International Journal of Information Security (2024) 23:577–608
https://doi.org/10.1007/s10207-023-00755-2

REGULAR CONTRIBUTION

Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities

Anjum[1] · Rahul Katarya[1]

Accepted: 2 September 2023 / Published online: 25 September 2023

Abstract

Information and communication technology has evolved dramatically, and now the majority of people are using internet and sharing their opinion more openly, which has led to the creation, collection and circulation of hate speech over multiple platforms. The anonymity and movability given by these social media platforms allow people to hide themselves behind a screen and spread the hate effortlessly. Online hate speech (OHS) recognition can play a vital role in stopping such activities and can thus restore the position of public platforms as the open marketplace of ideas. To study hate speech detection in social media, we surveyed the related available datasets on the web-based platform. We further analyzed approximately 200 research papers indexed in the different journals from 2010 to 2022. The papers were divided into various sections and approaches used in OHS detection, i.e., feature selection, traditional machine learning (ML) and deep learning (DL). Based on the selected 111 papers, we found that 44 articles used traditional ML and 35 used DL-based approaches. We concluded that most authors used SVM, Naive Bayes, Decision Tree in ML and CNN, LSTM in the DL approach. This survey contributes by providing a systematic approach to help researchers identify a new research direction in online hate speech.

Keywords Deep learning · Natural language processing (NLP) · Machine learning · Online hate speech (OHS) · Social media · Toxicity detection

1 Introduction

Social media sites like Facebook, WhatsApp, Instagram and Twitter are easy to use, a free source that provides advantages to people to air their voices. Now people can easily exchange their views and information from anywhere, anytime. According to a Global Digital Report [1], the world's total number of internet users in 2019 was 4.388 billion, among which 3.484 billion were online social media users. Also, according to the World Bank Report (2017), 241 million users on Facebook are Indians [1]. In Fig. 1, we summarize the total number of users on different online social media platforms with reference to the Global Social Networks [2]. Among all the social networking websites, Facebook has the maximum number of users. In today's scenario, massive amounts of data are shared online every day which makes social media the most significant medium of communication. Besides these excellent features, these sites, however, have downsides as well. In the absence of meaningful restrictions or procedures, anybody can make detrimental and untrue comments in abusive or offensive language against anybody with an intention to spoil one's image and status in the community. Also, since many people around the globe during the COVID-19 pandemic were working from home and staying indoors, internet usage has risen sharply. Though many people using social media platforms can communicate virtually with their friends and relatives, there is also a spread of frustration, anger and anxiety online. These negative feelings can easily lead to hatred toward someone else. So, it becomes a huge concern for the government and for all social media sites to detect hate content before it spreads into public in general.

Also, in the present scenario, more people are using social networking websites resulting in the generation of a massive amount of data. Handling such a large amount of information is a crucial and non-trivial task since there are several target

✉ Rahul Katarya
rahuldtu@gmail.com

Anjum
anjum_2792@yahoo.com

[1] Big Data Analytics and Web Intelligence Laboratory, Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India

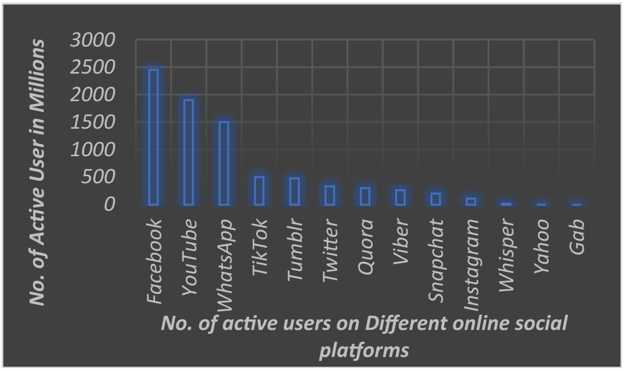


Fig. 1 Number of active users on social media in 2019

Table 1 Research questions

RQ1:	"What are the primary sources of articles for OHS detection?"
RQ2:	"What is hate speech and how it originated in online social media?"
RQ3:	"What are the available OHS datasets for different languages?"
RQ4:	"What are the extracted features and most used in the traditional machine learning algorithm for OHS?"
RQ5:	"What are the trends of Traditional machine learning for classifying an online hate speech?"
RQ6:	"What are the trends of Deep learning for classifying an online hate speech?"

groups and each group is exposed to particular hate-related words that complicate the task of automated classification [1]. Example: 1. "Queers are an abomination and need to be helped to go straight to hell." 2. "Wipe out the Muslims." Both sentences are hate speech toward a particular group. The primary reason for the increase in aggressive behavior and the generation of hate speech is the anonymity provided by the social media platforms [2]. Therefore, many social websites need to develop online hate speech detection tools to control the online circulation of toxic messages [5]. The social networking websites like Twitter, Facebook, etc., are developing artificial intelligence techniques to stop the dissemination of online hate speech and toxicity on their public network. For the detection of online hate/toxicity, there already exists a web browser plugin called "Hate Speech Blocker," which flags the user that the expression could be construed as hate speech [6].

1.1 Problem statement

The literature in computer science on online hate speech detection concentrates on a few languages: flaming, aggressive, offensive, toxicity and cyberbullying. All of these languages are compared, with a focus on their most prevalent manifestation. To increase the quality and applicability of automated solutions, we believe that a study on one language may be useful for research on another language. We also believe that precise and ordered terminology is necessary. We referred to the broad category of research papers and weblinks and google search that includes all of these forms: "Hate Speech, toxicity, flaming, cyberbullying, aggressive". We used the term “online hate speech (OHS)” as the phrase has never been used in linguistics or computer science before, to eliminate confusion and misinterpretation. Numerous social and computer disciplines, including psychology, political science and law, have examined the manifestation, dynamics and consequences of hate speech. The literature assessment reveals that a significant amount of study has been done on how to identify different types of hateful content. The reported publications have concentrated more on the many components of manual moderation and the difficulties that AI-based techniques should address. Fewer research articles concentrate on fully automated strategies for filtering harmful content on social networking sites. This article mainly focuses on the identification of hate speech using various artificial intelligence approaches because it offers precise definitions and solutions to the problem. Although some of the research issues (shown in Table 1) are addressed by our work, our study of the computer science literature enables us to provide additional recommendations and directions for future research.

This paper presents a survey of online hate speech identification using different Artificial Intelligence techniques. This review study looks into a number of research questions shown in Table 1 that will help us to learn about the most recent trends in online hate speech in the field of artificial intelligence. It also includes an overview of recently used machine learning and deep learning algorithms for evaluating data used by the proposed research problem.

This manuscript offers the following four contributions in greater detail:

1. Presented a framework of the online hate speech (OHS) manuscript given in Fig. 3.
2. Identified the most used traditional machine learning classifier with handcrafted features.
3. Compared different approaches of OHS detection including their advantages and disadvantages.
4. This paper provides an organized review to examine how hate speech and toxicity are incorporated into deep learning and machine learning algorithms.

This paper provides an organized review to examine how hate speech and toxicity are incorporated into deep learning and machine learning algorithms. In Sect. 1, we briefly explained the problem statement and the implication of the study. To answer **RQ1**: we presented the OHS methodology and paper organization in Sect. 2. The previous reviews of online hate speech in the domain of AI are discussed in Sect. 3. We answer the **RQ2**, by discussing the fundamentals of hate speech, how it is originated in online social media and laws that are adopted to combat it in Sect. 4. To answer **RQ3**, we compared and discussed all the available online datasets in Sect. 5. Section 6 aims to answer **RQ4** by discussing the types of features and those that are most used in the domain of hate speech. The traditional machine learning (ML) framework, models and earlier OHS work advantages and disadvantages are discussed in Sect. 7, which aims to answer **RQ5.** Section 7 holds the answer of **RQ6,** where deep learning framework and models and types of features used in OHS detection are presented. Section 8 covers all the evaluation metrics that are used by the researchers to evaluate the results of OHS. In Sect. 9 we concluded the findings of this survey, research opportunities and future steps.

2 Methodology and paper organization

This section outlines the processes taken to compile the prior contributions and to gather the computer science literature that will be the subject of our analysis.

In order to answer the RQ1: "What are the primary sources of articles for OHS detection?". We tried to find all the sources for the detection and analysis of OHS. We have found approximately 200 research papers and other documents from the Google search engine, ACM Digital Library, IEEE Xplore Digital Library, Springer Link, google scholar, Science Direct, Research Gate and Wiley Online Library. We shortlisted the most relevant 136 papers suitable for this research from the above set. The complete search methodology is shown in Fig. 2 using the PRISMA diagram [7].

We consistently gathered pertinent terms by scanning cited literature in order to discover the most detailed hate speech and other related surveys. Following that, we coined the terminology "hate," "hateful," "toxic," "aggressive," "abusive," "offensive," and "damaging speeches," as well as "cyberbullying," "cyberaggression," "flaming," "harassment," "denigration," "outing," "trickery," "exclusion," "cyberstalking," "flooding," "trolling". We utilize our proposed term, "online hate speech," to refer to the combination of all these concepts in the survey's remaining questions (abbreviated to OHS). We have also taken the papers which had "hate speech," "cyberbullying", "OHS detection using deep learning", “toxicity in online social media”, "OHS detection using machine learning" and "OHS detection using natural language processing" as the search keywords. The distribution of articles on online hate speech is shown in Table 2.

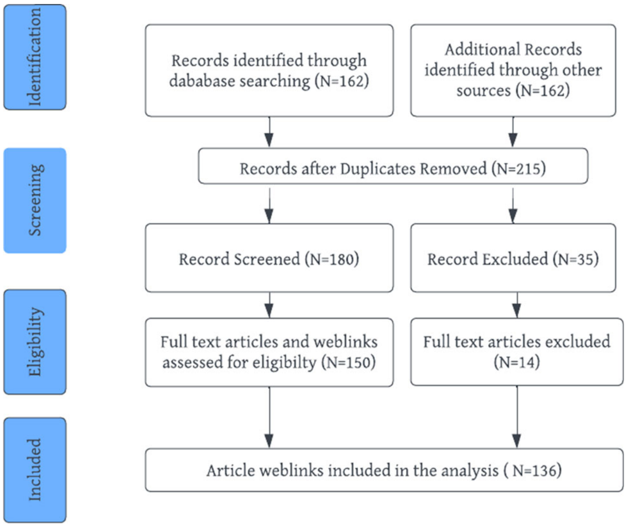

Identification
Screening
Eligibility
Included
Records identified through dababase searching (N=162)
Additional Records identified through other sources (N=162)
Records after Duplicates Removed (N=215)
Record Screened (N=180)
Record Excluded (N=35)
Full text articles and weblinks assessed for eligibilty (N=150)
Full text articles excluded (N=14)
Article weblinks included in the analysis (N=136)

Fig. 2 Evidence synthesis for the literature survey

This review considers a broad perspective of the researchers and our analysis of toxicity detection. The flow of information in this review is presented in Fig. 3.The yearwise classification of the online hate speech article is shown in Fig. 4a, and the content-wise distribution of the referred articles is shown in Fig. 4b.

It can be inferred from Fig. 4a that hate speech has been an area of focus (computer science and engineering) from 2016 onward and is now becoming a popular research area among researchers. Also, from Fig. 4b we can see that only four survey papers have been published on Online Hate Speech as a subject of research [4, 8] in computer science.

1. *Identification* We searched all the papers on online hate speech detection tasks, such as OHS datasets, different organization contributions, proposed OHS detection models and different feature extraction techniques by including each above-mentioned keyword as the search query. All the extracted papers were taken from several journals and websites, as mentioned in Table 2.
2. *Screening* After collecting all the related information. We removed duplicates and redundant searches.
3. *Eligibility* 46 records were present from psychology, law and social science backgrounds. So, in this phase we took only 15 relevant papers from them, which were required for the problem statement. Furthermore, only the relevant search concerning the research problem has been taken. We selected total 136 articles and weblinks on which we performed this survey.

Table 2 Amount of research contribution per source

S. no	Main source of articles	Journal Name	Number of articles
1	Elsevier	Information processing and management	1
		Expert system and application	2
		Data in brief	1
		Online Social Networks and Media Journal	1
		Computing	1
		Telematics and Informatics	1
		Applied Intelligence	1
		Computers in Human Behavior Journal	1
		Aggression and Violent Behavior	1
		Interacting with Computers	1
2	Springer Link	Multimedia tools and application	2
		Crime science	1
		SN Computer Science	1
		Human-centric Computing and Information Sciences	1
		Multimedia Systems	1
		Cognitive Computation	1
3	ACM Digital Library	ACM Transactions on Internet Technology	2
		Proceedings of the ACM on Human–Computer Interaction	1
		ACM Transactions on The Web	1
		ACM Transactions on Management Information Systems	1
4	IEEE Xplore Digital Library	IEEE Access	3
		IEEE Transactions on Computational Social Systems	1
5	Other Journals	Indonesian Journal of Electrical Engineering and Computer Science	1
		Journal of Artificial Intelligence Research	1
6	Google Scholar	-	11
7	Wiley Online Library	Periodicals (Policy and Internet)	2
8	Other	Total springer and including other journals proceedings like AIS eLibrary	92
9	Weblinks	-	20
	Total journal and weblinks		136

3 Previous review

In recent years, few survey papers have been published in the domain of OHS using artificial intelligence techniques. The authors of the paper [2–6] present the study of OHS. These works are mainly focused on the concept of online hate speech, techniques, features and datasets published in the area of OHS. In one of the paper [2], the authors establish the basic definition of hate speech by taking into consideration different connotations and concepts this phenomenon might occur. Then the authors provide a comparative analysis of the resources available for the research on hate speech and the pre-existing research from a computer science perspective. They deduce a lack of public datasets and metrics to establish and compare results in this field. But the author focused on the traditional machine learning approaches and did not compare different author work's limitations and advantages.

