Hate Speech Detection in Social Media (Twitter) Using Neural Network

Ara Zozan Miran and Hazha Saeed Yahia*

Department of Information Technology, Lebanese French University, Kurdistan Region, Iraq
E-mail: ara.zozan@lfu.edu.krd; hazha.yahya@lfu.edu.krd
**Corresponding Author*

Received 27 March 2022; Accepted 03 October 2022;
Publication 15 February 2023

Abstract

Hate speech recently became a real threat in social media, and almost all social media users are intended to in different ways. Hate speech is not limited to a group or society. It affects many people and can be classified as abusive, offensive, sexism, racism, political affiliation, religious hate, nationality, skin color, disability, gender-based, ethnicity, sexual orientation, immigrants, and others. Many researchers and authorities attempt to discover new procedures to sense hate speech in social media, especially on Facebook and Twitter, and many methods, models, and algorithms are used for this purpose. One of the most valuable models for detecting hate speech is Convolutional Neural Network (CNN). This review aims to assort academic studies on hate speech detection in Twitter using CNN-based models summarize the results of each model to expand the understanding of the recent circumstances of hate speech detection in Twitter. For this purpose, we implemented a broad, automated search using Boolean and Snowballing searching methods to find academic works in this area. Studies and papers have been distinguished, and the following information was obtained and aggregated from each article:

Journal of Mobile Multimedia, Vol. 19_3, 765–798.
doi: 10.13052/jmm1550-4646.1936

authors, publication's year, the journal name or the conference name, proposed model/method, the aim of the study, the outcome, and the quality of each study. According to the findings, the CNN and CNN-based models are standard models for hate speech detection. Besides, the findings show that other new models have a great compact on hate speech detection, and there is good progress in this field. However, the problems that still exist with hate speech detection models mainly are; most of the models cannot detect hate speech automatically. The methods are not suitable with all the languages, and they are working only with one language; most are best suited with the English language, and when they are used with datasets with other languages. Besides, the models are suffering from confusion in speech classification. Finally, most models are not considering a user-to-user speech in social media.

Keywords: Hate speech, Twitter, toxic, cyberbullying, convolutional neural network.

1 Introduction

Online social media enable spread humanities to be associated. However, one disadvantage of these social media is the ability for hateful and harmful content, or cyberhate, to be published and propagated [1, 2]. Hate speech refers to substances that aid violence and are intended to incite hatred towards persons or groups based on specified characteristics, for example, religion, disability, gender, age, veteran status, sexual orientation/gender identity. Nowadays, increasing the number of social media platforms has caused matters excess [3]. Unfortunately, not all substances are relevant; some might harm people, which causes terrible reflection once they use the media to propagate hate. Numerous studies focus on hate speech detection to show the visibility of the harm [4]. Because of the broad ascent in client-created content from online media, disdain discourse has additionally extended quickly [5, 6]. Hate speech, focusing on a specific individual or gathering, can cause personal injury, cyberbullying, fear in the public society, and segregation [7].

A unique multi-layer neural network for spatial information is Convolutional Neural Network (CNN or ConvNet), among various profound learning architectures [8]. The visual impression of living creatures inspires the architecture of CNN. However, it became famous after the record-breaking execution of AlexNet in 2012. It was started in 1980. Later in 2012, CNN

got the speed to take over various fields of computer vision, natural language processing, and many more [9, 10].

This paper aims to review previous works on the same domain and analyze each piece's used models and parameters to sense hate speech in social media, specifically (Twitter) platforms. Most of the reviewed papers enhanced the path planning according to different methods such as map-based methods, potential field methods, mathematical planning methods, and evolutionary-based methods [11]. From various accessible databases, a collection of (565) research papers has been collected with the expectation that the convened paper will provide a new vision of analyzing previous papers using one of the systematic review techniques such as Kitchenham for detecting hate speech in social media (Twitter) using neural networks. Many inquiries have been addressed to accomplish the review results; a few instances of the inquiries that have been responded to and dissected in the third section are kinds of hate speech utilized in comments by Twitter users, the kinds of hate speech data sets, the algorithms, models, and methods utilized for detecting hate speech on Twitter, the country's that has more works for detecting hate speech, deterge hate speech on Twitter, the languages of the datasets of hate speech [12]. This paper's organization was presented as follows: Section 2 presents the methodology used for this systematic review; Section 3 shows the discussions, results, and analysis; Section 4 shows the gaps in the literature related to this subject; and finally, the conclusion section.

2 Methodology

This review is primarily based on the Kitchenham method and Denyer and Tranfiel. This review has designed a specific protocol; the protocol consists of six phases: research question, search strategy, inclusion/exclusion criteria, selection criteria, data extraction, and data synthesis. In the subsequent section, each stage has been discussed in detail.

2.1 Research Question

We start the first phase of the review with the central question this research is trying to answer. The main question is "How to detect hate speech in Twitter using CNN?". For understanding the recent research about the topic, it is difficult to answer all the necessary questions with a broad question.

Accordingly, the question has been divided into many sub-questions, as the following:

QR1: Type of hate speech used by Twitter users in their comments?
QR2: What are the types of hate speech data sets?
QR3: What are the models that have been used for detecting hate speech?
QR4: Which country faces more hate speech?
QR5: What are the number of used datasets and their sizes?
QR6: What are the languages of the datasets of hate speech?

Each of the questions above has been addressed in the analysis section.

2.2 Search Strategy

Different databases and keywords were used to answer the research question and find the gaps in the search strategy phase. For this purpose, other popular online search databases were used, such as IEEE explorer,[1] Elsevier,[2] Google Scholar,[3] Springer,[4] ACM digital library,[5] and Hindawi.[6] For searching in these databases, we used two methods of searching; the Boolean search and Snowballing search methods. In the Boolean search, we converted the main questions and sub-questions to different keywords and used the keywords in the databases to retrieve related papers or research. The Boolean search method depends on (AND, OR, and NOT) [3]; the keywords have been connected through these Boolean words. For example, "detection of hate speech" OR "hate speech detection" OR "detecting hate speech")("hate speech detection" AND "Twitter" AND "CNN") ("hate speech detection" OR "Twitter" OR "CNN"). The following method that has been used for searching is Snowballing search method, for finding the highly cited works [4], beginning with backward snowballing, which entails using the reference list of each retrieved article and extracting or recovering the research that satisfies this review's inclusion and exclusion criteria, while omitting the papers studied previously. Hence, using Google Scholar to find additional works cited the reviewed research during the forward snowballing. Then, each research that cites the

[1] https://ieeexplore.ieee.org/
[2] https://www.elsevier.com/
[3] https://scholar.google.com.tw/
[4] https://www.springer.com/
[5] https://dl.acm.org/
[6] https://www.hindawi.com/

paper is scrutinized. Following the selection of citing research, the article passes through inclusion and exclusion for retrieval.

Different search keywords have been used for each search engine due to the databases' differences and capabilities. In the Boolean search process, other phrases and keywords related to the subject are used and categorized into three main categories; The first category includes keywords "detecting hate speech on Twitter using CNN" in general. Then the search keywords narrowed to "detecting hate speech on Twitter," "hate speech detection using CNN," "hate speech detection in Twitter using neural networks," "hate speech detection in Twitter using CNN," and "how to deterge hate speech on Twitter." The second part of the search has been done using forward and backward Snowballing. For the forward, we depend on the citations of the papers in Google Scholar. While, for the backward Snowballing, we rely on the reference lists of the research. The findings reached the two search categories (20 different research and conference papers).

Inclusion/Exclusion criteria

Inclusion criteria

1. Include articles from years (2010–2021)
2. Only papers written in English
3. Only using CNN model for hate speech detection and models that are based on CNN
4. Only hate speech in Twitter social media
5. Only papers published in journals or conferences

Exclusion criteria

1. Bias papers
2. Books, book chapters...
3. Using other models for hate speech detection rather than the CNN model
4. Comparison articles that are compared between different models of hate speech detection

2.3 Selection Criteria

The fourth phase involves manually evaluating the shortlisted researches to demonstrate their relevance to the aim and objective of this review. In the previous step, the article exclusion was based on the titles only, but it is based on titles and abstracts for the selected research in the last phase. The abstracts of each article are meticulously examined and analyzed to

Table 1 No. retrieved research after implementing the inclusion and exclusion criteria

Search Engines	No. of Regained Articles in Search Engines	No. of Regained Articles After Duplication	No. of Regained Articles After Implementing the Inclusion and Exclusion Criteria	No. of Regained Articles Based on Titles	No. of Regained Related Articles Based on Title and Abstract
IEEExplore	113	55	37	25	4
Elsevier	125	100	94	28	3
Springer	86	79	68	23	2
Hindawi	37	21	15	12	4
ACM	85	41	34	10	2
Google Scholars	119	108	96	15	5
Total	**565**	**404**	**344**	**113**	**20**

decide its relevance. The report is rejected if the abstracts or written text does not contain the relevant keywords. Only papers that perfectly suit the review's scope are kept following the analysis procedure. This procedure resulted in a condensed list of 20 relevant research publications cited in the current investigation. Table 1 shows the number of retrieved articles and the number of related retrieved researches after implementing the inclusion and exclusion criteria. Mainly the papers are retrieved from IEEExplore, followed by Elsevier, Springer, Hindawi, ACM, and Google scholar. Besides, the following diagram, Figure 1, shows the search process in databases for relevant research followed by the inclusion and exclusion process and indicates the number of accepted or rejected researches.

