelings and attitudes in users minds [36], and they are vital elements in delivering hate or offensive content over social media. Hence, pre-processing emojis text can be seen as a different area so that there can be an improvement in aggression detection.
- There are significantly fewer works on neutral tagged content [75]. So, devising a new method for handling neutral tagged contents in multi-label datasets in a better way can be considered a future job.
- In our systematic survey, we tracked that most work portrays techniques, separated features, and models utilized. In any case, it is uncommon to discover jobs with available public repositories. More sharing of code, calculations, measures for feature extraction, and stages can assist the area with advancing rapidly.
- In this article, some of the metaheuristic optimization approaches are also coined to solve hate speech detection. Apart from the mentioned metaheuristic approaches such as ALO and FMO, Parameter Optimization approach can also be implemented in future for solving the hate content detection.

Author contributions Anusha Chhabra: Software, Validation, Investigation, Data Curation, Writing – Original Draft, Visualization.Dinesh Kumar Vishwakarma: Conceptualization, Methodology, Formal Analysis, Resources, Writing – Review & Editing, Supervision, Project Administration, Funding Acquisition.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. M. Bose, "South Asia Journal," 2020. http://southasiajournal.net/india-senior-bjp-leaders-are-giving-india-a-free-tutorial-in-hate-speech/
2. R. E. Brannigan, J. L. Moss, and J. Wren, "The conversation," *Fertility and Sterility*, 2015. https://theconversation.com/hate-speech-is-still-easy-to-find-on-social-media-106020.
3. M. Suster, "Business Insider," *Amazon's Game-Changing Cloud Was Built By Some Guys In South Africa*, 2010. https://www.businessinsider.com/736-of-all-statistics-are-made-up-2010-2?r=US&IR=T%0Ahttp://www.businessinsider.com/amazons-game-changing-cloud-was-built-by-some-guys-in-south-africa-2012-3.
4. A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection using Natural Language Processing," *Soc. 2017 - 5th Int. Work. Nat. Lang. Process. Soc. Media, Proc. Work. AFNLP SIG Soc.*, no. 2012, pp. 1–10, 2017, doi: https://doi.org/10.18653/v1/w17-1101.
5. Cohen-Almagor, R.: Freedom of Expression v. Social Responsibility: Holocaust Denial in Canada. J. Mass Media Ethics Explor. Quest. Media Moral. **28**(1), 42–56 (2013). https://doi.org/10.1080/08900523.2012.746119
6. Delgado, R., Stefancic, J.: Images of the outsider in American law and culture: can free expression remedy deeply inscribed social Ills? Fail. Revolutions **77**(6), 3–21 (2019). https://doi.org/10.4324/9780429037627-2
7. Techterms.com, "Facebook Definition," 2008. http://www.techterms.com/definition/facebook.
8. Youtube, "YouTube hate policy," 2019. https://support.google.com/youtube/answer/2801939?hl=en.
9. Facebook, "What does facebook consider hate speech?," 2013. https://www.facebook.com/help/135402139904490.
10. Nockleby, J.T.: Hate Speech. In: Levy, L.W., Karst, K.L., et al. (eds.) Encyclopedia of the American Constitution, pp. 1277–1279. Macmillan, New York (2000)
11. Twitter, "Twitter_Hate Definition [online]," 2017. https://support.twitter.com/ articles/.
12. Davidson, T., Warmsley, D., Macy, M., Webe, I.: Automated hate speech detection and the problem of offensive language. Proc. 11th Int. Conf. Web Soc. Media, ICWSM **11**(1), 512–515 (2017)
13. de Gibert, O., Perez, N., García-Pablos, A., Cuadros, M.: Hate Speech Dataset from a White Supremacy Forum. arXiv preprint arXiv (2019). https://doi.org/10.18653/v1/w18-5102
14. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Comput. Surv. (2018). https://doi.org/10.1145/3232676
15. Chen, Y., Zhou, Y., Zhu, S., Xu, H.: Detecting offensive language in social media to protect adolescent online safety. Proc. - 2012 ASE/IEEE Int. Conf. Privacy, Secur. Risk Trust 2012 ASE/IEEE Int. Conf. Soc. Comput. Soc (2012). https://doi.org/10.1109/SocialCom-PASSAT.2012.55

16. Thompson, N.: Equality, Diversity and Social Justice. Sixth, PALGRAVE MACMILLAN (2016)
17. Guermazi, R., Hammami, M., Ben Hamadou, A.: Using a semi-automatic keyword dictionary for improving violent web site filtering. Proc. - Int. Conf. Signal Image Technol. Internet Based Syst. SITIS (2007). https://doi.org/10.1109/SITIS.2007.137
18. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: "Abusive language detection in online user content." 25th Int World Wide Web Conf. WWW **2016**, 145–153 (2016). https://doi.org/10.1145/2872427.2883062
19. Google and Jigsaw, "Perspective API," 2017. https://perspectiveapi.com.
20. Asia Centre, "Hate speech in Southeast Asia. New forms, old rules," 2020. [Online]. Available: https://asiacentre.org/wp-content/uploads/2020/07/Hate-Speech-in-Southeast-Asia-New-Forms-Old-Rules.pdf.