Similarly, the survey paper [4] explains the short, structured overview of hate speech using NLP. This survey compares different studies done on online hate speech from a natural language processing perspective. The review mainly focuses on comparing different types of features that are used

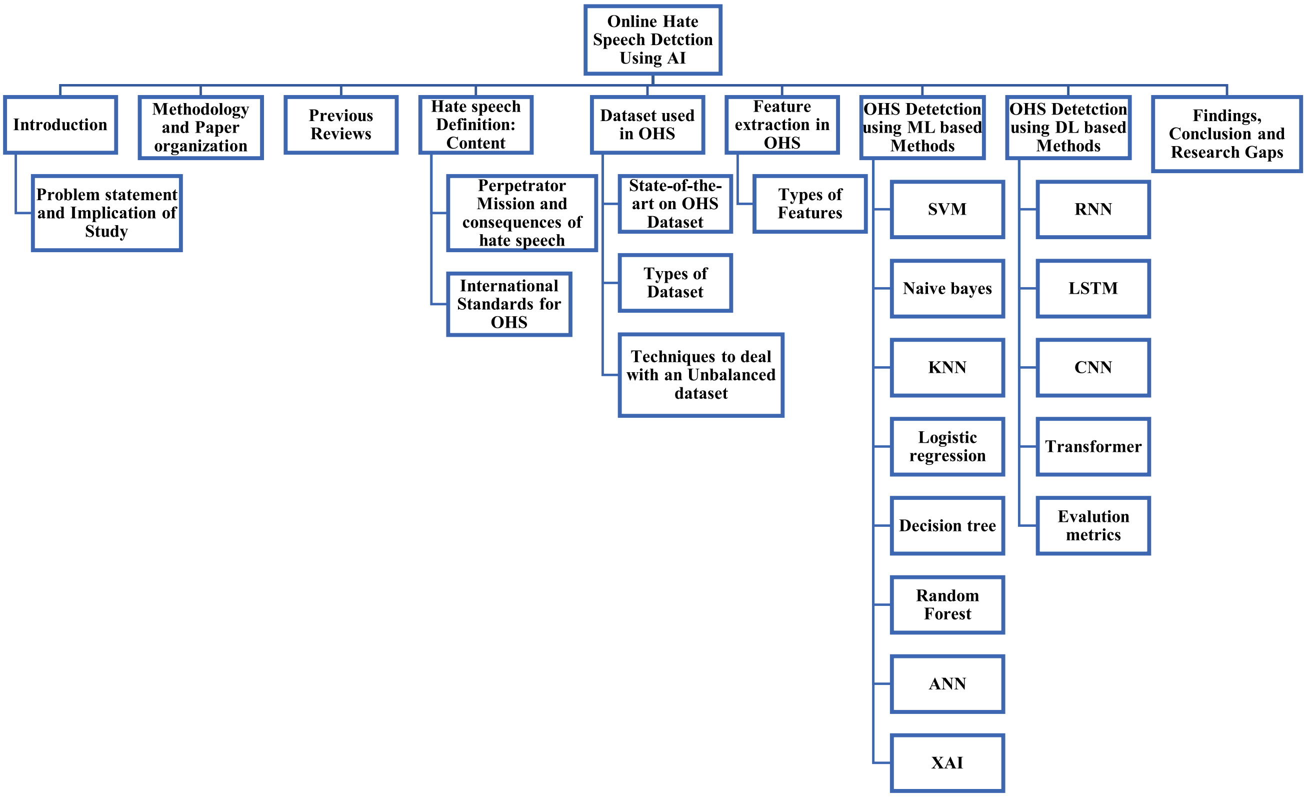

Online Hate Speech Detction Using AI
Introduction
Problem statement and Implication of Study
Methodology and Paper organization
Previous Reviews
Hate speech Definition: Content
Perpetrator Mission and consequences of hate speech
International Standards for OHS
Dataset used in OHS
State-of-the-art on OHS Dataset
Types of Dataset
Techniques to deal with an Unbalanced dataset
Feature extraction in OHS
Types of Features
OHS Detetction using ML based Methods
SVM
Naive bayes
KNN
Logistic regression
Decision tree
Random Forest
ANN
XAI
OHS Detetction using DL based Methods
RNN
LSTM
CNN
Transformer
Evalution metrics
Findings, Conclusion and Research Gaps

Fig. 3 Systematic representation of the manuscript

to classify hate speech. It compares features like basic syntactic features, character-level features, sentiment features and more. It argues that information from features based on the text may not alone be accurate enough and researchers shall also consider multimodal and meta information features for a more accurate result and judgment. It also addresses the issue of lack of public open sources resources like datasets. The survey paper [6] presented the meta-analysis of cyberbullying papers using soft computing techniques, but the author did not present the advantages and disadvantages of the previous literature. Furthermore, the survey was limited to the cyberbullying area only. This paper [7] aims to map different themes, concepts, stakeholders and research hotspots in the field of Online Hate Research. On the basis of this analysis, the authors deduce trends and patterns in OHR like what type of countries invest in it more and change of focus in the field with time. Moreover, they try to cluster the main focal points of the research field to understand what parts are predominantly taken up by researchers, namely cyberbullying, sexual solicitation and intimate partner violence, deep learning and automation and extremist. This study is constricted to the web of science core database and shall be expanded to more databases of papers. Very few survey papers have been seen in the area of online hate speech using artificial intelligence techniques, which covered all of the information in one place.

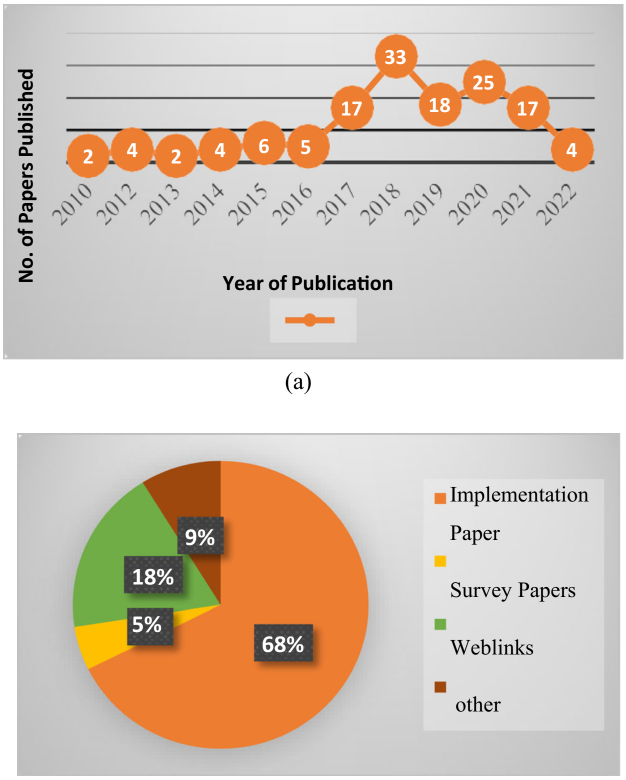
Our survey significantly differs from earlier efforts by examining the OHS problem using AI techniques. New conceptual elements that are crucial for autonomous detection tasks are brought to light, such as integrated definitions of OHS, datasets, various kinds of features and models that affect the outcomes. It also identifies deficiencies in the way detection tasks are currently designed, notably in terms of accounting for context and individual subjectivity.

The proposed review overcomes the shortcoming of the existing surveys by providing limitations of the existing techniques and a systematic review of the online hate speech problem.

4 Hate speech definition: Content

5 RQ2: "What is hate speech and how it originated in online social media?"

With the advent of social media and internet, we found OHS and toxicity present on every social networking website in the form of images, text and videos. With the recent advantage of mobile computing and the internet, social media provides a platform to share views and exchange information from anywhere anytime. Social media plays an essential role in the

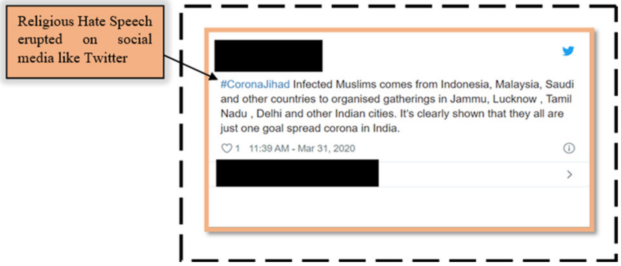


(a)

(b)

Fig. 4 **a**: Year-wise classification of the referred "related papers". **b**: Content-wise distribution of OHS article

origin of online hate speech. On sites like Facebook, Instagram and Twitter, users can hide their identity or can bully or use toxic thoughts without being noticed. The anonymity of the user on these social platforms provides the user to conceal their identity and say and do whatever atrocious they want [9]. The origin of OHS is the class of cybercrime. So, we proposed the Taxonomy of cyber-crime to understand the origin of OHS in a more transparent way. So, we classify the Hate problem in its various forms shown in Fig. 5. We have shown that hate speech is a part of the cybercrime and cyberbullying problem. Different authors define hate speech in different ways. The author [10] defines hate speech "The use of harsh and abusive words on online platforms to propagate immoral ideas such as communal or political polarity is called Online Hate Speech". In this paper [11] "The speech which use of offensive and hateful language to target specific characteristics of a person or a community is found to be hate speech". The author defines hate speech as when insulting and derogatory language is used to target certain people with the intend to humiliate them or condescend them [12]. Hate speech is an expression that vilifies and disparages a group of people or a person on the basis of the congregation in a social group recognized by attributes such as mental disability, race, religion, sexual orientation, or gender inequality and others [13]. Typically, hate speech promotes malevolent stereotypes and encourages savagery against people or a group. With this concept, we assume that "*hate speech* is any speech, which attacks an individual or a group intending to hurt or disrespect based on the identity of a person". For example, in the COVID-19 pandemic, the communal harmony between Hindus and Muslims got deteriorated due to a maligning campaign carried out on Twitter shown in Fig. 6,


Religious Hate Speech erupted on social media like Twitter

[redacted]

#CoronaJihad Infected Muslims comes from Indonesia, Malaysia, Saudi and other countries to organised gatherings in Jammu, Lucknow , Tamil Nadu , Delhi and other Indian cities. It's clearly shown that they all are just one goal spread corona in India.

1 11:39 AM - Mar 31, 2020

[redacted]

Fig. 6 Hate speech content on Twitter

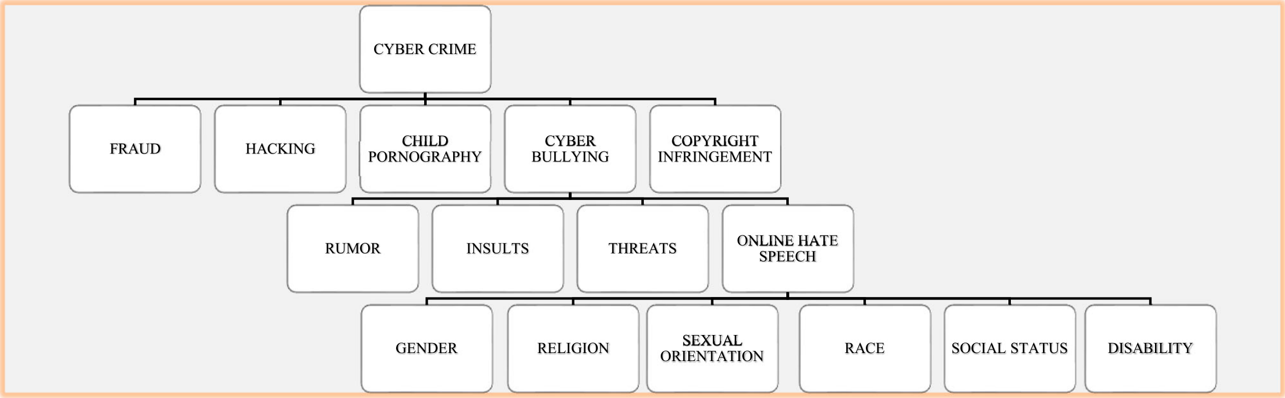


Fig. 5 Taxonomy of cyber crime

which describes the religious hate speech content and anti-social elements that exist in our society. Certain applications of detecting hate speech content are in politics, terrorism, casteism, religion. Various types of hate speech content are shown in Fig. 7. Most of the work in OHS using artificial intelligence has been done in *racism, sexism and religious* areas. Other areas of hate speech are untouched or either classified in the field of hate or non-hate category. We also surveyed five practical ways to deal with OHS in online social networking platforms like Instagram, Twitter, Facebook, that is:

- o *Report it* Hate speech violates most site's terms of service; people can report it anonymously.
- o *Block it* Block abusive users
- o *Do not share it* Forwarding any type of hate speech is wrong because offensive content can be traced back to them.
- o *Call it out* Understand how other people feel, and find ways to nurture empathy and compassion.
- o *Learn more* Hate often stems from ignorance, so learn from other's experiences.

The *consequences* of OHS can be low self-esteem, anxiety, depression, and in some cases, a victim can commit suicide. Therefore, the analysis and detection of online hate speech in social media is an area of concern.

5.1 Perpetrator mission and consequences of hate speech: a brief analysis

In the USA, the Federal Bureau of Investigation finds that almost all crimes, including hate speech crimes, are based on four factors [14, 15], explained in Table 3. In the manual of Ontario [16], they identify some consequences of hate crimes. Also, adolescents play an important role for being a bystander who does not participate in online hate, but they observe all things, by being a victim who suffer from online hatred and being perpetrators who do hate crimes by posting, replying and forwarding toxic content [17].

Studies show if offline aggression increases, then online hate crime also increases. There can be various consequences of online hate speech for a victim and others as well. A victim can experience anxiety, depression and in the worst case can commit suicide [18]. We categorize the repercussion of hate speech on society in Fig.8. Hate speech impacts the victim and sometimes the whole community. A person can be inflicted with psychological harm like low self-esteemed. Sometimes, it also affects the target group from which the victim belongs to and makes the group or community vulnerable.

Table 3 Perpetrator motive

Thrill-seeking	Where some people do hate crimes to make themselves happy, or they were enjoying themselves by seeing their victim sensitive to their religion, ethnicity, gender, or background
Defensive	Hate crimes arise when perpetrators are defensive about their community and to protect their society
Retaliatory	The motive of the perpetrators here is revenge
Mission offenders	Ideological reasons of the criminal such as "terrorism" where innocent people prey to perpetrators

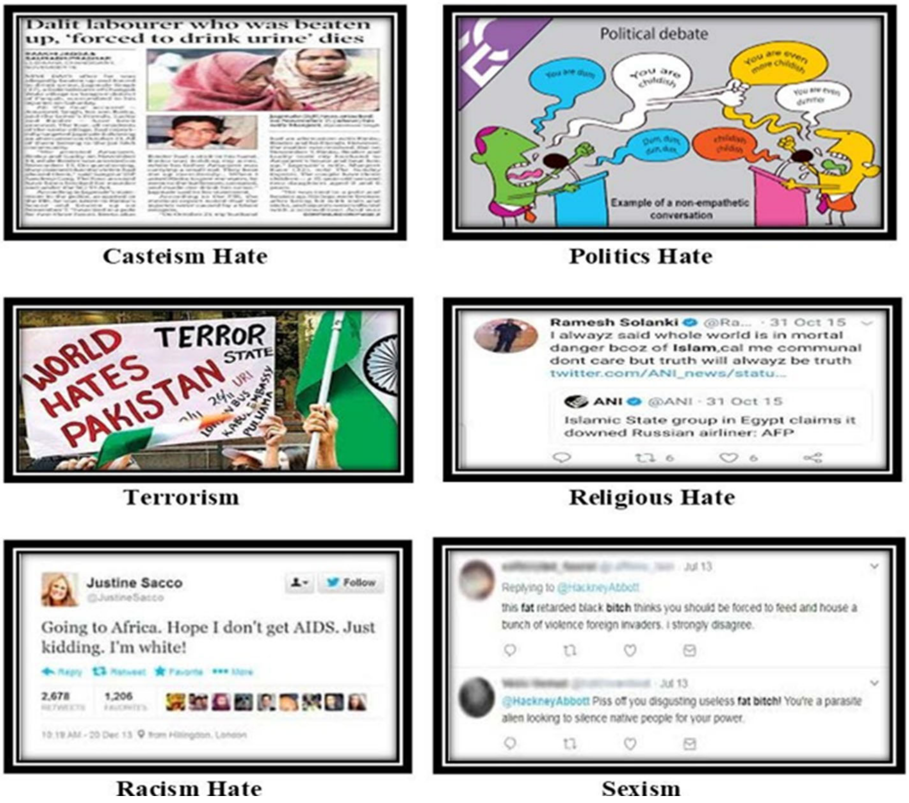
5.2 International standards for OHS

We found cyberbullying has been a long-studied terminology that threatens the individual, whereas hate speech is an unpleasant language addressed to the individual or a group of people. Figure 9 shows registered cyberbullying cases along with the country of origin. Because of these high number of cases on online social media like Twitter, Facebook, etc., needs to share the responsibility to intercede and quarantine the toxic content, which is widespread on their platforms [19], hate speech on online platforms can lead to violence and is a general threat to peace and social harmony. To discourage the use of toxic language, some popular social media websites like Facebook, Twitter, Instagram and YouTube have framed new policies and guidelines [19–22]. From Fig. 9, we can conclude that India has the maximum number of reported cyberbullying cases [23] in 2019, then Brazil and the USA.