2.4 Data Extraction

In the fourth phase, the information collection addresses the review question and study criteria. The data extraction for this review is defined as a set of specific values extracted from each article. These values include; the aim and objective of the study, proposed methods and models, and whether the used model is new or based on other existing models? Type, number, and size of the datasets. Then the set of parameters that the models are based on and summarize each study's results. Data extraction must include answers to the review question. Data collection forms have been created to provide standard information, including the article's title and author/s, date of publishing with the journal's name, country, proposed algorithm, classification, and many other fields. For more information, see Appendix A, the data extraction form.

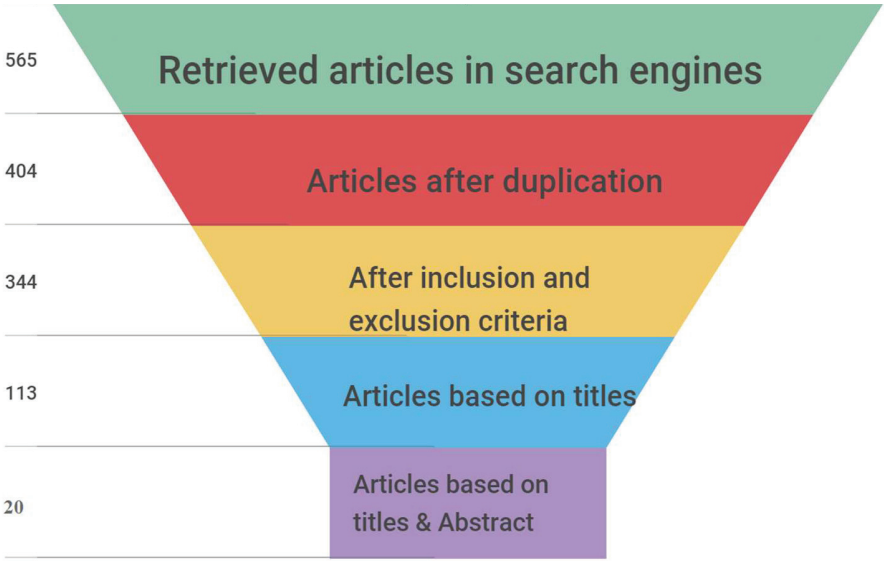


Figure 1 Article retrieval process.

2.5 Data Synthesis

In this phase, compiling and summarizing the results of included studies have been done. Usually, the data synthesis can be descriptive or quantitative, so we based on a descriptive summary for the results in this systematic review. The extracted information is organized in two tables for showing the experimental details used and the advantages and limitations of each study.

The experimental table includes the following details:

1. Used model or method
2. Dataset
3. Hate speech vocabulary used in the dataset
4. Classification of the dataset

The results table includes the following:

1. F1 score
2. Recall
3. Precision
4. Accuracy
5. Advantages of the used model or method
6. Limitation of the study

3 Results and Analysis

3.1 Results

Among the selection of (565) papers, (20) papers have been finalized after the Inclusion and Exclusion Criteria. Based on this classification, the retrieved researches were categorized, and the common factor between all of the studies is CNN. Moreover, each article describes and proposes a new model of CNN for detecting hate speech in different languages. Some of the used models are CNN (Convolutional neural network), LTSM (Long-short term memory network), Deep hate, and GRU (Gated Recurrent Units). Figure 2 shows the retrieved research article distribution based on the publishing years, from 2018 to 2022. The graph shows a remarkable increment since 2018 in detecting hate speech using the neural network CNN. As shown, most of the research papers have been published in 2020, and in 2019 and 2021, there is an equivalent number of publishing research articles. In the last few years, there has been a massive rise in using hate speech in social media, and many

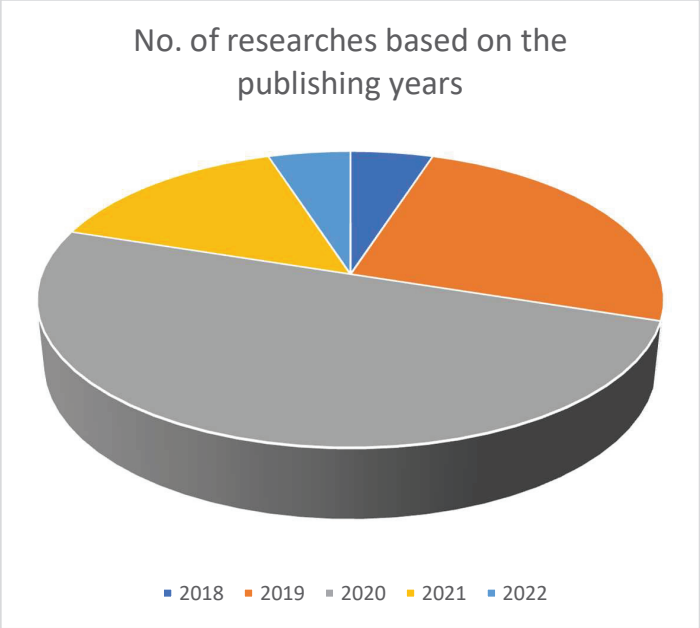


Figure 2 Spreading of research based on the publishing year.

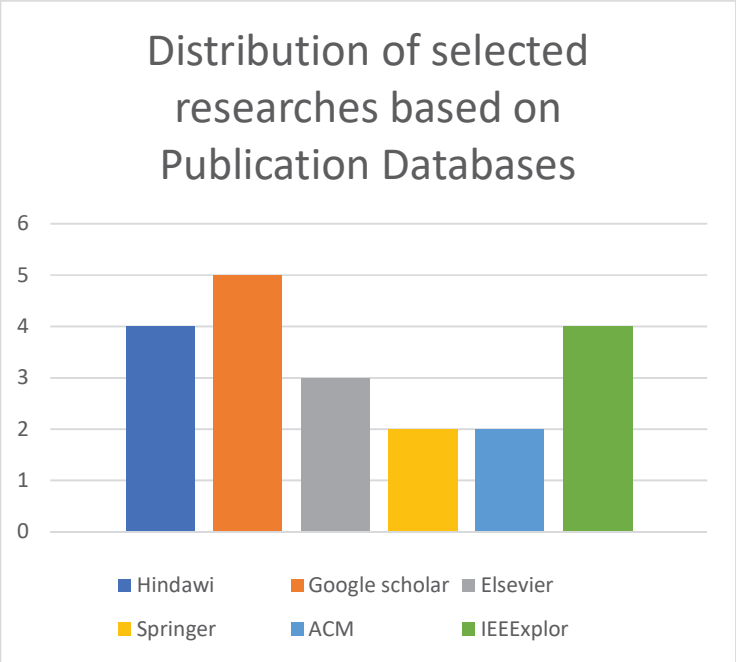


Figure 3 Number of selected researches based on publication databases.

models have been used to detect and demonstrate the harm of using social media as a cyberbullying in platforms.

Many types of research have been published in Google scholar (%25), Hindawi and the IEEE Explore database almost have the same rate with (%20) for each. ACM and Springer publisher have the same rate with (%10). Finally, Elsevier has (%15) as shown in Figure 3.

Figures 4 show the distribution of research based on countries. It is demonstrated that the UK has the most significant number of publications after that India, then China and Jorden come with equivalent numbers. Finally, the rest of the countries have the same number of publications.

Figure 5 shows the distribution of top countries that use CNN models for detecting hate speech on social media. As shown, the UK has the most significant number of publications and is shown in dark green color; then the other countries come with different shades of green color.

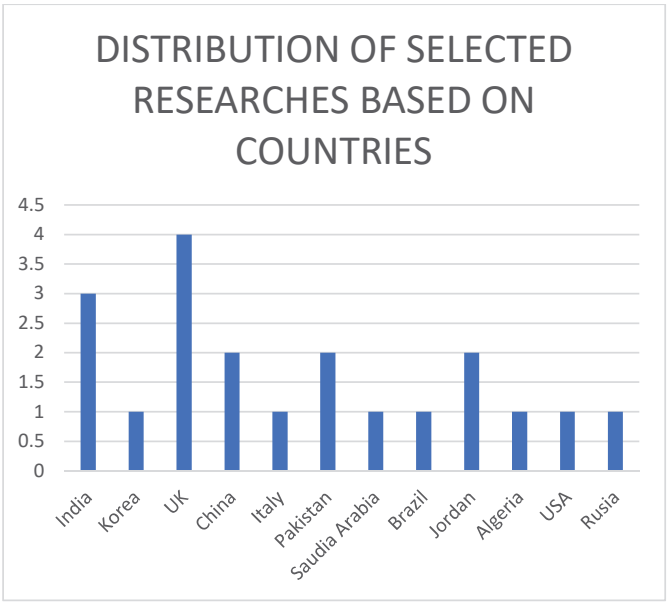


Figure 4 Distribution of studies based on countries.