21. Lomborg, S., Bechmann, A.: Using APIs for data collection on social media. Inf. Soc. **30**(4), 256–265 (2014). https://doi.org/10.1080/01972243.2014.915276
22. Facebook, "Facebook [Online]," 2022. https://www.facebook.com/about/privacy/update.
23. Lindsey, "Instagrams-Api," 2022. https://rapidapi.com/blog/how-to-navigate-and-connect-to-instagrams-api/ (accessed Mar. 09, 2022).
24. Twitter_Rules, "https://help.twitter.com/en/rules-and-policies/twitter-api," 2022. https://help.twitter.com/en/rules-and-policies/twitter-api.
25. M. S. Jahan and M. Oussalah, "A systematic review of Hate Speech automatic detection using Natural Language Processing," arXiv:2106.00742v1, 2021, [Online]. Available: http://arxiv.org/abs/2106.00742.
26. Dhanya, L.K., Balakrishnan, K.: "Hate speech detection in asian languages: a survey", ICCISc 2021–2021 Int. Conf. Commun. Control Inf. Sci. Proc. (2021). https://doi.org/10.1109/ICCISc52257.2021.9484922
27. Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., Patti, V.: Resources and benchmark corpora for hate speech detection: a systematic review. Lang. Resour. Eval. **55**(2), 477–523 (2021). https://doi.org/10.1007/s10579-020-09502-8
28. N. Naaz, Y. Malik, and K. P. Adhiya, "Hate Speech Detection in Twitter-A Survey," *Int. J. Manag. Technol. Eng.*, vol. 9, no. 1, pp. 1272–1277, 2019, [Online]. Available: http://www.ijamtes.org/gallery/147-jan19.pdf.
29. Robinson, D., Zhang, Z.: Detection of hate speech in social networks: a survey on multilingual corpus. Comput. Sci. Inf. Technol. (2020). https://doi.org/10.5121/csit.2019.90208
30. Alrehili, A.: Automatic hate speech detection on social media: A brief survey. Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA (2019). https://doi.org/10.1109/AICCSA47632.2019.9035228
31. Mohiyaddeen and Dr: Shifaulla Siddiqui, "Automatic hate speech detection: a literature review." Int. J. Eng. Manag. Res. **11**(2), 116–121 (2021). https://doi.org/10.31033/ijemr.11.2.17
32. Araque, O., Iglesias, C.A.: An Ensemble Method for Radicalization and Hate Speech Detection Online Empowered by Sentic Computing. Cognit. Comput (2022). https://doi.org/10.1007/s12559-021-09845-6
33. Burnap, P., Williams, M.L.: Us and them: identifying cyber hate on Twitter across multiple protected characteristics. EPJ. Data Sci. (2016). https://doi.org/10.1140/epjds/s13688-016-0072-6
34. Kwok, I., Wang, Y.: Locate the hate: Detecting tweets against blacks. Proceedings of the 27th AAAI Conference on Artificial Intelligence, AAAI (2013). https://doi.org/10.1609/aaai.v27i1.8539
35. Baydoğan, V.C., Alatas, B.: Çevrimiçi Sosyal Ağlarda Nefret Söylemi Tespiti için Yapay Zeka Temelli Algoritmaların Performans Değerlendirmesi. Fırat Üniversitesi Mühendislik Bilim. Derg. **33**(2), 745–754 (2021). https://doi.org/10.35234/fumbd.986500
36. Husain, F., Uzuner, O.: "Investigating the Effect of Preprocessing Arabic Text on Offensive Language and Hate Speech Detection", *ACM Trans*. Asian Low-Resource Lang. Inf. Process. **21**(4), 1–20 (2022). https://doi.org/10.1145/3501398
37. Chowdhury, A.G.: ARHNet - Leveraging Community Interaction For Detection Of Religious Hate Speech In Arabic". Proc. 57th Annu. Meet. te Assoc. Comput. Linguist. **2019**, 273–280 (2019)
38. Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," pp. 88–93, 2016, doi: https://doi.org/10.18653/v1/n16-2013
39. Liu, S., Forss, T.: Combining N-gram based similarity analysis with sentiment analysis in web content classification. KDIR 2014 - Proc. Int. Conf. Knowl. Discov. Inf. Retr (2014). https://doi.org/10.5220/0005170305300537
40. Greevy, E., Smeaton, A.F.: Classifying racist texts using a support vector machine. Proc. Sheff. SIGIR - Twenty-Seventh Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr (2004). https://doi.org/10.1145/1008992.1009074
41. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. 26th Int. World Wide Web Conf. 2017, WWW 2017 Companion **2**, 759–760 (2017). https://doi.org/10.1145/3041021.3054223
42. Katona, E., Buda, J., Bolonyai, F.: Using N-grams and Statistical Features to Identify Hate Speech Spreaders on Twitter. CEUR Workshop Proc. **2021**, 2025–2034 (2021)
43. Mehdad, Y., Tetreault, J.: "Do Characters Abuse More Than Words? Dialogue. SIGDIAL 2016 - 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference (2016). https://doi.org/10.18653/v1/w16-3638
44. Mulki, H., Ali, C.B., Haddad, H., Babao, I.: Tw-StAR at SemEval-2019 Task 5: N-gram embeddings for Hate Speech Detection in Multilingual Tweets. Proc. 13th Int Work. Semant. Eval. **2019**, 503–507 (2019)