We found two bodies that make laws for the OHS: UDHR, Universal declaration of human rights, is an international body for human rights that stands for freedom of speech and expression given in article 19. To use this law appropriately, Article 29(2) established some restrictions [24]. Similarly, the European Convention on Human Rights, the International Covenant on Civil and Political Rights [25], broadens the restriction on hate speech. The government has a right and responsibility to intercede when there is a high probability of imminent harm and then take preventive policing.

6 Datasets used in OHS

Input data play an essential role in machine learning; therefore, it is important to use the relevant and correctly annotated data.



Dalit labourer who was beaten up, 'forced to drink urine' dies

Ramesh Solanki @Ra... · 31 Oct 15
I alwayz said whole world is in mortal danger bcoz of **Islam**,cal me communal dont care but truth will alwayz be truth twitter.com/ANI_news/statu...

ANI @ANI · 31 Oct 15
Islamic State group in Egypt claims it downed Russian airliner: AFP

Justine Sacco
@JustineSacco

Going to Africa. Hope I don't get AIDS. Just kidding. I'm white!

Reply Retweet Favorite More

2,678 RETWEETS 1,206 FAVORITES

10:19 AM - 20 Dec 13 from Hillingdon, London

Jul 13
Replying to @HackneyAbbott
this **fat** retarded black **bitch** thinks you should be forced to feed and house a bunch of violence foreign invaders. i strongly disagree.

Jul 13
@HackneyAbbott Piss off you disgusting useless **fat bitch!** You're a parasite alien looking to silence native people for your power.

Fig. 7 Types of hate speech on online social media

6.1 State of the art on OHS dataset

RQ4: "What are the available OHS datasets for different languages?"

In this study, we collected datasets from various reliable sources, and almost all the available datasets are explained in Table 4. Many researchers have used different types of hate speech datasets which are based on language, race, ethnicity, etc. Most of the datasets are available on the GitHub website. To collect data from Twitter, many researchers have used Twitter's Streaming API for analysis of hate speech as a data source, where researchers can have free access to 1% of all the data. The collected data always have metadata and are downloaded in the JSON format. Later, we need to convert it into a CSV file. The author provides an unbalanced 16 k annotated dataset collected from Twitter [8], which classify as racist, sexist and neither. In paper [9] a Facebook crawler was used to retrieve the comment from the Facebook post and five volunteered students annotated 6502 comments as no hate, strong hate or weak hate. In this [10] author used the Tumblr search APIs to get the data from Tumblr. Two–three experienced annotators performed the annotation of 2456 posts as racist, radicalized or unknown. HatEval dataset is available from collab website [11]. Whisper is an anonymous app that does not store old data, so the author [12] collected the data in real time using a distributed web crawler. Most of the authors used the kappa and Interrater agreement to capture the quality of the dataset. Cohen kappa is a statistical measure of inter-rater agreement of the agreement between the two raters for categorical items. Suppose we have a bunch of people and two and more raters have to find out whether each individual in this group is able to his job not. So the experts have to evaluate the group of people independently and to find out whether each individual is able to perform the job[13]. In Table 4 we have also discussed relevant details of the given datasets.

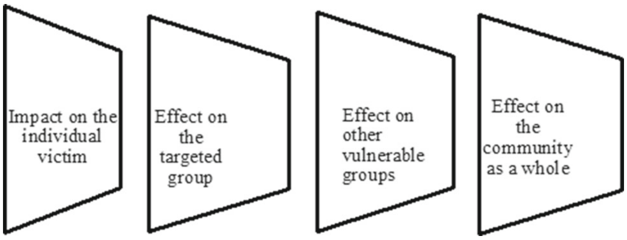

Impact on the individual victim
Effect on the targeted group
Effect on other vulnerable groups
Effect on the community as a whole

Fig. 8 Repercussion of hate speech

Only a few prior surveys included an in-depth examination of OHS databases. We attempted to cover practically all of the accessible datasets in our work, and scholars can also refer to the hate speech databases for extra information.[1]

We investigated most of the datasets that are used in the detection of OHS is imbalanced. So, to use these datasets for

[1] https://hatespeechdata.com/.

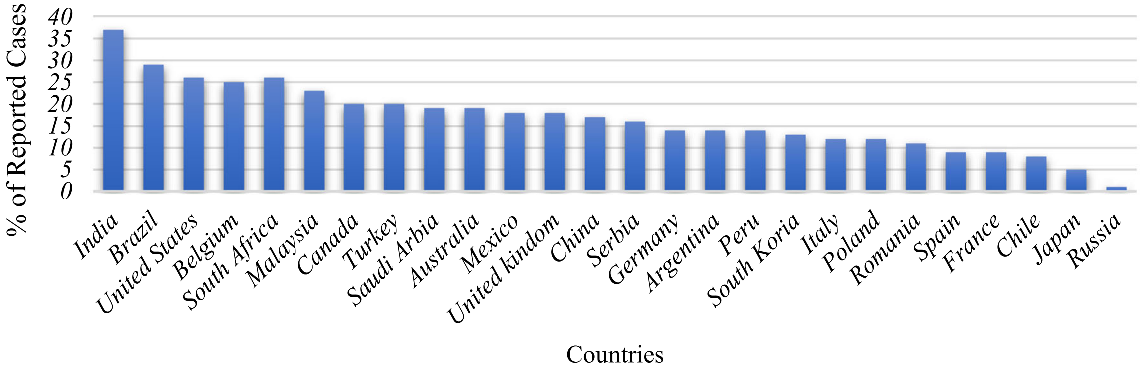


Fig. 9 Registered cyberbullying case

classification, the researcher adopted oversampling or undersampling techniques. In the next section, we have discussed a few techniques for the sampling purpose and their associated advantages and disadvantages.

6.2 Types of datasets

This section discussed the datasets used in the previous papers for OHS detection. In supervised machine learning, we deal with the labeled dataset, but in unsupervised machine learning, we deal with the unlabeled dataset. Few labeled data with a high amount of unlabeled data are used in semi-supervised learning. The labeling of data is labor-intensive and high-cost associated work. So, in this section, we explored the type of dataset which can be further classified as balanced and unbalanced dataset and we found that mostly all the given datasets in Table 4 are an unbalanced form. Therefore, for better results, different sampling techniques are taken into consideration by the authors.

- *Labeled dataset and unlabeled dataset* The labeled datasets are the one in which we have both the parameters that are input and output. The author [49] collected the unlabeled multilingual data from Twitter. Thereafter keyword-based approach is used to annotate the data and then, transfer learning is used to cluster the data into hate and non-hate. To manually tag the dataset is a very time-consuming and labor-intensive job. Therefore, tools development that can automatically label the text is a very interesting area to work on. On the other hand, in the unlabeled data, we do not have the output parameter, which means that the tag is not attached to the data. We only have raw data that we fed into the classifier, which finds the hidden parameters within a dataset. The author [27] has used labeled and unlabeled dataset for training and testing the classifier, respectively. To work with an unlabeled dataset is less costly as compared to labeled dataset and is therefore used in unsupervised machine learning.
- *Balanced dataset and unbalanced dataset*: When all the dataset are almost equally distributed among all the classes, then it is known as a balanced dataset. Example: suppose we have two classes as hate and non-hate, and the dataset contains 10 k tweets. Hate: 4.5 k and non-hate:5.5 k. But in a real-time scenario, we have some degree of imbalance like medical diagnosis, fraud detection, etc. If this degree of imbalance is low, then it is still called a balanced dataset. However, if this degree of imbalance is high, then this will impact the performance of the model [55]. So, when almost all the dataset belongs to one class only, it is called an imbalanced dataset. Example: From the total 10 k tweets, we have 2000 for hate and 8000 for non-hate. The author [56] used an imbalanced dataset in their work, but the classifier falsely classifies new observations to the majority class. In Sect. 2.2.1, we explore some majorly used sampling algorithms that are used in the previous work.

6.2.1 Techniques to deal with an unbalanced dataset

The term "class imbalance problem" in machine learning refers to categorization issues where groups of data are not separated equally. Sometimes considerable skew in the classification process of a binary or multi-class classification task is indicated by the nature of the problem in many application areas.

Under-sampling To mitigate the effect of an imbalanced dataset, the author [57] has used the under-sampling technique, in which random samples have been chosen from the majority class data present in training set to balance with the minority class. But this technique might discard some crucial information because it reduces the samples from the majority class, which can lead to the loss of some relevant information. Becoming more selective with the examples from the majority class that are eliminated can be an extension of under-sampling strategy. The Heuristics approaches [32] are frequently used in this process, which tries to find redundant

Table 4 A detail list of online hate speech dataset

References	Source	Language	Size	Summary
[14] [15] [16] [17]	GitHub [by zeerak W][a]	English	6655 Tweet Dataset Distribution: NAACL_SRW_2016 None: 11,559 Racism: 1969 Sexism: 3378	The dataset is annotated by only 1 Expert and three amateur annotators The author found k = 0.57 Cohen kappa value for the proposed dataset
[9]	GitHub [keras-team][b] And WaCky corpora[c]	English	17,567 Comments three classes as strong hate, weak hate, and No hate	Three different annotators are used to annotate the dataset, and the comment is collected from the Facebook pages To find the level of an agreement, the author computed the Fleiss' kappa(k = 0.19) inter-annotator agreement
[18]	GitHub [by zeerakW][d]	English	6909 Tweets Dataset Distribution: NLP + CSS_2016 Neither: 5263 Racism: 207 Sexism: 1269 Both: 52 Link: 118	Twitter API is used to collect data To find the reliability of the dataset, the author calculated the Fleiss' kappa(k = 0.74)
[19] [20]	GitHub [by T Davidson][e]	English	25,000 Tweets Dataset Distribution: Hate: 1430 Offensive: 19,190 Neither: 4163	Three crowdflower workers coded the tweets manually The author used Flesch Reading Ease scores and Flesch-Kincaid Grade Level to capture the quality of each tweet The author found a 92% intercoder-agreement score
[21]	WebScope Dataset[f]	English	2000 Comments Two classes as clean or abusive	All the comments were collected from yahoo's new posts page The agreement rate of annotated data is 0.922, and Fleiss's Kappa is 0.843
[22]	Stormfront And crowdflower[g]	English	10,568 Two classes as Racist and religion	Sentence-level annotator from Stormfront and to find the hate or non-hate they used the crowdflower dataset

Table 4 (continued)

References	Source	Language	Size	Summary
[10]	Tumblr dataset[h]	Arabic	5,569 comments. Dataset Distribution: Hate: 2512 Non_hate: 3057	To annotate the data, they arrange tasks on crowd flower websites only who speak Arabic Two annotators were participated to annotate the data The proposed dataset found highly imbalanced, and the inter-annotator agreement and Cohen's Kappa coefficient was 0.95
[11]	HatEval[i]	English and Spanish	9 k Dataset Distribution: English_train Non_ hate: 5217 Hate: 3783	The data are collected from the Twitter page Hate against immigrants and women taken into consideration only
[23]	Kaggle[j]	English	Dataset Distribution: Neutral: 2898 Insulting: 1049	The data are collected from Twitter
[24]	TRAC(Facebook)[k]	English and Hindi	Non-aggressive:69% Overtly aggressive: 16% Covertly aggressive: 16%	The data are collected using Facebook API from Facebook
[25]	Hatebase database[l]	All languages	N/A	Hatebase consists of all the hate words that are present in almost all languages. Example: Gender Sexual-orientation, disability class
[1]	HASOC (2019)[m]	Hindi, German and English	5983- Hindi 7005-English 4649-German	The dataset is classified into non-hate, offensive and hate and offensive. Also, the data were collected from Twitter and Facebook websites
[26]	Zenodo[n]	English, German, Spanish, French and Greek	Approx 90 k-English 62 k-German 38-Spanish 39 k-French 62 k-greek	Each dataset contains a tweet id and their annotation To access the dataset pre-request to the zenodo is required

Table 4 (continued)

References	Source	Language	Size	Summary
[27, 28]	Github[o]	Arabic	Total 6000 text Tweets 2,526- hate. Which is divided as: [Jews-33% Shia-32% Christians-25% Atheists-24% Muslims-9% Sunnis-7%]	The author collected the data from Twitter The dataset is classified into hate and non-hate class and contains religious hate speech The dataset gives an accuracy of 0.79 while experimenting on GRU-based RNN with pre-trained embeddings
[29]	Github[p]	English, French, and Arabic tweets	English-5647 French -4014 Arabic-3353	The dataset was collected based on Directness, Hostility, Target, Group and Annotator attributes The annotator agreement scores for labeling the dataset are 0.153, 0.244, and 0.202 for English, French, and Arabic, respectively,
[30]	Github[q]	Arabic	Total 5,846 tweets Abusive-1728 Normal- 3650 Hate-468	The author collected the data from Twitter, which was Group-directed and Person-directed Tweets
[31]	File[r]	Arabic	Total 1,100 tweets Percentage abusive: 0.59	The Twitter platform is used for the dataset collection
[31]	Github[s]	English	Total 33,776 posts Hate-14,614 Non-hate-19,162	The author collected the dataset from the Gap website

[a] https://github.com/ZeerakW/hatespeech/
[b] https://github.com/keras-team/keras
[C] WaCky corpora
[d] https://github.com/ZeerakW/hatespeech, https://github.com/AkshitaJha/NLP_CSS_2017
[e] https://github.com/t-davidson/hate-speech-and-offensive-language
[f] https://webscope.sandbox.yahoo.com/?guccounter=1
[g] Stormfront database, crowdflower, github.com/aitor-garcia-p/hate-speech-dataset, https://data.world /crowd flower/hate-speech-identification
[h] https://data.mendeley.com/datasets/hd3b6v659v/2, https://github.com/nuhaalbadi/Arabic_hatespeech
[i] https://competitions.codalab.org/competitions/19935
[j] https://kaggle.com/c/detecting-insults-in-social-commentary
[k] http://trac1-dataset.kmiagra.org
[l] https://hatebase.org/recent_sightings/
[m] https://hasocfire.github.io/hasoc/2019/dataset.html
[n] https://zenodo.org/record/3520152#.XcL0OnUzY5k
[o] https://github.com/nuhaalbadi/Arabic_hatespeech
[p] https://github.com/HKUST-KnowComp/MLMA_hate_speech
[q] https://github.com/Hala-Mulki/L-HSAB-First-Arabic-Levantine-HateSpeech-Dataset
[r] http://alt.qcri.org/~hmubarak/offensive/TweetClassification-Summary.xlsx
[s] https://github.com/jing-qian/A-Benchmark-Dataset-for-Learning-to-Intervene-in-Online-Hate-Speech

examples that should be deleted or beneficial examples that should not be deleted.

Over-sampling Class imbalance decreases the predictive power of the classification systems. These algorithms frequently attempt to maximize classification accuracy, a parameter that benefits the dominant class. A classifier can nevertheless achieve high classification accuracy even if it cannot accurately anticipate even one instance of a minority class. In this technique, we increase the number of minority class data in the training set. Each point in the minority class tries to increase, to balance with the majority class. It is much more efficient than under-sampling because, in under sampling, we lose some amount of data. However, oversampling is prone to overfitting because we try to duplicate the example of the minority class in the training dataset [58]. In order to address the overfitting problem in oversampling for the binary classification, this research [33] offers combining the k-means clustering algorithm with SMOTE. The proposed over sampler may locate and focus on input space regions where the creation of false data is most efficient by using clustering.