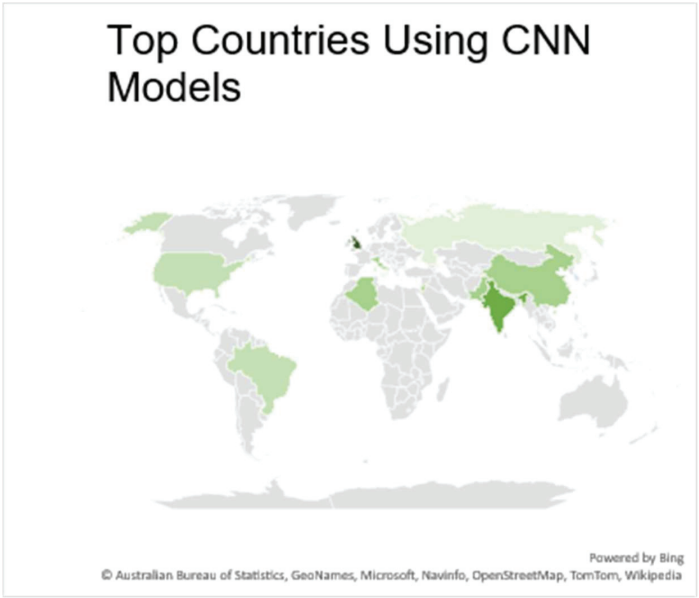


Figure 5 Top countries using CNN models on world map.

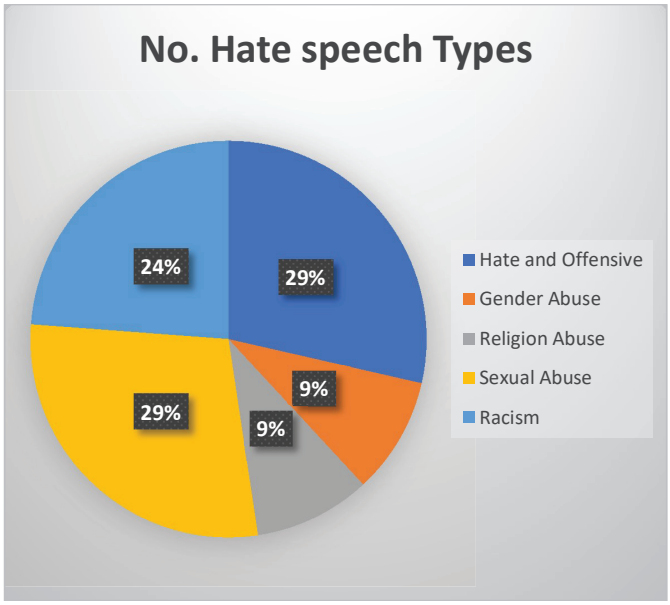


Figure 6 Number of hate speech types distributed.

3.2 Analysis

The information removed and gathered has already been totaled for addressing the examination questions. In the below sections, each examination question has been responded to based on the consequences of the information extraction process.

RQ1: What types of hate speech are used in comments by Twitter users?

Many hate speeches have been detected, but the common ones are (sexual abuse, religious violence, national extremism, gender abuse, bias, radicalism) [15]. Figure 6 shows the distribution of hate speech phrases used in comments on Twitter. As demonstrated, Sexual abuse, Hate, and offensive have the most significant number compared to other used words.

RQ2: What are the types of hate speech data sets?

Regarding this record, the term hate speech is seen as any correspondence in talk, creating, or lead of those assaults or uses shriveling or unreasonable language concerning an individual or a gathering in light of their personality [16]. Hate speech can be detected on social media in different types (Text, Images, or Videos). The research papers shown below in Table 2 demonstrate the Text type of hate speech detection.

Table 2 List of hate speech types data sets

Malik, Pranav, Aggrawal, Aditi, and Vishwakarma, Dinesh K. (2021)	Ziqi Zhang, David Robinson, and Jonathan Tepper (2018)
Modha, Sandip, Majumder, Prasenjit, Mandl, Thomas, and Mandalia, Chintak (2020)	Faris, Hossam, Aljarah, Ibrahim, Habib, Maria, and Castillo, Pedro A (2020)
Jihyung Moon, Won Ik Cho, Junbum Lee (2020)	Hammad Rizwan, Muhammad Haroon Shakeel, Asim Karim (2020)
Ziqi Zhang and Lei Luo (2019)	Abdullah Aref, Rana Husni Al Mahmoud, Khaled Taha, and Mahmoud Al-Sharif (2020)
Neeraj Vashistha, and Arkaitz Zubiaga (2021)	Ibrahim Abu-Farha and Walid Magdy (2020)
Yanling Zhou, Yanyan Yang, Han Liu, Xiufeng Liu, and Nick Savage (2020)	Amrutha B R, Bindu K R (2019)
Patricia Chiril, Endang Wahyu Pamungkas, Farah Benamara, Voronique Moriceau, and Viviana Patti(2022)	Elouali, Aya Elberrichi, Zakaria, Elouali, Nadia (2019)
Muhammad Sajjad, Fatima Zulifqar, Muhammad Usman Ghani Khan, and Muhammad Azeem (2019)	Matthew Beatty (2020)
Alshalan, Raghad, and Al-Khalifa, Hend (2020)	Rui Cao, Roy Ka-Wei Lee, Tuan-Anh Hoang (2020)
Samuel C. Silva, Adriane B. S. Serapião, and Ivandré Paraboni1 (2019)	Ekaterina Pronoza, Polina Panicheva, Olessia Koltsova, Paolo Rosso (2021)

RQ3: What are the models that have been used for detecting hate speech?

The spread of hate speech in social media recently led to many algorithms for detecting them. The standard model that has been used is CNN; the result of the selected research papers shows that there is a mix between the models such as CNN and LTSM (Long Short-Term Memory), CNN and Deep learning, CNN and BERT model, CNN and GRU (Gated recurrent units). Figure 7 shows that the most used models are CNN and LSTM for detecting hate speech in social media platforms such as Twitter.

RQ4: Which country faces more hate speech?

As social media plays a significant role in our daily life and hate speech became a part of everyday speech on social media platforms, this growth that expands to many different countries such as the UK, China, India. . . . etc. to control this expansion, many researchers proposed many different kinds of

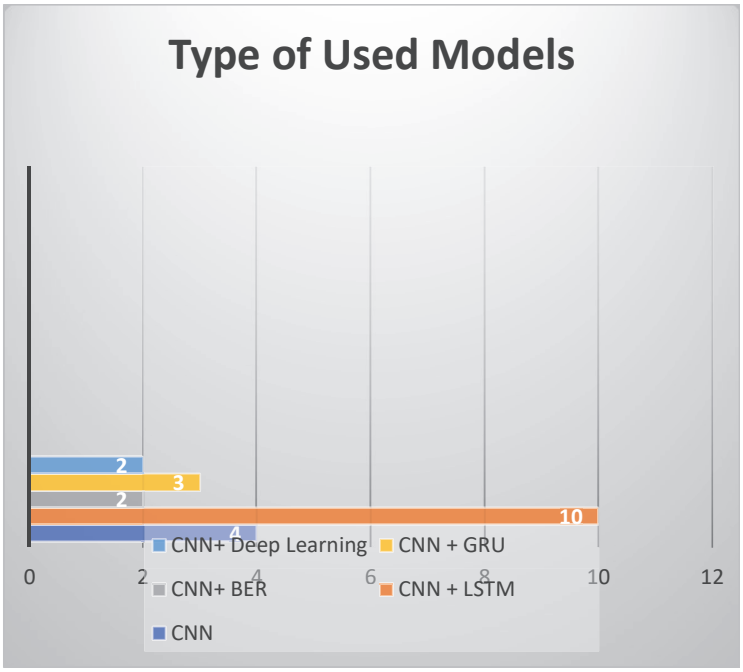


Figure 7 No. used models.

algorithms and modules to detect hate speech [17], as Figure 8 shows the top three countries that have worked more for detecting hate speech.

RQ5: What are the number of used datasets and their sizes?

Extracting hate speech in social media will result in massive data ; in this case, each selected paper used different numbers of datasets with different sizes. Figure 9 shows the number of used datasets in the selected papers, while Figure 10 shows the dimensions of the used datasets.