45. S. N. Group, "Stanford NLP Group," 2005. https://nlp.stanford.edu/.
46. Wang, C., Day, M., Wu, C.: Political Hate Speech Detection and Lexicon Building : A Study in Taiwan. IEEE Access **10**, 44337–44346 (2022). https://doi.org/10.1109/ACCESS.2022.3160712
47. Liu, S., Forss, T.: "New classification models for detecting hate and violence web content Knowl. IC3K 2015 - Proc. 7th Int. Jt. Conf. Knowl. Discov. Knowl. Eng. Knowl. Manag **1**, 487–495 (2015). https://doi.org/10.5220/0005636704870495
48. Gitari, N.D., Zuping, Z., Damien, H., Long, J.: A lexicon-based approach for hate speech detection. Int. J. Multimed. Ubiquitous Eng. **10**(4), 215–230 (2015). https://doi.org/10.14257/ijmue.2015.10.4.21
49. S. Agarwal and A. Sureka, "Characterizing Linguistic Attributes for Automatic Classification of Intent Based Racist/Radicalized Posts on Tumblr Micro-Blogging Website," 2017, [Online]. Available: http://arxiv.org/abs/1701.04931.
50. Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., Tesconi, M.: Hate me, hate me not: Hate speech detection on Facebook. CEUR Workshop Proc. **1816**, 86–95 (2017)
51. Ali, M.Z., Rauf, S., Javed, K., Hussain, S.: Improving hate speech detection of urdu tweets using sentiment analysis. IEEE Access **9**, 84296–84305 (2021). https://doi.org/10.1109/ACCESS.2021.3087827
52. Baydogan, C., Alatas, B.: Sentiment analysis in social networks using social spider optimization algorithm. Teh. Vjesn. **28**(6), 1943–1951 (2021). https://doi.org/10.17559/TV-20200614172445

53. Pablo, J., Jiménez, J.: Topic modelling of racist and xenophobic YouTube comments. Analyzing hate speech against migrants and refugees spread through YouTube in Spanish. TEEM'21 Ninth Int. Conf. Technol. Ecosyst. Enhancing Multicult **2021**, 456–460 (2021)
54. Liu, H., Alorainy, W., Burnap, P., Williams, M.L.: Fuzzy multi-task learning for hate speech type identification. Web Conf. 2019 - Proc. World Wide Web Conf. WWW (2019). https://doi.org/10.1145/3308558.3313546
55. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**(4–5), 993–1022 (2003). https://doi.org/10.1016/b978-0-12-411519-4.00006-9
56. V. Mujadia, "IIIT-Hyderabad at HASOC 2019 : Hate Speech Detection," *CEUR Workshop Proc.*, 2019.
57. Kumar, P.M.A., Pradesh, A.: Hate Speech Detection using Text and Image Tweets Based On Bi-directional Long Short-Term Memory. 2021 Int. Conf. Disruptive Technol. Multi-Disciplinary Res. Appl **2021**, 158–162 (2021)
58. Dinakar, K., Reichart, R., Lieberman, H.: Modeling the detection of textual cyberbullying. AAAI Work. - Tech. Rep. **11–02**, 11–17 (2011)
59. Gaydhani, A., Doma, V., Kendre, S., Bhagwat, L.: Detecting Hate Speech and Offensive Language on Twitter using Machine Learning : An N-gram and TFIDF based Approach. IEEE Int. Adv. Comput. Conf **1809**, 08651 (2018)
60. Gambino, G., Pirrone, R., Ingegneria, D.: CHILab @ HaSpeeDe 2: Enhancing Hate Speech Detection with Part-of-Speech Tagging. CEUR Workshop Proc. **2020**, 165 (2020)
61. Erizal, E., Setianingsih, C.: "Hate Speech Detection in Indonesian Language on Instagram Comment Section Using Maximum Entropy Classification Method." 2019 Int Conf. Inf. Commun. Technol. **2019**, 533–538 (2019)
62. Bilal, M., Khan, A., Jan, S., Musa, S.: Context-Aware Deep Learning Model for Detection of Roman Urdu Hate Speech on Social Media Platform. IEEE Access **10**, 121133–121151 (2022). https://doi.org/10.1109/ACCESS.2022.3216375
63. Zhou, X., et al.: "Hate Speech Detection based on Sentiment Knowledge Sharing." Proc. 59th Annu. Meet. Assoc. Comput. Linguist. 11th Int Jt. Conf. Nat. Lang. Process. **2021**, 7158–7166 (2021)
64. Plaza-del-Arco, F.M., Molina-González, M.D., Ureña-López, L.A., Martín-Valdivia, M.T.: Comparing pre-trained language models for Spanish hate speech detection. Expert Syst. Appl. **166**, 114120 (2021). https://doi.org/10.1016/j.eswa.2020.114120
65. W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in *Proceeding LSM '12 Proceedings of the Second Workshop on Language in Social Media*, 2012, no. Lsm, pp. 19–26, [Online]. Available: http://dl.acm.org/citation.cfm?id=2390374.2390377.