Simple oversampling does not add any new information to the model because it is just duplicating the existing examples, making it vulnerable to overfitting, which can also lead to low bias and high variance results. Therefore, in order to tackle the problem of oversampling, SMOTE was introduced by the author [59] in 2002. SMOTE works on the principle of nearest neighbor and evaluates the average of it by considering the examples that are close in the feature space without duplicating the data points. By using this technique, we can create synthetic examples using skew and rotation in the feature space rather than duplicating them [60].

7 Feature extraction IN OHS

Detection of hate speech using machine learning is a prominent approach. The accuracy of traditional machine learning algorithms mainly depends on feature extraction. In this section, we will discuss all the handcrafted features of the machine learning algorithm. In the feature selection process, with the increase in the number of features, the threshold value increases, which in turn may decrease the accuracy of the model. Therefore, whenever we give large feature data, our model gets confused because it is learning too much information. In order to resolve this situation, we do not select all the features from the particular dataset; instead, we use some specific type of features only, which increases the accuracy of the model. In Sect. 6.1, we have discussed the types of features that play an important role in classifying the text as hate or non-hate.

8 RQ5: "What are the extracted features in the Traditional machine learning algorithm for OHS?"

8.1 Types of features

Simple surface-level features In order to classify the text in the different classes, these types of features are basic things to be performed first. The majority of the authors have used BOW, N-gram, char-n-gram, frequency of URL, punctuation, and capitalization in the given sentence. Bow and TF-IDF approach does not store the semantic information because there is a chance of overfitting. The author [61] used a multi-task learning approach, where different features like BOW, N-gram and sub-word embeddings were used. BOW technique [62] is employed to make the dictionary of the misogynistic and non- misogynistic. However, researchers used these features with other high-level features in order to increase the efficiency of the model [3, 56, 58, 60, 63–66]. We conclude that the performances of these features are very predictive.

Word generalization Most of the authors yields good classification results using Bow, meaning, in training and testing datasets, these predictive words will appear. If the dataset contains small sentences, then our model can suffer from the data sparsity. Therefore, by using the word generalization technique, this issue can be addressed. To achieve the task [63], the clusters of words are taken as additional features and brown clustering can be used to do so. If new words come up, then, based on some degree of similarity, we assign any one of the clusters to that word. In the paper [67] Word embeddings using gensim's word2vec model had been used which was found to be useful when compared to simple BOW and TF-IDF. The author [27] provides a short survey on OHS using NLP. According to the author, token-level approaches as compared to character-level approaches perform better. Word embedding and paragraph embeddings use the same concept [42, 57].

Sentiment analysis Hate speech itself is a negative word. If a sentence is negative in polarity, then it may be a case of hate speech or offensive speech. By taking this assumption in mind, several approaches of sentiment analysis are taken into consideration. The author has two different approaches: a multi-step approach or a single-step approach [68]. In the multistage approach, the author used sentiment analysis in the first step to finding the negative polarity, and then these negative features are further used to find the exact dictionary of the hateful words. On the other hand, in a single-step approach [39], only features are exacted using the sentiment analysis and are classified as hate or non-hate based on the polarity of the word. High variation in the degree of the polarity, such as highly negative words, also plays an important role in the classification. The SentiStrength algorithm can also be used as a feature extraction algorithm to find the type of polarity of the document [69].

Lexical-based approach Generally, the hate speech consists of hate words, so the authors use the general assumption that hate speech contains hate words or negative words (like insulting words, slurs, etc.). In the lexicon approach, hateful words are taken into consideration [70]. If the word is present in the dictionary, then only classifier predict is as hate; otherwise, it will classify the sentence into the non-hate category. Hatebase1 is popularly used to find all the hate or negative words that are present in all the languages. Apart from all the list of hate words, the author focuses on the list of some specific classes of hate like racism, sexism or ethnic hate-related words. Some authors also try to identify the hate words by manual inspection tasks. In paper [71] author used the rule-based approach for subjectivity detection and to develop the hate speech classifier. For the sentiment analysis, subjectivity analysis plays a vital role, and multi-perspective question answering is used for subjective clues. They applied the bootstrapping algorithm to augment the lexicon. The author considers mostly blog and Israel-Palestinian conflict datasets for race, nationality and religion target groups. Most of the authors [8, 55, 58, 72–77] used the lexical approach in addition to other features or as some baseline features.

Linguistic features Sometimes, the classifier often confuses between the offensive or hate speech. Identifying the semantics of the sentences plays an essential role in hate speech detection [68] as language often comes both in the form of slurs and insults. Hence, tagging POS (part of speech) information adds some semantic information into the classifier [73]. But POS alone cannot improve the performance; therefore, some authors add more information about the data like type dependency relationship [33]. Example 1: Wipe out the Muslims. Here, the term (wipe out, Muslims) has a typed dependency between both the words. The dictionary-based approach [42] is not very useful for context-specific mapping of the offensive words. Hence, to capture the opinion, the author has used a domain-based corpus approach.

Knowledge base To identify the statement as hate or non-hate is not an easy task, not even by using linguistic features. Sometimes, to classify the sentence, we need some background knowledge or domain knowledge [63]. Example: "Put wig and lipstick and behave as who you really are." In the given statement, hate is directed toward a boy and comments about the sexuality (LGBT) or gender of the boy. Therefore, in order to classify, one needs to have world knowledge. The author [78] introduced some world knowledge using automated reasoning, but that requires a lot of manual coding.

Multimodal information Modern social media is very popular for publishing multimodal information like audio, video, images and text. The hate does not come in the form of only texts. Lots of other content is circulated every day on social media platforms. To extract the information from the images, the author uses predictive features like user comments to find the semantics of the image. Also, the author [79] works on text and acoustic speech, but it does not yield very satisfactory results.

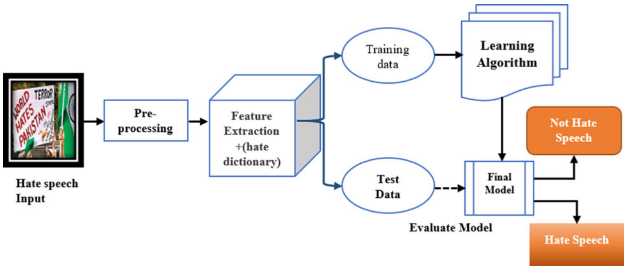


Fig. 10 Traditional framework For OHS

We analyzed all the features that are used in the various research on a different algorithm for OHS detection. Finding the best features in traditional machine learning is a very important task. Therefore, we have discussed all the features in table 5 that are used in the previous papers of OHS and we found that the most extracted features are surface-level features, linguistic features and lexicon features which outperformed the other existing features when used with the AI techniques.

.

9 OHS detection using traditional machine learning-based methods

This survey covered the various methods that have been adopted for solving the problem of OHS. The general framework of the OHS detection methodology is shown in Fig.10. The data are first pre-processed by removing punctuation, tokenization, stopwords and stemming or lemmatization so that they can be made fit for mining and feature extraction. To train the model, features are then extracted using various techniques like Bow, TF-IDF, word embeddings, etc. After pre-processing, the features are extracted from the pre-processed data. The next step is to pass the processed data in our trained classifier which classifies them into positive or negative class.

To answer the RQ6 from Table 1 We explored the various papers of OHS using machine learning to deal with online hate speech.

9.1 Support vector machine

The support vector machine (SVM) was invented back in the ’90 s by Vladimir Vapnik. SVM makes use of kernel trick to model nonlinear decision boundaries. It draws a decision boundary near the extreme points in the dataset. Therefore, SVM algorithm is essentially a frontier that best segregates the two classes. The author [56] has used SVM to

Table 5 List of handcrafted features used for the detection of OHS

S. no	Feature class	References
1	*Surface-level feature* A1: Bag-of-word A2: Negation A3: unigram A4: n-gram A5: Frequency of URL mention A6: Token Length and Capitalization A7: Non- English Words	[77, 80, 81, 82, 83, 84, 85, 79, 56, 65, 78, 86, 87, 3, 63, 88, 89, 55, 90–92]
2	*Word-generalization* B1: Set of words (Clustering) B2: Word Embeddings B3: TF-IDF	[3, 42, 56, 58–60, 63, 65, 66]
3	*Sentiment analysis* S1: Positive and Negative polarity S2: Neutral words	[58, 93, 57, 82, 65, 73]
4	*Lexical resources* L1: General Hate-Related terms L2: Contextual Information	[94, 33, 55, 65, 76, 91, 94, 80, 95, 32, 65, 79, 73, 55, 96, 97, 32, 66]
5	*Linguistic feature* F1: n-gram + POS information F2: Dependency Relationships F3: Syntactic Feature and Semantic feature	[44, 40, 98, 99, 100, 65, 101, 66]
6	*Knowledge-based feature* K1: Heteronormative Context	[99, 73, 102, 99]
7	*Meta- information* M1: Background information about the user of the Post M2: No of Post by User M3: No of reply by user M4: Location M5: Correlation between the number of post and hate speech	[41, 103, 72, 73]
8	*Multimodal information* C1: Images C2: Audio C3: Video and Audio Content	[104]

find the racist text using different kernel functions on the Bow, bigrams and pos in order to find the best effective technique. The highest accuracy was achieved on Bow using the polynomial function, but Pos performed worse than bow and bigram. It has been observed [73] that the SVM performed best on the surface-level features. On the binary classification [73], the SVM classifier gives the highest results in terms of accuracy. In paper [105] author collected data from yahoo newsgroup posts and the American Jewish Congress. A template-based strategy is used to generate features from the corpus. The author took the problem as word-sense disambiguation and used SVM light classifier as a linear kernel function. The proposed result using this classifier on the dataset was not accurate. Also, the bi-gram and tri-gram degraded the performance of the classifier.

Furthermore, long linguistic pattern was not detected and also resulted in a low recall and precision value. This paper [102] presented the annotation framework for hate speech of tweets that were collected during the Kenyan election. They developed the framework for the extracted text and employed bootstrapping and n-gram technique to obtain the hateful tweets from the 394 k collected data. For the reliability of annotated tweets, the author used Krippendorff's alpha. The same concept described in the duplex theory of hate (i.e., passion, distance and commitment feature for the hate speech framework) was used in the paper [26]. Out of 394 k tweets, 94% of tweets labeled ethnic. The authenticity of the data are not cared about, i.e., fake news and propaganda. Also, this framework is applicable only for short messages. SVM is one of the major adopted techniques by the researchers [3, 42, 65].

9.2 Naive Bayes

It is a supervised learning algorithm that is used for binary and multiclass classification problems. It is based on the Bayes theorem given by Thomas Bayes: the algorithm makes naïve assumption that the features are independent of each other, which makes the algorithm simple and effective.

$$P(A|B) = (P(B|A)P(A))/P(B) \tag{1}$$

P(A|B): The probability of finding the event A, when event B is true.

P(A): Prior probability that is the probability of an event before event B.

P(B): Prior probability that is the probability of an event before event A.

P(B|A): The probability of finding the event B, when event A is true.

In the detection of hate speech, the author [58] used naïve Bayes by extracting the surface-level features and lexicon features and found that the voting classifier gives the best results compare to the lexicon-based approach for the classification. The author [3] took at least three annotators to annotate the hate words and compared the results also Standard Pre-processing TF-IDF and n-gram is used after that Naïve Bayes gives the same accuracy as other classifiers. By using the hard ensemble, the author [8] achieved the highest accuracy of 78.3% with naïve compared to other classifiers on the unbalanced dataset.

9.3 k-nearest neighbor

It is one of the simplest and most used classification algorithms. This algorithm is used when data points are separated into several classes to predict the classification of a new sample point. KNN captures the idea of similarity. It is used to solve nonlinear classified data points means if the data points are distributed in a nonlinear manner, where we cannot just draw a straight line, there we can use KNN. In order to find the similarity between the data points, Euclidean distance, Manhattan distance is calculated. Then an object is classified based on the number of votes of its neighbors with the object being assigned to the class most common among its nearest neighbors. To find the prominent pages on Facebook, the author [58] used Betweenness Centrality. Very few works have been identified in the field of hate speech detection.

9.4 Logistic regression

Logistic regression (LR) is used to solve binary classification and multiclass classification problems, i.e., output $y \in \{0,1\}$. Regression estimates the relationship between the dependent and independent variables. Hence, LR is most widely used when the dependent variable or the output is in binary format or categorical format. The author [42] implemented a logistic regression with the surface-level features, which gives comparable results. We did not find much work on word generalization and knowledge-based features in logistic regression. Furthermore, very few works have been seen by considering different features set to classify the sentences shown in Table 6.

9.5 Decision tree

Decision tree (DT) is a flow chart-like structure in which each internal node represented a "test" on an attribute, and each branch represents the outcome of the test, and each leaf node represents a class label. DT is used to map nonlinear relationship, means if data are not easily separable, then we draw or split it into different classes. DT is used by the authors [42], and surface-level features were the first choice of the research to use in the classification process.

9.6 Random forest

It creates DT on data samples and then gets the predictions from each of them and finally selects the best solution by means of voting. It is an ensemble method that is better than a single DT because it reduces the overfitting by averaging the result. The author [70] used the ensemble of DT to work on the video platform to find the hatred on the multimodal data. The author finds the maximum accuracy of 0.94% with a weighted-vote ensemble. Author [106] detects the hateful content on Twitter and Whisper. As whisper is an anonymous mobile application, they collected nearly one-year data from the whisper app and 1% random sample from Twitter, which is available to all the users. They present the computational method to detect hate speech in which they divide the sentence into four parts, i.e., I, Intensity, user intent and hate target. Also, there is a possibility of biases as the collected data are from the online social network.

9.7 Artificial neural networks

It is an interconnection of assembly of nodes to form structures using a directed link. A simple artificial neural network (ANN) consists of only one hidden layer. Perceptron is a simple neural network which can be further classified as a single layer and multilayer. Multilayer perceptron consists of hidden layers and hidden networks. The author [60] fed extracted features into the simple ANN classifier and followed a genetic-based approach to detect the hate speech in the Albanian language.

