Neurocomputing 546 (2023) 126232

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Survey paper

A systematic review of hate speech automatic detection using natural language processing ☆

Md Saroar Jahan *, Mourad Oussalah

University of Oulu, CMVS, BP 4500, 90014 Finland

ARTICLE INFO

Article history:
Received 14 May 2021
Revised 20 December 2022
Accepted 22 April 2023
Available online 09 May 2023
Communicated by Zidong Wang

Keywords:
Hate speech detection review
Systematic review
PRISMA hate speech
NLP deep learning review

ABSTRACT

With the multiplication of social media platforms, which offer anonymity, easy access and online community formation and online debate, the issue of hate speech detection and tracking becomes a growing challenge to society, individual, policy-makers and researchers. Despite efforts for leveraging automatic techniques for automatic detection and monitoring, their performances are still far from satisfactory, which constantly calls for future research on the issue. This paper provides a systematic review of literature in this field, with a focus on natural language processing and deep learning technologies, highlighting the terminology, processing pipeline, core methods employed, with a focal point on deep learning architecture. From a methodological perspective, we adopt PRISMA guideline of systematic review of the last 10 years literature from ACM Digital Library and Google Scholar. In the sequel, existing surveys, limitations, and future research directions are extensively discussed.

1. Introduction

In the era of social computing, the interaction between individuals becomes more striking, especially through social media platforms and chat forums. Microblogging applications opened up the chance for people worldwide to express and share their thoughts instantaneously and extensively. Driven, on one hand, by the platform's easy access and anonymity. And, on the other hand, by the user's desire to dominate debate, spread/ defend opinions or argumentation, and possibly some business incentives, this offered a fertile environment to disseminate aggressive and harmful content. Despite the discrepancy in hate speech (HS) legislation from one country to another, it is usually thought to include communications of animosity or disparagement of an individual or a group on account of a group characteristic such as race, color, national origin, sex, disability, religion, or sexual orientation [100]. Benefiting from the variation in national hate speech legislation, the difficulty to set a limit to the constantly evolving cyberspace, the increased need of individuals and societal actors to express their opinions and counter-attacks from opponents and the delay in manual check by internet operators, the propagation of hate speech online has gained new momentum that continuously challenges both policy-makers and research community. With the development in natural language processing (NLP) technology, much research has been done concerning automatic textual hate speech detection in recent years. A couple of renowned competitions (e.g., SemEval-2019 [191] and 2020 [192], GermEval-2018 [183]) have held various events to find a better solution for automated hate speech detection. In this regard, researchers have populated large-scale datasets from multiple sources, which fueled research in the field. Many of these studies have also tackled hate speech in several non-English languages and online communities. This led to investigate and contrast various processing pipelines, including the choice of feature set and Machine Learning (ML) methods (e.g., supervised, unsupervised, and semi-supervised), classification algorithms (e.g., Naives Bayes, Logistic Regression (LR), Convolution Neural Network (CNN), LSTM, BERT deep learning architectures, and so on). The limitation of the automatic textual-based approach for efficient detection has been widely acknowledged, which calls for future research in this field. Besides, the variety of technology, application domain, and contextual factors require a constant up-to-date of the advance in this field in order to provide the researcher with a comprehensive and global view in the area of automatic hate text (HT) detection. Extending existing survey papers in this field, this paper contributes to this goal by providing an updated systematic review of literature of automatic textual hate speech detection with a

☆ This work is supported by the European Young-sters Resilience through Serious Games, under the Internal Security Fund-Police action: 823701-ISFP-2017-AG-RAD grant, which is gratefully acknowledged.

* Corresponding author.
E-mail addresses: mjahan18@edu.oulu.fi (M.S. Jahan), Mourad.Oussalah@oulu.fi (M. Oussalah).

https://doi.org/10.1016/j.neucom.2023.126232
0925-2312/

special focus on machine learning and deep learning technologies. We frame the problem, its definition and identify methods and resources employed in HT detection. We adopted a systematic approach that critically analyses theoretical aspects and practical resources, such as datasets, methods, existing projects following PRISMA guidelines [90]. In this regards, we have tried to answer the following research questions:

- Q1: What are the specificities among different HS branches and scopes for automatic HS detection from previous literature?
- Q2: What is the state of the deep learning technology in automatic HS detection in practice?
- Q3: What is the state of the HS datasets in practice?The above-researched questions will examine barriers and scopes for the automatic hate speech detection technology. A systematic review-based approach is conducted to answer Q1 and Q2, where we will try to depict and categorize the existing technology and literature. The third research question Q3, will be answered by critically examining the scope and boundaries of the dataset identified by our literature review, highlighting the characteristics and aspects of the available resources.

This review paper is organized as follows: Section 2 will include a brief theoretical definition of HS. Section 4 examines the previously identified review papers of HS detection. Section 5 details the systematic literature review document collection methodology. Section 6 presents the results of this literature review, including the state of deep learning technology. Section 7 emphasizes on the available resources (datasets and open-source projects). After that, in Section 8, an extensive discussion is carried out. Finally, we have highlighted future research directions and conclusions at the end of this paper.

2. Background

2.1. What is hate speech?

Deciding if a portion of text contains hate speech is not simple, even for human beings. Hate speech is a complex phenomenon, intrinsically associated with relationships between groups, and relies on language nuances. Different organization and authors have tried to define hate speech as follow:

1. **Code of Conduct between European Union Commission and companies:** "All conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, color, religion, descent or national or ethnic" [184].
2. **International minorities associations (ILGA):** 'Hate crime is any form of crime targeting people because of their actual or perceived belonging to a particular group. The crimes can manifest in a variety of forms: physical and psychological intimidation, blackmail, property damage, aggression and violence, rape'[1].
3. **Academia**- Nobata et al. [99]defined HS as an act that attacks or demeans a group/individual based on race, ethnic origin, religion, disability, gender, age, disability, or sexual orientation/-gender identity. Similarly, Nockleby [100] defined HS as "any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic." Warner and Hirschberg [177] distinguished the occurrence of hate speech related wording with user's intention to harm an individual or a group. Otherhand, Waseem and Hovy (2016) [179] viewed hate speech in the form of racist and sexist remarks
4. **Facebook:** "We define hate speech as a direct attack against people on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease. We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing, and calls for exclusion or segregation. We consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants, and asylum seekers from the most severe attacks, though we do allow commentary and criticism of immigration policies. Similarly, we provide some protections for characteristics like occupation, when they're referenced along with a protected characteristic[2]."
5. **Twitter:** 'You may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease' [3]. Examples from Twitter hate-speech are: (See Table 1)
   - "I'm glad this [violent event] happened. They got what they deserved [referring to persons with the attributes noted above]."
   - "[Person with attributes noted above] are dogs" or "[person with attributes noted above] are like animals."
6. **YouTube:** 'We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/-gender, sexual orientation, victims of a major violent event and their kin, and veteran Status'. [4]

To better understand the definitions, we consider different terms analysis from the above definitions. Table 2 summarizes key components that characterize HS in these definitions. For instance, all the definitions point out that HS has specific targets (person, group, nationality, etc.). Furthermore, most of these definitions refer to religion, gender discrimination, race, color, ethnicity, and violence. While less common criteria for measuring HS include curse, disability, property damage, age, and serious disease.

2.2. Other Related Concepts

From the above definitions and contents analysis, it is clear that some elements are highly related to hate speech (e.g., racism, violence, gender discrimination, etc.). Moreover, we have found several previous works that have presented significant branches of HS. The analysis of HS's different branches helps to reach insights from different perspectives. This expects to contribute to spotting and recognizing the interrelationships among these terminologies. Here we will discuss some of the essential categories that are found relevant in most HS studies:

Cyberbullying: Chen et al. [31], Dinakar et al. [37] defined the electronic form of traditional bullying, also, referred to cyberbullying, as the aggression and harassment targeted to an individual who is unable to defend himself. [41] opined that bullying is known for its repetitive act to the same individual, unlike hate speech which is more general and not necessarily intended to hurt a specific individual [41].

[1] https://www.ilga-europe.org/what-we-do/our-advocacy-work/hate-crime-hate-speech

[2] https://www.facebook.com/communitystandards/hate_speech

[3] https://help.twitter.com/en/rules-and-policies/twitter-rules#hateful-conduct

[4] https://support.google.com/youtube/answer/2801939?hl=en

Table 1
Content Analysis of Hate Speech Definitions.

Source	Target	Religion	Gender Discrimination	Race/ Ethnicity/ Color	Radicalization/ Violence	Others
EU Code of conduct	yes	yes	yes	yes	yes	-
ILGA	yes	no	no	no	yes	intimidation, blackmail, property damage, rape
Scientific paper [99]	yes	yes	yes	yes	no	disability, age
Facebook	yes	yes	yes	yes	yes	curse, serious disease, Occupation
Twitter	yes	yes	yes	yes	no	age, disability, serious disease
YouTube	yes	yes	yes	yes	yes	age

Table 2
Sentence classification into different HS categories.

Sentence	General HS	Cyber-bulling	Abusive Offensive	Sexisim/ Gender Dicsrimination	Racism	Radica-lization	Insult/ Swear word
1. John doe is < religion name > from < nationality > dirty is a bisexual and violent	yes	yes	yes	yes	yes	yes	exist
2. John doe is dirty	yes	yes	yes	no	no	no	exist
3. Asylum seekers are not good	yes	no	yes	no	no	no	not-exist
4. John Doe is not good person	no	yes	yes	no	no	no	not-exist
5. John Doe is not bad	no	no	no	no	no	no	exist
6. Group for blacks only!	no	no	no	no	no	no	exist
7. This is bad, but John Doe is lucky	no	no	no	no	no	no	exist
8. John doe working hard. Ugly	no	yes	yes	no	no	no	exist

Racism: This category includes racial offense or tribalism, regionalism, xenophobia (especially for migrant workers) and nativism (hostility against immigrants and refugees), and any prejudice against a particular tribe, region, color, or physical posture of an individual. For instance, offending an individual because he belongs to a specific tribe, region, or country [8].

Sexism, Gender discrimination: O'Brien [101] defined sexism as a prejudice or a discrimination based on a person's sex or gender. Sexism can affect anyone, although, it primarily affects women and girls. It has also been linked to stereotypes and gender roles. Matsumoto [86] stated that in many types of hate, there could be the existence of sexual harassment contents. Moreover, Jha and Mamidi [66] reported that sexism might come in two different forms: Hostile (which is an explicit negative attitude) and Benevolent (which is more subtle).

Radicalization: This concept is usually referred to as a motive towards violent extremism. Radicalization and hate speeches are closely related and sometimes used equivalently. Some authors link radicalization to religious-based hate speech. Wadhwa and Bhatia [175] referred to radical groups as "cyber-extremists.".

Abusive language, Offensive language: The term abusive language refers to hurtful language and includes hate speech, derogatory language, and profanity. However, many researchers referred to the abusive language as offensive language [99].

Religious hate speech: This includes any religious discrimination, such as Islamic sects, calling for atheism, Anti-Christian, and their respective denominations or anti-Hinduism groups. [13] reported that religious hate speech is considered a motive for crimes in countries with the highest social crimes.

2.3. Relationship of HS Concepts and Example

From the above definitions of general HS and other related concepts, we have drawn a relationship diagram shown in Fig. 1.

All other related concepts, cyberbullying, racism, sexism, abusive language, and radicalization, have been derived from the HS concept, which acts as a parent node in this hierarchical construction. In the second level, cyberbullying and abusive/offensive concepts are distinguished. Next, other components like sexism, racism, and radicalization are discerned.

A critical claim advocated by some social science and psychiatry scholars stipulates that cyberbullying cases must include both Insult/Swear wording and a second person/person's name [108]. If there is no second person/person's name available, it may not be considered cyberbullying other than general HS. This means that all cyberbullying cases can also be cast into the HS category, while some HS may not be classified as cyberbullying.

However, the above categorization is not always faithful, as exemplified in Table 2. For example, sentence-7 ("This is bad, but John Doe is lucky") includes both Insult word and Person name; however, it is not an HS, cyberbullying, nor abusive case as the relationship between the two is not established. Similarly, sentence-5, "John Doe is not bad," contains both Person name and Insult word with a clear link between the two entities, but it is not connected to HS, cyberbullying, nor abusive case due to the presence of negating form. Likewise, sentence-4, 'John Doe is not a good person,' does not contain Insult words, but it is considered cyberbullying and offensive. In sentence-3, 'Asylum seekers are dirty,' has an Insult word and target; it falls into HS but is not categorized as cyberbullying because its target is not a Person/ second person.

All these examples show the requirements mentioned above for HS, cyberbullying, and other cases are potential conditions for the occurrence of HS but not sufficient due to the complexity of the natural language modifiers expression that could negate or shift the meaning.

Besides, recognition of HS may be boosted when multiple sentences were put together as in sentence-8 ("John doe working hard. Ugly"), which is considered as a Cyberbullying and Offensive case even though the second sentence "Ugly" contains only an Insult word without any second person/Person entity.

There could be examples of sentences that may comprise multiple HS concepts at the same time. Sentence-1 shows that it contains all categories, HS, cyberbullying, abusive, gender discrimination, racism, and radicalization.

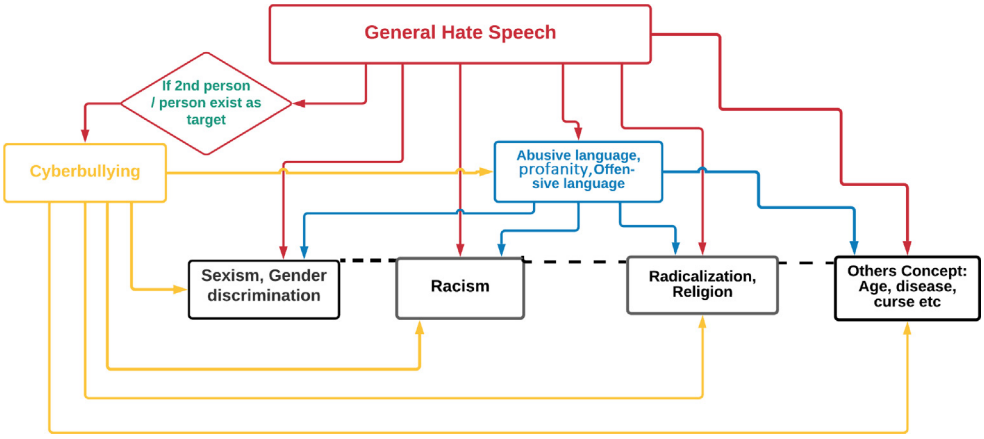


Fig. 1. Relational diagram between different type of hate speech concepts.