RQ6: What are the languages of the datasets of hate speech?

Over the last decade, a considerable body of work and extensive work has been done in the space of program of instinctive distinguishing proof and arrangement of disdain discourse and related displays classes such as racism, sexism, hate speech, etc [18]. The phenomenon of hate speech increases widely over the countries almost all the countries face that problem, below Figure 11 shows the number of languages of the datasets for hate speech. As demonstrated, the most extensive dataset is for the English language.

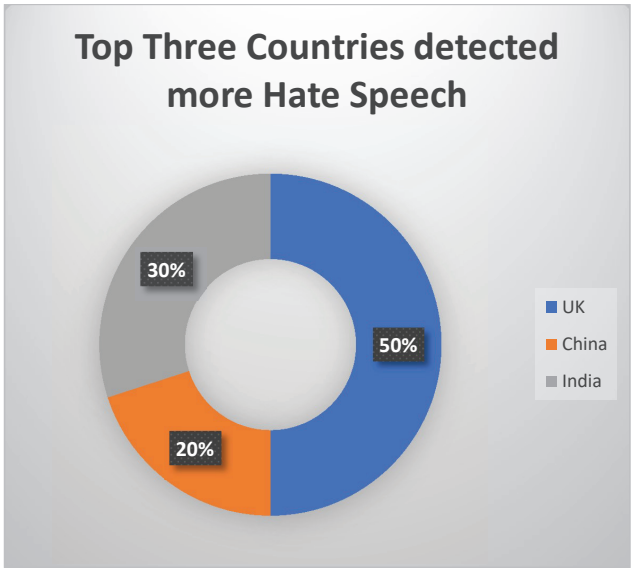


Figure 8 Top three countries detected more hate speech.

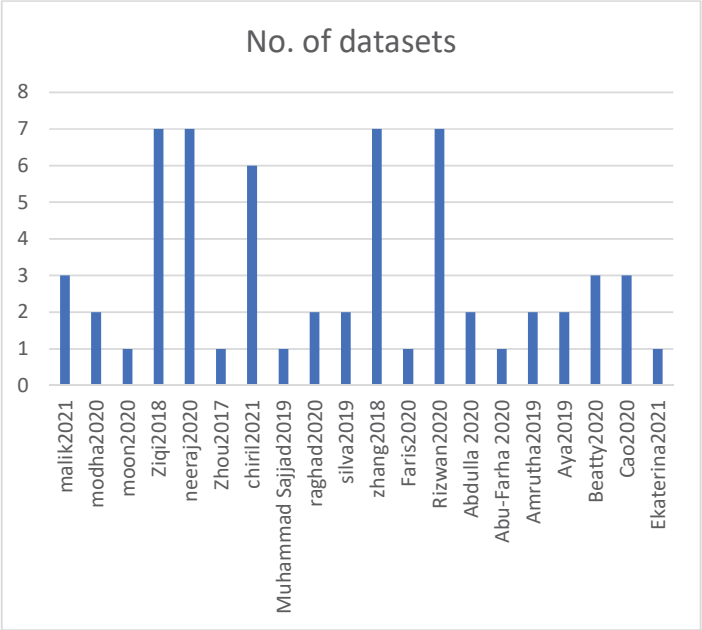


Figure 9 No. of used datasets.

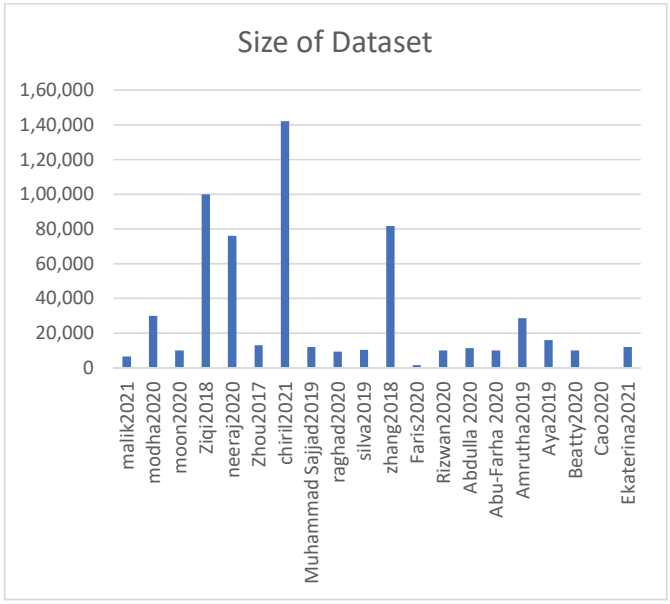


Figure 10 Size of used datasets.

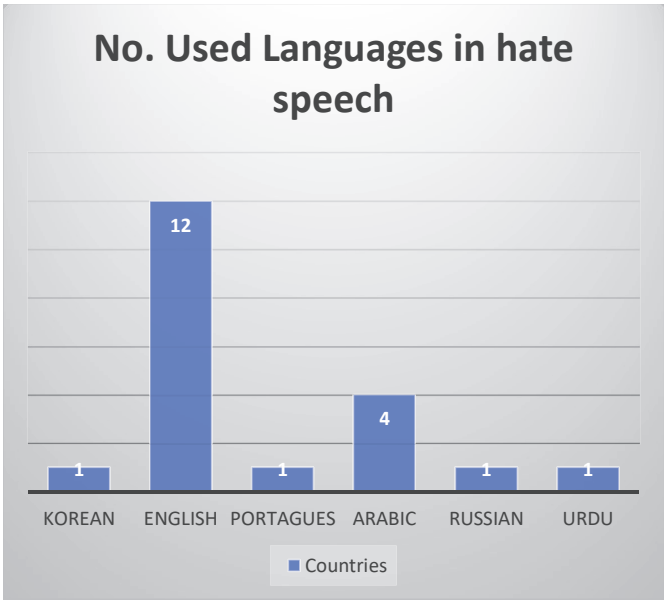


Figure 11 Distribution of languages in hate speech.

4 Discussion

CNN is a particular type of multilayer neural network or deep learning architecture, as the traditional Convolutional Neural Network consists of single or multiple blocks of convolution and pooling layers [19], followed by one or multiple fully connected (FC) layers and an output layer, and each separate neural works in its responsive field and is associated with different neurons such that they cover the whole visual area, different types of layers were described in the consequent section [5]. Detecting hate speech in social media requires other mechanisms and models to find and evaluate the abused words; many research articles have been written specifically for this subject below show the discussion of used models.

First: Using (CNN) Convolutional Neural Network and (GRU) Gated Recurrent Units models

Convolutional Neural Network was first developed for recognizing handwriting [20]. It was used for reading zip codes and postal codes in the postal sector and was recently used in many computer vision tasks like image classification, object detection, object localization, and segmentation domains [21]. CNN consists of three main layers; the Convolutional, pooling, and Fully-connected (FC) layers. The convolutional layer is followed by the pooling layer or additional convolutional layers, and the fully-connected layer is the final layer. The CNN increases its complexity with each layer, identifying more essential parts of the image. Earlier layers concentrate on simple features of the images, such as colors and edges. In contrast, the other layers of CNN recognize more significant elements or shapes of the object until it finally identifies the intended object [22, 23].

GRU has been proposed as an improved version of the standard recurrent neural network (RNN) to make each recurrent unit adaptively capture dependencies of different time scales [24, 25]. To execute machine learning tasks such as image recognition and speech recognition, GRU is a good choice for connecting the node sequences [26, 27].

In [28], the authors used CNN and GRU models, and the research article aimed to see hate and non-hate words focusing on refugees, Muslims, sexism, and racism. They used around (100,000) datasets for testing the models. The models are structured to find implied highlights that can be valuable for recognizing disdain tweets in the long tail. The only limitation that they faced was the lack of training data sets. In the same area, [29] used the same models for detecting abused Arabic keywords in the Twitter platform; they

designed to explore several neural network models based on a convolutional neural network (CNN). The journal article used precisely (9316) data sets for evaluating the models; the main targeted areas of detecting keywords were Racism, Religious, Ideological, Tribal, and Inter-religious. The main goal expands the explore and assess the proposed models on other oppressive language datasets. After that, the journal article used (81688) data sets for detecting Arabic hate speech; they used the models to see racism, sexual, refuge, and Muslim words. The analysis shows that the presence of dynamic ideas, for example, sexism, racism, or hate, is undeniably challenging to identify if exclusively founded on text-based content. In any case, the errand may profit from information regarding gatherings and correspondence modes. Finally, the authors in [30] used (2.6 M) datasets for detecting hate and non-hate twits in the Russian language using the same models. The models naturally distinguishing hate speech should be utilized to defame the creators in no way, shape, or form. The used tools should be applied aside, not initially substituting the used datasets.