9.8 Explainable artificial intelligence

Explainable artificial intelligence (XAI) is technology which decodes the reason behind the neural networks and presents it in form understandable by humans [107]. With neural networks becoming more and more complex with many more parameters and feature engineering becoming a thing of the past, making deep learning models justifiable is the need of the hour. XAI has already gained significance in the domain of computer vision with visualizations like class activation maps becoming more and more popular. Class Activation maps are made by overlaying the features of a layer in DNN on the image to classified signifying the importance a model places on a particular region or pixel. Class activation maps help data scientist design a model which uses relevant features to make a decision, making the model more reliable. The adoption of XAI has been low though recently sudden interest has been seen. The author [107] released a benchmark dataset in which each tweet has a class-label (hate, offensive, normal), a target community and the rationale behind its class-labels. The author further shows that it is not necessary that the models performing best according to traditional

Table 6 ML classifier used with general features of OHS

ML classifier	Feature							
	Surface-level feature	Word-generalization	Sentiment analysis	Lexical resources	Linguistic feature	Knowledge-based feature	Meta-information	Multimodal information
SVM	[42, 3, 40 [98, 100, 103, 82, 81, 50, 4 [77, 3, 74, 4]	[32, 65]	[82, 65]	[95, 32, 65]	[44, 40] [98, 99] [100, 65] [101]	[102, 99]	[102, 103]	N/A
NB	[42, 40, 108, 103, 109, 81] [77, 92, 104, 3, 74] [79]	[99, 65] [79, 56]	[65]	[95, 65] [79]	[40, 99] [65, 33]	[102, 99]	[102, 103] [72]	[104]
RF	[42, 40, 100, 82, 77, 3]	[65]	[82, 65]	[65]	[40, 100] [65, 33] [79]	[102]	[102, 13] [72]	N/A
KNN	[40, 74]	[58]	N/A	N/A	[40]	N/A	[72]	N/A
DT	[42, 40, 77, 74, 79]	[79]	N/A	[94] [79]	[40, 94, 33]	[94]	[72]	N/A
Logistic regression	[39, 40, 100, 31, 110, 82, 50, 77, 74]	N/A	[82]	[39]	[100, 111]	N/A	[31, 72]	N/A
Adaptive boosting	[82]	N/A	N/A	N/A	N/A	N/A	N/A	N/A
NLP	[93, 104]	[64]	N/A	N/A	N/A	N/A	N/A	[104]
J48	[3, 73, 55]	[73, 55]	[73]	[73, 55]	[99, 73] [55]	[99, 73]	[72, 73]	[73]

**N/A: Authors did not perform the task

Table 7 ML algorithms used in the research papers

Algorithms	No of frequencies used in the paper
SVM	26
Naïve Bayes	21
Random forest	13
Decision tree	12
Logistic regression	10
ANN	3
KNN	1
XAI	1

metrics such as accuracy, macro-F1 score and AUROC score will necessarily perform well on explainability metrics such as Plausibility, comprehensiveness and sufficiency.

Based on the total 95 articles in OHS, approximately 40 research papers used the traditional machine learning approach. SVM, Naive Bayes and decision tree are the most common approaches used in the papers of OHS in computer science background as shown in Table 7.

As part of the practical work that has been done, hate speech is being explored in relation to other pertinent concepts, including social media and machine learning. Machine learning techniques are being used to classify hate speech and automatically identify it.

According to the aforementioned literature, 136 research publications provided a variety of strategies for locating online hate speech in social networks. Unsupervised machine learning was discovered to be a relatively recent subject of study. Some researchers combined various techniques, such as sentiment analysis, emotional analysis and text mining, to effectively categorize the hate texts. As a result, each study has a unique perspective and understanding of online hate speech detection. In a nutshell, we have highlighted the following common flaws and limitation with current approaches.

1. From the study, it has been observed that the existing research covers mostly lexicon (simple keywords)-based hate speech analysis. As a result, the outcome of those models would not be able detect semantic of the text.
2. Facebook, Twitter and other social media platforms including the research papers that we have studied do not have a real-time hate speech detecting system and the corrective measures are taken only after the expression is posted online. So, real-time detection system can be made so that the corrective measures can be taken on time.
3. The majority of the methods are quite complex, including deep logical structures, complex equations, derivatives and formulas. Algorithms also required an excessive amount of computational time to execute. Straightforward and less complex model should be implemented so that the computational cost can be reduced.
4. Most of the researchers worked on highly imbalanced dataset, which would result in an inaccurate result. So, to deal with the class imbalance problem authors should adopt some strategies some of them are already listed in Sect. 5.2.1.
5. We also invested that majority of the study only used supervised learning and none of the author explored the area of unsupervised ML.

In Table 8, we have shown a comparison related to various traditional machine learning approaches and their associated advantages and disadvantages.

Considering the fact that online hate speech can occur in different formats, where the word, sentence, semantic and pragmatic knowledge of the language are significant. So, from the study, it has been observed that ngram and word embeddings can be a suitable approach to achieve better accuracy with machine learning models. Furthermore, LR and SVM often performed well when experimented with different approaches. We can see in Table 6 that surface-level features and linguistic features are most used with different traditional machine learning classifiers. Very little work has been done using other handcrafted features except for surface level and lexical resource. Moreover, some of the areas are not even explored. (Marked as ’NA’, Table 6). In the OHS, there is further scope to work on KNN, Adaptive Boosting classifier, "cleaning and stemming" and annotation of the data using automatic machine learning tools.

10 OHS detection using traditional deep learning-based methods

Traditional machine learning and deep learning, both offer ways to train models and classify data. In traditional machine learning, we manually extract features, but in deep learning, we skip the manual step of extracting features; instead, we put data directly into the deep learning algorithm like a convolutional neural network (CNN), which then further predicts the object. Therefore, deep learning is a subtype of machine learning which deals directly with data (like images) and is often more complicated. In this section, we have covered the various methods of deep learning that have been adopted for solving the problem of OHS. Figure 11 shows how deep learning model classifies the text as hate speech or not hate speech by taking some inputs. A deep neural network is a type of artificial neural network which has more than one hidden layer that helps to extract higher-level features from the dataset. At each level, the input is slightly transformed, and it gives more details of the data. Deep learning behaves

Table 8 Traditional frameworks of OHS

References	Approach	Language	Dataset	Methodology	Merits	Limitation
Raufi et al. [60]	ANN	Albanian	3620 words from Albanian forums	The author used Standard Pre-processing, feature extraction using a Bag of Words Extracted Featured Fed into ANN and Classified, and then new words are added to the hate vocabulary	The highest accuracy achieved is 94%, with a 60–30 spilled	In the Long Run, many word features will become irrelevant. Their current system is developed on "per word"-based detection, where deeper language constructs are not in their scope
Martins et al.[65]	RF, NB, and SVM	English	Davidson and Warmsley Total tweets are 24,782 Hate-1430 Offensive-19190 non-hate-4162	The author used lexicon-based and machine learning approaches to predict hate speech contained in a text, using an emotional approach through sentiment analysis	Finds the best accuracy with SVM compared to NB and RF i.e., 80.56%	The author gives emphasis on emotional features only Uses fixed vocabulary found on hatebase, Semantic features are not considered
Sharma et al. [78]	Machine Learning and NLP	English	Dataset available on Kaggle **2235** number of samples from various sites	Standard Pre-processing such as stop word removal stemming, etc., is done followed by labeling and adding the time comment was made	Real-time tweets are extracted from multiple online sites and created a new dataset	Prepares only data but does not build a classifier
Pelzer et al. [64]	NLP and Automated reasoning	Swedish	Collected 17,176 comments from open forums i.e., Avpixlat and Samha llsnytt	The author compared their developed automatic hate speech method with the manual analysis Build a (NLP + Automated reasoning) approach The author took six politicians, three males, and three females	NLP + AR approach is more easily adapted to other languages by modifying the underlying dependency recognition rules	NLP + AR technique finds very small hateful comments compared to manual inspection Hatecategory dictionaries do not capture all hate expressions
Davidson et al. [42]	Logistic regression, SVM and Sentiment Lexicon	English	Collected 24,782 annotated Tweets named as Davidson and Warmsley dataset	Standard Pre-processing, unigram, bigram, trigram features are extracted with POS tagging To classified a text as hate author used NLP. TF-IDF is used to find the most relevant word, and BOW is used to find the most frequent word	Overall F1 score of 90 is achieved with SVM and LR The dataset is highly skewed, not able to classify hate speech with high accuracy, will not be able to handle unseen vocabularies	40% of hate speech is misclassified:the precision and recall scores for the hate class are 0.44 and 0.61, respectively
Diwhu et al.[3]	SVM, J48, Naïve Bayes, Random Forest, Random Tree	Turkish	Collected 1288 Tweetsfrom Twitter. Where 159 was classified as hate and, 1129 classified as non-hate	The author took at least three annotators to annotate the hate words and compared the results They used Standard Pre-processing TF-IDF, and n- gram is used after that	Accuracy is in the range of 60 on almost all models	The author has adopted a complete lexical approach. This model will fail if new vocabulary is observed in the data

Table 8 (continued)

References	Approach	Language	Dataset	Methodology	Merits	Limitation
Rodriguez et al.[58]	Sentiment Analysis (VADER) Emotional Analysis (JAMIN), K-means clustering	English	Collected 1000 comments from each page from the Facebook using FB graph API	The proposed framework intends to identify prominent pages in social media where potential hate speech promoters may exist To find the prominent pages on Facebook, they used Betweenness Centrality	The author proposes a new way of dealing with hate speech, rather than building a classifier that classifies each tweet into hate and non-hate	The method is not completely automated since a man will be required to inspect the words near each centroid
Watanabe et al.[57]	Unigram and pattern classification, J48graft	English	Three different datasets from two from Crowdflower and one from GitHub. Divide dataset into three categories as hate, offensive, and clean	The author proposed an approach that collects words and expression in a pragmatic way and uses them with patterns, along with other sentiment-based features to detect hate speech	The proposed approach reaches an accuracy equal to 87.4% for the binary classification of tweets into offensive and non-offensive and an accuracy equal to 78.4% for the ternary classification of tweets into, hateful, offensive and clean	Richer dictionary of hate speech patterns can be used for the better classification
Greevy and Smeaton [56]	SVM	English	3 million words collected from Yahoo The phrases and a list of words have collected from the Yahoo dictionary	Applied SVM model on BOW, Bigram and POS feature Experiment conducted using the default linear, polynomial, radial basis function and sigmoid tanh as kernel functions The author used differentkernel functions on Bow, bigrams and pos in order to find the best effective technique	Polynomial proved to be the most effective kernel function for both BOW and POS The highest accuracy was achieved on Bow using the polynomial function	The author used the linear kernel function and sigmoid tanh for BOW and POS. But they are computationally expensive Pos performed worse than Bow and bigram Better results can be achieved if experimented with bow + bigram or bow + bigram + pos bow + bigram + pos
Smeelakshmi et al.[67]	Facebook pre-trained word embeddings, SVM-radial bias, Random Forest, SVM-linear	Hindi English code mixed data	10,000 data from different sources	It is found that character-level features give more compared to doc information for code-mixed classification The author used doc2vec and word2vec and FastText library for the feature selection	In the first experiment, the author finds that the RF gives high accuracy of 0.6415% compared to SVM- RBF and simple-linear when incorporating Doc2vec features In the second experiment, the author found that the SVM-RBF gives high accuracy of 0.7511% compared to RF and simple linear by incorporating word2vec embeddings In the 3rd experiment, the author found that SVM-RBF performed better than previous techniques by incorporating FastText embeddings with an accuracy of 0.8581	Classification of the tweets has been not done on multi-classification

**N/A: Authors did not perform the task

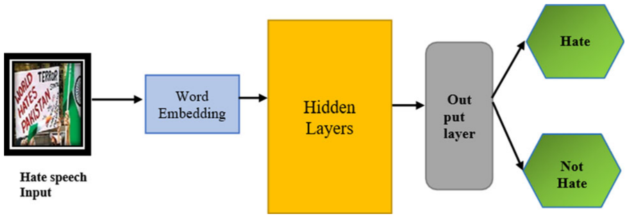


Fig. 11 Deep learning framework For OHS

like a black box for some researchers because it does not require feature engineering. We found that as compared to ML, very little research has been done in the area of deep learning for hate speech detection till 2019. The reasons for the less amount of research in DL can be label data scarcity and unavailability of the high-performance GPU. However, the trend has shifted to deep learning in 2020. According to our findings, we found the majority of research papers from the year 2020 in deep learning as compared to traditional machine learning. In upcoming sections, we will discuss different types of deep learning models that have been used in the previous literature.

10.1 Recurrent neural network

ANN cannot capture the sequence of information, which means it does not have an account for the memory. On the other hand, RNN is a type of neural network that captures information about the sequence or time-series data. It can take variable size input and give variable size output and works very well with time-series data. They are a class of artificial neural networks where connections between nodes form a directed graph that allowing information to flow back into the previous parts of the network. Thus, each model in the layers depends on past events, allowing information to persist. RNN works on the given recursive formula in equation 2. In order to detect sentences as hate or not, the author implements tests with RNN, data-partition, epoch, learning rate and batch size. All these parameters affect the system performance. The author [112] used UTFPR models in order to process the text. Then character embeddings fed into the RNN layers. The proposed system is based on the compositional RNN. The proposed model is robust, even when the input data are noisy, but the dataset that is used to feed the RNN is very small, and the performance of the classifier can be affected if a large dataset is taken.

$$S_t = F_w(S_t - 1, X_t) \qquad (2)$$

Xt —input at time step *t; St*—state at time step *t; Fw* —Recursive function

Social media such as Facebook, Twitter and Instagram are becoming a ubiquitous platform for people to share and express their opinion toward something [113]. Online Social network, especially Twitter, has a prodigious influence on the success or demolition of a person's image [114]. The author [84] used an RNN DL-based approach to detect the hate speech text on Twitter data. Thereafter, 1235 posts were analyzed using case folding, tokenization, cleansing and steaming. The data are collected from the Twitter accounts by the Twitter API. Using RNN (recurrent neural network) and LSTM (long short-term memory), it can process not only single data but also an entire sequence of data at a time. word2vec is used to convert sentences into vector value or to find the semantic meaning. Test the data with epoch, which resultant in high precision of 91% and recall 90% and an accuracy of 91%. The author [115] represents machine learning with a hybrid NLP approach where killer NLP with ensemble deep learning is used to examine the data, which gives 98.71% accuracy of the system. The authors [50] address the problem of identifying speech promoting religious hatred in the Arabic Twitter. They created an Arabic dataset of 6000 tweets annotated for the task of hate speech detection and Arabic lexicon with scores representing their polarity and strength. They also developed the various classification model using a lexicon-based, *n*-gram and deep learning-based approach. But the author used GRUs rather than LSTMs because GRUs can be trained faster and may achieve the best performance on datasets that have a limited number of training examples. GRU (gated recurrent unit)-based RNN model produced the best results for the evaluation metrics. The study [134] demonstrates how psychologists have looked into the connection between hate and personality. The author used a text-mining strategy that completely automates the personality inference process. A deep learning algorithm called PERSONA has been developed to identify hate speech online.

10.2 Long short-term memory

LSTMs are a modified version of a recurrent neural network capable of learning long-term dependencies, usually used for time series analysis. They can process images, speech and video. It is made up of gates viz. input, output and forget which have the function of, respectively, receiving the data, outputting it and deciding what to pass and what not to In RNN, we suffer from the vanishing gradient problem, which is as we propagate the error back through all the multiple layers of the RNN. Hence, LSTM solves the problem of vanishing gradient and gives much better accuracy than RNN because RNN fails to establish the long-term dependencies. To classify the OHS, the author [85] used the LSTM classifier and FastText library and found that the binary classifier obtained comparable results as that of sentiment analysis. The author [38] used GloVe embedding-based method and LSTM classifier, in which embeddings learned from the model, and that leads to high accuracy. The author [79] used two models one Textual model and the second Acoustic model. LSTM model performs better on textual data rather than on acoustic data. To determine the hateful or neural [43], the author used NLP classifiers with paragraph2vec. The performance of the experiment has increased as the number of hidden layers increased, also the author experiment with the five hidden connected units and two hidden layers which gives the 0.99 AUC over 200 iterations. An ensemble of the LSTM classifier improves the classification [115]; also, the author used a combination of various features, which gives the high F score 0.9320. To work in the Hinglish language, author [116] found that the LSTM classifier calculated a maximum recall value of 0.7504 on specific hyperparameter settings.