66. Haralambous, Y., Lenca, P.: Text classification using association rules, dependency pruning and hyperonymization. CEUR Workshop Proc. **1202**, 65–80 (2014)
67. Abro, S., Shaikh, S., Ali, Z.: Automatic Hate Speech Detection using Machine Learning : A Comparative Study. Int. J. Adv. Comput. Sci. App. **11**(8), 484–491 (2020)
68. Baydogan, C., Alatas, B.: Deep-Cov19-Hate: A Textual-Based Novel Approach for Automatic Detection of Hate Speech in Online Social Networks throughout COVID-19 with Shallow and Deep Learning Models. Teh. Vjesn. **29**(1), 149–156 (2022). https://doi.org/10.17559/TV-20210708143535
69. Chiril, P., Wahyu, E., Farah, P., Véronique, B., Viviana, M., Patti, V.: Emotionally Informed Hate Speech Detection : A Multi - target Perspective. Cognit. Comput. (2022). https://doi.org/10.1007/s12559-021-09862-5
70. Mullah, N.S., Zainon, W.M.N.W.: Advances in machine learning algorithms for hate speech detection in social media: a review. IEEE Access **9**, 88364–88376 (2021). https://doi.org/10.1109/ACCESS.2021.3089515
71. Naseem, U., Razzak, I., Eklund, P.W.: "A survey of pre-processing techniques to improve short-text quality : a case study on hate speech detection on twitter. Multimed Tools Appl (2020). https://doi.org/10.1007/s11042-020-10082-6
72. P. Burnap and M. Williams, "Hate Speech, Machine Classification and Statistical Modelling of Information Flows on Twitter: Interpretation and Communication for Policy Decision Making," in *Internet, Policy & Politics*, 2014, pp. 1–18, [Online]. Available: http://orca.cf.ac.uk/id/eprint/65227%0A.
73. Khan, M.M., Shahzad, K., Malik, M.K.: "Hate speech detection in Roman Urdu", *ACM Trans*. Asian Low-Resource Lang. Inf. Process. **20**(1), 1–19 (2021). https://doi.org/10.1145/3414524
74. Hua, T., Chen, F., Zhao, L., Lu, C.-T., Ramakrishnan, N.: STED: semi-supervised targeted-interest event detection". Knowledge Discov. Data Mining **2013**, 1466–1469 (2013)
75. Ali, R., Farooq, U., Arshad, U., Shahzad, W., Omer, M.: Computer speech & language hate speech detection on Twitter using transfer learning. Comput. Speech Lang. **74**, 101365 (2022). https://doi.org/10.1016/j.csl.2022.101365
76. Ma, C., Du, X., Cao, L.: Improved KNN algorithm for fine-grained classification of encrypted network flow. Mdpi Electron (2020). https://doi.org/10.3390/electronics9020324
77. Ferreira, P.J.S., Cardoso, J.M.P.: k NN prototyping schemes for embedded human activity recognition with online learning. Mdpi Comput (2020). https://doi.org/10.3390/computers9040096
78. Kumar, P., Bhawal, S.: Computer speech & language hate speech and offensive language detection in Dravidian languages using deep ensemble framework. Comput. Speech Lang. (2022). https://doi.org/10.1016/j.csl.2022.101386
79. Alfina, I., Mulia, R., Fanany, M.I., Ekanata, Y.: 2018 "Hate speech detection in the Indonesian language: a dataset and preliminary study. 2017 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS **2018**, 233–237 (2017). https://doi.org/10.1109/ICACSIS.2017.8355039
80. Bosco, C., Orletta, F.D., Poletto, F., Tesconi, M.: Overview of the EVALITA 2018 hate speech detection task. CEUR Workshop Proc. (2018). https://doi.org/10.4000/books.aaccademia.4503
81. Bai, X., Merenda, F., Zaghi, C., Caselli, T., Nissim, M.: RuG EVALITA 2018: hate speech detection in Italian social media. CEUR Workshop Proc. **2263**, 1–5 (2018)
82. Chen, H., McKeever, S., Delany, S.J.: Abusive text detection using neural networks. CEUR Workshop Proceedings, 2086(2), 258–260.ction using neural networks. CEUR Workshop Proc. **2086**(2), 258–260 (2017)
83. M. Wiegand, J. Ruppenhofer, A. Schmidt, and C. Greenberg, "Inducing a lexicon of abusive words ? a feature-based approach," NAACL HLT 2018 - 2018 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, pp. 1046–1056, 2018, doi: https://doi.org/10.18653/v1/n18-1095.