The above few cases explain the complications of detecting HS cases and determining their exact categorization, which involves examining all the paragraph's textual information.

2.4. Generic Pipeline of Automatic HS Detection

Kowsari et al. [71] stipulated four essential parts for any text classification task, which still are valid for the HS classification as well. Fig. 2 highlights the generic pipeline of the HS detection task as a text classification system-based approach. Its main components are described below:

(i) Dataset collection and preparation: is the first step in HS detection pipeline. Often, datasets are collected from social media platforms (Facebook, Youtube, Twitter, etc.). Preprocessing is performed according to dataset structure and quality. Typically, this involves filtering and normalization aspect of textual inputs, which include tokenization, stopwords removal, misspelling correction, noise removal, stemming, lemmatization, among others. We shall also notice that the dataset maybe provided initially so that no collection is required. As part of data preparation, training and testing parts of the dataset should be distinguished for the subsequent machine learning step.

(ii) Feature Engineering: is the next phase of the analysis where appropriate features are extracted from the textual inputs so that unstructured text sequences are converted into structured features. Common techniques for feature extractions are TF-IDF, semantic, lexical, topic modeling, sentiment, BOW, word embedding (FastText, GloVe, Word2Vec).

Sometimes, dimensionality reduction is applied to reduce the time and memory complexity. Examples of dimension reduction methods are principal component analysis (PCA), linear discriminant analysis (LDA), non-negative matrix factorization (NMF), random projection, autoencoders, and t-distributed stochastic neighbor embedding (t-SNE) [71].

(iii) Model Training: is one of the most crucial step of the text classification pipeline where a machine learning/deep learning model is trained on the training dataset. Several classifiers can be tailored based on task requirements: RF, NB, LR, CNN, RNN, BERT, etc. Commonly word embedding can be jointly used in a neural network model as an embedding layer which helps to enhance deep learning performance. The output of the machine-learning/deep-learning model can be either binary decision (e.g., hate versus non-hate speech) or multi-class output where the model discriminates various type of hate speech and non-hate speech.

(iv) Evaluation: is the final part of the text classification pipeline where the performance of the machine learning/deep learning model is estimated. Several evaluation metrics are used for this purpose: accuracy, F1 score, precision, Matthews Correlation Coefficient (MCC), receiver operating characteristics (ROC), area under the ROC curve (AUC).

3. List of Acronyms

In order to ease the readability and maintain the coherence of the various notations employed, we list in Table 3 the various acronyms and their complete form cited throughout this paper. This concerns mainly the machine learning, deep learning, and feature sets of techniques reviewed in this paper.

4. Related work

From a computer science point of view, the scientific study of hate speech is comparatively a new topic, for which the number of review papers in the field is limited. We found only a few survey/review articles during the process of literature review. These were obtained using a systematic review-based approach where we adopted PRISMA framework [90] and conducted a brief systematic review of previous reviews in HS as will be detailed in the following section.

4.1. Methodology for collecting related review papers

4.1.1. Keywords

To collect related review papers, we first selected the best keywords to retrieve relevant information from the search database (Google scholar[5] and ACM[6]). Hate speech is a new concept that became popular recently; therefore, we considered other relevant terms referring to particular hate speech types (e.g., cyberbullying, sexism, racism, and homophobia). The search keywords were: review/survey hate speech detection, review/survey Offensive Or abusive language detection, review/survey sexism detection, review/survey sexism detection, and review/survey cyberbullying detection.

4.1.2. Search for documents

We utilized Google Scholar and ACM digital library in order to search for systematic reviews containing the term "review/survey' with the keywords mentioned above in their titles and abstracts. At the same time, no date and language restrictions were imposed. The reason for not selecting "systematic" as a search keyword is

[5] https://scholar.google.com/
[6] https://dl.acm.org/

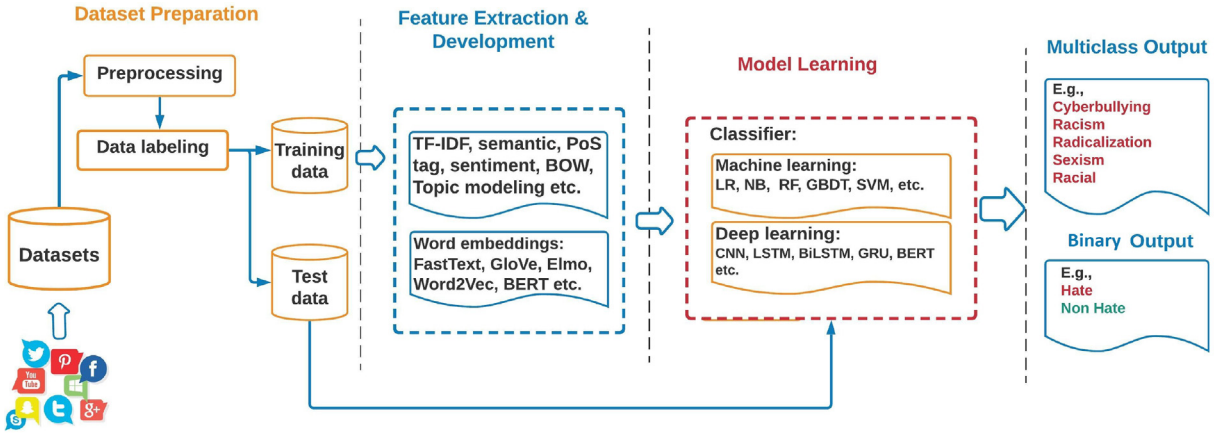


Fig. 2. Generic pipeline of automatic HS detection system.

Table 3
Acronyms and their full forms.

Acronym	Full form	Acronym	Full form
HS	Hate Speech	HT	Hate Text
CNN	Convolutional neural network	LSTM	Long short-term memory
GRU	Gated Recurrent Units. Its a gating mechanism in RNNN	BiLSTM	Bidirectional Long Short-Term Memory
BGRU	Bidirectional gated recurrent unit network	RNN	Recurrent neural
BERT	Bidirectional Encoder Representations from Transformers	RoBERTa	A Robustly Optimized BERT Pretraining Approach
DistilBERT	A distilled version of BERT: smaller, faster	ALBERT	A Lite BERT for Self-supervised Learning
mBERT	Multilingual BERT	TWilBert	A specialization of BERT architecture both for the Spanish and the Twitter domain
LR	Logistic Regression	NB	Naive Bayes
RF	Random forest	SVM	Support Vector Machin
TF-IDF	Term frequency-inverse document frequency	SG	Skip gram
GBDT	Gradient Boosting Decision Tree	ELMO	Embeddings from Language Models
BOW	Bag of word	CBOW	Continiuos bag of word
PoS Tag	Part-of-speech tagging	GHSOM	The Growing Hierarchical Self-Organizing Map

that we wanted to collect all reviews that were not only based on systematic methodology but also on narrative methods. The last search was run on 10 December 2022. The title, abstract, authors' names and affiliations, journal name, and year of publication of the identified records were exported to an MS Excel spreadsheet for further analysis.

4.1.3. Review of related review papers.

The above approach identified seven review papers. The study selection process is summarized in Fig. 3. While the initial literature search resulted 2312 records, 2263 were eliminated because either those were not review/survey documents related to HS and computer science and engineering (CSE) or duplicate from both databases. The full texts of the remaining 48 reviews were carefully screened, and 35 articles were excluded because those did not have enough information to consider as review/survey documents or not related to HS/CSE domains. The remaining thirteen review papers, which passed the eligibility test, were divided into two main categories: narrative and systematic. Typically, narrative reviews do not reveal the methodology of data collection, in contrast to systematic reviews. (See Fig. 4).

The survey of Schmidt and Wiegand [133] is the most cited one (more than 1000 citations) [7]. The paper follows a narrative method of analysis without revealing the data collection approach. The authors provided a short, comprehensive, structured, and critical overview of the field of automatic HS detection in NLP, highlighting key terminology and focusing on feature engineering relevant to HS/bullying identification, and finally, reviewing the current dataset and societal challenges.

Fortuna and Nunes [46] is the second most cited review paper and followed a systematic review-based approach. The authors presented a critical overview of how the textual automatic detection of hate speech evolved over the past few years. Their analysis proposed a unified and a clearer definition of the HS concept that can help build a model for the automatic detection of HS from machine learning perspective. Additionally, a comparison of performance of various HS detection algorithms is reported. Especially, they found that due to the lack of standards in the dataset, which could make the result biased. We found this review in term of its approach very relevant to our study. However, since this review was conducted at the end of 2017, an update literature is needed.

The survey in Tsapatsoulis and Anastasopoulou [173] focused on cyberbullying on Twitter where an emphasis was given on identifying Twitter abusers and indicated steps required to develop practical applications for the detection of Cyberbullers.

Al-Hassan and Al-Dossari [8] presented a summary of several discussed papers, organized according to their publication date. Their review covered i) English Anti-social behaviors; ii) English hate speech, and finally, iii) Arabic Anti-social behaviors. The provided tables can serve as a quick reference for all the key works performed on automatic HS detection on social media. The approaches and their respective experimental results were listed concisely. They also provided a summary of multilingual contributions directly related to hate speech, with a special focus to the Arabic language in social media platforms. Mishra et al. [87] exam-

[7] All citations mentioned in this paper were last searched on 10-December-2022

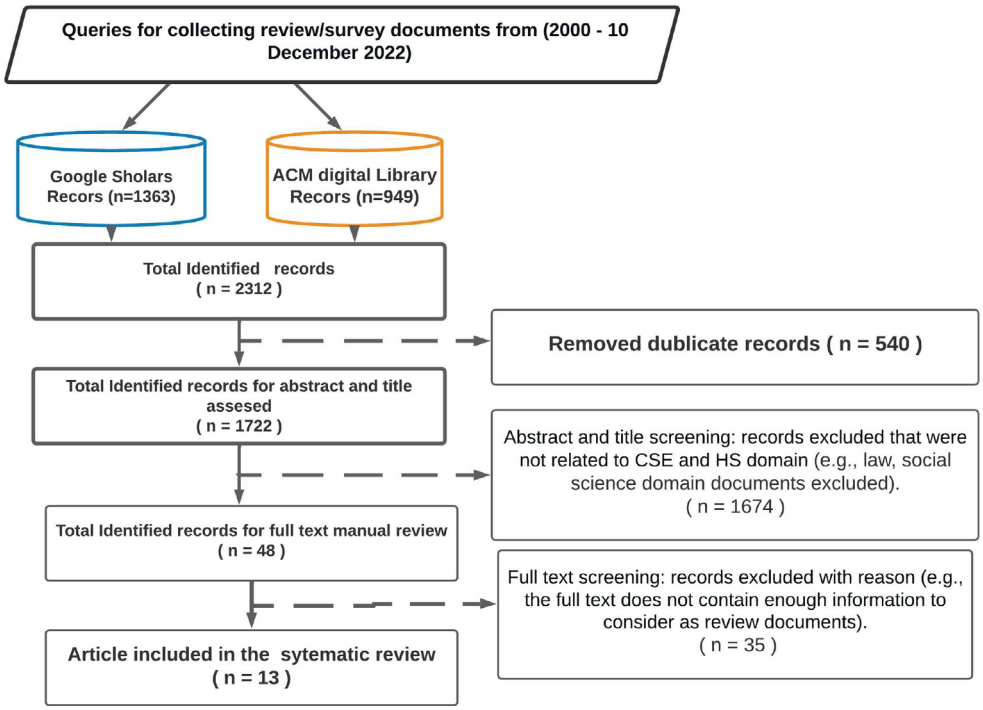


Fig. 3. PRISMA flowchart for selection of previous related review/survey documents.

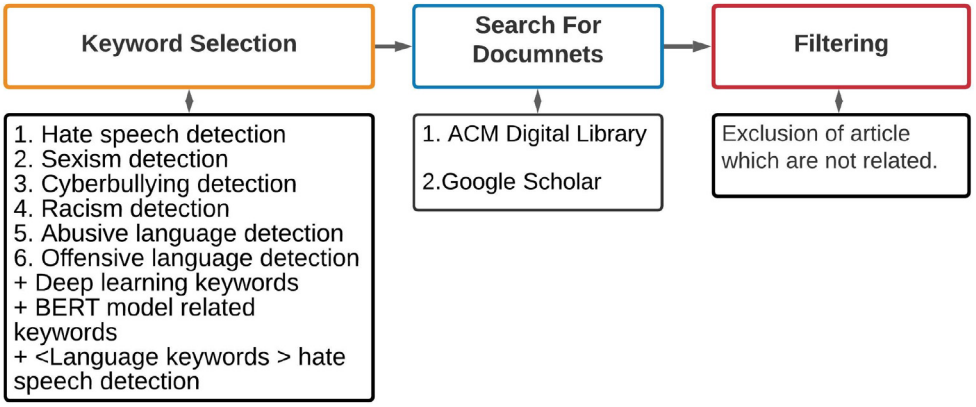


Fig. 4. Methodology for document collection.

ined the existing HS datasets and reviewed the computational approaches to abuse detection, analyzing their strengths and weaknesses, discussing the emerging prominent trends, while highlighting the remaining challenges and outlining possible solutions.

Poletto et al. [114] followed a systematic review approach to analyze existing HS resources including their development methodology, topical focus, language coverage, and other factors. The results of their analysis highlighted a heterogeneous, growing landscape marked by several issues and room for improvement.

Pradhan et al. [117] reviewed strategies for tackling the problem of identifying offense, aggression, and hate speech in user's textual posts as well as comments, micro-blogs, from Twitter and Wikipedia. Although the coverage of the dataset investigated was quite narrow.