10.3 Convolutional neural network

Convent or CNN, it is a subclass of DNN (deep neural networks). CNN mostly used in the area of analyzing visual imagery. The three layers of an image are converted into a vector of suitable size, and then a DNN is trained on it. Their other applications include video understanding, speech recognition and understanding natural language processing. The author [39] used CNN in order to find racism and sexism speech. The proposed model is tested by tenfold cross-validation and gives a 78.3% f score. The author [68] employed text features, i.e., surface-level features, linguistic features and sentiment features in deep learning classifiers, and then implemented an ensemble-based novel approach. The author finds the accuracy of 0.918 with the novel approach. Batch size, epoch and learning rate affect the system performance. Also, the studies show that a larger training dataset produces better results [27]. To visualize the online aggression on Twitter and Facebook, the CNN-based web browser plugin had been presented by the authors [117].

10.4 Transformer methods

The transformer [118] is the latest innovation that has taken the domain of natural language processing by storm. Transformer like its predecessor has the ability to account long-term dependencies, but unlike LSTMs transformers do not process data sequentially as done in the case of RNN's and LSTMS. Instead, to account for the position of each word is added to its embedding. The transformer was first introduced for machine translations (Sequence to Sequence Model), and thus it has two components, an encoder and a decoder. Though only the encoder is relevant in text classification tasks such as Hate speech detection. It is vital to understand to study transformer in totality. In an encoder the inputs are first fed into a self-attention layer which generates an embedding taking into account all other words in sentence and depicts the relevance of each word with respect to a particular word. The embeddings obtained from self-attention layer are fed into neural network. This process is repeated many times, i.e., many layers of self-attention and neural networks are stacked to form the encoder. The decoder of a transformer is very similar to the encoder except for an encoder–decoder attention layer, which is added to find the inputs relevant to a particular output [118]. In the context of hate speech detection embedding obtained from the pre-trained model such as BERT (Bidirectional Encoder Representations from Transformers) has been widely used. BERT is a transformer trained using the masked LLM technique. Masked LM technique [119] requires 15% of the words in the sentence to be masked, and the transformer then attempts to predict these words from context during the training process. In the paper [120] the author showed the efficacy of finetuning the Bert in the context of hate speech detection. Comparing pre-trained models for hate speech detection explores and compares various multilingual transformers such as Mbert, Beto.

In this paper [121], the authors argue that, for the multi-class classification problem of online hate speech, transformers must be used over basic traditional machine learning, basic RNN-based deep learning or even attention-based RNN models to achieve the state-of-the-art accuracy. They propose a streamlined version of BERT, called DistilBERT, which has half the number of parameters with no loss in performance. On comparison and experimentation with various LSTM and BERT-based models, DistilBERT outperforms all the models given on various metrics. This paper [122] provides us with a comparative analysis of three different types of models, namely baseline traditional machine learning models, Deep Learning models and Transfer Learning-based models for Hate Speech classification in the Spanish language. This comparison shows how Transfer learning models outperform traditional machine learning models, which are used as the baseline. They evaluate the performance of pre-trained Language models. The authors showcase that the pre-trained monolingual language model (BETO) outperforms pre-trained multilingual models like Bert and XLM, concluding the requirement of hate speech models to be language-specific. In the paper [123], the author uses GPT- 2; it is a language modeling transformer released by open AI. It was trained on a massive dataset of Web text, which required storage space of 40GB and contained parameters ranging from 117 million to 1500 million. Though both BERT and transformers are transformers a stark difference can be observed between these two in their usage; while BERT finds its usage in creating embedding that incorporates the context of whole sentence GPT-2 is widely to generate sentences. The architecture of these transformers presents a stark difference as well, while BERT is entirely made of encoders and GPT-2 is entirely made up of decoders. Further, GPT-2 relies on autoregression that is GPT-2 produces

Table 9 Classification of type of input to deep learning model

DL classifier	Input type					
	BOW, N-gram, Char n-gram, Skip-gram	Word embeddings, TF-IDF	Positive, negative words	Hate-related terms	POS information	Image, audio, video
ANN	[104, 79]	[43, 104, 79, 84, 80, 79, 78, 77, 75, 74, 73, 72, 64, 65, 46, 45] [44, 43, 42, 41] [40, 39, 35, 38] [37, 36]	N/A	[43, 79]	N/A	[104]
CNN	[37, 110] [8]	[37, 86]	[8]	N/A	[8, 101]	N/A
DNN	[129, 45, 81]	[129]	N/A	[129]	[45]	N/A
RNN	[50]	[84, 130]	N/A	N/A	[101]	N/A
LSTM	[110, 79]	[86, 85]	N/A	[96, 97, 32]	[101]	N/A
Dense NN	N/A	N/A	N/A	N/A	N/A	N/A

**N/A: Authors did not perform the task

tokens sequentially and once one token is produced, it is included as input for the next token. Though the technique of autoregression has its cons since on using auto-regression, the model loses the ability to utilize the context on both sides. It has been proven that GPT-2 achieves excellent results. The authors of this paper [124] propose a novel solution to the hate speech binary classification problem statement by scaling up the small public datasets available using a Deep Generative model, here GPT-2[125] to produce large datasets for the training of Deep Learning-based classifiers and satisfy their extensive data requirements. In the paper, the GPT-2 was finetuned according to the public datasets for the generation of data points. Then they test these models intra-dataset and cross-dataset among the public ones to compare the increase in accuracy and generalization across different probability distributions of datasets. In the paper [126], the author used the transfer learning and Compact Bert variants in a pipeline model. The pre-processed data are loaded into batches of text and true labels and tokenized with a pre-trained BERT tokenizer. The final layer is removed and a dense layer of size 3 is added, because of three different classes then the dense layer SoftMax is used, to get probability scores for each class where maximum probability results in predicted label. Also, Focal loss is used as a cost function. It is beneficial with a class imbalance problem. In order to improve the overall accuracy of the system the author [127] used the ensemble of different features and study the effects of TF-IDF and sentiment bases features. The author also presented the criterion for the selection of computational complexity and classification performance among the existing methods. To detection of hate speech in Spanish language different pretrained models were analyzed [128], where SVM and logistic regression was used for text categorization and Bert model was fine-tuned with input of 512 tokens, output vector has dimension of 768. However, the transfer learning models outperformed the traditional machine learning approaches for the Spanish vocabulary.

In Table 9, we have analyzed the types of inputs that can be provided to the deep learning algorithms so that model can perform better by taking low computation resource. However, we did not get satisfactory results as word embedding is the first choice of the researchers for the input parameter and other methods of DL with varied input parameters were not explored. Most of the fields in Table 9 are NA (not applicable), which means that no work has been done using these inputs in the specific type of classifier. In Table 11, we concluded each DL paper's merits and limitations, but it is not very clear in the papers which approach performed better. Also, some recent studies show that deep learning gives better results than a traditional framework, but again these results are not very consistent. Based on the selected 111 papers, we found that most of the authors used SVM, Naive Bayes, Decision Tree in ML and CNN, LSTM in the DL approach also shown in Tables 7 and 9. From the recent trend, we have also all seen that the transformer-based techniques are the most used approaches among the researcher.

From Table 10, we found that most of the authors used SVM, Naive Bayes, Decision Tree in ML and CNN, LSTM in the DL approach. From the recent trend, we have also all seen that the transformer-based techniques are the most used approaches among the researcher (Table 11).

Table 10 DL algorithms used in the research papers

Algorithm	No of frequencies used in the paper
CNN	11
LSTM	10
RNN	8
DNN	6
Transformer	8

10.5 Different organization contribution toward OHS

In this section, we have discussed the various workshops and competitions which contributed to the online hate speech problem.

o **SemEVAL**

o It is a research workshop that works to advance the SOTA on semantic analysis and offers different NLP tasks based on semantic analysis to build efficient systems for these problems. Through these challenges, it aims to build datasets that can be publicly used for further research.[2]

o HASOC (hate speech and offensive content identification in Indo-European Languages)

o It is a forum that provides datasets in multiple languages for two Hate Speech subtasks for different classification. Participants are expected to use these datasets and create systems as solutions to these subtasks. These datasets comprise ten thousand annotated tweets.[3]

o **GermEVAL**

o This is a series of Natural Language Processing tasks in the German language that are released for people to build efficient systems on. The datasets are provided by the forum and are an amalgamation of German tweets.[4]

o **TRAC**

o This workshop aims to use NLP and related methods for the detection of online aggression, trolling, cyberbullying and related phenomena in text and speech present on social media platforms to deal with inflammatory content. It has two subtasks, each pertaining to a different set of classes and to solve these problems, it gives 5000 annotated data from social media in Bangla, Hindi and English.[5]

o Hateful meme challenge

This challenge is organized by Facebook AI, wherein they provide a dataset of memes containing text and images. The goal is to create a system wherein the model is able to accurately identify hate speech in this multimodal dataset and perform classification. The dataset contains 10000+ examples of memes which are annotated.[6]

OSACT4 Shared Task on Offensive Language Detection (Subtask A)

This challenge uses the Arabic SemEVAL dataset for binary classification problem statement of Arabic Hate Speech. The goal is to create a system which is capable of classifying Arabic tweets into offensive or non-offensive.

MEX-A3T

The goal of this community is to improve the further research in misinformation and aggressive speech by improving the research in NLP-related task. This research group provides different tracks to the researchers in the same domain only.

10.6 Evaluation metrics

Evaluation metrics are the mathematical functions that provide constructive feedback and are used to measure the quality of a traditional machine learning model. Most of the state-of-the-art online hate speech detection used an F1 score [31, 73, 99], precision [105, 131], recall [43, 131] and accuracy[43] for measuring the effectiveness of the parameters. We have discussed some most used evaluation metrics in the literature. With XAI becoming more and more relevant in artificial Intelligence, it is important to discuss the metrics used to measure the explainability of a model.

1.*Precision* The piece of relevant information from the total information.

$$P = \text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

2. *Recall* The percentage of total relevant information correctly classified by the classifier.

$$R = Recall = \frac{TP}{TP + FN} \tag{2}$$

3. *F1 score*: An F1 score is defined as the harmonic mean of precision and recall. F1 score has become the preferred choice of measuring the performance of machine learning models. This can be attributed to the fact that F1 score gives equal weightage to both precision and recall and it punishes models that lack even in one of them.

$$F1Score = \frac{(2 * P * R)}{P + R} \tag{3}$$

[2] https://semeval.github.io/SemEval2021/tasks.html.

[3] https://hasocfire.github.io/hasoc/2020/index.html.

[4] https://swisstext-and-konvens-2020.org/shared-tasks/.

[5] https://sites.google.com/view/trac2/live?authuser=0.

[6] https://ai.Facebook.com/blog/hateful-memes-challenge-and-dataset/.

Table 11 Deep learning methods for OHS

References	Approach	Language	Dataset	Methodology	Merits	Limitation
Saksesi et al.[84]	RNN	Indonesian	The author collected 1235 Words from Twitter	The author used word2vec embeddings to get the text matrix In order to detect sentences as hate or not, they implement tests with RNN	The author tested the result on a small dataset over which they get PR-91%,RC-90% and, Accuracy-91%	Better results can be achieved if the size of the data increased
Albadi, Kurdi, and Mishra [50]	RNN plus GRU (gated recurrent unit)	Arabic	Collected 600 Arabic tweets from Twitter	The author used MADAMIRA 2.1 to lemmatize the data For a lexicon-based approach features selection, the author used AraHate-PMI (pointwise mutual information), AraHate-Chi, and AraHate-BNS (Bi-Normal Separation	As compared to lexicon-based and SVM and LR-based model, GRU-Based RNN performs the best with an accuracy of 0.79	The author did not compare the earlier state of the arts with the developed model; also, error analysis is missing
Sazany et al. [85]	LSTM, FastText algorithm	Indonesian	713Twitter political posts from Twitter	The author used two types of datasets i.e., riopolitics and okkyabusive In this study, two types of word embeddings have been used, i.e., word2vec and FastText library	By using FastText embeddings with riopoliticsthey achieved the highest results, i.e., 97.39 f1 score	In the proposed research, the model configuration, such as the classifier, number of layers, training batch size, is not analyzed Also, the training and testing dataset size is very small for the desired purpose
Vigna et al. [32]	LSTM SVM	Italian	17567comments collected from Facebook	To increase the system accuracy, the author also used the sentiment polarity lexicon and wordembeddings lexicons	The author found that the binary classifier obtained comparable results as that of sentiment analysis	In the given study, both SVM and LSTM are not able to discriminate between the three classes
Badjatiya et al. [38]	CNN, LSTM, FastText	English	16 K tweets collected on sexism and racism and neither from Twitter publicly available data	In the given study GloVepre-trained embedding and 10-Fold Cross-Validation have been used The author applied Adam for CNN and LSTM and RMS-Prop for FastText as an optimizer	In this study, the author found that CNN is performed better than LSTM, which was better than FastText We also learned that Embeddings learned from deep neural network models when combined with gradient boosted decision trees led to the best accuracy Values	Not applicable

Table 11 (continued)

References	Approach	Language	Dataset	Methodology	Merits	Limitation
Park and Fung [39]	HybridCNN	English	Waseem and Hovy 2016 English dataset (20 k)	The author implements three CNN-based models to classify sexist and racist abusive language i.e., CharCNN, WordCNN and HybridCNN Max pooling is performed after the convolution to capture the feature that is most significant to the output	The proposed model is tested by tenfold cross-validation and gives a 78.3% f score	More precise results can be explored if training the two-step classifiers on separate datasets (larger dataset)
Paetzold et al. [112]	RNN	English and Spanish	HatEval website	The author used UTFPR (minimalistic Recurrent Neural Networks) models in order to process the text The character embeddings are used in the RNN layers The proposed system is based on the compositional RNN	The proposed model is robust, even when the input data are noisy	More reliable ways of re-using pre-trained compositional models can be tested
Sutejo and Lestari [79]	LSTM	Indonesian	Two types of data, text- 2273 and audio- 2469. Collected from different social media websites	The author used two models one Textual model and the second Acoustic model Word n-gram features are employed for the classification For the acoustic model author used low-level descriptor features and, Uni-bi-bow features which give the high F1 score for the textual modeland MFCC_E_D_A features (of an acoustic model)	The author found the textual model gives the best result as that of the acoustic model	CBOW (87.98%) performed better than word n-gram and their combination (the highest achieved 83.91%) In the study, there are several incorrect results due to the bias of some people
Andreou et al.[68]	Ensemble-based classification using CNN DNN RNN	English	Davidson et al	The author employed text features, i.e., surface-level features, linguistic features and sentiment features The author implemented novel ensemble-based classification	Mandola processes the information in real The author finds the accuracy of 0.918 with the novel approach Mandola is the first system that provides a systematic and integrated approach for detecting hate speech	The proposed system is not compatible with any cross-lingual

In multiclass classification there are mainly two methods of calculating F1 score, namely microaveraged F1 score and macroaverage F1 score.

A) *F1 microaveraged* This metric is simply calculated by taking the harmonic mean of microprecision and microrecall. An important feature of this metric is that it assigns equal value to each label, the repercussion of which is the not enough attention is given to minority classes in case of imbalanced datasets. Since Imbalanced datasets are seen in abundance in the domain of hate speech detection, the use of microaveraged F1 score should be minimized.

$$\text{Micro Avereged Precision} = \frac{\sum TP}{\sum TP + \sum FP} \tag{4}$$

$$\text{Micro averaged recall} = \frac{\sum TP}{\sum TP + \sum FN} \tag{5}$$

B) *F1 macroaveraged* This is calculated by simply taking the mean of F1 scores obtained on each class individually. This metric assigns equal value to each class and thus should be the preferred metric in the context hate speech detection where datasets are generally imbalanced and models are expected to be proficient in detecting all classes.