84. Pawar, R., Agrawal, Y., Joshi, A., Gorrepati, R., Raje, R.R.: "Cyberbullying detection system with multiple server configurations. IEEE Int. Conf. Electro Inf. Technol. (2018). https://doi.org/10.1109/EIT.2018.8500110
85. N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate speech detection with comment embeddings," in WWW 2015 companion—proceedings of the 24th international conference on World Wide Web, 2015, pp. 29–30, doi: https://doi.org/10.1145/2740908.2742760.
86. W. Z, "Are you a racist or am i seeing things? Annotator influence on hate speech detection on twitter.," in proceedings of the first

workshop on NLP and computational social science. Association for computational linguistics, 2016, pp. 138–142.
87. Malmasi, S., Zampieri, M.: Detecting hate speech in social media. Int. Conf. Recent. Adv. Nat. Lang. Process. RANLP. **2017**, 467–472 (2017). https://doi.org/10.26615/978-954-452-049-6-062
88. M. R. Jha A, "When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In:," In proceedings of the second workshop on NLP and computational social science. Association for Computational Linguistics, 2017, pp. 7–16.
89. Santosh, T.Y.S.S., Aravind, K.V.S.: Hate speech detection in Hindi-English code-mixed social media text. ACM Int. Conf. Proc. Ser. (2019). https://doi.org/10.1145/3297001.3297048
90. Özel, S.A., Akdemir, S., Saraç, E., Aksu, H.: "Detection of cyberbullying on social media messages in Turkish", *2nd Int.* Conf. Comput. Sci. Eng. UBMK **2017**, 366–370 (2017). https://doi.org/10.1109/UBMK.2017.8093411
91. M. Fernandez and H. Alani, "Contextual semantics for radicalisation detection on Twitter," *CEUR Workshop Proc.*, vol. 2182, 2018.
92. Abozinadah, E.A., Mbaziira, A.V., Jones, J.H.J.: Detection of abusive accounts with Arabic tweets. Int. J. Knowl. Eng. **1**(2), 113–119 (2015). https://doi.org/10.7763/ijke.2015.v1.19
93. Abozinadah, E.A., Jones, J.H.: A statistical learning approach to detect abusive twitter accounts. ACM Int. Conf. Proceeding Ser (2017). https://doi.org/10.1145/30932413093281
94. Alakrot, A., Murray, L., Nikolov, N.S.: Dataset construction for the detection of anti-social behaviour in online communication in Arabic. Procedia Comput. Sci. **142**, 174–181 (2018). https://doi.org/10.1016/j.procs.2018.10.473
95. Alakrot, A., Murray, L., Nikolov, N.S.: Towards accurate detection of offensive language in online communication in Arabic. Procedia Comput. Sci. **142**, 315–320 (2018). https://doi.org/10.1016/j.procs.2018.10.491
96. A. A. E. M. B. N. H. Alhuzali and M. Abdul-Mageed, "Think Before Your Click: Data and Models for Adult Content in Arabic Twitter," *Proc. Elev. Int. Conf. Lang. Resour. Eval. (LREC 2018)*, 2018.
97. Haidar, B., Chamoun, M., Serhrouchni, A.: A multilingual system for cyberbullying detection: Arabic content detection using machine learning. Adv. Sci. Technol. Eng. Syst. **2**(6), 275–284 (2017). https://doi.org/10.25046/aj020634
98. Magdy, W., Darwish, K., Weber, I.: Failed revolutions: using Twitter to study the antecedents of ISIS support. First Monday (2016). https://doi.org/10.5210/fm.v21i2.6372
99. Xiang, G., Fan, B., Wang, L., Hong, J., Rose, C.: "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. ACM Int. Conf. Proceeding Ser (2012). https://doi.org/10.1145/23967612398556
100. V. Nahar, S. Al-maskari, X. Li, and C. Pang, "Databases Theory and Applications - 25th Australasian Database Conference, {ADC} 2014, Brisbane, QLD, Australia, July 14–16, 2014. Proceedings," vol. 8506, 2019, 2014, doi: https://doi.org/10.1007/978-3-319-08608-8.
101. Agarwal, S., Sureka, A.: 'Using KNN and SVM based one-class classifier for detecting online radicalization on twitter.' Int. Conf. Distributed Comput. Internet Technol. (2015). https://doi.org/10.1007/978-3-319-14977-6_47
102. Kaati, L., Omer, E., Prucha, N., Shrestha, A.: Detecting multipliers of Jihadism on Twitter. Proc 15th IEEE Int. Conf. Data Min. Work. ICDMW (2015). https://doi.org/10.1109/ICDMW.2015.9
103. Di Capua, M., Di Nardo, E., Petrosino, A.: Unsupervised cyber bullying detection in social networks. Proc. Int. Conf. Pattern Recognit. (2016). https://doi.org/10.1109/ICPR.2016.7899672
104. Abdelfatah, K.E., Terejanu, G., Alhelbawy, A.A.: "Unsupervised Detection of Violent Content in Arabic Social Media. Comput. Sci.Info Technol. (2017). https://doi.org/10.5121/csit.2017.70401
105. Pitsilis, G.K., Ramampiaro, H., Langseth, H.: Effective hate-speech detection in Twitter data using recurrent neural networks. Appl. Intell. **48**(12), 4730–4742 (2018). https://doi.org/10.1007/s10489-018-1242-y
106. S. Suryawanshi, B. R. Chakravarthi, M. Arcan, and P. Buitelaar, "Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text," *Proc. Second Work. Trolling, Aggress. Cyberbullying*, vol. 2020-Decem, no. May, pp. 32–41, 2020, [Online]. Available: https://www.aclweb.org/anthology/2020.trac-1.6.