Table 4 shows the result of previous reviews study. From 2000 to 2020 Only two of the identified reviews were found to follow a systematic-review based approach [46,114] in this field. Unfortunately, one of the review [46] was performed in late 2017 does not cover more recent work in the field. On the other hand, the work by [114] is mainly focused on HS datasets corpora. The rest of the identified review papers were not systematic. Some of them reviewed a small number of documents or targeted a specific part of this field (e.g., [173] focused on Twitter cyberbullying). However, we also do see a rise in the trend of systematic review papers published after 2020, and four systematic review papers have been added to the list in only two years of time. For instance, [167] reviewed 63 publicly available abusive training datasets and future direction, focusing mainly on HS dataset resources similarly to [114], while insightfully discussing some open issues in HS detection in overall as well as identifying the best practices for creating abusive content training datasets. Another work by [155] focused on resource handling, investigating how preprocessing techniques impact model learning. Authors in [146] surveyed work on deep learning methods (CNN, RNN, Transformmer etc) whiletargeting code-mixed Indian language in HS. Their Empirical results showed that pretrained multilingual transformer models with selective translation and transliteration achieved the best results compared to other models. Shruthi and Anil-Kumar performed bibliometrics

Table 4
Review of identified survey papers on automatic hate speech detection

Paper Title	Authors, Year	Publisher Name	Review Focus	Citation	Review Type
A Survey on Hate Speech Detection using Natural Language Processing	Schmidt and Wiegand [133] 2017	ACL	Features for HS Detection, anticipating alarming societal Changes, sata annotation, classification methods, etc.	1020	Narrative
A Survey on Automatic Detection of Hate Speech in Text	Fortuna and Nunes [46] 2018	ACM	How HS work evolved from past, definition of HS, and classification method.	683	Systematic
Cyberbullies in Twitter A focused review	Tsapatsoulis and Anastasopoulou [173] 2019	IEEE	Mainly for the cyberbullying on Twitter; emphasis was given to identifying Twitter abusers.	3	Narrative
Detection of Hate Speech in social networks: A survey on multilingual corpus	Al-Hassan and Al-Dossari [8] 2019	COSIT	Different category HS detection, multilingual HS detection, and HS in Arabic language.	112	Narrative
Tackling Online Abuse: A Survey of Automated Abuse Detection Methods	Mishra et al. [87] 2019	ACL	Describe the existing datasets and review the computational approaches of HS detection.	53	Narrative
Resources and benchmark corpora for hate speech detection: a systematic review	Poletto et al. [114] 2020	Springer	Primarily focused on HS datasets.	117	Systematic
A Review on Offensive Language Detection	Pradhan et al. [117] 2020	Springer	Finding best strategies for HS detection.	14	Narrative
Directions in abusive language training data, a systematic review: Garbage in, garbage out	Vidgen, Bertie and Derczynski, Leon [167] 2020	PlosOne (open access)	Reviews of 63 publicly available abusive training datasets and future direction.	159	Systematic
Towards generalisable hate speech detection: a review on obstacles and solutions	Yin, Wenjie and Zubiaga, Arkaitz [171] 2021	PeerJ CS	Summarise how generalisable existing hate speech detection models are and the reasons why hate speech models struggle to generalise.	59	Systematic
A Literature Review of Textual Hate Speech Detection Methods and Datasets	Alkomah, Fatimah and Ma, Xiaogang [140] 2022	MDPI	Study HS detection models and Dataset	7	Systematic
A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter	Naseem, Usman and Razzak, Imran and Eklund, Peter W [155] 2021	Springer	Case study of different pre-processing steps.	43	Narrative
Thirty years of research into hate speech: topics of interest and their evolution	Tontodimamma, Alice and Nissi, Eugenia and Sarra, Annalina and Fontanella, Lara [164] 2021	Springer	Bibliometrics and social network analysis of HS research	47	Systematic
A Survey of Recent Neural Network (NN) Models on Code-Mixed Indian Hate Speech Data	Dowlagar, Suman and Mamidi, Radhika [146] 2021	ACM	HS detection NN model performance analysis.	2	Narrative

and social network analysis of HS publications [164] and provided an overall HS domain conceptual map, although its lacks categorization and detailed content analysis. The lack of coverage in the aforementioned reviews and the need for literature up-to-date in the existing systematic reviews, together with a focus on the machine and learning perspective, were the main motivation for the current review work, which aims to fill in this gap.

Our approach is complementary to that of [46] in the sense that it follows a systematic review based approach as well but it presents a more up-to-date literature and bears also some key differences. First, we adopt PRISMA protocol for systematic review-based analysis, with two search databases (Google scholar and ACM digital library). Second, a special care during the scrutinizing phase is devoted to track machine learning and deep learning methods with their associated evaluation performance and dataset employed. Third, we track the timely evolution of records and methods. We also enumerate existing data collections, including multilingual more comprehensively than in previous studies. Fourth, we summarize existing open-source projects and valuable resources in the field, and finally, we highlight the key challenges and open research agenda.

5. Systematic literature review methodology for collecting hate speech documents

The previous Section 4 presented a brief systematic review of past review papers related to HS automatic detection. This section details our systematic literature review regarding HS automatic detection. For this purpose, we adopted PRISMA framework as Moher et al., [90], highlighting the keyword selection, Search sources, and filtering process.

5.1. Keyword selection

The first phase conducted was the keywords selection. Since hate speech is a concept that comprises broad hate categories, our search criteria were partitioned into six categories: hate speech, sexism, racism, cyberbullying, abusive, and offensive. This provides us the best chance of retrieving a significant number of relevant work. Besides, as we wanted to pay special attention to machine learning and deep learning-based methods, several related abbreviations and keywords have been accommodated and added to the keyword search (i.e., CNN, LSTM, RNN, BERT, etc.). Furthermore, 20 top-speaking languages [8] added in search keywords to retrieve multilingual works. A selected list of keywords is shown in Table 5:

5.2. Search sources

We used two different databases (ACM Digital Library and Google Scholar), aiming to gather the most significant number of records in the areas of computer science and engineering (CSE).

[8] The Ethnologue https://www.ethnologue.com/, one of the trusted platforms regarding language information based on more than a thousand bibliography references.

Table 5
Keywords list for the query search.

Different type HS keywords	Deep learning Keywords	Deep learning BERT model related keywords	Language keywords
1. Hate speech detection 2. Sexism detection 3. Cyberbullying detection 4. Racism detection 5. Abusive language detection 6. Offensive language detection	1. Deep learning hate speech 2. CNN hate speech 3. LSTM hate speech 4. RNN hate speech 5. Deep learning Cyberbullying 6. CNN Cyberbullying 7. LSTM Cyberbullying 8. Deep Learning Offensive 9. CNN Offensive 10. LSTM Offensive	1. BERT hate speech 2. BERT Cyberbullying 3. BERT Abusive 4. BERT Rasism 5. BERT Sexism 6. BERT Offensive	Chinese (Mandaren, Yue, Wu), Hindi, Spanish, Arabic, Bengali, French, Russian, Portuguese, Urdu, Indonesian, German, Japanese, Marathi, Telugu, Turkish, Tamil, Korean.

This is motivated by the availability of search through an API, allowing the application of simple NLP modules to identify duplication and check string matching as well as record statistical trends. On the other hand, our desire to focus on computer science aspect of HS detection makes ACM library as an ideal candidate for search database, while Google scholar expects to identify all other relevant and high impact results outside ACM community.

The input database search consists of an OR-logical combination of keywords of individual categories (hate speech OR sexism OR racism. . .etc.), see Table 5 for detailed listing of such keywords. Besides, this process has been automated by utilizing the beautifulsoup python web crawler for both databases (ACM and Google Scholar) by monitoring the API output, which consists of the paper's title, publication year, abstract, author's name, publisher information, citation, and link to the full article. We have made our scraping code publicly available for the community [9]. Initially, we collected papers from 2000 to 2021, and no language restrictions were imposed since we wanted to gather multilingual work associated with hate speech detection. The last search for article collection was run on 18 March 2021.

The title, publication year, abstract, author's names, publication venue, and link of full papers were exported to an MS Excel spreadsheet for further analysis.

5.3. Filtering Documents

The PRISMA flow chart highlighting the inclusion and exclusion criteria for the document search and inclusion in the database is summarized in Fig. 5. Initially, 44,030 documents were collected from 2000 to 2021. Since we have collected data from two different databases, duplicate papers have been removed automatically from the system, leaving 33670 records for further title and abstract scrutinizing. In this respect, most documents that were not related to CSE fields and not associated with different hate speech categories (general hate, cyberbullying, abusive, offensive, sexism, racism, etc.) were excluded after the title and abstract screening.

The remaining 1329 papers were considered for full-text review. Two independent reviewers with knowledge in this field have carefully performed this manual scrutinizing task. Those articles that were not related to hate speech or CSE fields (e.g., LAW, Psychology fields) were discarded. Similarly, descriptive and conceptual papers that fail to produce any solid results were also ignored. During this phase, disagreements between the two reviewers were discussed and resolved by consensus. If no agreement could be reached, the views of a third reviewer would have been taken into consideration.

Finally, 463 articles were considered for the final systematic review and analysis. Among those 463 papers, 96 papers have been found to follow deep learning methods.

6. Systematic review results

6.1. Number of publications per year

As we can see in Fig. 6, a total of 463 papers were identified from 2000 to 2021 (including deep learning and all other methods). Before 2010, we have found only 1 document related to hate speech. From 2010 to 2016, only 25 papers associated with HS detection were found, yet there was no work related to deep learning. However, since 2017 the number of published documents raised rapidly with a steady increase of deep learning based HS detection approach. A total of 96 documents were found from 2017 to 2021 using deep-learning HS detection, indicating a trend of almost doubling the number of deep learning approach each year. The relatively small value in 2021 is due to the fact that the collection of new documents stopped in March 18, 2021.

6.2. Publication Venue

We have scrutinized the obtained records with respect to publication venues in an attempt to identify any dominating trend. From the total of 463 identified documents in textual HS automatic detection, we have found 72 different venues. The publication venues with more than 4 occurrences in our collection are presented in Fig. 7. The most common platforms for publication of hate speech documents were ACLWEB[10], ArXiv[11], IEEE[12], Springer[13], and ACM[14]. The Association for Computational Linguistics (ACL) is

[9] https://github.com/saroarjahan/Google_sholars_ACM_digital_library_crawler

[10] https://www.aclweb.org/portal/
[11] https://arxiv.org/
[12] https://www.ieee.org/
[13] https://www.springer.com/gp
[14] https://www.acm.org/

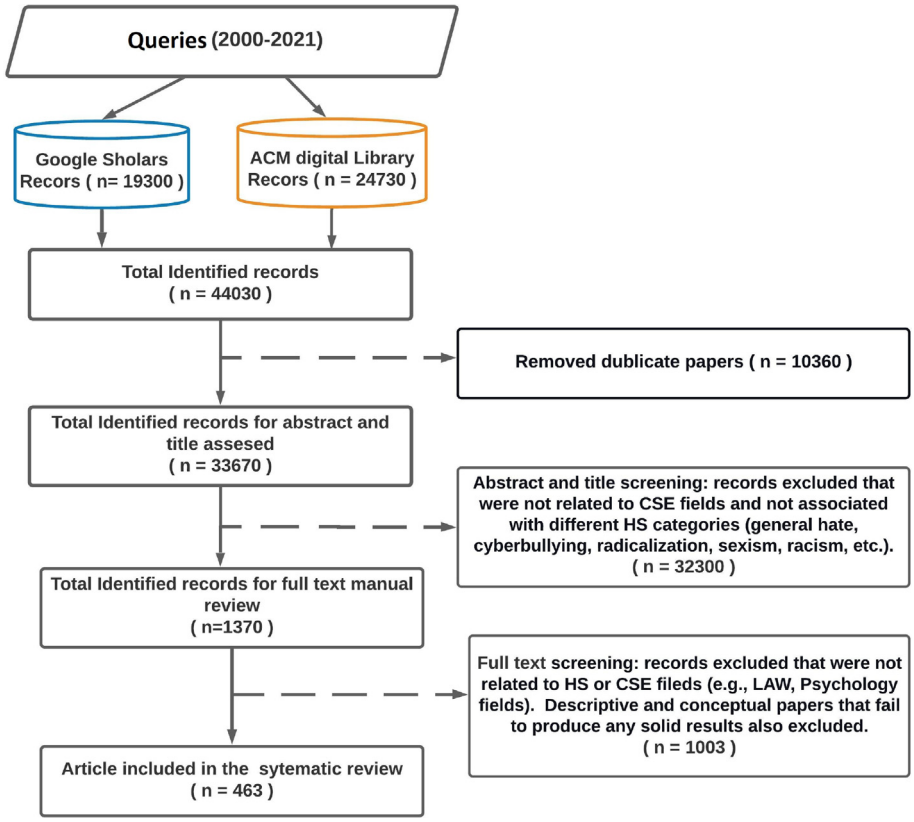


Fig. 5. PRISMA flowchart for the selected studies for systematic review.

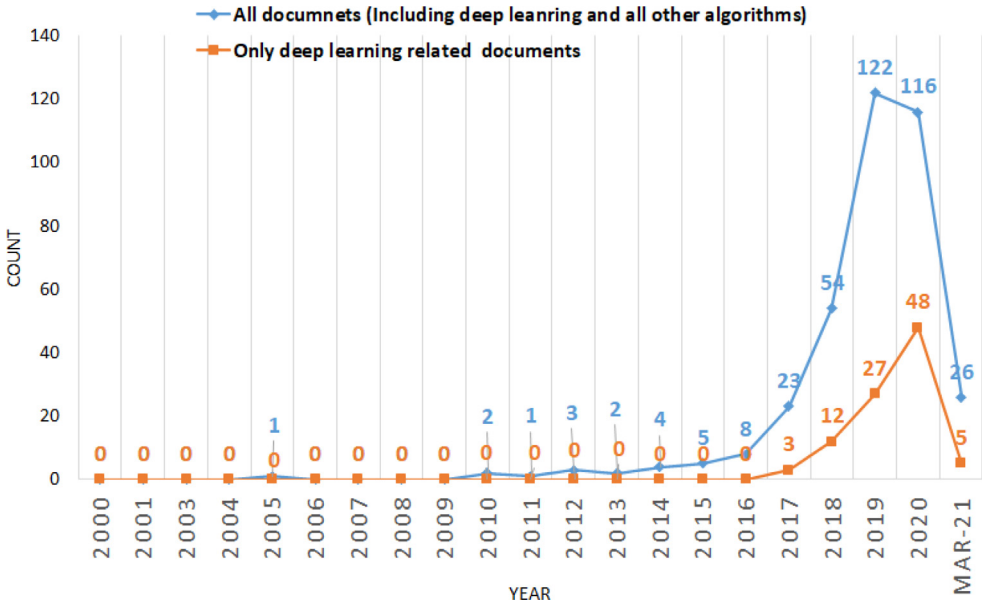


Fig. 6. Number of publications per year from 2000–2021 related automatic hate speech detection in NLP (blue line represent all 463 documents including deep learning and other ML approach, and yellow line represent 96 documents related to Deep-learning method).

the premier international scientific and professional society working on NLP's computational problems. Therefore, a vast number of papers were published in ACL-WEB forums. The second most popular source was Arvix, an open-access repository of electronic preprints. This can partially be explained by the fact that the hate speech detection area has become popular, with many autonomous and exploratory work being conducted. Additionally, this high number of publication venues reveals that HS automatic detection is not limited to a few sources of publication venue and testifies the field's multidisciplinary nature.