4. *Confusion matrix* It is a performance measurement matrix comparing the actual and predicted observations through the values of False Positives (FP), True Negatives (TN), True Positive (TP) and False Negative (FN) labels (Matrix 1).

Confusion Matrix: $= \begin{vmatrix} TP & FN \\ FN & TN \end{vmatrix}$ (6).

5. *Accuracy* Is the measure which tells how efficiently the classification models produce the results correctly.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

6. *Comprehensiveness*: In XAI, we essentially try to predict the factors which led to a model's decision. To calculate the comprehensiveness, the factors predicted by the XAI model are first removed from the datapoint. In the context of hate speech detection, the equivalent of this is removing the words predicted by the XAI model. Now, this new modified datapoint is then fed into the model. The change in the model's confidence in prediction is noted. A change implies that the factors predicted by the model indeed contributed to the model's decision[132].

7. *Sufficiency*: This metric measures how important the extracted rationales(words or phrases in the context of Hate speech detection) for the model to make a prediction[132].

8. Matthews correlation coefficient (MCC): It tries to find the relation between the true and predicted values. Higher value of the coefficient shows the better results. Whenever the given dataset is highly imbalance in that case it is found that MCC has given best results compared to the accuracy [133]. Its value always lies between -1 and 1. The given formula is shown in Eq. 7.

MCC= $\frac{TP\times TN-FP\times FN}{\sqrt{}(TP+FN)(TP+FN)(TN+FP)(TN+FN)}$ (8).

Both precision and recall are very important and the most used evaluation metrics in traditional machine learning and deep learning classification. We can calculate the accuracy by providing the given values to TN, TP, FP, and FN. By getting the values of precision and recall from equations 1 and 2, we can calculate the F1 score that is used to test the accuracy of the parameter. Some authors also used AUC (area under the curve) to compute the performance of the model. The aforementioned metric evaluation formulas were used by mostly all other authors mentioned in related works to evaluate the performance of their machine learning model.

11 Findings, conclusion and research gaps

The growth of social media has been exponential and people are sharing information, expressing opinions like never before. However, research on hate speech has not been able to keep pace with the multiplicity of social media platforms and their associated problems. Our goal was to cover all the aspects that play an essential role in the field of OHS detection. But our study is limited to computer science background, and we have not considered the culture-specific ways of communication in a different language for detecting OHS. In this survey, we presented a systematic approach that investigates the types of features and classifiers that are most used in OHS detection. From the survey, we found that SVM, Naïve Bayes, Decision Tree, CNN and LSTM are the most used algorithm, and surface-level features are the first choice of the researcher. We learned the concept of hate speech and laws to limit hate speech. Additionally, we presented an application of hate speech. We concluded that very limited studies and papers had been published in the OHS detection from the computer science perspective. We also found that most of the authors used self-generated datasets which are not available online so to find the credibility of these dataset and results achieved with these datasets is also a problem in itself. Finally, we identified some challenges in the field of OHS, the availability of open-source code and the self-generated dataset link, which leads to the lack of comparative studies that can evaluate the existing approaches.

Based on our study, we found several research gaps which can be considered in future work.

- From the study, it has been observed that the existing research covers mostly lexicon (simple keywords)-based features for the hate speech analysis, which restricted the results because the models will not be suitable if whole meaning of the sentence is needed. So, knowledge-based feature, semantic features can be taken into consideration with lexicon-based features. By this, the accuracy of the model can be increased.

- Facebook, Twitter and social media platforms do not have a real-time hate speech detecting system, and the corrective measures are taken only after the expression is posted online. So, the hate speech detecting plugin can be made, which can analyze hate speech in real time.
- We also invested that hate speech does not only come in the form of text but can take the form of audio, video, picture, etc. But in the area of hate speech detection multimodal OHS detection is very less unexplored.
- The research work has been limited to spotting hate in the English language and few pieces of research in Arabic, Indonesian, Italian, Turkish, Swedish, Albanian Language and hate content in the rest of the languages like Hindi goes unfiltered.
- Another limitation that we found is to get the balanced dataset for the OHS. A very limited and less skewed dataset is available online.
- To lubricate the online hate speech detection and analysis, the unlabeled data should be examined for the unsupervised machine learning model as the labeling of data is a very time-consuming task. Therefore, to address hate speech problems, further study of the deep learning model is essential and advantageous.
- In order to furnish research in the field, a multimodal and multilingual dataset should be developed.
- Some cultures may represent anger and hate in linguistically distinct ways, which can be taken into consideration while building the online hate speech model.

Implication of study

This study is highlighting the need for interdisciplinary collaboration between computer science and other fields, such as linguistics, sociology and psychology, to develop more comprehensive approaches to OHS detection that take into account language and cultural differences.

Academics can benefit from this study by understanding the current state of the art in OHS detection, the most commonly used algorithms and surface-level features. This study's limitations can help researchers identify gaps in the field and focus on exploring culture-specific ways of communication for detecting OHS. Practitioners in the field of social media moderation can use this study to inform their strategies for identifying and removing hate speech from social media platforms. This research's findings can help them determine which algorithms and features are most effective in OHS detection. Policymakers and politicians can use this study to inform legislation and regulations around hate speech and social media. The study's presentation of hate speech and the laws that limit it can help policymakers better understand the issue and take informed actions to address it. The challenges identified in the study, such as the lack of open-source code and self-generated datasets, can inform future research and development efforts in OHS detection. Addressing these challenges can lead to the development of better approaches to OHS detection and more reliable datasets, enabling more comparative studies to evaluate existing approaches. In summary, this study on OHS detection in the context of social media can provide valuable insights for various stakeholders and inform future research, policymaking and social media moderation strategies.

Data availability statements Data generated or analyzed during this study are included in this published article.

Declarations

Conflict of interest All the authors of this paper declare that he/she has no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Newman, N., Fletcher, R., Kalogeropoulos, A. et al.: Reuters Institute Digital News Report 2018 (2018)
2. Global social media ranking (2019). https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/
3. Diwhu, G., Ghdwk, W.K.H., Ihpdoh, R.I.D., Vwxghqw, X.: Automated detection of hate speech towards woman on Twitter. In: International Conference On Computer Science And Engineering. pp 7–10 (2018)
4. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Comput Surv (2018). https://doi.org/10.1145/3232676
5. bbc Facebook launches initiative to fight online hate speech. In: bbc. ps://www.bbc.com/news/technology-40371869
6. Organisation International Alert (2016) A plugin to counter hate speech online. https://europeanjournalists.org/mediaagainsthate/hate-checker-plugin-to-counter-hate-speech-online/
7. Salminen, J., Guan, K., Jung, S.G. et al.: A literature review of quantitative persona creation. In: Conf Hum Factors Comput Syst - Proc 1–15 (2020). https://doi.org/10.1145/3313831.3376502
8. Biere, S., Analytics, M.B.: Hate speech detection using natural language processing techniques. VRIJE Univ AMSTERDAM 30 (2018)
9. DePaula, N., Fietkiewicz, K.J., Froehlich, T.J. et al.: Challenges for social media: misinformation, free speech, civic engagement, and data regulations. In: Proceedings of the Association for Information Science and Technology, pp. 665–668 (2018)
10. Varade, R.S., Pathak, V.: Detection of hate speech in hinglish language. In: ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers) (2020)
11. Djuric, N., Zhou, J., Morris, R. et al.: Hate speech detection with comment embeddings. In: Proceedings of the 24th International Conference on World Wide Web. Association for Computing Machinery, New York, NY, USA, pp. 29–30 (2015)

12. Davidson, T., Warmsley, D,. Macy, M., Weber, I.: Automated Hate Speech Detection and the Problem of Offensive Language. (2017). arXiv170304009v1 [csCL] 11 Mar 2017 Autom
13. Miró-Llinares, F., Moneva, A., Esteve, M.: Hate is in the air! But where? Introducing an algorithm to detect hate speech in digital microenvironments. Crime Sci. **7**, 1–12 (2018). https://doi.org/10.1186/s40163-018-0089-1
14. Daniel Burke The four reasons people commit hate crimes. In: CNN. https://edition.cnn.com/2017/06/02/us/who-commits-hate-crimes/index.html
15. Equality and Diversity Forum (2018) Hate Crime: Cause and effect | A research synthesis. Equal Divers Forum
16. ONTARIO PO, GENERAL MOA: CROWN POLICY MANUAL (2005). https://files.ontario.ca/books/crown_prosecution_manual_english_1.pdf
17. Räsänen, P., Hawdon, J., Holkeri, E., et al.: Targets of online hate: examining determinants of victimization among young finnish Facebook users. Violence Vict. **31**, 708–725 (2016)
18. Contributors, W.: Hate crime. In: Wikipedia (2020). https://en.wikipedia.org/wiki/Hate_crime
19. twitter Twitter policy against Hate speech. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
20. facebook Hate speech. https://www.facebook.com/communitystandards/hate_speech
21. Instagram Instagram policy for hate speech. https://help.instagram.com/477434105621119
22. Youtube YouTube hate policy. https://support.google.com/youtube/answer/2801939?hl=en
23. Dr. Amarendra Bhushan Dhiraj: Countries Where Cyber-bullying Was Reported The Most In 2018 (2018)
24. United nations: Universal Declaration of Human Rights (1948)
25. Nations S-G of the U: European Convention on Human Rights, the International Covenant on Civil and Political Rights (1966)
26. Gagliardone, I., Patel, A., Pohjonen, M.: Mapping and analysing hate speech online. In: SSRN Electronic Journal. p 41 (2015)
27. Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Pp. 1–10 (2017)
28. Nastiti, F.E., Prastyanti, R.A., Taruno, R.B., Hariyadi, D.: Social media warfare in Indonesia political campaign: a survey. In: Proceedings - 2018 3rd International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2018. IEEE, pp 49–53 (2019)
29. Kumar, A., Sachdeva, N.: Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis. Multimed. Tools Appl. (2019). https://doi.org/10.1007/s11042-019-7234-z
30. Waqas, A., Salminen, J., Jung, S., et al.: Mapping online hate: a scientometric analysis on research trends and hotspots in research on online hate. PLoS ONE **14**, 1–21 (2019). https://doi.org/10.1371/journal.pone.0222194
31. Waseem, Z., Hovy, D.: Hateful symbols or hateful people ? Predictive features for hate speech detection on Twitter. In: Association for Computational Linguistics Proceedings of NAACL-HLT. pp 88–93 (2016)
32. Vigna, F. Del, C. A., Orletta, F.D. et al.: Hate me , hate me not : Hate speech detection on Facebook. In: In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy. pp 86–95 (2017)
33. Agarwal S, Sureka A (2017) But I did not mean it! - Intent classification of racist posts on tumblr. In: Proceedings - 2016 European Intelligence and Security Informatics Conference, EISIC 2016. IEEE, pp 124–127
34. CodaLab Competition. https://competitions.codalab.org/competitions/19935.
35. Wang, G., Wang, B., Wang, T. et al: Whispers in the dark: Analysis of an anonymous social network. In: Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC. pp 137–149 (2014)
36. Ziai, A.: cohen kappa. In: Medium (2017). https://towardsdatascience.com/inter-rater-agreement-kappas-69cd8b91ff75
37. Gambäck. B,, Sikdar, U.K.: Using Convolutional Neural Networks to Classify Hate-Speech. In: Proceedings ofthe First Workshop on Abusive Language Online. pp 85–90 (2017)
38. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep Learning for Hate Speech Detection in Tweets. In: arXiv:1706.00188v1 [cs.CL]. p 2 (2017)
39. Park, J.H., Fung, P.: One-step and two-step classification for abusive language detection on Twitter. In: Association for Computational Linguistics Proceedings of the First Workshop on Abusive Language Online, pages 41–45, Vancouver, Canada, July 30. pp 41–45 (2017)
40. Waseem, Z.: Are you a racist or am i seeing things ? Annotator influence on hate speech detection on Twitter. In: Proceedings of2016 EMNLP Workshop on Natural Language Processing and Computational Social Science. pp 138–142 (2016)
41. Jha, A: When does a Compliment become Sexist ? Analysis and Classification of Ambivalent Sexism using Twitter Data. In: Proceedings ofthe Second Workshop on Natural Language Processing. pp 7–16 (2017)
42. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language ∗. In: arXiv (2017)
43. Alorainy, W., Burnap, P., Liu, H.A.N., Williams, M.L.: " The Enemy Among Us ": detecting cyber hate speech with threats-based othering language embeddings. ACM Trans. Web 13 (2019)
44. Nobata, C., Tetreault, J.: Abusive language detection in online user content. In: International World Wide Web Conference. Pp. 145–153 (2016)
45. Al, Z., Amr, M.: Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach. Springer Comput. (2019) https://doi.org/10.1007/s00607-019-00745-0
46. Detecting Insults in Social Commentary. https://www.kaggle.com/c/detecting-insults-in-social-commentary
47. MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N.F.O. (2019) Hate speech detection: challenges and solutions. PLoS ONE 14(8): e0221152. https://doi.org/10.1371/journal.pone.0221152. https://sites.google.com/view/trac1/shared-task
48. Timothy Quinn: Hatebase database. (2017). https://www.hatebase.org/
49. Charitidis, P., Doropoulos, S., Vologiannidis, S., et al.: Towards countering hate speech against journalists on social media. Online Soc. Netw. Media **17**, 10 (2020). https://doi.org/10.1016/j.osnem.2020.100071
50. Albadi, N., Kurdi, M., Mishra, S.: Are they our brothers? Analysis and detection of religious hate speech in the Arabic Twittersphere. In: Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018. IEEE, (pp. 69–76) (2018)
51. Al-Hassan, A., Al-Dossari, H.: Detection of hate speech in Arabic tweets using deep learning. Multimed. Syst. (2021). https://doi.org/10.1007/s00530-020-00742-w
52. Ousidhoum, N., Lin, Z., Zhang, H. et al.: Multilingual and multi-aspect hate speech analysis. EMNLP-IJCNLP 2019 - 2019 Conf Empir Methods Nat Lang Process 9th Int Jt Conf Nat Lang Process Proc Conf 4675–4684 (2020). https://doi.org/10.18653/v1/d19-1474
53. Mulki, H., Haddad, H., Bechikh Ali, C., Alshabani, H.: L-HSAB: A Levantine Twitter dataset for hate speech and abusive language, pp. 111–118 (2019). https://doi.org/10.18653/v1/w19-3512