107. T. Deshpande and N. Mani, *An Interpretable Approach to Hateful Meme Detection*, vol. 1, no. 1. Association for Computing Machinery, 2021.
108. D. Kiela *et al.*, "The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes," pp. 1–17, 2020, [Online]. Available: http://arxiv.org/abs/2005.04790.
109. N. Muennighoff, "Vilio: State-of-the-art Visio-Linguistic Models applied to Hateful Memes," arXiv:2012.07788v1, pp. 1–6, 2020, [Online]. Available: http://arxiv.org/abs/2012.07788.
110. Yuan, S., Wu, X., Xiang, Y.: A two phase deep learning model for identifying discrimination from tweets. In Adv. Database Technol. EDBT (2016). https://doi.org/10.5441/002/edbt.2016.92
111. Gambäck, B., Sikdar, U.K.: "Using Convolutional Neural Networks to Classify Hate-Speech. In Proceed First Workshop Abusive Language Online (2017). https://doi.org/10.18653/v1/w17-3013
112. F. P. Park JH, "One-step and two-step classification for abusive language detection on twitter.," 2017.
113. Zhang, Z., Robinson, D., Tepper, J.: "Detecting hate speech on twitter using a convolution gru based deep neural network. In Lecture Note. Comput. Sci. **10843**, 745–760 (2018)
114. Erico, C., Salim, R., Suhartono, D.: A systematic literature review of different machine learning methods on hate speech detection. Int. J. Info. Vis. **4**, 213–218 (2020)
115. Ayo, F.E., Folorunso, O., Ibharalu, F.T., Osinuga, I.A., Abayomi-Alli, A.: "A probabilistic clustering model for hate speech classification in twitter. Expert Syst. Appl. (2021). https://doi.org/10.1016/j.eswa.2021.114762
116. Baydogan, C., Alatas, B.: Metaheuristic ant lion and moth flame optimization-based novel approach for automatic detection of hate speech in online social networks. IEEE Access **9**, 110047–110062 (2021). https://doi.org/10.1109/ACCESS.2021.3102277
117. Asiri, Y., Halawani, H.T., Alghamdi, H.M., Abdalaha Hamza, S.H., Abdel-Khalek, S., Mansour, R.F.: enhanced seagull optimization with natural language processing based hate speech detection and classification. Appl. Sci. (2022). https://doi.org/10.3390/app12168000
118. Y. G. and X. L. Pengfei Du,: Towards an intrinsic interpretability approach for multimodal hate speech detection. Int. J. Pattern Recognit. Artif. Intell. (2022). https://doi.org/10.1142/S0218001422500409
119. N. Albadi, M. Kurdi, and S. Mishra 2018 "Are they our brothers? Analysis and detection of religious hate speech in the Arabic Twittersphere. *Proc. 2018 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM. Doi:* https://doi.org/10.1109/ASONAM.2018.8508247.
120. Miok, K., Škrlj, B., Zaharie, D., Robnik-Šikonja, M.: To ban or not to ban: bayesian attention networks for reliable hate speech detection. Cognit. Comput. **14**(1), 353–371 (2022). https://doi.org/10.1007/s12559-021-09826-9
121. Wullach, T., Adler, A., Minkov, E.: Towards hate speech detection at large via deep generative modeling. IEEE Internet

Comput. **25**(2), 48–57 (2021). https://doi.org/10.1109/MIC.2020.3033161
122. Gomez, R., Gibert, J., Gomez, L., Karatzas, D.: "Exploring hate speech detection in multimodal publications. Proc. - 2020 IEEE Winter Conf. Appl. Comput. Vision, WACV (2020). https://doi.org/10.1109/WACV45572.2020.9093414
123. A. Das, J. S. Wahi, and S. Li, "Detecting Hate Speech in Multimodal Memes," 2020, [Online]. Available: http://arxiv.org/abs/2012.14891.
124. Zhou, Y., Yang, Y., Liu, H., Liu, X., Savage, N.: Deep learning based fusion approach for hate speech detection. IEEE Access **8**, 128923–128929 (2020). https://doi.org/10.1109/ACCESS.2020.3009244
125. Muhammad, I.Z., Nasrun, M., Setianingsih, C.: Hate speech detection using global vector and deep belief network algorithm. 2020 1st Int. Conf. Big Data Anal. Pract. IBDAP (2020). https://doi.org/10.1109/IBDAP50342.2020.9245467
126. Le-hong, P.: Knowledge-based systems diacritics generation and application in hate speech detection on vietnamese social networks. Knowledge-Based Syst. (2021). https://doi.org/10.1016/j.knosys.2021.107504
127. P. Vijayaraghavan, H. Larochelle, and D. Roy, "Interpretable Multi-Modal Hate Speech Detection," *Int. Conf. Mach. Learn.*, 2021.