6.3. Categorization according to methods employed and detection performance

This section first reviews the results (identified documents) in terms of machine learning and feature employed, the platform

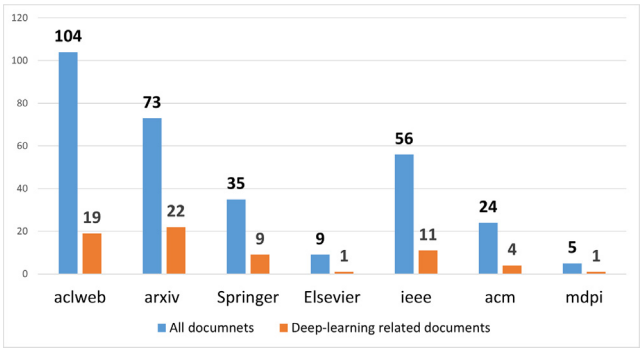


Fig. 7. Most Used Data Platforms (blue histogram represent all 463 documents including deep learning and other ML approach, and yellow histogram represent 96 documents related to Deep-learning method).

used for dataset collection, category of hate speech investigated, and bibliometric data of the publication[15] as well as performance metrics (Accuracy (Acc), Precision (P), Recall (R)) claimed by the authors. After that, we evaluated the state-of-the-art of deep-learning methods related to HS automatic detection.

6.3.1. Statistical trends of results

Tables 6 and 7 present some characteristics of selected highly cited papers, organized according to their publication date. The tables can serve as a quick reference for all the key works performed in textual automatic HS detection. The employed methods and related experimental results in terms of Precision, Recall, or F-measure are listed concisely. Table 6's attributes include data platform, type of hate speech, type of machine learning, features, and performance results. One notices, for instance, the dominance of Twitter and YouTube as the main source of data and supervised machine learning algorithms as the dominant machine learning type. Whereas Table 7 repeats the previous process for non-English and multilingual textual data. Furthermore, we have plotted statistical figures regarding all the 463 documents to understand the overall trend according to language, types of HS, data collection platform, ML approach, features, and algorithm used for HS detection.

Fig. 8 illustrates the proportion of the various languages of the textual input analyzed by HS automatic detection algorithms in the identified records. As expected, English textual source is by far the most investigated. This is rationally justified by the fact that initial work in the field has started with English textual dataset as in Dinakar et al. [37], but also due to the maturity of the natural language Parsers developed for English language as well as the abundance of benchmarking dataset that enable researchers to carry out useful comparative results. Nevertheless, with the advances in multilingual parsers and deep learning technology, together with increasing pressures from policy-makers to handle hate speech issues at local resources, non-English HS detection toolkits have seen a steady increase. The figure indicates that about 51% of all works in this field are performed on English dataset, with an increase of proportion of other languages as well where Arabic (13%) [93,60,12,176], Turkish (6%) [176,104], Greek (4%) [176,6,136], Danish (5%) [106,176], Hindi (4%) [121,22,88], German (4%) [73,120], Malayalam (3%) [130,109], Tamil (3%) [130,20], Chinese (1%) [138,139,188], Italian (2%) [116], Urdu (1%) [126,95,7], Russian(1%) [17], Bengali (1%) [63,127,70], Korean (1%) [91], French (1%) [16,102,51], Indonesian (1%) [14], Portuguese (1%) [14], Spanish (1%) [57] and Polish (1%) [118] seem to dominate the rest of the languages in this field.

[15] All citations last updated on 03-Mar-2021

It is worth noticing that despite Chinese being the 2nd largest speaking language globally, it has been much less investigated in HS detection community. One reason could be the lack of Chinese language in HS competition workshop such as SemEval2020 [192] and Hasoc2020 [82] where multilingual tasks included English, Danish, Greek, Arabic, Tamil, Malayalam, Hindi, German and Turkish, which encouraged many participants all over the world to work in these languages.

Fig. 9 depicts the percentage of different HS categories in the identified records. We can see that publications related to 'general hate' (36%) are a dominant trend followed by 'abusive language' 23% of total records. Cyberbullying and Radicalization categories share the same percentage of (15%) each. While relatively a small percentage is assigned to religion (5%), racial (3%), and sexism (3%) associated hate speech categories.

From Fig. 10, one can notice that 47% of dataset employed in HS detection was collected from Twitter social network, followed by Facebook (12%), Youtube (9%) and Wikipedia (5%). This indicates a growing trend in research community to tackle the occurrence of hate speech in social media forums as social media platforms constitute by far the dominant agora of hate speech because of easy access, fast spread and societal impact. The rest of the dataset sources, with low occurrence rate, were mainly investigated for comparative purpose and benchmarking. For instance, [110] used Formspring dataset for cyberbullying detection and developed methods that enhance question–answer systems. In Section 7, we discuss in detail the current state of the art of HS dataset.

Fig. 11 provides a global trend in terms of types of machine learning approaches employed in our identified records. Among the ML approaches, we distinguish supervised, semi-supervised, and unsupervised like approaches. The analysis revealed that most of the works adopted supervised methods (73%). From Tables 6 and 7, we can observe that any of these three methods can achieve high-performance accuracy, and there is no substantial evidence to favor one over another one, whereas only the context of data (e.g., availability and quality of training samples) can play a role in deciding about the suitability of one category over another one. For example, [31] used an unsupervised method and lexical & syntactic features to achieve 98% accuracy. Similarly, several works were based on supervised and semi-supervised methods that have shown close or better performance [21,2] . Nevertheless, it is worth mentioning the popularity of the supervised like approach over other ML approaches, possibility due to the multiplication of benchmarking dataset and machine learning/ deep learning platforms that promote supervised approach.

Fig. 12 and 13 depict the percentage of ML algorithms and features employed in the identified records, respectively. The SVM method emerges as the most popular HS detection model covering 29% of total records. The use of deep learning models to HS started to rise from 2017 [21,107] to quickly cover about 22% of total identified records. On the other hand, LR (20%) and NB (14%) were also among popular ML methods investigated by the researchers. We also noticed that in many deep-learning related methods, non-deep learning models were often employed as baseline to compare the performance of the investigated deep-learning model [40,9,64,21].

Fig. 13 shows that TF-IDF based features cover 29% of the total records. However, word embedding models, which have widely being used in deep learning embedding layers, cover 33% of the entire records. The PoStag (3%), topic modeling (3%), and sentiment (3%) features were the least used features. This suggests that the deep-learning models and embedding features seem comparatively popular and widely used by the community. Next, we explored the popularity of the various deep-learning architectures (e.g., CNN, LSTM, BiLSTM, etc.) for HS automatic detection. The results are outlined in the next Section 6.3.2.

Table 6
Summary of key contributions for English HS detection and their performance in terms of Precision (P), Recall (R), F1-Score (F), Citation (C)

Author, Year	Platform	Type	ML Approach	Features Representation	Algorithm	P	R	F	C
Chen et al. [31] 2012	YouTube	Abusive	Un-Supervised	Lexical and syntactic	Match Rules	0.98	0.94	-	679
Xiang et al. [187] 2012	Twitter	Abusive	Semi-Supervised	Topic modelling	LR	-	-	0.84	298
Dinakar et al. [37] 2012	YouTube	Cyberbullying	Supervised	Tf-idf, lexicon, PoS tag, bigram	SVM	.66	-	-	448
Warner and Hirschberg [177] 2012	Yahoo, news-group	Radicalization	Supervised	Template-based, PoS tagging	SVM	.59	.68	.63	684
Wadhwa and Bhatia [175] 2013	Twitter	Radicalization	Un-Supervised	Topic identification, N-grams	Topic-entity mapping	-	-	-	40
Kwok and Wang [76] 2013	Twitter	Racism	Supervised	Unigram	Naïve Bayes	-	-	-	445
Nahar et al. [96] 2014	Myspace, Slashdot	Cyberbullying	Semi-Supervised	Linguistic features	Fuzzy SVM	.69	.82	.44	92
Burnap and Williams [28] 2014	Twitter	Hate	Supervised	BOW, Dependencies, Hateful Terms	Bayesian LR	.89	.69	.77	112
Agarwal and Sureka [5] 2015	Twitter	Radicalization	Semi-Supervised	Linguistic,Term Frequency	KNN, SVM	-	-	.83	146
Gitari et al. [54] 2015	Blog	Hate, Weakly hate, Strongly hate	Semi-Supervised	Lexicon, Semantic, theme-based features	Rule based	0.73	0.68	0.70	414
Djuric et al. [39] 2015	Yahoo	Finance Hate,	Supervised	Paragraph2vec,	CBOW, LR	-	-	-	643
Waseem and Hovy [179] 2016	Twitter	Hate	Supervised	Character n-grams	LR	0.72	0.77	0.73	1333
Di Capua et al. [36] 2016	YouTube,Form-Spring, Twitter	Cyberbullying	Un-Supervised	Semantic and syntactic features	GHSOM network and K-mean	.60	.94	.74	83
Park and Fung [107] 2017	Twitter	Abusive	Supervised	Character and Word2vec	Hybrid CNN	0.71	0.75	0.73	2347
Chen et al. [30] 2017	Youtube, Myspace, SlashDot	Abusive	Supervised	Word embeddings	FastText	-	0.76	-	21
Badjatiya et al. [21] 2017	Twitter	Sexist, Racist	Supervised	FastText, GloVe Random Embedding,Tf-IDF, BOW	LR, SVM CNN, LSTM and GBDT	0.93	0.93	0.93	992
Wiegand et al. [182] 2018a	Twitter, Wikipedia, UseNet	Abusive	Supervised	Lexical,linguistics and word embedding	SVM	.82	.80	0.81	124
Pawar et al. [110] 2018	Form-spring	Cyberbullying	Supervised	BOW	M-NB and Stochastic Gradient Descent	-	-	.90	15
Watanabe et al. [180] 2018	Twitter	Hate, Offensive	Supervised	Sentiment-Based, Semantic, Unigram	J48graft	0.79	0.78	0.78	254
Malmasi and Zampieri [81] 2018	Twitter	Hate, offensive	Supervised	N-grams, Skip-grams, hierarchical, word clusters	RBF kernel, SVM	0.78	0.80	0.79	205
Pitsilis et al. [113] 2018	Twitter	Racism or Sexism	Supervised	Word-based frequency, vectorization	RNN and LSTM	0.90	0.87	0.88	161
Fernandez and Alani [45] 2018	Twitter	Radicalization	Supervised	Semantic Context	SVM	.85	.84	.85	27
Ousidhoum et al. [102] 2019	Twitter	Sexual orientation, Religion, Disability	Supervised	BOW	LR, biLSTM	-	-	94	157
Zhang and Luo [193] 2019	Twitter	Racism, Sexism	Supervised	Word embeddings	CNN + GRU	-	-	0.94	224

6.3.2. Overview of Deep-learning records

Our systematic review has identified 96 documents related to the application of deep-learning technology/models to the task of automatic hate speech detection. In the sequel, we have analyzed two crucial aspects: the architecture of the deep learning model and the features employed.

Fig. 14 summarizes the finding in terms of the percentage of the various deep-learning algorithms employed. Notice that BERT (33%) becomes prevalent though it was only introduced recently in 2019. The next most popular deep-learning models are LSTM and CNN, which covered 20% and 12% of total identified records. Hybrid models (combination of multiple models) depicted in the plot are exemplified by BERT + CNN (2%), LSTM + CNN(9%), LSTM + GURU(1%) and BERT + LSTM(2%).

In Table 8, we tracked different architectures of deep learning employed in the identified records. Surprisingly a total of 24 different deep learning models and feature combinations were found, possibly highlighting the diversity of research's attempts to produce novel architectures building on existing HS detection architectures. Most of the architectures used two steps: (i) word embedding layer employing models such as Word2Vec, FastText, GloVe; (ii) deep learning layer, where one distinguishes, among others, CNN, LSTM, GRU architectures. (See Table 9).

During the course of this analysis, four significant insights are distinguished:

I) **Comparison between non deep-learning and deep-learning models** revealed that deep-learning models outperformed popular classifiers (NB, LR, RF, SVM) in most studies [40,9,64,21]. An early work by Badjatiya et al. [21] has compared HS detection with different ML models (LR, SVM, GBDT) showed that deep learning models using either CNN

Table 7
Summary of key contributions for non-English language in HS detection.

Author, Year	Platform	Language	Class	ML Approach	Features Representation	Algorithm	P	R	F	C
Abozinadah et al. [4] 2015	Twitter	Arabic	Abusive	Supervised	Profile and tweet-based features, bag of words, N-gram, TF-IDF	Naïve Bayes	0.85	0.85	0.85	63
Magdy et al. [79] 2015	Twitter	Arabic	Terrorism (Pro-ISIS and Anti-ISIS)	Supervised	Temporal patterns, Hashtags	SVM	0.87	0.87	0.87	132
Kaati et al. [67] 2015	Twitter	Arabic	Terrorism (Support or Oppose Jihadism)	Semi-Supervised	Data dependent features and data independent features.	AdaBoost	0.56	0.86	0.86	61
Abozinadah [2] 2016	Twitter	Arabic	Abusive	Un-Supervised	Lexicon, bag of words (BOW), N-gram	SVM	0.96	0.96	0.96	25
Abozinadah and Jones Jr [3]	Twitter	Arabic	Abusive	Supervised	PageRank (PR) algorithm, Semantic Orientation (SO) algorithm, statistical measures.	SVM	0.96	0.96	0.96	28
Mubarak et al. [93] 2017	Twitter, Arabic News Site	Arabic	Abusive, Offensive	Un-supervised	unigram and bigram, Log Odds Ratio (LOR), Seed Words lists None.	Just performed extrinsic evaluation	0.98	0.45	0.60	236
Haidar et al. [59] 2017	Facebook, Twitter	Arabic	Cyber- bullying (Yes, No)	Supervised	Tweet to SentiStrength, Feature Vector	SVM	0.93	0.94	0.92	67
Abdelfatah et al. [1] 2017	Twitter	Arabic	Violent (Violent, Nonviolent)	Un-supervised	Sparse Gaussian process latent variable model, morphological features Vector Space Model	K-means clustering	0.56	0.60	0.58	16
Alfina et al. [14] 2017	Twitter	Indonesian	Hate (Hate, Non-hate)	Supervised	BOW and n-gram	Random Forest	-	-	0.93	138
Özel et al. [105] 2017	Twitter, Instagram	Turkish	Hate	Supervised	BOW	M-Naïve Bayes	-	-	0.79	61
Alakrot et al. [11] 2018	YouTube	Arabic	Offensive, In-offensive	Supervised	N-gram	SVM	0.88	0.80	0.82	71
Albadi et al. [13] 2018	Twitter	Arabic	Religious hate, Not hate	Supervised	Word embeddings(AraVec)	GRU-based RNN	0.76	0.78	0.77	136
Alshehri et al. [15] 2018	Twitter	Arabic	Adult, Regular user	Supervised	Lexicon, N-grams, bag-of-means (BOM)	SVM	0.70	0.93	0.78	22
Jaki and De Smedt [65] 2019	Twitter	German	Radicalization (Muslim, Terrorist, Islamo fascistoid)	Un-Supervised	Skip grams and Character trigrams	K-means, single-layer averaged Perceptron	0.84	0.83	0.84	45
Alami et al. [12] 2020	Twitter	Arabic	-	Supervised	-	AraBERT	90	-	-	11
Sai and Sharma [130] 2020	Twitter	Tamil-English, Malaylam-English (Code-mix)	-	Supervised	-	XLM-RoBERTa + mBERT	TENG 90, MENG 77	-	-	16
Polignano et al. [116] 2019	Twitter	Italian	-	Supervised	-	AlBERTO	90	94	-	10
Wang et al. [176] 2020	Twitter	English Turkish Arabic Danish Greek	-	Supervised	-	XLM-RoBETa base and lagre	92 82 90 81 83	-	-	24

or LSTM model performed on average 13–20% better that LR, SVM or GBDT models. Another recent work by Al-Hassan and Al-Dossari [9] compared SVM as a baseline model to CNN, CNN + LSTM, GRU, CNN + GRU. The authors found that, in all cases, CNN outperformed the baseline model by at least 7% in terms of accuracy.