Second: Using (CNN) Convolutional Neural Network and (LSTM) Long Short-Term Memory models

The Long Short-Term Memory (LSTM) network model is a developed type of recurrent neural network (RNN) [31, 32] capable of learning order dependence in sequence prediction problems and complex problems such as machine translation, speech recognition, and more. LSTM has four interacting layers with a unique method of communication [33]. A standard LSTM network contains memory blocks called cells. Moreover, it has two states (the cell state and the hidden state) transferred to the next cell. LSTM is designed to avoid the long-term dependency problem [34].

Besides the first section of used models, some research articles used CNN and LSTM models to detect hate speech in social media [35], the journal article proposes a model that can see seven different languages in a single Twitter comment using the mentioned models. The proposed model used (16 K) datasets for testing and evaluating the model [36]. Moreover, the authors in [37] represent the models using around (6517) datasets to detect toxic hate speech; the results demonstrate that CNN can adopt it selves with other models such as LSTM comprehend and efficiently designs if there should arise an occurrence of short words and commotion in datasets. The research article in [38] demonstrates the use of seven different datasets to detect hate and abused comments on Twitter. Using CNN and LSTM models

will give the advantages of outperforming other models in almost all used datasets, and it works quickly with loading all the data. Finally, in [39], they propose an innovative deep learning approach for the automatic detection of cyber-Arabic hate speech on Twitter. The result of used (1634) datasets show the increased number of used times the proposed model performs better than other models.

Third: Using Char (CNN), (BILSTM) Bidirectional LSTM, and Bidirectional Encoder Representations from Transformers (BERT) Models

There are many other deep learning models that are used for detecting hate speech in social media, such as Char CNN, BiLSTM, and Bert models. CharCNN is a model that is based on CNN and used mostly for text classification [40, 41]. **Bidirectional LSTM (biLSTM)**, is a series of processing models that consists of two LSTMs: the first LSTM takes the input in a forward direction, and the second LSTM takes input in a backward direction [42]. BiLSTMs effectively expand the amount of information available to the network. At the same time, BERT is an advanced as well as a more realistic technique since it accepts the fact that a document can simultaneously belong to multiple classes. BERT is known to have achieved exceptional results in eleven natural language understanding (NLU) tasks [43]. Moreover, in [44], the research article explained the hate speech detection in Korean entertainment news and comments on social media platform Twitter using Char (CNN), (BILSTM) and (BERT) models. They have used around (10000) datasets from Online news platforms in Korea; the results show that BERT achieves the best performance compared with Char CNN and BILSTM, the only concern that they faced as their data and layers of the used algorithms especially CNN was not clear.

Fourth: Using Classifier Fusion Model

Using different classification methods to generate a better result [45], the journal article of [46] applied several fusion models on deep learning methods. It combined the classifiers to further develop the general grouping execution. They have used almost (13000) datasets to test, train, and evaluate the detected hateful or not hateful words against women on the Twitter platform. The results demonstrate that fusion processing is a suitable method for detecting hate speech. It is considered sensible to accomplish the commonsense meaning of execution at additional expense.

Fifth: Using Combination Models

The objective things grow ceaselessly with the passage of endlessly time is a significant element that can't be overlooked during the time spent evolving things [47, 48]. In [49] the research article describes the way of using deep learning models for classification of tweets combined with simple machine learning models detecting hate speech on Pakistanis social media such as sexism and racism. They have used around (12000) datasets for testing, training, and evaluating the hate speech words. They ran classification trials with different classifiers and identified the top three classifiers that outperformed in almost all circumstances. Logistic Regression, Random Forest, and Support Vector Machine are examples of classifiers.

Sixth: Using CNN Models

One of the unique types of ANN design is that of the Convolutional Neural Network (CNN). CNN's are fundamentally used to tackle troublesome picture-driven design acknowledgment assignments and, with their exact yet straightforward design, offer a worked-on strategy for getting begun with ANNs [50, 51]. The following journal articles used other models for detecting hate speech words on the Twitter platform. In [52] explained the issues relating to detecting hate speeches, as they address how the detection results can be displayed in an online environment. They have used Trolling, Aggregation, and cyberbullying (TRAC) and posted on political leaders' Twitter accounts to detect racism keywords. In [53], they have used a robust system to simplify haste speech towards different targets; the main focus was detecting Sexism, misogyny, ethnicity, religion, and race words. Moreover, in [54], the authors explained how to detect hate speech on social media in portages language. They have used almost (10366) datasets Using the CNN model for detecting Racism, Sexism, religious intolerance, cursing, and not offensive words. After that, in [55] and [56] the research articles demonstrated the detection of hate and non-hate keywords on social media platforms using CNN models. In the same area [57], and [58] used Deep learning, Transfer learning, multitask learning, and GRU for detecting Hate, Not Hate, Offensive, Non-Offensive speeches on the social media platform. The results show that using a multitask learning setting was very helpful because of the significant connection between two assignments. Finally, in [59] and [60] they have highlighted the detection of the hate non-hate words using CNN, deep learning, and LSTM models using various numbers of datasets.

Table 3 Experimental table

#	File Name	Proposed Model	Dataset	Hate Speech Vocabulary	Classification of the Dataset
1	malik2021	Multi-layer perception (MLP), CNN and LTSM	ALONE (AdoLescents On twittEr), HASOC dataset, and ALONE-HASOC-Mixed	profane, hate and Offensive	toxic: 2379 and non-toxic: 4138
2	modha2020	Traditional SVM	Trolling, Aggregation and cyberbulling (TRAC) and commented posted on political leader's Facebook page and Twitter accounts	Racism	75% Racist and 25% non-racist
3	moon2020	CharCNN, BiLSTM and BERT	Online news platforms in Korea	sexism, political affiliation, religion, nationality, skin colour, and disability	Gender: 15,710, Others: 18,160, and none hate speech: 65,700
4	Ziqi2018	CNN + GRU	7 datasets as: WZ, WZ-S.amt, WZ-S.exp, WZ-S.gb, WZ.pj, DT, and RM	hate and non-hate focusing on: refugee and Muslims, sexism and racism	Racism: 22,468, Hate: 18,696, Sexism: 40,836, and Others: 18,000
5	neeraj2020	CNN LTSM model	7 datasets: HASOC2019-EN, HASOC2019-Hi, Tdavidser et al, ElSharif et al, Ousidhoum et l, SemEval 2019, and Pmathur et al.	sexual orientation, religion, nationality, gender, and ethnicity	Hate: 61,840 and Abuse: 4,550
6	Zhou2017	classifer fussion	SemEval 2019 Task 5	hateful or not hateful against women and immigrants	Hate: 5,460, and non-hate: 7,540
7	chiril2021	different models as: Baseline, Elmo, LSTM, LASTM fastText, CNN fastText, and BERT	Davidson, Founta, Waseem, AMIcorpora, HatEval, and IbertEval	Sexism, misogyny, ethnicity, religion, race (xenophobia and racism)	hate speech: 1430, offensive: 19190, abusive: 27037, hateful: 4948, spam: 14024, racism: 1957, sexism: 3216, misogyny: 4096, women: 2608
8	Muhammad Sajjad2019	Combination of CNN, Random forest, Logistic regression, LTSM, SVM, Glove, random embedding, and Baseline features	Wassem and Hovy	Sexism, racism, or neither	sexism: 2901, racism:1302, neither: 7704

9	raghad2020	CNN, GRU, CCC+GRU, and BERT	GHSD and RHSD	Racism, Religious, Ideological, Tribal, and Inter-religions	Hate tweets: 2539none-hate tweets: 6425
10	silva2019	CNN	OffComBr-3	Racism, sexism, homopholia, xenophebia, religious intolerance, cursing, and not offensive	Hate speech: 1228, and Non hate speech: 4440
11	zhang2018	CNN+GRU	WZ-L, WZ-S.amt, WZ-S.exp, WZ-S.gb, WZ-LS, DT, and RM	Racism, sexism, refugee and Muslim	WZ-L: 16, 093, WZ-S.amt: 6, 594, WZ-S.exp: 6, 594, WZ-S.gb: 6, 594, WZ-LS: 18, 625, DT: 24, 783, and RM: 2, 435
12	Faris2020	Smart deep learning approach that combined CNN with LSTM	Twitter streaming Application Programming Interface (API) and "rtweet" package were used to collect data	Islam and terrorism, Freedom, media, refugee, racism, Homeland and others	Alwahadat sport club: 14, The Jordanian league: 44, Faisaly Jordan: 24, Islam and terrorism and Damage Islam: 100, Racism: 1193, Refugees: 240, Freedom, Media, Homeland, Nahed, Hattar, and Extreme: 19
13	Rizwan2020	CNN-gram Combines with BERT, XLM-RoBRTa, FastText, RomUrEm	Roman Urdu Hate-Speech and Offensive Language Detection (RUHSOLD) Dataset	Abusive/Offensive, sexism, religious Hate, profane	Abusive/Offensive: 2402, Sexism: 839, Religious Hate: 782, Profane: 640, and Normal: 5349
14	Abdulla 2020	Random Forest, Complement NB, DecisionTree,	Twitter dataset on Sunnah and Shia (SSTD)	Hate, Not Hate	Hate 642, Not Hate 2590
15	Abu-Farha 2020	Deep learning, Transfer learning and Multitask learning	SemEval 2020 Arabic offensive language	Hate, Not Hate, Offensive, Non-Offensive	Hate 361, Not Hate 6639, Offensive 1410, Non-Offensive 5590