128. G. Sahu, R. Cohen, and O. Vechtomova, "Towards A Multi-agent System for Online Hate Speech Detection," *Proc. 20th Int. Conf. Auton. Agents Multiagent Syst.*, 2021.
129. A. Jiang, Aiqi; Zubiaga, *Cross-lingual Capsule Network for Hate Speech Detection in Social Media*, vol. 1, no. 1. Association for Computing Machinery, 2021.
130. Perifanos, K.: Multimodal hate speech detection in greek social media. Mdpi Multimed. Technol. Interact. (2021). https://doi.org/10.3390/mti5070034
131. Aldjanabi, W., Dahou, A., Al-qaness, M.A.A., Elaziz, M.A., Helmi, A.M., Damaševi, R.: Arabic offensive and hate speech detection using a cross-corpora multi-task learning model. Mdpi inform. (2021). https://doi.org/10.3390/informatics8040069
132. Al-Makhadmeh, Z., Tolba, A.: Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach. Computing **102**(2), 501–522 (2020). https://doi.org/10.1007/s00607-019-00745-0
133. Corazza, M., Menini, S., Cabrio, E., Tonelli, S., Villata, S.: A multilingual evaluation for online hate speech detection. ACM Trans. Internet Technol. (2020). https://doi.org/10.1145/3377323
134. Kapil, P., Ekbal, A.: A deep neural network based multi-task learning approach to hate speech detection. Knowledge-Based Syst. (2020). https://doi.org/10.1016/j.knosys.2020.106458
135. Aulia, N., Budi, I.: Hate speech detection on Indonesian long text documents using machine learning approach. ACM Int. Conf. Proceeding Ser. (2019). https://doi.org/10.1145/33304823330491
136. Badjatiya, P., Gupta, M., Varma, V.: Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. Web Conf. 2019 - Proc World Wide Web Conf. WWW **10**(1145/3308558), 3313504 (2019)
137. G. Nascimento, F. Carvalho, A. M. Da Cunha, C. R. Viana, and G. P. Guedes, 2019 "Hate speech detection using Brazilian imageboards," *Proc. 25th Brazillian Symp. Multimed. Web, WebMedia.* https://doi.org/10.1145/3323503.3360619.
138. A. S. Saksesi, M. Nasrun, and C. Setianingsih, "Analysis Text of Hate Speech Detection Using Recurrent Neural Network," *Proc. - 2018 Int. Conf. Control. Electron. Renew. Energy Commun. ICCEREC 2018*, pp. 242–248, 2018, doi: https://doi.org/10.1109/ICCEREC.2018.8712104.
139. Mozafari, M., Farahbakhsh, R., Crespi, N.: A BERT-Based transfer learning approach for hate speech detection in online social media. Conf. Comp. Net, Their Appl Int (2020). https://doi.org/10.1007/978-3-030-36687-2_77
140. V. Basile *et al.*, "SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter," *NAACL HLT 2019 - Int. Work. Semant. Eval. SemEval 2019, Proc. 13th Work.*, pp. 54–63, 2019, doi: https://doi.org/10.18653/v1/s19-2007.
141. N. Mehrabi, F. Morstatter, N. Saxena, and L. G. Jan, "A Survey on Bias and Fairness in Machine Learning," arXiv:1908.09635v3, 2022.
142. Ahmed, Z., Vidgen, B., Hale, S.A.: Tackling racial bias in automated online hate detection : towards fair and accurate detection of hateful users with geometric deep learning. EPJ Data Sci. (2022). https://doi.org/10.1140/epjds/s13688-022-00319-9
143. Watanabe, H., Bouazizi, M., Ohtsuki, T.: Hate speech on twitter: a pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE Access **6**, 13825–13835 (2018). https://doi.org/10.1109/ACCESS.2018.2806394
144. Mollas, I., Chrysopoulou, Z., Karlos, S., Tsoumakas, G.: ETHOS: a multi-label hate speech detection dataset. Complex Intell. Syst. (2022). https://doi.org/10.1007/s40747-021-00608-2
145. A. Velankar, H. Patil, A. Gore, S. Salunke, and R. Joshi, "L3Cube-MahaHate: A Tweet-based Marathi Hate Speech Detection Dataset and BERT models," arXiv:2203.13778v2, pp. 1–12, 2022, [Online]. Available: http://arxiv.org/abs/2203.13778.
146. Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P., Mukherjee, A.: "HateXplain: a benchmark dataset for explainable hate speech detection",. 35th aaai conf Artif. Intell. AAAI **17A**, 14867–14875 (2021)
147. K. Yang, W. Jang, and W. I. Cho, "APEACH: Attacking Pejorative Expressions with Analysis on Crowd-Generated Hate Speech Evaluation Datasets," 2022, [Online]. Available: http://arxiv.org/abs/2202.12459.
148. J. A. Leite, D. F. Silva, K. Bontcheva, and C. Scarton, "Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis," arXiv:2010.04543, 2020, [Online]. Available: http://arxiv.org/abs/2010.04543.