I) **Comparison between CNN, LSTM, BiLSTM, and GRU models** revealed mixed results in terms of which deep-learning architecture performs most. For instance, a comparison between CNN and LSTM architecture showed a better CNN performance results in [64], whereas Badjatiya et al. [21] found that LSTM architecture performs better than CNN. Besides, the difficulty in data full reconstruction, the difficulty to reproduce exact preprocessing stages or use of distinct embedding can render such comparison more challenging. Yin et al. [189] conducted a comparative study between two deep-neural network architectures: "RNN (with GRU and LSTM layers) and CNN". They found that RNN is more suited for the long-ranged context dependencies, while CNN sounds better in extracting local features. They also revealed that GRU performs better in case of long sentences. However, several other papers suggested that the

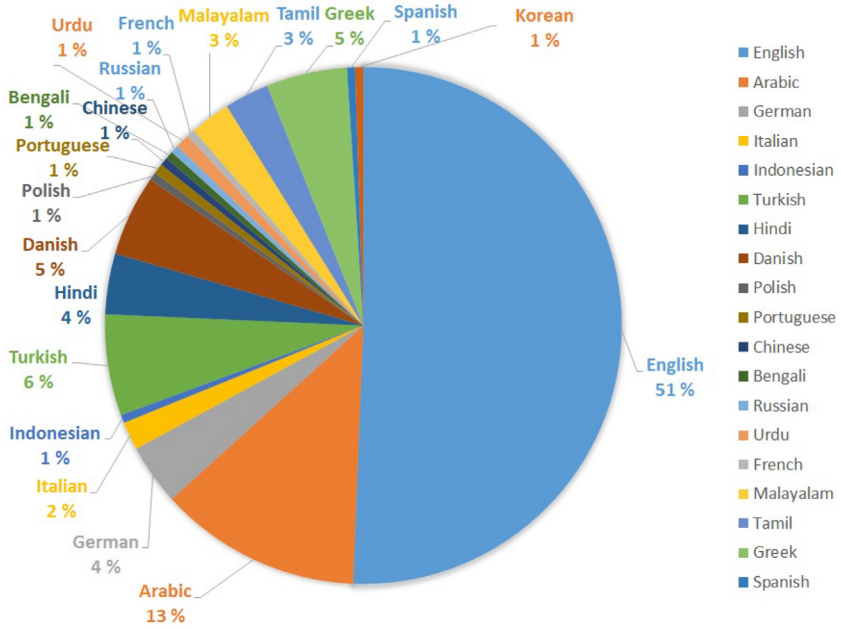


Fig. 8. Statistics of the percentage of previous HS wrok in different language.

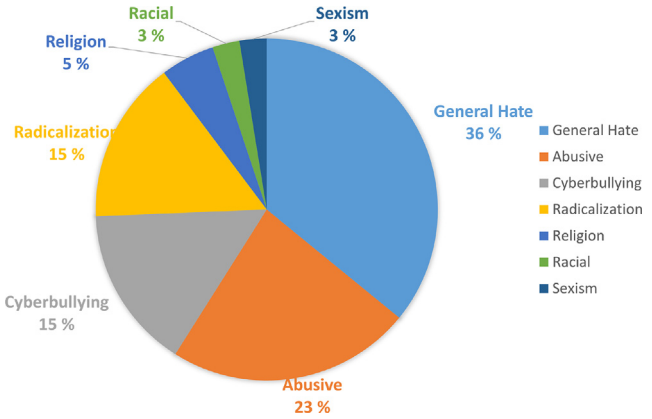


Fig. 9. Statistics of the percentage of HS work in different categories.

concatenation of two or more deep learning models perform better than a single deep learning model [9,194,69,134,72]. For example, CNN + LSTM and CNN + GRU both performed better than the single application of LSTM and CNN [9]. Zhou et al. [194] suggested a fusion of three CNN models with different parameters as a viable way to improve the performance of hate speech detection. Kapil and Ekbal [69] compared CNN, LSTM, and LSTM + GRU architectures for HS detection in five commonly used datasets, and concluded that in all cases, LSTM + GRU outperform a single LSTM and CNN model by 2–3%. Similarly, Shruthi and KM [134] proposed a hybrid fusion architecture BiLSTM + Random Embedding + TF-IDF + LR that achieves at least a 12% increase in performance compared to a single architecture.

III) **Comparison between word embedding** revealed a lack of comparative studies in this area as well. However, from Table 8, we can see that Word2vec and FastText were regularly used with different deep learning architectures. Although this popularity does not entail systematically a better performance score. For instance, Saleh Alatawi et al. [131] compared word2Vec, GloVe, and Google NewsVec, and showed that word2Vec performed only 1% better than others. However, when compared to BERT model, BERT-Large showed the best accuracy and F1 score. Another work by Rizos et al. [125] comparing FastText, GloVe and Google-NewsVec revealed no significance accuracy among these embeddings.

On the other hand, since Word2Vec, FastText, GloVe use a vector representation to represent words in a way that captures semantic or meaning-related relationships as well as syntactic or grammar-based relationships, this also bears inherent limitation in the sense that this cannot capture polysemy relationship. That is, for the same word, even if it has different meanings in different contexts, the corresponding vector representation is unchanged. We shall mention the merits of the recently introduced ELMO word embedding model, which was designed to overcome the aforementioned shortcoming. A recent work by Zhou et al. [194] using ELMO embedding showed a better performance compared to CNN. However, since ELMO is a relatively new, the in-depth comparison with other embedding model is still in its infancy. This leaves the door wide for future experiments.

IV) **The rise of BERT** is a striking trend that can be seen in Fig. 15, which testifies of its popularity in hate speech detection community (38% share of deep-learning models) in the past five years. This demonstrates the importance of this model as a key state-of-the-art method in the field. Several works explored BERT performance in HS detection [98,122,131,40] where almost all authors who compared BERT model to other deep learning models concluded on the superiority of BERT architecture. For instance, Ranasinghe et al. [122] compared BERT to FastText, CNN, and LSTM. Similarly, Saleh Alatawi et al. [131] compared BERT to BiLSTM and LSTM where BERT showed a significance performance increase. Also, BERT model achieved top performance in the multilingual tasks in [40,116,12,130,176].

Some research performed a comparison among different BERT models. For instance, Wang et al. [176] found that BERT-large model outperformed BERT-base model for HS detection. In addition, two different BERT architecture con-

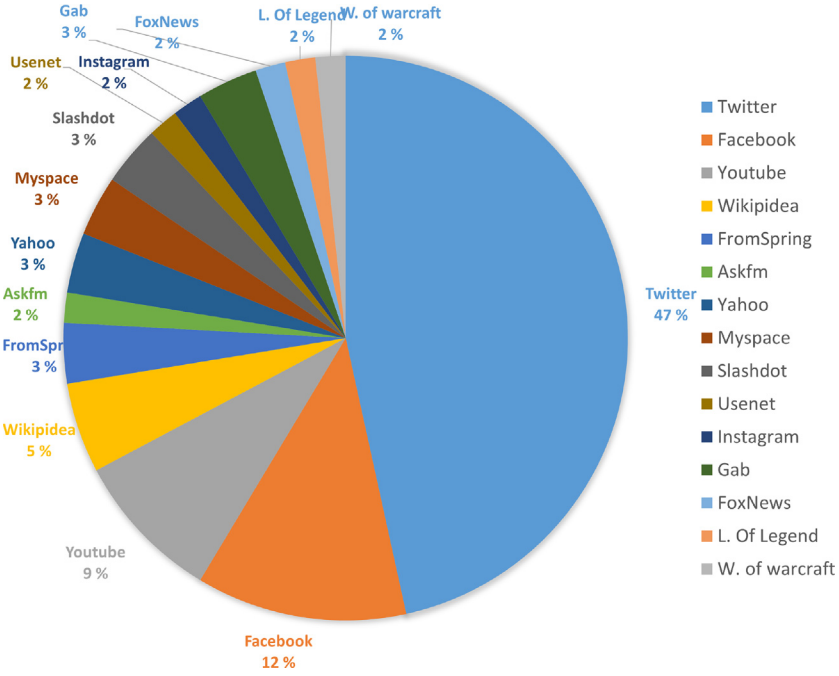


Fig. 10. Statistics of platforms used for data collection.

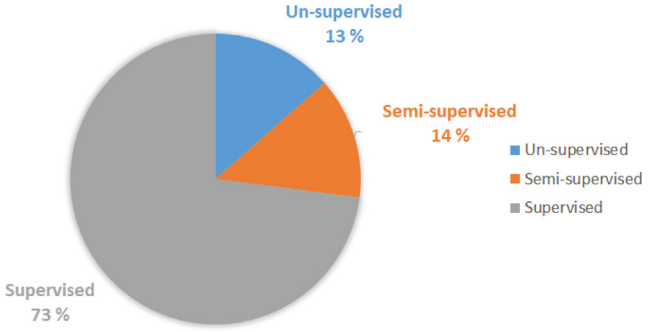


Fig. 11. Statistics on types of ML approach used for HS detection (e.g., supervised, semi-supervised or unsupervised).

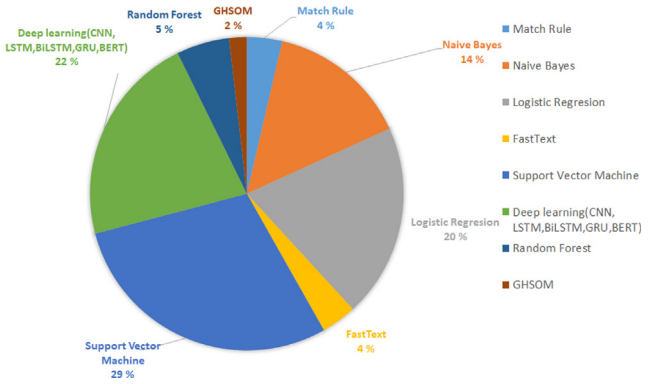


Fig. 12. Statistics of algorithm types used for HS detection.

catenations (e.g., XLMR-B + mBERT) are found to perform better than a single pretrained BERT[130]. Furthermore, BERT architecture can be trained in a such a way that it can perform a specific task. For example, Gonzalez et al. [57] proposed TWiLBERT, a specialization of BERT architecture for the Spanish and Twitter domain. They performed an extensive evaluation of TWilBERT models on 14 different text classification tasks, such as irony detection, sentiment analysis and emotion detection. The results obtained by TWilBERT outperformed the state-of-the-art systems and mBERT. Another work by Caselli et al. [29] introduced HateBERT, a re-trained BERT model for abusive language detection in English. The model was trained on a large-scale dataset of Reddit comments in English from communities banned for being offensive, abusive, or hateful. In all cases, HateBERT outperformed the corresponding general BERT model.

Besides, some other language-specific BERTs models developed over time for monolingual outperformed multilingual model mBERT: AraBERT (Arabic) [18], AlBERTo (Italian) [115], FinBERT (Finnish) [19], CamemBERT(French) [84], Flaubert (French [77]), BERT-CRF (Portuguese) [137], BERTje (Dutch) [174], RuBERT (Russian) [75] and BERTtweet (A pre-trained language model for English Tweets) [97]. However, to best of our knowledge, not every model has yet been tested for HS domain except AraBERT [12,38] and AlBERTo [116] which shown better performance for HS detection.

Table 10 summarizes selected key works from three major recent HS detection competitions SemEval-2019 [191], SemEval-2020 [192], and Hasoc-2020 [82].

In SemEval-2019, Task A (offensive language detection) was the most popular sub-task with 104 participating teams. Among the top-10 teams, seven used BERT with variations in the parameters and in the pre-processing steps. The top-performing team Liu et al. [78] used BERT-base-uncased with default-parameters, but with a max sentence length of 64 and trained for 2 epochs and achieved 82.9% F1 score which was 1.4 points better than Nikolov and Radivchev [98]. Although, the difference between the next five systems, ranked 2–6, is very marginal (less than one percent (81.5%-80.6%)). The top nonBERT model by Mahata et al. [80] was ranked fifth. They used an ensemble of CNN and BLSTM + BGRU, together with Twitter word2vec embeddings and token/hashtag normalization.

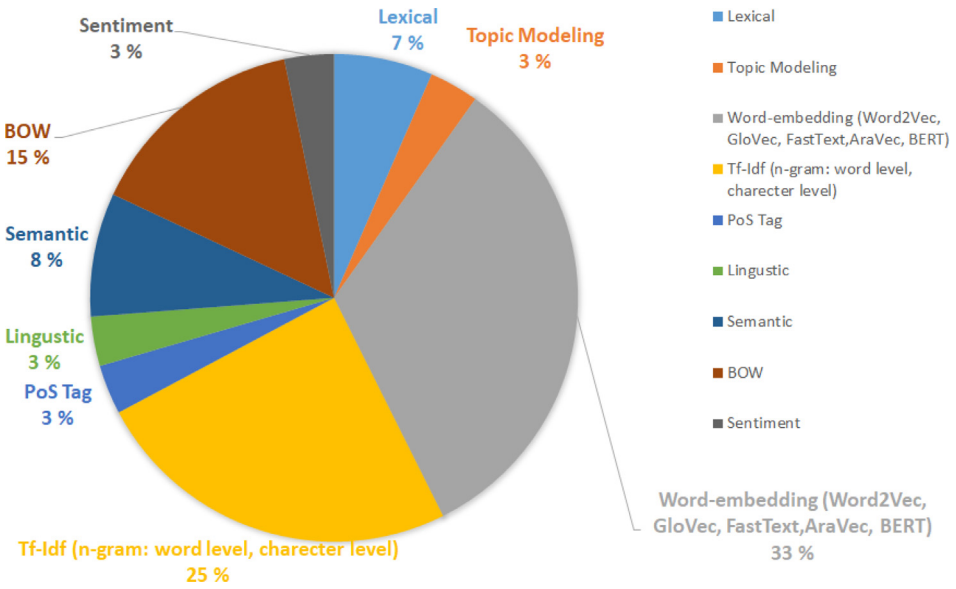


Fig. 13. Statistics of features employed by the ML/deep learning algorithms.