54. Ljubešić, N., Erjavec, T., Fišer, D.: Datasets of Slovene and Croatian moderated news comments, pp. 124–131 (2019). https://doi.org/10.18653/v1/w18-5116
55. Dinakar, K.: Modeling the detection of textual cyberbullying. In: 2011, Association for the Advancement of Artificial Intelligence, pp 11–17 (2011)
56. Greevy, E., Smeaton, A.F.: Classifying racist texts using a support vector machine. In: ACM Proceeding, pp 468–469 (2004)
57. 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
58. Rodriguez, A., Argueta, C., Chen, Y.L.: Automatic detection of hate speech on facebook using sentiment and emotion analysis. In: 1st International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2019. Pp. 169–174 (2019)
59. Hall, L.O., WPKNVCKWB,: snopes.com: Two-striped Telamonia Spider. J Artif Intell Res **2009**, 321–357 (2006). https://doi.org/10.1613/jair.953
60. Raufi, B., Xhaferri, I.: Application of machine learning techniques for hate speech detection in mobile applications. In: 2018 International Conference on Information Technologies, InfoTech 2018 - Proceedings. IEEE, pp 1–4 (2018)
61. Waseem, Z., Thorne, J., Bingel, J.: Bridging the gaps: multi task learning for domain transfer of hate speech detection. In: Online Harassment, Human–Computer Interaction Series, pp 29–55 (2018)
62. Lynn, T., Endo, P.T., Rosati, P., et al.: Data set for automatic detection of online misogynistic speech. Data Br. **26**, 104223 (2019). https://doi.org/10.1016/j.dib.2019.104223
63. Plaza-Del-Arco, F.-M., Molina-González, M.D., Ureña-López, L.A., Martín-Valdivia, M.T.: Detecting Misogyny and Xenophobia in Spanish Tweets using language technologies. ACM Trans. Internet Technol. **20**, 1–19 (2020). https://doi.org/10.1145/3369869
64. Pelzer, B., Kaati, L., Akrami, N.: Directed digital hate. In: 2018 IEEE International Conference on Intelligence and Security Informatics, ISI 2018, pp. 205–210 (2018)
65. Martins, R., Gomes, M., Almeida, J.J. et al.: Hate speech classification in social media using emotional analysis. In: Proceedings - 2018 Brazilian Conference on Intelligent Systems, BRACIS 2018, pp. 61–66 (2018)
66. Basak, R., Sural, S., Ganguly, N., Ghosh, S.K.: Online public shaming on Twitter: detection, analysis, and mitigation. IEEE Trans. Comput. Soc. Syst. **6**, 208–220 (2019). https://doi.org/10.1109/TCSS.2019.2895734
67. Sreelakshmi, K., Premjith, B., Soman, K.P.: Detection of hate speech text in Hindi-English Code-mixed Data. Procedia Comput. Sci. **171**, 737–744 (2020). https://doi.org/10.1016/j.procs.2020.04.080
68. Andreou, A., Orphanou, K., Pallis, G.: MANDOLA : A Big-Data Processing and Visualization. ACM Trans. Internet Technol. 20 (2020)
69. Zimbra, D., Abbasi, A., Zeng, D., Chen, H.: The state-of-the-art in Twitter sentiment analysis. ACM Trans. Manag. Inf. Syst. **9**, 1–29 (2018). https://doi.org/10.1145/3185045
70. Mariconti, E., Suarez-Tangil, G., Blackburn, J., et al.: "You know what to do": proactive detection of YouTube videos targeted by coordinated hate attacks. Proc ACM Hum.-Comput. Interact (2019). https://doi.org/10.1145/3359309
71. Gitari ND, Zuping Z, Damien H, Long J (2015) A Lexicon-based approach for hate speech detection a Lexicon-based approach for hate speech detection. Int. J. Multimed. Ubiquitous Eng. https://doi.org/10.14257/ijmue.2015.10.4.21
72. Lima, L., Reis, J.C.S., Melo, P. et al.: Inside the right-leaning echo chambers: characterizing gab, an unmoderated social system. In: Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ASONAM 2018. pp 515–522 (2018)
73. 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, pp. 13825–13835 (2018)
74. Ruwandika, N.D.T., Weerasinghe, A.R.: Identification of hate speech in social media. In: 2018 International Conference on Advances in ICT for Emerging Regions (ICTer) : Identification. IEEE, pp. 273–278 (2018)
75. Alorainy W, Burnap P, Liu H, et al.: Suspended accounts : a source of tweets with disgust and anger emotions for augmenting hate speech data sample. In: Proceeding of the 2018 International Conference on Machine Learning and Cybernetics. IEEE (2018)
76. Setyadi, N.A., Nasrun, M., Setianingsih, C.: Text analysis for hate speech detection using backpropagation neural network. In: The 2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC). IEEE, pp 159–165 (2018)
77. Alfina, I., Mulia, R., Fanany, M.I., Ekanata, Y.: Hate speech detection in the Indonesian language: A dataset and preliminary study. In: 2017 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2017. pp 233–237 (2018)
78. Sharma, H.K., Singh, T.P., Kshitiz, K., et al.: Detecting hate speech and insults on social commentary using NLP and machine learning. Int. J. Eng. Technol. Sci. Res. **4**, 279–285 (2017)
79. Sutejo, T.L., Lestari, D.P.: Indonesia hate speech detection using deep learning. In: International Conference on Asian Language Processing. IEEE, pp 39–43 (2018)
80. Lekea, I.K.: Detecting hate speech within the terrorist argument : a greek case. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, pp 1084–1091 (2018)
81. Liu, H., Burnap, P., Alorainy, W., Williams, M.L.: A fuzzy approach to text classification with two-stage training for ambiguous instances. IEEE Trans. Comput. Soc. Syst. **6**, 227–240 (2019). https://doi.org/10.1109/TCSS.2019.2892037
82. Wang, J., Zhou, W., Li, J., et al.: An online sockpuppet detection method based on subgraph similarity matching. In: Proceedings - 16th IEEE International Symposium on Parallel and Distributed Processing with Applications, 17th IEEE International Conference on Ubiquitous Computing and Communications, 8th IEEE International Conference on Big Data and Cloud Computing, 11t. IEEE, pp. 391–398 (2019)
83. Wu, K., Yang, S., Zhu, K.Q.: False rumors detection on Sina Weibo by propagation structures. In: Proc - Int Conf Data Eng 2015-May:651–662 (2015). https://doi.org/10.1109/ICDE.2015.7113322
84. Saksesi, A.S., Nasrun, M., Setianingsih, C.: Analysis text of hate speech detection using recurrent neural network. In: The 2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC) Analysis. IEEE, pp. 242–248 (2018)
85. Sazany, E.: Deep learning-based implementation of hate speech identification on texts in Indonesian : Preliminary Study. In: 2018 International Conference on Applied Information Technology and Innovation (ICAITI) Deep. IEEE, pp 114–117 (2018)
86. Son, L.H., Kumar, A., Sangwan, S.R., et al.: Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network. IEEE Access **7**, 23319–23328 (2019). https://doi.org/10.1109/ACCESS.2019.2899260

87. Salminen, J., Hopf, M., Chowdhury, S.A., et al.: Developing an online hate classifier for multiple social media platforms. Human-centric Comput. Inf. Sci. **10**, 1–34 (2020). https://doi.org/10.1186/s13673-019-0205-6
88. Coste, R.L. (2000) Fighting speech with speech: David Duke, the anti-defamation league, online bookstores, and hate filters. In: Proceedings of the Hawaii International Conference on System Sciences. p 72
89. Gelber, K.: Terrorist-extremist speech and hate speech: understanding the similarities and differences. Ethical Theory Moral Pract. **22**, 607–622 (2019). https://doi.org/10.1007/s10677-019-10013-x
90. Zhang, Z.: Hate speech detection: a solved problem ? The challenging case of long tail on Twitter. Semant WEB IOS Press **1**, 1–5 (2018)
91. Hara, F.: Adding emotional factors to synthesized voices. In: Robot and Human Communication - Proceedings of the IEEE International Workshop, Pp. 344–351 (1997)
92. Fatahillah, N.R., Suryati, P., Haryawan, C.: Implementation of Naive Bayes classifier algorithm on social media (Twitter) to the teaching of Indonesian hate speech. In: Proceedings—2017 International Conference on Sustainable Information Engineering and Technology, SIET 2017, pp. 128–131 (2018)
93. Ahmad Niam, I.M., Irawan, B., Setianingsih, C., Putra, B.P.: Hate speech detection using latent semantic analysis (LSA) method based on image. In: Proceedings - 2018 International Conference on Control, Electronics, Renewable Energy and Communications, ICCEREC 2018. IEEE, pp. 166–171 (2019)
94. Gitari, N.D., Zuping, Z., Damien, H., Long, J.: A lexicon-based approach for hate speech detection. Int. J. Multimed. Ubiquitous Eng. **10**, 215–230 (2015)
95. Chen, Y., Zhou, Y., Zhu, S., Xu, H.: Detecting offensive language in social media to protect adolescent online safety. In: Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012. IEEE, pp. 71–80 (2012)
96. Pitsilis, G.K., Ramampiaro, H., Langseth, H.: Effective hate-speech detection in Twitter data using recurrent neural networks. Appl. Intell., Pp. 4730–4742 (2018)
97. Pitsilis, G.K., Ramampiaro, H., Langseth, H.: Detecting offensive language in Tweets using deep learning (2018). arXiv:180104433v1 1–17. https://doi.org/10.1007/s10489-018-1242-y
98. Warner, W., Hirschberg, J.: Detecting hate speech on the World Wide Web. In: Association for Computational Linguistics Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012), pp. 19–26 (2012)
99. Dinakar, K., Jones, B., Havasi, C., Lieberman, H.: Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Trans. Interact. Intell. Syst. **2**, 30 (2012). https://doi.org/10.1145/2362394.2362400
100. 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**, 223–242 (2015). https://doi.org/10.1002/poi3.85
101. Garc, A: Hate speech dataset from a white supremacy forum. In: Proceedings of the Second Workshop on Abusive Language Online, pp. 11–20 (2018)
102. Ombui, E., Karani, M., Muchemi, L.: Annotation framework for hate speech identification in Tweets : Case Study of Tweets During Kenyan Elections. In: 2019 IST-Africa Week Conference (IST-Africa). IST-Africa Institute and Authors, pp. 1–9 (2019)
103. Hosseinmardi, H., Mattson, S.A., Rafiq, R.I. et al.: Detection of cyberbullying incidents on the Instagram Social Network. In: arXiv:1503.03909v1 [cs.SI] 12 Mar 2015 Abstract (2015)
104. Raufi, B., Xhaferri, I.: Application of machine learning techniques for hate speech detection in mobile applications. In: 2018 International Conference on Information Technologies (InfoTech-2018), IEEE Conference Rec. No. 46116 20–21 September 2018, St. St. Constantine and Elena, Bulgaria. IEEE (2018)
105. Warner, W., Hirschberg, J.: Detecting hate speech on the World Wide Web. In: 19 Proceedings of the 2012 Workshop on Language in Social Media (LSM. pp 19–26) (2012)
106. Wang, G., Wang, B., Wang, T. et al.: Whispers in the dark : analysis of an anonymous social network categories and subject descriptors. ACM 13 (2014)
107. Mathew, B., Saha, P., Yimam, S.M. et al.: HateXplain: a benchmark dataset for explainable hate speech detection. In: ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers). p 12 (2020)
108. Kiilu, K.K., Okeyo, G., Rimiru, R., Ogada, K.: Using Naïve Bayes Algorithm in detection of Hate Tweets. Int. J. Sci. Res. Publ. 8:99–107. https://doi.org/10.29322/ijsrp.8.3.2018.p7517 (2018)
109. Sanchez, H.: Twitter Bullying Detection, pp. 1–7 (2016). In: https://www.researchgate.net/publication/267823748
110. Gröndahl, T., Pajola, L., Juuti, M. et al.: All you need is "love": Evading hate speech detection. In: Proceedings of the ACM Conference on Computer and Communications Security. pp 2–12 (2018)s
111. Correa, D., Silva, L.A., Mondal, M., et al.: The many shades of anonymity : characterizing anonymous social media content. Assoc Adv. Artif. Intell. 10 (2015)
112. Paetzold, G.H., Malmasi, S., Zampieri, M.: UTFPR at SemEval-2019 Task 5: Hate Speech Identification with Recurrent Neural Networks. In: arXiv:1904.07839v1. p 5 (2019)
113. Miro-Llinares, F., Rodriguez-Sala, J.J.: Cyber hate speech on twitter: analyzing disruptive events from social media to build a violent communication and hate speech taxonomy. Int. J. Design Nat. Ecodyn. pp 406–415 (2016)
114. Rizoiu, M.-A., Wang, T., Ferraro, G., Suominen, H.: Transfer learning for hate speech detection in social media. arXiv:190603829v1 (2019)
115. Pitsilis, G.K., Ramampiaro, H., Langseth, H.: Effective hate-speech detection in Twitter data using recurrent neural networks. Appl. Intell. **48**, 4730–4742 (2018). https://doi.org/10.1007/s10489-018-1242-y
116. Varade, R.S., Pathak, V.B.: Detection of hate speech in hinglish language. Adv. Intell. Syst. Comput. **1101**, 265–276 (2020). https://doi.org/10.1007/978-981-15-1884-3_25
117. Modha, S., Majumder, P., Mandl, T., Mandalia, C.: For surveillance detecting and visualizing hate speech in social media: a cyber watchdog for surveillance. Expert Syst. Appl. (2020). https://doi.org/10.1016/j.eswa.2020.113725
118. Maxime: What is a Transformer?No Title. In: Medium (2019). https://medium.com/inside-machine-learning/what-is-a-transformer-d07dd1fbec04
119. Horev R BERT Explained: State of the art language model for NLP Title. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270
120. Mozafari, M., Farahbakhsh, R., Crespi, N.: A BERT-based transfer learning approach for hate speech detection in online social media. Stud. Comput. Intell. 881 SCI:928–940 (2020). https://doi.org/10.1007/978-3-030-36687-2_77
121. Mutanga, R.T., Naicker, N., Olugbara, O.O. (2020) Hate speech detection in twitter using transformer methods. Int. J. Adv. Comput. Sci. Appl.; 11, 614–620 . https://doi.org/10.14569/IJACSA.2020.0110972
122. 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 (2021)

123. Pandey, P.: Deep generative models. In: medium. https://towardsdatascience.com/deep-generative-models-25ab2821afd3
124. Wullach, T., Adler, A., Minkov, E.M.: Towards hate speech detection at large via deep generative modeling. IEEE Internet Comput. (2020). https://doi.org/10.1109/MIC.2020.3033161
125. Dugas, D., Nieto, J., Siegwart, R., Chung, J.J.: NavRep : Unsupervised representations for reinforcement learning of robot navigation in dynamic human environments (2021)
126. Behzadi, M., Harris, I.G., Derakhshan, A.: Rapid cyber-bullying detection method using compact BERT models. In: Proc - 2021 IEEE 15th Int Conf Semant Comput ICSC 2021 199–202. (2021) https://doi.org/10.1109/ICSC50631.2021.00042
127. Araque, O., Iglesias, C.A.: An ensemble method for radicalization and hate speech detection online empowered by sentic computing. Cognit. Comput. (2021). https://doi.org/10.1007/s12559-021-09845-6
128. 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
129. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: 26th International World Wide Web Conference 2017, WWW 2017 Companion (2019)
130. Mossie, Z., Wang, J.H.: Vulnerable community identification using hate speech detection on social media. Inf. Process Manag. **57**, 102087 (2020). https://doi.org/10.1016/j.ipm.2019.102087
131. Magu, R., Joshi, K., Luo, J.: Detecting the hate code on social media. In: Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017. pp 608–611 (2017)
132. Qian, J., Bethke, A., Liu, Y., et al.: A benchmark dataset for learning to intervene in online hate speech. In: EMNLP-IJCNLP 2019 - 2019 Conf Empir Methods Nat Lang Process 9th Int Jt Conf Nat Lang Process Proc Conf 4755–4764 (2020). https://doi.org/10.18653/v1/d19-1482
133. Chicco, D., Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genom. **21**, 1–13 (2020). https://doi.org/10.1186/s12864-019-6413-7
134. Lee, K., Ram, S.: PERSONA: Personality-based deep learning for detecting hate speech. In: International Conference on Information Systems, ICIS 2020 - Making Digital Inclusive: Blending the Local and the Global. Association for Information Systems (2021)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.