(*Continued*)

Table 3 Continued

#	File Name	Proposed Model	Dataset	Hate Speech Vocabulary	Classification of the Dataset
16	Amrutha2019	The Gated Recurrent Unit (GRU) is useful for recording sequence orders, the Convolution Neural Network (CNN) is useful for feature extraction, and the Universal Language Model Fine-tuning (ULMFiT) model is based on the transfer learning technique.	publicly available dataset AND Newly created dataset	Hate Speech	–
17	Aya2019	CNN and LSTM	Word-level representation only and word-level representation and Feelings	Racism and Sex	1943 racist tweets, 3166 sexist tweets, 10889 neutral tweets and tweets that belong to more than one class
18	Beatty2020	t graph convolutional networks	Classification of retweet networks and tweet texts while applying adversarial attacks	Abusive, Hateful, and neither	Abusive: 2154, Hateful: 967, and Neither: 6953
19	Cao2020	Utilizes different types of feature embeddings to represent a post p. The feature embeddings are subsequently fed into neural network models to learn three types of textual information latent representations, namely, semantic, sentiment, and topic.	WZ-LS, DT, FOUNTA, and COMBINED	Racism, sexism, both, and neither	racism: 82, sexism: 3,332, neither: 9.767, hate: 5337, offensive: 19,190, abusive: 19,232, spam: 13,840, normal: 119952, inappropriate: 47,194, and neither: 4,163
20	Ekaterina2021	CNN + GRU + Word2Vec Skip-gram embeddings Ensemble of ALBERT models	Word2Vec-Ethno	Hate, No Hate	Positive: 2630, Negative: 4080, and natural: 2630

Table 4 Table of results

#	File Name	F1 Score	Precision	Recall	Accuracy	Advantages	Limitations
1	malik2021	MLP (0.64), CNN(0.81), and LTSM (0.78)	MLP (0.63), CNN(0.83), and LTSM (0.80)	MLP (0.64), CNN (0.89), and LTSM (0.79)	MLP (0.64), CNN(0.82), and LTSM (0.78)	When dealing with small sequences of words and noise in datasets, CNN model can adapt better than the LSTM and it has better performance	This model is monomodal and designed only for English language
2	modha2020	0.7593	0.8116	0.7386	Low accuracy	The model has features to visualize hate and sentiment with the ability to identify and block hateful individuals.	Not considering the nature of user-to-user speech in social media
3	moon2020	F1 for bias: 0681 F1 for gender: 0.633	–	–	–	With bias detection, CharCNN performed better than other models. But with Hate speech detection, BERT outperformed the other two methods	The used layers of CNN and its parameters are not clear
4	Ziqi2018	0.92	0.93	0.93	–	the skipped structure of CNN and GRU can find implicit features in the hate tweets	CNN performs better
5	neeraj2020	Average: 0.71	average: 0.88	average: 0.87	average: 0/71	The CNN LSTM performance is high due to the use of test-driven approaches.	the CNN LSTM model needed moderate GPU processing and inferred a single sample instance in near-real times
6	Zhou2017	0.689	–	–	0.741	The accuracy and F1 score can be improved when using several different classifiers with different parameters	Although the results of the F1 score are good, the degree of integration is not deep enough, and the primary word vector expression in CNN must replace.

(*Continued*)

Table 4 Continued

#	File Name	F1 Score	Precision	Recall	Accuracy	Advantages	Limitations
7	chiril2021	0.685	0.68	0.53	0.68	Its advantages to combining several abusive languages in datasets. It can assist in detecting abusive language in non-generalizable (unseen) problems.	Building a robust system needs improvement for generating different topical focuses and targets.
8	Muhammad Sajjad2019	0.967	0.974	0.96		Logistic Regression, Random Forest and Support Vector Machine give better classification results	CNN method gives better results when it has been used with Glove Embedding and used for training with Logistic Regression and for tagging features with POS.
9	raghad2020	0.79	0.81	0.78	0.83	All the proposed models used in this paper, show good results but CNN outperformed other models.	Some Arabic tweets cause confusion, which affects the classification and may classify these tweets as hateful. Besides, the misclassified tweets are questionable, and it is difficult to consider them as hate or non-hate. Thus, hate speech detection is difficult and depends highly on the context.
10	silva2019	0.89			87%	the proposed model is a low-cost computational model	The Portuguese dataset that used in this work had a lack of available robust datasets for analysis with some irrelevant topics in the users' comments
11	zhang2018	0.89				The used dataset lacks some hate speech, such as against religious (Muslim) and refugees. Therefore, they have been added to the dataset. Besides, some hate speech vocabularies are challenging to detect based on textual contents, such as; sexism, racism, and hate are challenging to see.	Tweets that contained sexist hate speech were difficult to detect because these types of speech need to understand the implication of the language.

12	Faris2020	71.688	68.965	79.768	66.564	The used method had better performance when its number of epochs increased	The study requires large benchmark datasets, especially for the comprehensive lexical resource of abusive, offensive Arabic expressions.
13	Rizwan2020	BERT + CNN-gram: 0.90, XLM-RoBERTa + CNN-gram: 0.88, FastText + CNN-gram: 0.80, RomUrEm + CNN-gram: 0.89	BERT + CNN-gram:0.90, XLM-RoBERTa + CNN-gram: 0.88, FastText + CNN-gram: 0.81, RomUrEm + CNN-gram: 0.89	BERT + CNN-gram:0.90, XLM-RoBERTa + CNN-gram: 0.88, FastText + CNN-gram: 0.80, RomUrEm + CNN-gram: 0.89	BERT + CNN-gram:0.90, XLM-RoBERTa + CNN-gram: 0.88, FastText + CNN-gram: 0.81, RomUrEm + CNN-gram: 0.89	The BERT+CNNgram outperforms the other models with the highest F1 score. Also, findings of the coarse-grained classification tasks indicate that instead of training embeddings, using existing pre-trained embeddings by fine-tuning them on the task is a more wise choice.	The proposed models used in this paper confuse between abusive/offensive and profanes compared to other labels. Besides, the models have limitations concerning the intricacies of human language for subtle differences between profane language and targeted abuse or offensive language.
14	Abdulla 2020	RF(0.87) CNB(0.78) DT(0.87) SVM(0.88) CNN(0.88)	RF(0.78) CNB(0.85) DT(0.84) SVM(0.81) CNN(0.83)	RF(0.99) CNB(0.72) DT(0.46) SVM(0.95) CNN(0.39)	RF(0.78) CNB(0.68) DT(0.74) SVM(0.8) CNN(0.8)	FastText can integrate sub-word information into the embedding learning process	It can be implemented only on Arabic Tweets
15	Abu-Farha 2020	CNN-BiLSTM Offensive(0.901) CNN-BiLSTM Hate (0.702)				The multitask learning that used in the paper was useful	The multitask learning used in this paper must implement on resources that contain lexicons and experiments.
16	Amrutha2019	GRU(130.2) CNN(128.16) ULM-FIT(194.01)			GRU(191.25) CNN(188.5) ULMFIT(194.2)	ULMFiT model outperforms the GRU and CNN models with better F1 and accuracy scores.	If Considering only the last hidden layer, the data may get lost. Therefore, there should be Concatenate pooling must be used with max-pooled and mean pooled
17	Aya2019	0.6			0.8835	–	For improving the method, the architecture of the LSTM layer must improve.