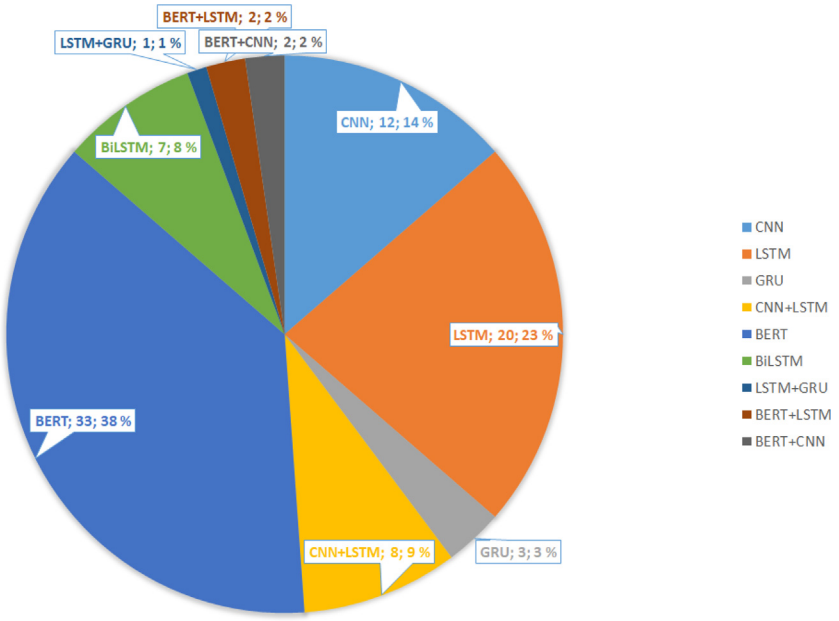


Fig. 14. Statistics of previous work based on the percentage of different deep learning algorithms used (e.g., CNN, LSTM, BERT etc).

In semEval-20, 145 teams submitted official runs on the test data and 70 teams submitted system description papers. The best team teamWiedemann et al. [181] achieved an F1 score of 0.9204 using an ensemble of ALBERT models of different sizes. The second team Wang et al. [176] achieved an F1 score of 0.9204, and it used RoBERTa-large that was fine-tuned with the dataset in an unsupervised way. The third team Dadu and Pant [33], achieved an F1 score of 0.9198, using an ensemble that combined XLM-RoBERTa-base and XLM-RoBERTa-large trained on Subtask A data for all languages. The top-10 teams were close to each other and employed BERT, RoBERTa or XLM-RoBERTa models; sometimes CNNs and LSTMs were also been mentioned either for comparison or hybridization purpose.

Over 40 research groups participated in Hasoc-2020 competition. The top ranked submission for Hindi-hate speech detection, used a CNN with FastText embeddings as input [121]. The best performance for German hate speech detection task was achieved using a fine-tuned versions of BERT, DistilBERT and RoBERTa [73]. Similarly, the top performance in English-language HS detection was based on a LSTM architecture with GloVe embeddings as input [89].

7. Resources for hate speech detection

In the conducted literature review, several useful resources were identified. In this section, we represent the datasets and open source projects.

7.1. Hate speech available datasets

Regarding the datasets, we found 69 datasets in 21 different languages. In this section, we summarize the most used dataset

Table 8
Algorithms and feature used in the papers related to deep-learning analysis.

Architecture name	Frequencies	Architecture nam	Frequencies
Word2Vec + LSTM	6	Word2Vec + CNN	4
RandomEmbedding + LSTM	4	RandomEmbedding + CNN	3
Word2Vec + BiLSTM	2	Word2Vec + CNN + LSTM	4
FAstText + LSTM	4	FastText + CNN	3
FAstText + GRU	4	FastText + GRU	1
FAstText + GRU	4	FastText + GRU	1
AraVec + LSTM	4	AraVec + CNN	3
AraVec + CNN + LSTM	1	Skip-gram + CNN + LSTM	1
ELMO + CNN	1	BERT + CNN	3
ELMO + BERT	1	SKIP-GRAM + CNN	2
BERT Base	7	BERT Large	8
GloVe + CNN	2	GloVe + GBDT + CNN	1
CNN + CNN + CNN	2	CNN + BiGRU	1

Table 9
Summary of some key contributions in HS detection by using Deep learning method and their respective results, in the metrics: Precision (P), Recall (R), F1-Score (F), Citation (C).

Author, Year	Platform	Type	ML Approach	Features Representation	Algorithm	P	R	F	C
Badjatiya et al. [21] 2017	Twitter, 16 k	Hate speech	Supervised	FastText, Random embedding, GloVe	CNN, LSTM, GBDT	.93	.93	.93	992
Yin et al. [189] 2017	-	-	Supervised		CNN, GRU and LSTM	.94	-	-	932
Rizos et al. [125] 2019	Twitter, 24 k	Hate speech	Supervised	FastText, Word2Vec, GloVe	CNN, LSTM, GRU	-	-	.69	65
Kamble and Joshi [68] 2018	Twitter, 3.8 k	Hate speech	Supervised	Word2Vec	LSTM, BiLSTM, CNN	.83	.78	.80	60
Ranasinghe et al. [122] 2019	Twitter,	Hate speech	Supervised	FastText	LSTM, GRU, BERT	-	-	.78	111
Faris et al. [44] 2020	Twitter, (Arabic)	Hate speech	Supervised	Word2Vec, Aravec	CNN + LSTM	.65	.79	.71	25
Al-Hassan and Al-Dossari [9] 2021, Springer	Twitter, 11 k (Arabic)	Hate, Racism, Sexism	Supervised	Keras word embeding	LSTM,GURU, CNN + GRU, CNN + LSTM	.72	.75	.73	19
Duwairi et al. [42] 2021, Springer	Twitter, 9 k,2 k (Arabic)	Hate, Hate, Abusive,Misogyny, Racism, Religious Discrimination,	Supervised	SG, CNN, CBOW	CNN, CNN-LSTM, and BiLSTM-CNN	.74	-	-	19
Dowlagar and Mamidi [40] 2021, arXiv	Twitter, (English, German, Hindi)	Hate, Offensive	Supervised	ELMO	BERT, Multilingual-BERT	.83	-	83	10
Sigurbergsson and Derczynski [135] 2019	Twitter, Reddit, newspaper comments (Danish)	Hate, Offensive, Not, Target, Individual, Group	Supervised	-	AUX-Fast-BiLSTM	-	-	.67	112
Ousidhoum et al. [102] 2019	Twitter	Sexual orientation, Religion, Disability, Target Group	Supervised	BOW	LR, biLSTM	-	-	80	157
Mulki et al. [94] 2019	Twitter	Abusive, Hate	Supervised	-	CNN and BiLSTM-CNN	.74	-	-	94

attributes and statistics in Tables 11 and 12. This includes dataset names (some names are based on papers title), publication year, dataset source link [16], dataset sizes, the ratio of offensive contents, the class used for annotation, and datasets' language. We noticed that many authors collected their datasets from social media and then annotated them manually based on task requirements. Several annotations have been carried out with experts [179,193,74], native speakers [50], volunteer [177], or through crowd-sourcing from anonymous users [34,186,191]. Below we present the primary findings of this analysis.

I) **Datasets language and platform:** Among 69 datasets of 21 different languages, see Fig. 16], English dominates by far others, representing 26 datasets alone. However, Arabic, German, Hind-English, Indonesian and Italian are represented in a total of 6, 3, 4, 4 and 5 open datasets, respectively. The rest of the languages have low presence in this set of open dataset. All datasets were collected from different social media platforms (Twitter, Facebook, Youtube, etc.), with exception to Chung et al. [32]'s dataset where some portion were synthetically produced. Twitter is shown to be the most popular platform for collecting hate speech datasets (45% of total datasets were collected from Twitter). Facebook is the second most popular source. The rest of the SM has only been used few times.

An interesting code-mixed dataset by [85] would be an example for targeting users who are active in SM and use the mixed form of language. This dataset has highlighted the predominance of Hindi–English code–mixed data representing the large spread of mixed forms and Hindi words written in Latin script in a non-formal online communication among Indians SM users. Similar code-mixed dataset work is also done in Tamil-English and Malayalam-English using BERT, which achieved a 90% F1 score [130].

(II) **Datasets sources:** Most of the dataset source repositories are available on GitHub. Therefore, nearly all datasets were publicly available. However, those dataset collected from Twitter have only Twitter Id instances which should be used to retrieve the full tweet messages. Since many tweets might be deleted over time, one may expect that the reconstruction of the full dataset may not be possible.

[16] All datasets' link last access was on 03-March-2021

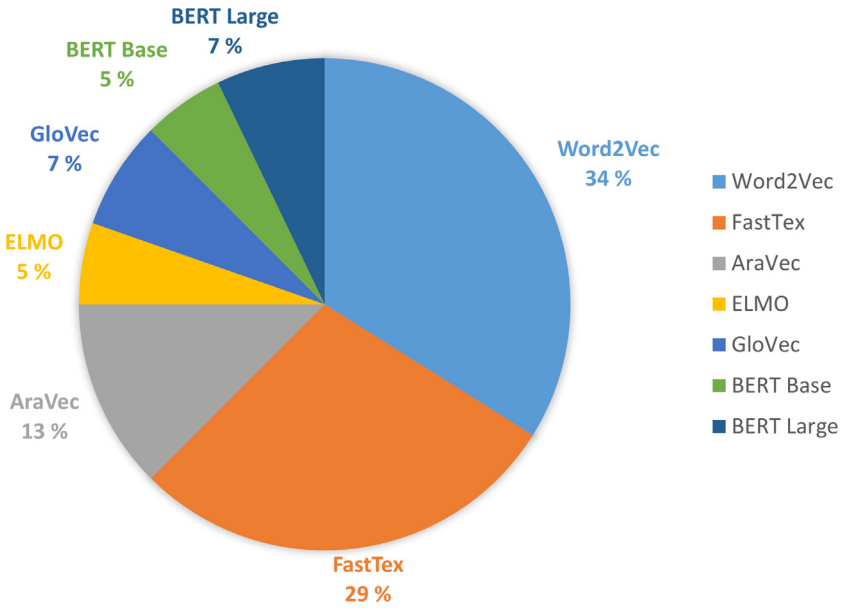


Fig. 15. Statistics of word embedding (e.g., Word2Vec, EELMO, FastText, BERT etc) in Deep learning related records.

Table 10
Best architecture from previous competition and their respective results, in the metrics: Precision (P), Recall (R), F1-Score (F), Citation (C).

Organiser	Author, Year	Platform	Language	Features Representation	Algorithm	F1 (%)	C
SemEval-19 Task 6	Liu et al. [78], 2019	Twitter 14 k	English	-	LSTM, BERT	82.9	149
SemEval-19 Task 6	Nikolov and Radivchev [98], 2019	Twitter 14 k	English	-	NB, CNN, LR, SVM, BERT-Large	81.5	73
SemEval-19 Task 6	Mahata et al. [80], 2019	Twitter 14 k	English	Word2vec	CNN, BLSTM + BGRU	80.6	22
SemEval-20 Task 12	Wiedemann et al. [181], 2020	Twitter, 14 k	English	-	BERT-base, BERT large, RoBERTa, XLM-RoBERTa, ALBERT	92	32
SemEval-20 Task 12	Wang et al. [176], 2020	Twitter, 14 k	English	-	XLM-RoBERTa base, XLM-RoBERTa large	91.9	24
SemEval-20 Task 12	Dadu and Pant [33], 2020	Twitter, 14 k	English	-	XLM-RoBERTa	91.8	11
HASOC2020	Mishraa et al. [89], 2020	Twitter	English	GloVe	LSTM	51	12
HASOC2020	Kumar et al.[73], 2020	Twitter	German	-	BERT, DistilBERT and RoBERTa	52	9
HASOC2020	Raja et al.[121], 2020	Twitter	Hindi	FasText	BiLSTM, CNN	52	8

(III) **Annotation classes:** Fig. 17 illustrates the diversity of annotations in the original datasets (e.g., hate, offensive, race, gender, sexism, misogyny, toxicity, group. target, political, etc.). This diversity of annotations translates the variety of hate speech categories and the academic willingness to explore this rich panorama. However, from Table 11 and 12, one notices that most of these annotations were based on binary classification (e.g., hate versus non-hate, racism versus non-racism, etc.) [13,10,132,25,124,119,43, 26,128,50]. Ternary class levels as well (e.g., Hate, Abusive, Normal) are explored in [94,93,179,186,102]. Some authors used a larger number of classes and sub-classes (up to six) as in [103,191,83]. In summary, one can distinguish three strategies of annotation scheme. The first one is a binary scheme: two mutually exclusive events (typically yes/no) to mark the presence or absence of HS (or a category of HS). The second one advocates a non-binary scheme with a fixed number of mutually exclusive classes, accounting either for different shades of a given HS category, such as strong hate, weak hate, no hate [111], overtly aggressive, covertly aggressive, not aggressive [74], or for several classes at the same time, such as racism, sexism, racism and sexism, none [179].

The third strategy features multi-level annotation, with finer-grained schemes accounting, for instance, for the type of hate speech, its severity, and the target group. This is the most complex annotation scheme and typically involves several different traits and a scale of variation. For example, [56] distinguish between racist, sexist, homophobic, religion-based attack, as well as the community targeted by the attack in the annotation process. Nobata et al. [99] discriminate between clean and abusive language, where abusive is labeled as hate speech, derogatory or profane. Basile et al. [24] adopt a three-layer binary annotation for HS, aggressiveness, and nature of the target (individual or group).

Table 11
Datasets for Hate Speech Detection for English.