149. C. S. de Alcântara, D. Feijó, and V. P. Moreira, "Offensive video detection: Dataset and baseline results," *Lr. 2020 - 12th Int. Conf. Lang. Resour. Eval. Conf. Proc.*, no. May, pp. 4309–4319, 2020.
150. T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar, "ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection," 2022, [Online]. Available: http://arxiv.org/abs/2203.09509.
151. Evkoski, B., Pelicon, A., Mozetič, I., Ljubešić, N., Novak, P.K.: Retweet communities reveal the main sources of hate speech. PLoS ONE (2022). https://doi.org/10.1371/journal.pone.0265602
152. H. R. Kirk, B. Vidgen, P. Röttger, T. Thrush, and S. A. Hale, "Hatemoji: A Test Suite and Adversarially-Generated Dataset for Benchmarking and Detecting Emoji-based Hate," 2021, [Online]. Available: http://arxiv.org/abs/2108.05921.
153. N. Romim, M. Ahmed, M. S. Islam, A. Sen Sharma, H. Talukder, and M. R. Amin, "HS-BAN: A Benchmark Dataset of Social Media Comments for Hate Speech Detection in Bangla," arXiv:2112.01902v1, pp. 1–8, 2021, [Online]. Available: http://arxiv.org/abs/2112.01902.
154. M. R. Karim, B. R. Chakravarthi, J. P. McCrae, and M. Cochez,2020 "Classification Benchmarks for Under-resourced Bengali Language based on Multichannel Convolutional-LSTM Network. *Proc. - 2020 IEEE 7th Int. Conf. Data Sci. Adv. Anal. DSAA*. https://doi.org/10.1109/DSAA49011.2020.00053.
155. Wu, C.S., Bhandary, U.: Detection of hate speech in videos using machine learning. Int. Conf. Comput. Sci. Comput. Intell. CSCI Proc (2020). https://doi.org/10.1109/CSCI51800.2020.00104

156. M. Zampieri *et al.*, "SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020)," *Proc. Int. Work. Semant. Eval.*, no. OffensEval, 2020.
157. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Predicting the type and target of offensive posts in social media. Proc. NAACL-HLT **2019**, 1415–1420 (2019)
158. Haddad, H., Mulki, H., Oueslati, A.: T-HSAB: a tunisian hate speech and abusive dataset. Commun. Comput. Inf. Sci. (2019). https://doi.org/10.1007/978-3-030-32959-4_18
159. H. Mulki, H. Haddad, C. Bechikh Ali, and H. Alshabani, "L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language," pp. 111–118, 2019, doi: https://doi.org/10.18653/v1/w19-3512.
160. Elsherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., Belding, E.: Peer to peer hate: hate speech instigators and their targets. AAAI Conf. Web Soc. Media (2018). https://doi.org/10.1609/icwsm.v12i1.15038
161. A. M. Founta *et al.*, "Large scale crowdsourcing and characterization of twitter abusive behavior," 12th *Int. AAAI Conf. Web Soc. Media, ICWSM 2018*, no. Icwsm, pp. 491–500, 2018
162. J. Moon, W. I. Cho, and J. Lee, "BEEP ! Korean Corpus of Online News Comments for Toxic Speech Detection," pp. 25–31, 2017.
163. H. Mubarak, K. Darwish, and W. Magdy, "Abusive Language Detection on Arabic Social Media," *Proc. First Work. Abus. Lang. Online*, pp. 52–56, 2017, doi: https://doi.org/10.18653/v1/w17-3008.
164. Toutenburg, H.: Mathematical statistics with applications. Computational Statistics & Data Anal (1992). https://doi.org/10.1016/0167-9473(92)90162-9
165. C. A. Goodfellow I, Bengio Y, *Deep Learning*. MIT Press, Cambridge, 2016.
166. T. Mikolov, A. Deoras, S. Kombrink, L. Burget, and J. H. Černocký, "Empirical evaluation and combination of advanced language modeling techniques," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, no. August, pp. 605–608, 2011.
167. De Souza, G.A., Da Costa-Abreu, M.: Automatic offensive language detection from Twitter data using machine learning and feature selection of metadata. Proc. Int. Jt. Conf. Neural Networks (2020). https://doi.org/10.1109/IJCNN48605.2020.9207652
168. M. Polignano *et al.*, "A L BERT O : Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets," *CEUR Workshop Proc.*, 2019.
169. Webb, G.I.: Decision tree grafting from the all-tests-but-one partition. IJCAI Int. Jt. Conf. Artif. Intell. **2**, 702–707 (1999)
170. S. Tulkens, L. Hilte, E. Lodewyckx, B. Verhoeven, and W. Daelemans, "A Dictionary-based Approach to Racism Detection in Dutch Social Media," 2016, [Online]. Available: http://arxiv.org/abs/1608.08738.
171. Burnap, P., Williams, M.L.: Cyber hate speech on twitter: an application of machine classification and statistical modeling for policy and decision making. Policy Internet **7**(2), 223–242 (2015). https://doi.org/10.1002/poi3.85