(*Continued*)

Table 4 Continued

#	File Name	F1 Score	Precision	Recall	Accuracy	Advantages	Limitations
18	Beatty2020	0.67	0.74			Graph convolutional networks have a better F1 score in detecting the hate speech tweets on Twitter	The model is not analyze the specific accounts associated with hate speech diffusion
19	Cao2020	WZ-LS 78.19, DT 89.92, FOUNTA 79.09, COMBINED 92.43	WZ-LS 77.95, DT 89.97, FOUNTA 78.95, COMBINED 92.48	WZ-LS 79.48, DT 90.39, FOUNTA 80.43, COMBINED 92.45		When dealing with word embedding, CNN and LSTM models perform better than their performance on character-level embedding inputs. At the same time, the character-level bi-gram embedding is worse among the three types of input feature embeddings. Besides, the basic CNN and LSTM with word embedding models could outperform HybridCNN and CNN-GRU models	They incorporate non-textual features into the Deep hate model and improve the posts' sentiment and topic representations with more advanced techniques
20	Ekaterina2021	NB 0.76, LSTMÞGRU 0.864, Convers-RuBERTÞDense 0.833				The Convers-RuBERT model outperformed both classical machine learning and LSTMÞGRU models.	There should be models that detect hate speech automatically and not used as a replacement for expert judgment

5 Conclusion

The systematic review presented in this work provides the state of the art of scientific literature about hate speech detection in Twitter using the Convolutional Neural networks (CNN) based models. For this purpose of reviewing the recent works, (20) related works have been selected among (565) different works. The selection is dependent on the inclusion and exclusion criteria and also based on the main question and the sub-questions that this review tends to answer. Data extraction from the selected study was collected and available online. The extracted data contain all the required information about each study, including article name, authors, publishers, and more detailed information. The analysis of the information obtained from each study allowed finding progress and gaps in this research area. In particular, this systematic review has found many gaps in the part of hate speech detection. Many models have been used, most are CNN-based models, and each has added advantages to the recent models for hate speech detection [61]. Besides, still, these models and methods have many limitations and gaps. Including; most of the models cannot detect the hate speech automatically, not suitable with all the languages, they are working only with one language, most are best suited with the English language, and when they used with datasets with other languages, the models are suffering from confusion in speech classification. Finally, most models are not considering a user-to-user speech in social media. In conclusion, although the hate speech detection in social media has improved recently, this systematic review has found many gaps in the part of hate speech detection, and at the same time, hate speech threats are increasing rapidly that need immediate and real solutions.

References

[1] W. Alorainy, P. Burnap, H. Liu, and M. Williams, "The Enemy Among Us: Detecting Hate Speech with Threats Based 'Othering' Language Embeddings," 2018, [Online]. Available: http://arxiv.org/abs/1801.07495.

[2] S. T. Luu, K. Van Nguyen, and N. L. T. Nguyen, "A Large-Scale Dataset for Hate Speech Detection on Vietnamese Social Media Texts," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12798 LNAI, pp. 415–426, 2021, doi: 10.1007/978-3-030-79457-6_35.

[3] L. Ketsbaia, B. Issac, and X. Chen, "Detection of hate tweets using machine learning and deep learning," *Proc. – 2020 IEEE 19th Int. Conf. Trust. Secur. Priv. Comput. Commun. Trust. 2020*, pp. 751–758, 2020, doi: 10.1109/TrustCom50675.2020.00103.

[4] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection using Natural Language Processing," *Soc. 2017 – 5th Int. Work. Nat. Lang. Process. Soc. Media, Proc. Work. AFNLP SIG Soc.*, no. 2012, pp. 1–10, 2017, doi: 10.18653/v1/w17-1101.

[5] S. Ahammed, M. Rahman, M. H. Niloy, and S. M. M. H. Chowdhury, "Implementation of Machine Learning to Detect Hate Speech in Bangla Language," *Proc. 2019 8th Int. Conf. Syst. Model. Adv. Res. Trends, SMART 2019*, pp. 317–320, 2020, doi: 10.1109/SMART46866.2019.9117214.

[6] S. Malmasi and M. Zampieri, "Detecting hate speech in social media," *Int. Conf. Recent Adv. Nat. Lang. Process. RANLP*, vol. 2017-Septe, pp. 467–472, 2017, doi: 10.26615/978-954-452-049-6-062.

[7] M. Polignano, P. Basile, M. de Gemmis, and G. Semeraro, "Hate speech detection through Alberto Italian language understanding model," *CEUR Workshop Proc.*, vol. 2521, 2019.

[8] A. Alotaibi and M. H. Abul Hasanat, "Racism Detection in Twitter Using Deep Learning and Text Mining Techniques for the Arabic Language," *Proc. – 2020 1st Int. Conf. Smart Syst. Emerg. Technol. SMART-TECH 2020*, pp. 161–164, 2020, doi: 10.1109/SMART-TECH49988.2020.00047.

[9] M. A. Carlin and M. Elhilali, "A Framework for Speech Activity Detection Using Adaptive Auditory Receptive Fields," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 12, pp. 2422–2433, 2015, doi: 10.1109/TASLP.2015.2481179.

[10] B. Gambäck and U. K. Sikdar, "Using Convolutional Neural Networks to Classify Hate-Speech," no. 7491, pp. 85–90, 2017, doi: 10.18653/v1/w17-3013.

[11] P. Mayr, I. Frommholz, and G. Cabanac, "Bibliometric-enhanced information retrieval: 7th international BIR workshop," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10772 LNCS, pp. 827–828, 2018, doi: 10.1007/978-3-319-76941-7.

[12] N. Albadi, M. Kurdi, and S. Mishra, "Investigating the effect of combining GRU neural networks with handcrafted features for religious hatred

detection on Arabic Twitter space," *Soc. Netw. Anal. Min.*, vol. 9, no. 1, pp. 1–19, 2019, doi: 10.1007/s13278-019-0587-5.

[13] M. B. Aliyu, "American Journal of Engineering Research (AJER) Efficiency of Boolean Search strings for Information Retrieval," *Am. J. Eng. Res.*, vol. 6, no. 11, pp. 216–222, 2017.

[14] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," *ACM Int. Conf. Proceeding Ser.*, 2014, doi: 10.1145/2601248.2601268.

[15] A. Kumar and N. Sachdeva, "Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis," *Multimed. Tools Appl.*, vol. 78, no. 17, pp. 23973–24010, 2019, doi: 10.1007/s11042-019-7234-z.

[16] W. Alorainy, P. Burnap, H. Liu, and M. L. Williams, "'The Enemy Among Us,"' *ACM Trans. Web*, vol. 13, no. 3, pp. 1–26, 2019, doi: 10.1145/3324997.

[17] M. Bani Yassein, S. Aljawarneh, and Y. Wahsheh, "Hybrid Real-Time Protection System for Online Social Networks," *Found. Sci.*, vol. 25, no. 4, pp. 1095–1124, 2020, doi: 10.1007/s10699-019-09595-7.

[18] P. Fortuna, J. Soler-Company, and L. Wanner, "How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?," *Inf. Process. Manag.*, vol. 58, no. 3, p. 102524, 2021, doi: 10.1016/j.ipm.2021.102524.

[19] Z. L. Chia, M. Ptaszynski, F. Masui, G. Leliwa, and M. Wroczynski, "Machine Learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection," *Inf. Process. Manag.*, vol. 58, no. 4, 2021, doi: 10.1016/j.ipm.2021.102600.

[20] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998, doi: 10.1109/5.726791.

[21] T. Sercu, C. Puhrsch, B. Kingsbury, I. B. M. T. J. Watson, and Y. Heights, "Very deep multilingual convolutional neural networks for LVCSR Center for Data Science, Courant Institute of Mathematical Sciences, New York University," *Icassp 2016*, pp. 4955–4959, 2016.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd Int. Conf. Learn. Represent. ICLR 2015 – Conf. Track Proc.*, pp. 1–14, 2015.

[23] B. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Cnn实际训练的," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2012.

[24] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," *Proc. SSST 2014 – 8th Work. Syntax. Semant. Struct. Stat. Transl.*, pp. 103–111, 2014, doi: 10.3115/v1/w14-4012.

[25] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," pp. 1–9, 2014.

[26] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Light Gated Recurrent Units for Speech Recognition," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 2, no. 2, pp. 92–102, 2018, doi: 10.1109/TETCI.2017.2762739.

[27] J. Liu, C. Wu, and J. Wang, "Gated recurrent units based neural network for time heterogeneous feedback recommendation," *Inf. Sci. (Ny).*, vol. 423, pp. 50–65, 2018, doi: 10.1016/j.ins.2017.09.048.

[28] Z. Zhang and L. Luo, "Hate speech detection: A solved problem? The challenging case of long tail on Twitter," *Semant. Web*, vol. 10, no. 5, pp. 925–945, 2019, doi: 10.3233/SW-180338.

[29] R. Alshalan and H. Al-Khalifa, "A deep learning approach for automatic hate speech detection in the saudi twittersphere," *Appl. Sci.*, vol. 10, no. 23, pp. 1–16, 2020, doi: 10.3390/app10238614.

[30] E. Pronoza, P. Panicheva, O. Koltsova, and P. Rosso, "Detecting ethnicity-targeted hate speech in Russian social media texts," *Inf. Process. Manag.*, vol. 58, no. 6, p. 102674, 2021, doi: 10.1016/j.ipm.2021.102674.

[31] À. A. Carracedo and R. J. Mondéjar, "Profiling Hate Speech Spreaders on Twitter," *CEUR Workshop Proc.*, vol. 2936, no. August, pp. 1801–1807, 2021.