Name	Year	Link	Platform	Size	Ratio	Class	Ref., Citation
Hate Speech and Offensive Language	2017	GitHub	Twitter	24,802	.06	Binary (hate speech, offensive but not hate speech, or neither offensive nor hate speech)	Davidson et al. [34], 1990
Hate Speech Dataset	2018	GitHub	Stormfront	9,916	.11	Hate, Relation, Not	de Gibert et al. [52], 259
Predictive Features for Hate Speech Detection	2016	GitHub	Twitter	16,914	.32	Sexist, Racist, Not	Waseem and Hovy [179], 1333
Hate Speech Detection Fox news comments	2017	GitHub	Fox News	1528	0.28	Binary (Hate/ not)	Gao and Huang [50], 186
Hate Speech Twitter annotations	2016	GitHub	Twitter	4,033	0.16	Racism, Sexism	Waseem [178], 474
Sexism using twitter data	2016	GitHub	Twitter	712	1	Sexism	Jha and Mamidi [66], 123
Misogyny Identification at IberEval 2018	2016	GitHub	Twitter	3,977	0.47	Sexism	Jha and Mamidi [66], 123
CONAN Multilingual Dataset of Hate Speech	2019	GitHub	Synthetic, Facebook	1,288	1	Islamophobia	Chung et al. [32], 112
Characterizing and Detecting Hateful Users	2018	GitHub	Twitter	4,972	0.11	Hate, Not hate	Ribeiro et al. [124], 183
Online Hate Speech (Gab)	2019	GitHub	GAB	33,776	0.43	Hate, Not hate	Qian et al. [119], 100
Online Hate Speech (Reddit)	2019	GitHub	Reddit	22,324	0.24	Hate, Not hate	Qian et al. [119], 100
Multilingual and Multi-Aspect Hate Speech	2019	GitHub	Twitter	5,647	0.76	Gender, Sexual orientation, Religion, Disability	Ousidhoum et al. [102], 157
HS Detection in Multimodal Publications	2020	GitHub	Twitter	149,823	0.25	No attacks to any community, Racist, Sexism, Homophobia, Religion-based attack, Attack to other community	Gomez et al. [56], 124
SemEval-2019 Task 6	2019	Link	Twitter	14,100	0.33	Offensive, Not Offensive, Target, Not Target, Individual, Group, Othe	Zampieri et al. [190], 518
Multilingual Detection of HS Against Immigrants and Women	2019	Link	Twitter	13,000	0.4	Hate, Not Hate, Group, Individual, Aggression, Not Aggression	Basile et al. [23], 160
Peer to Peer Hate	2019	GitHub	Twitter	27,330	0.98	Hate, Not Hate	ElSherief et al. [43], 129
HASOC-2019 (English)	2019	GitHub	Twitter, Facebook	7,005	0.36	Hate, Offensive, Neither, Profane, Targeted, Not Targeted	Mandl et al. [83], 286
Twitter Abusive Behavior	2018	Link	Twitter	80,000	0.18	Abusive, Hateful, Normal, Spam	Founta et al. [49], 422
Online Harassment	2017	Not available	Twitter	35,000	0.16	Harassment, Not Harassment	Golbeck et al. [55], 153
Personal Attacks	2017	GitHub	Wikipedia	115,737	0.12	Personal attack, Not Personal attack	Wulczyn et al. [186], 655
Toxicity	2017	GitHub	Wikipedia	100,000	NA	very toxic, neutral, very healthy	Wulczyn et al. [186], 655
Cyberbullying (World of Warcraft)	2016	GitHub	World of Warcraft	16,975	.01	Harassment, Not Harassment	Bretschneider and Peters [26], 17
Cyberbullying (League of Legends)	2016	GitHub	League of Legends	17,354	0.01	Harassment, Not Harassment	Bretschneider and Peters [26], 17
Lexicon for Harassment	2018	GitHub	Twitter	24189	0.01	Racism, Sexism, Appearance-related, Intellectual, Political	Rezvan et al. [123], 55
Aggression and Friendliness	2017	GitHub	Wikipedia	160,000	NA	Very aggressive, Neutral, Very friendly	Wulczyn et al. [186], 655

(IV) **Dataset size and ratio of abusive contents:**
Fig. 18 and 19 show the statistic of dataset size and the ratio of offensive content, respectively. We can see 41% of datasets are relatively of a small size (only (0–5) k posts). 14% have (5–10) k sentences. Therefore, most of the dataset (55%) can be cast into a very small size category, indicating the challenges behind acquiring large-scale labeled data for hate speech detection purpose. On the other hand, we have not found any balance dataset. This can lead to overfitting and harm generalisability, especially for deep learning models [58].
Another critical factor that may affect the training process of the model is the ratio of classes. In this regards, we noticed a high number (37%) of the datasets contain significantly less than 20% of offensive content. However, the rest of the datasets (63%) have more than 20% offensive content, which could be considered as decent for training purposes.

(V) **Number of Citations**[17]**:** We collected the number of citations for each dataset source document in Google Scholar and concluded that most dataset sources were cited more than 50 times (Fig. 20). Davidson et al. [34] is the most cited papers that have presented a dataset of 24802 tweets, which are manually annotated by CrowdFlower (CF) workers. Workers were asked to label each tweet as one of three categories: hate speech, offensive but not hate-speech, or neither offensive nor hate speech. Three or more people participated in each tweet annotation process. They have used the majority decision for each tweet to assign a label. This dataset has only 6% hate speech, 76% offensive but non-hate, and the rest of the tweets were neutral.

[17] **All citations last updated on 10-December-2022**

Table 12
Hate speech datasets and Corpus for Arabic, German, Danish, French, Indonesian, Italian, Bengali, Urdu, Russian and Hindi.

Name	Year	Link	Platform	Size	Ratio	Class	Lang.	Ref.
Abusive Language Detection on Arabic Social Media	2017	Link	Twitter	1,100	0.59	Obscene, Offensive but not obscene, Clean	Arabic	Mubarak et al. [93], 236
Abusive Language Detection on Arabic Social Media	2017	Link	AlJazeera	32,000	0.81	Obscene, Offensive but not obscene, Clean	Arabic	Mubarak et al. [93], 236
Religious Hate Speech in the Arabic	2018	GitHub	Twitter	16,914	0.45	Hate, Not hate	Arabic	Albadi et al. [13], 136
Anti-Social Behaviour in Online Communication	2018	Link	YouTube	15,050	.39	Offensive, Not Offensive	Arabic	Alakrot et al. [10], 54
Multi-Aspect Hate Speech Analysis	2019	GitHub	Twitter	3,353	0.64	Gender, Sexual orientation, Religion, Disability	Arabic	Ousidhoum et al. [102], 157
Arabic Levantine HateSpeech Dataset	2019	GitHub	Twitter	5,846	.38	Hate, Abusive, Normal	Arabic	Mulki et al. [94], 91
European Refugee Crisis	2017	GitHub	League of Legends	469	NA	Anti-refugee hate, Not Hate	German	Ross et al. [129], 383
Offensive Statements Towards Foreigners	2016	GitHub	Facebook	5,836	0.11	slightly offensive, explicitly offensive. targets (Foreigner, Government, Press, Community, Other, Unknown)	German	Bretschneider and Peters [27], 44
GermEval 2018	2016	GitHub	Twitter	8,541	0.34	Offense, Other, Abuse, Insult, Profanity	German	Wiegand et al. [183], 258
HASOC-2019 (German)	2019	GitHub	Twitter, Facebook	4,669	0.24	Hate, Offensive, neither, Hate, Offensive, or Profane	German	Mandl et al. [83], 286
Offensive Language and Hate Speech Detection for Danish	2019	GitHub	Twitter, Reddit, newspaper comments	3,600	.12	Offensive, Not, Within Offensive (Target, Not), Within Target (Individual, Group, Other)	Danish	Sigurbergsson and Derczynski [135], 112
CONAN HS French	2019	GitHub	Synthetic, Facebook	17,119	1	Islamophobic, not Islamophobic	French	Chung et al. [32], 112
MLMA hate speech	2019	GitHub	League of Legends	4,014	0.72	Gender, Sexual orientation, Religion, Disability	French	Ousidhoum et al. [102], 157
Offensive Language Identification in Greek	2020	GitHub	Twitter	4779	0.01	Offensive, Not, Target, Not, Individual, Group, Other	Greek	Pitenis et al. [112], 111
HS in Indonesian Language	2017	GitHub	Twitter	713	0.36	Hate, Not Hate	Indonesian	Alfina et al. [14], 138
HS and Abusive Language in Indonesian Twitter	2019	GitHub	Twitter	13,169	0.42	No hate speech, No hate speech but abusive, Hate speech but no abuse, Hate speech and abuse, Religion/creed, Race/ethnicity, Physical/disability, Gender/sexual orientation, Other invective/slander, within hate, strength (Weak, Moderate and Strongt)	Indonesian	Ibrohim and Budi [62], 113
Preliminaries Study for Abusive Language	2016	GitHub	Twitter	2,016	0.54	Not abusive, Abusive but not offensive, Offensive	Indonesian	Ibrohim and Budi [61], 83
An Italian Twitter Corpus	2018	GitHub	Twitter	1,827	0.13	Immigrants, Not	Italian	Sanguinetti et al. [132], 177
Hindi-English Code-mixed Data	2018	GitHub	Facebook	21,000	0.27	None, Covert Aggression, Overt Aggression, Physical threat, Sexual threat, Identity threat, Non-threatening aggression, Attack, Defend, Abet	Hindi, English	Kumar et al. [74], 132
Hindi-English Code-mixed Data	2018	GitHub	Facebook	18,000	0.06	None, Covert Aggression, Overt Aggression, Physical threat, Sexual threat, Identity threat, Non-threatening aggression, Attack, Defend, Abet	Hindi, English	Kumar et al. [74], 132
Offensive Tweets in Hinglish Language	2018	GitHub	Twitter	3,189	0.65	Not Offensive, Abusive, Hate	Hindi, English	Mathur et al. [85], 88
A Dataset of Hindi-English Code-Mixed	2018	GitHub	Twitter	4,575	0.36	Hate, Not	Hindi, English	Bohra et al. [25], 164
HASOC-2019 (Hindi)	2016	GitHub	Twitter, Facebook	5,983	0.51	Hate, Offensive or Neither, Profane, Targeted or Untargeted	Hindi, English	Mandl et al. [83], 286
Bengali HaS Dataset	2020	GitHub	Facebook, YouTube, Wikipedia, news-articles	-	-	Personal, Political, Religious, Geopoitical and Gender abusive hate	Bengali	Karim et al. [70], 26

(continued on next page)

Table 12 (*continued*)

Name	Year	Link	Platform	Size	Ratio	Class	Lang.	Ref.
Russian Dataset	2020	GitHub	-	-	-	-	Russian and Ukrainian	Andrusyak et al. [17], 17
Roman Urdu	2020	GitHub	-	-	-	-	Urdu	Rizwan et al. [126], 25

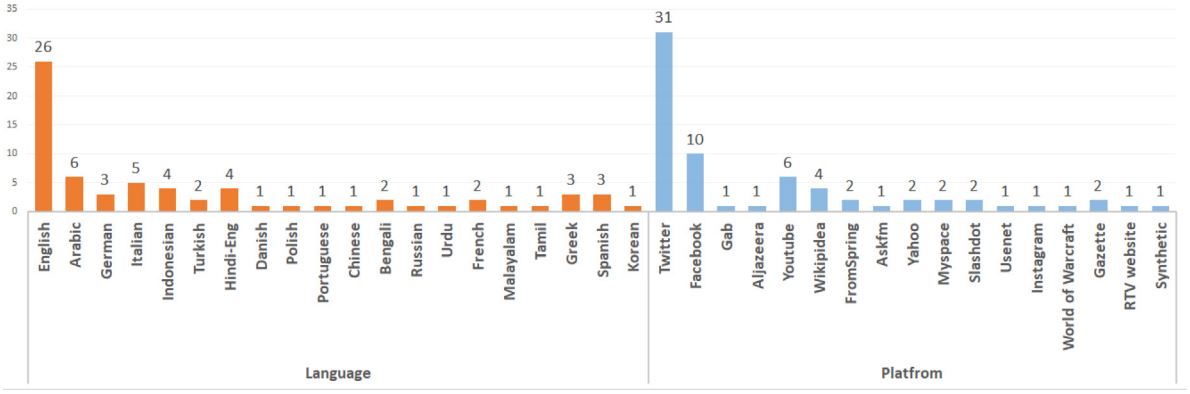


Fig. 16. Number of Datasets from different Language and Social Media.

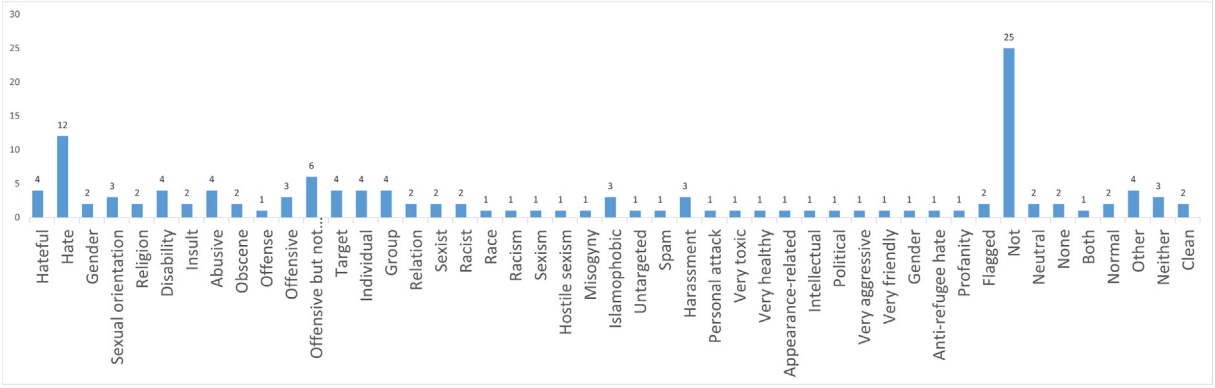


Fig. 17. Number of Classes used for annotation.

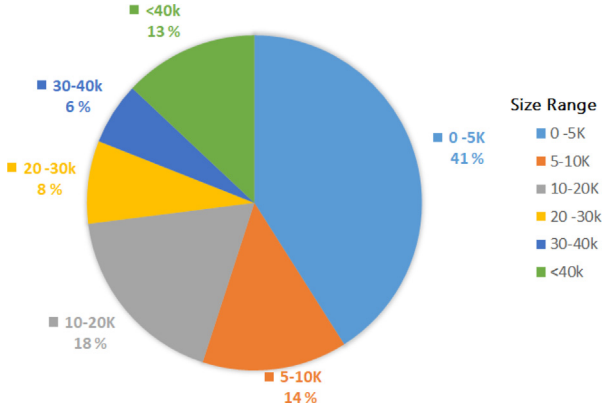


Fig. 18. Statistics of dataset size. k represent thousand.