[32] A. Bisht, A. Singh, H. S. Bhadauria, J. Virmani, and Kriti, *Detection of hate speech and offensive language in twitter data using LSTM model*, vol. 1124. Springer Singapore, 2020.

[33] H. T.-T. Do, H. D. Huynh, K. Van Nguyen, N. L.-T. Nguyen, and A. G.-T. Nguyen, "Hate Speech Detection on Vietnamese Social Media Text using the Bidirectional-LSTM Model," pp. 4–7, 2019.

[34] O. Levy, K. Lee, N. FitzGerald, and L. Zettlemoyer, "Long short-term memory as a dynamically computed element-wise weighted sum," *ACL 2018 – 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 2, pp. 732–739, 2018, doi: 10.18653/v1/p18-2116.

[35] A. Elouali et al., "Hate speech detection on multilingual twitter using convolutional neural networks," *WebSci 2020 – Proc. 12th ACM Conf. Web Sci.*, vol. 34, no. 4, pp. 1–6, 2020, doi: 10.18280/ria.340111.

[36] C. Parada, M. Dredze, D. Filimonov, and F. Jelinek, "Contextual information improves OOV detection in speech," *NAACL HLT 2010 – Hum. Lang. Technol. 2010 Annu. Conf. North Am. Chapter Assoc. Comput. Linguist. Proc. Main Conf.*, no. June, pp. 216–224, 2010.

[37] P. Malik, "Toxic Speech Detection using Traditional Machine Learning Models and BERT and fastText Embedding with Deep Neural Networks," no. Iccmc, pp. 1254–1259, 2021.

[38] N. Vashistha and A. Zubiaga, "Online multilingual hate speech detection: Experimenting with hindi and english social media," *Inf.*, vol. 12, no. 1, pp. 1–16, 2021, doi: 10.3390/info12010005.

[39] H. Faris, I. Aljarah, M. Habib, and P. A. Castillo, "Hate speech detection using word embedding and deep learning in the Arabic language context," *ICPRAM 2020 – Proc. 9th Int. Conf. Pattern Recognit. Appl. Methods*, no. March, pp. 453–460, 2020, doi: 10.5220/0008954004530460.

[40] Q. Hua, S. Qundong, J. Dingchao, G. Lei, Z. Yanpeng, and L. Pengkang, "A Character-Level Method for Text Classification," *Proc. 2018 2nd IEEE Adv. Inf. Manag. Commun. Electron. Autom. Control Conf. IMCEC 2018*, no. Imcec, pp. 402–406, 2018, doi: 10.1109/IMCEC.2018.8469258.

[41] J. Wang, Z. Wang, D. Zhang, and J. Yan, "Combining knowledge with deep convolutional neural networks for short text classification," *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 0, pp. 2915–2921, 2017, doi: 10.24963/ijcai.2017/406.

[42] T. Chen, R. Xu, Y. He, and X. Wang, "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN," *Expert Syst. Appl.*, vol. 72, pp. 221–230, 2017, doi: 10.1016/j.eswa.2016.10.065.

[43] S. González-Carvajal and E. C. Garrido-Merchán, "Comparing BERT against traditional machine learning text classification," no. Ml, 2020.

[44] J. Moon, W. I. Cho, and J. Lee, "BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection," no. October, pp. 25–31, 2020, doi: 10.18653/v1/2020.socialnlp-1.4.

[45] D. Ruta and B. Gabrys, "An Overview of Classifier Fusion Methods An Overview of Classifier Fusion Methods," no. January 2000, 2016.

[46] Y. Zhou, Y. Yang, H. Liu, X. Liu, and N. Savage, "Deep Learning Based Fusion Approach for Hate Speech Detection," *IEEE Access*, vol. 8, pp. 128923–128929, 2020, doi: 10.1109/ACCESS.2020.3009244.

[47] T. Li, Y. Zhang, and T. Wang, "SRPM–CNN: a combined model based on slide relative position matrix and CNN for time series classification," *Complex Intell. Syst.*, vol. 7, no. 3, pp. 1619–1631, 2021, doi: 10.1007/s40747-021-00296-y.

[48] M. M. Ahsan, T. E. Alam, T. Trafalis, and P. Huebner, "Deep MLP-CNN model using mixed-data to distinguish between COVID-19 and Non-COVID-19 patients," *Symmetry (Basel).*, vol. 12, no. 9, 2020, doi: 10.3390/sym12091526.

[49] M. Sajjad, F. Zulifqar, M. U. G. Khan, and M. Azeem, "Hate Speech Detection using Fusion Approach," *2019 Int. Conf. Appl. Eng. Math. ICAEM 2019 – Proc.*, pp. 251–255, 2019, doi: 10.1109/ICAEM.2019.8853762.

[50] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," no. December, 2015, [Online]. Available: http://arxiv.org/abs/1511.08458.

[51] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," *Proc. 2017 Int. Conf. Eng. Technol. ICET 2017*, vol. 2018–Janua, no. August, pp. 1–6, 2018, doi: 10.1109/ICEngTechnol.2017.8308186.

[52] S. Modha, T. Mandl, P. Majumder, and D. Patel, "Tracking Hate in Social Media: Evaluation, Challenges and Approaches," *SN Comput. Sci.*, vol. 1, no. 2, 2020, doi: 10.1007/s42979-020-0082-0.

[53] P. Chiril, E. W. Pamungkas, F. Benamara, V. Moriceau, and V. Patti, *Emotionally Informed Hate Speech Detection: A Multi-target Perspective*, vol. 14, no. 1. Springer US, 2022.

[54] S. C. Silva, A. B. S. Serapião, and I. Paraboni, "Hate-speech detection in Portuguese using CNN and psycho-linguistic dictionary," no. September, 2019.

[55] H. Rizwan, M. H. Shakeel, and A. Karim, "Hate-speech and offensive language detection in Roman Urdu," *EMNLP 2020 – 2020 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 2512–2522, 2020, doi: 10.18653/v1/2020.emnlp-main.197.

[56] A. Aref, R. Husni Al Mahmoud, K. Taha, and M. Al-Sharif, "Hate Speech Detection of Arabic Shorttext," pp. 81–94, 2020, doi: 10.5121/csit.2020.100507.

[57] I. Abu Farha and W. Magdy, "Multitask Learning for {A}rabic Offensive Language and Hate-Speech Detection," *Proc. 4th Work. Open-Source Arab. Corpora Process. Tools, with a Shar. Task Offensive Lang. Detect.*, no. May, pp. 86–90, 2020, [Online]. Available: https://www.aclweb.org/anthology/2020.osact-1.14.

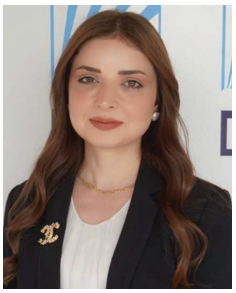
[58] S. Tructures, Z. Deng, Y. Luo, J. Zhu, and B. Zhang, "B Ayesian L Earning of D Eep N Eural N Etwork," *2019 Int. Conf. Intell. Comput. Control Syst.*, no. 2, pp. 1–20, 2019.

[59] M. Beatty, "Graph-Based Methods to Detect Hate Speech Diffusion on Twitter," *Proc. 2020 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2020*, pp. 502–506, 2020, doi: 10.1109/ASONAM 49781.2020.9381473.

[60] R. Cao, R. K. W. Lee, and T. A. Hoang, "DeepHate: Hate Speech Detection via Multi-Faceted Text Representations," *WebSci 2020 – Proc. 12th ACM Conf. Web Sci.*, pp. 11–20, 2020, doi: 10.1145/3394231.3397890.

[61] I. Shahin, A. B. Nassif, and M. B. Alsabek, "COVID-19 Electrocardiograms Classification using CNN Models," *Proc. – Int. Conf. Dev. eSystems Eng. DeSE*, vol. 2021–December, pp. 448–452, 2021, doi: 10.1109/DESE54285.2021.9719358.

Biographies



Ara Zozan Miran, received a bachelor's degree in Computer science from the University of Kurdistan – Hawler in 2016 and a master's degree in Software Engineering from the University of Kurdistan – Hawler in 2018. She is currently working as an assistant lecturer at the Department of Information Technology, Faculty of Computer Engineering and Science, Lebanese French University Erbil, Iraq. Her research areas include enhancing AODV routing protocols, The 3D face mask recognition to minimize COVID19, and Evaluating e-governments.

Hazha Saeed Yahia, received a bachelor's degree in Information Technology from the University of Kurdistan – Hawler in 2010 and a master's degree in Computer System Engineering from the University of Kurdistan – Hawler in 2016. She is now a Ph.D. candidate in Information and Communication Technologies at Duhok Polytechnic University. She is currently working as an assistant lecturer at the Department of Information Technology, Faculty of Computer Engineering and Science, Lebanese French University. Her research areas include artificial intelligence, meta-heuristic optimizations, e-services, and e-governments.