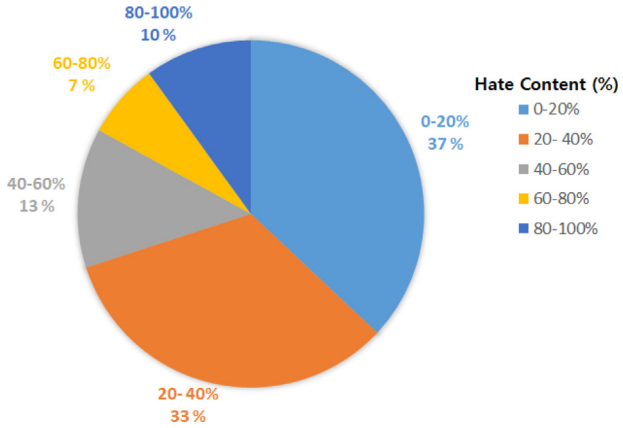


Fig. 19. Statistics of the proportion of abusive content contained in hate speech dataset.

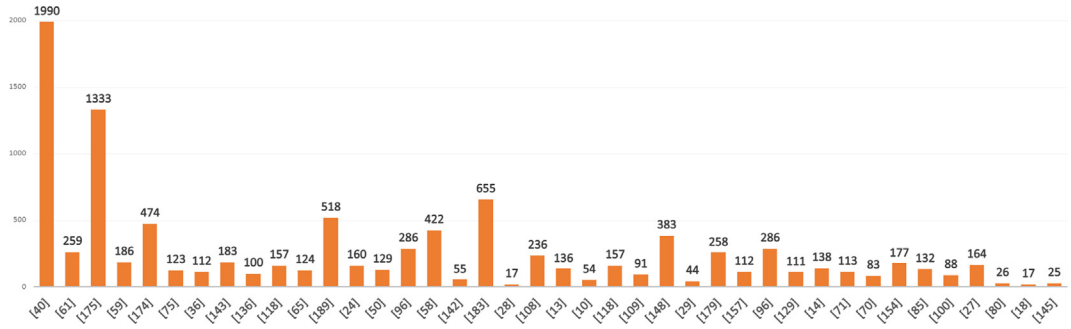


Fig. 20. Number of citations of reference papers of dataset.

The second most cited dataset created by Waseem and Hovy [179] contains 16,914 tweets where 3,383 are related to sexist, 1,972 to racist, and 11,559 to neither sexist nor racist. The authors manually annotated their dataset, after which it was sent to an outside annotator (a 25 year old woman studying gender studies and a nonactivist feminist) to review their annotation.

The above two popular datasets provide some insights into the dataset annotation process and criteria. We also noticed that the citation index does not reflect the quality of the dataset itself, but rather its ease and simplicity, which motivated other researchers to test such dataset in their proposals. There could be many possible factors that may represent the quality of a dataset (e.g., annotation criteria, label definitions, understanding perception, dataset size, the ratio of classes). However, we do not have sufficient experimental evidence to compare these dataset quality indices across various HS domains. To the best of our knowledge, only one study found in our search attempted to estimate the quality metrics and the similarity between datasets. In this respect, Fortuna et al. [48] compared the categories across the annotated dataset with respect to both similarity to other categories and homogeneity. For this purpose, average FastText embedding pretrained on Wikipedia was used to represent each category, while intra-dataset class homogeneity index has been put forward to assess category homogeneity. One of their observations is that Davidson's [34] "hate speech" is very different from Waseem's [179] "hate speech", "racism", "sexism", while being relatively close to Basile's HS dataset [24]. One of the main conclusions of their experiments is that many different definitions are being used for equivalent concepts, which makes most of the publicly available datasets incompatible.

7.2. Open source projects

We checked if there are any open-source projects available for hate-speech automatic detection or can be used as examples or sources for annotated data. For this, we carried out a search on GitHub repository with the search query "hate speech" in the available search engine. We found 1039 repositories, and only 53 were regularly forked and updated. Since this is a large number of repositories, it was challenging to include all of them in this paper and comment on them individually. Therefore, we have restricted to the 15 top-ranked one. Furthermore, we have exported the project repository names and descriptions into a CSV file for word cloud representation, which may help us understand the content of these open source projects in terms of the provided description.

Table 13 shows some highly cited HS detection papers source code. For example, [34] used Twitter dataset with TF-IDF, n-gram feature and LR-SVC model architecture. Furthermore, we have found the source code of [21] which used FastText and CNN, and LSTM models, achieving 78% F1 score and 85% accuracy. Furthermore, a new Korean dataset found in highly 'forked' GitHub repository claimed to be the first human-annotated Korean corpus for toxic speech detection and sizeable unlabeled corpus (Table 13, Index 2).

Another interesting repository named "Hate_sonar" used the BERT approach and the dataset in [34]. It created an easily installable python library, which anyone can use for their test project without having any coding skill.

Furthermore, some highly 'started' and 'forked' works appeared mainly relevant to sentiment analysis; namely TextBlob, VaderSentiment and Transformer. Here, the Transformer provides thousands of pre-trained models (mainly BERT) to perform tasks on texts such as classification, information extraction, question answering, summarization, translation, text generation, and sentiment analysis [185].

Fig. 21 shows the word cloud representation of the identified GitHub project descriptions. The illustration indicates the high proportion of words such as hate speech, hateSpeech, speech hate, hateful meme, cyberbullying offensive language, and toxic comment in the repositories related to HS detection. Furthermore, Twitter, tweet, Twitter hate, Reddit, and other social media were also present in this illustration, although with less intensity compared to hate detection related wording. Similarly, deep-learning methods such as BERT, LSTM, CNN, where the increased popularity of BERT is more emphasized. Surprisingly, this is in a full agreement with the literature review as well, where BERT was found to be a dominant trend in deep-learning methods. Among the fifteen top forked repositories, four projects were linked to BERT based model. The connection between hate speech and sentiment analysis is also stressed in Fig. 21, as well as the growing interest in non-English HS detection through wordings like multilingual, Indonesian, Hindi, indicating repositories for multilingual HS detection tasks.

Regarding the programming languages employed, we noticed that most of the projects (88%) were developed in Python language, while others used JavaScript (4%), Java (2%), HTML and CSS (4%), and three projects with GO (Fig. 22).

8. Research challenges and opportunities

The above literature review for deep learning and non-deep learning and resource analysis summarized the main research in the field of HS automatic detection from textual inputs. At the same time, we have also identified several challenges and research gaps (Table 14) from previous research.

8.1. Open Source Platforms or Algorithms:

There are indeed many open-source projects available related to HS. However, only few project source codes are available from well-known publications. From the 1039 projects in GitHub, we have only found 53 projects regularly maintained and forked, which may question the usability and source code quality of the

Table 13
Fifteen most popular GitHub open source projects.

Repository Name	Github Link	Focus	Publication Ref., Year	Features Representation	Algorithm	Star	Fork
1. Hate speech and offensive language	Link, Source code given	Twitter 25 k dataset used for HS detection.	Davidson et al. [35]	TF-IDF, n-gram, bi-gram	LR, SVC	656	293
2. Korean HateSpeech Dataset	Link, Dataset source code given but project source code not available.	The first human-annotated Korean corpus for toxic speech detection and the large unlabeled corpus. The data is comments from the Korean entertainment news aggregation platform.	Moon et al. [92]	-	-	308	36
3. Twitter hatespeech	Link, source code available.	Implementation of paper - "Deep Learning for Hate Speech Detection"	Badjatiya et al. [21]	Fasttext, BOW	CNN, LSTM	204	79
4. Hate sonar	Link, Install $ pip install hatesonar	HateSonar allows you to detect hate speech and offensive language in text, without the need for training. There's no need to train the model. You have only to fed text into HateSonar. It detects hate speech with the confidence score.	Davidson et al. [34]	-	BERT	156	36
5. Hate speech dataset	Link, Source code available	These files contain text extracted from Stormfront, a white supremacist forum. A random set of forums posts have been sampled and n manually labelled as containing hate speech or not.	de Gibert et al. [53]	-		120	51
6. Dataset for Learning to Intervene	Link, Only dataset available	HS intervention along with two fully-labeled datasets collected from Gab and Reddit. Distinct from existing hate speech datasets, their datasets retain their conversational context and introduce human-written intervention responses.		-		60	11
7. HateXplain	Link, dataset and soruce code available, Install: pip install -r requirements.txt	Multilingual multi-aspect hate speech analysis dataset	Mathew, Binny and Saha, Punyajoy and Yimam, Seid Muhie and Biemann, Chris and Goyal, Pawan and Mukherjee, Animesh [154]	-	BERT	138	49
8. MLMA hate speech	Link, dataset and soruce code available	Multilingual multi-aspect hate speech analysis dataset	Ousidhoum et al. [103]	-	LR	49	5
9. Korean Hate Speech	Link, dataset and soruce code available	This hate speech detection model trained on kocohub/korean-hate-speech	2020	-	BERT	42	5
10. likers-blocker	Link, Browser addon	Block hate promoter twitter suer	2021	-		301	13
11. Twitter Hate Speech Detection	Link, dataset and soruce code available	project analyzed a dataset CSV file from Kaggle containing 31,935 tweets.	2020	BOW, TF-IDF	LR, NB, RF	38	25
12. Korean hate speech language modeling	Link, dataset and soruce code available	Recurrent Neural Network based Hate Speech Language Model for Korean Hate Speech Detection	2020		RNN	24	5
13. TextBlob	Link, soruce code available, install: pip install -U textblob	TextBlob is a Python library provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more	2021			8.4 k	1.1 k
14. vaderSentiment	Link, soruce code available, install: pip install vaderSentiment	VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.	2021			3.8 k	921
15. Transformers	Link, soruce code available, install: pip install transformers	Transformers provides thousands of pre-trained models to perform tasks on texts such as classification, information extraction, question answering, summarization, translation, text generation, sentiment analysis	Wolf et al. [185,185]		BERT, ALBERT	76 k	17.2 k

rest of the projects. More sharing of code with a clear documentation, algorithms, processes for feature extraction, and open-source datasets can help the discipline evolves more quickly.

8.2. Language and System Barriers

Language evolves quickly, particularly among young populations that frequently communicate in social networks, demanding continuity of research for HS datasets. For instance, online platforms are removing hate contents manually and automatically [18] [19]. However, those who spread HS content will always try to develop a new way to evade and by pass any system imposed restriction. For

[18] https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/
[19] https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.html

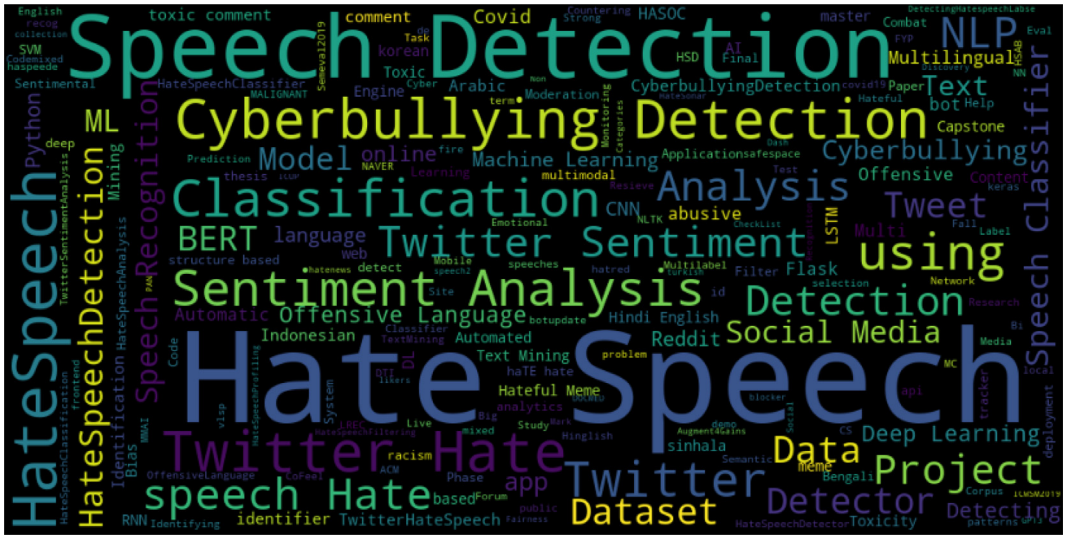


Fig. 21. Word cloud representation of GitHub open source projects.

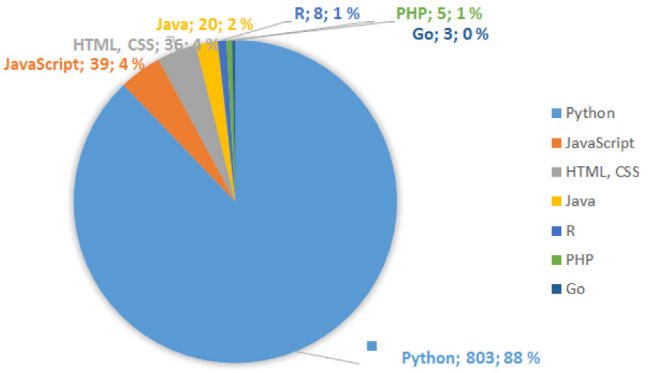


Fig. 22. GitHub open source projects programming language.

example, some users do post HS content as images containing the hate text, which circumvent some basis automatic HS detection. Although image to text conversion might solve some particular problem, still several challenges arise due to limitation of such conversation as well as existing automatic HS detection. Besides, changing the language structure could be another challenge, for example, through usage of unknown abbreviations and mixing different languages, e.g., i) Writing part of a sentence in one language and the other part in another language; (ii) Writing sentence phonetics in another language (e. g., writing Hindi sentences using English).

8.3. Dataset

There are no commonly accepted datasets recognized as ideal for automatic HS detection task. Authors annotate dataset differently based on their understanding and tasks requirement. For instance, Fig. 17 highlights 47 different annotation labels from 69 datasets, which stress on the diversity of the existing datasets. Besides, 55% of the available datasets' sizes are small and contain a tiny portion of hate content. Many datasets were annotated through crowd-sourcing, which may also question the knowledge of the annotator.

Clear label definitions. There is a prerequisite to have a clear label definition, separating HS from other types of offensive languages [34,49]. Indeed, dataset can cover a broader spectrum targeting multiple fine-grained HS categories (e.g., sexism, racism, personal attacks, trolling, cyberbullying). This can be performed through either multi-labelling approach, although one notices the presence of ambiguous cases as in [178]'s racism and sexism labels, or in a hierarchical manner as in [24]'s and [74]'s work on subtypes of HS and aggression, respectively.

Annotation quality. The offensive nature of hate speech and abusive language makes the grammatical structure and cross-sentence boundaries loose, leading to challenging annotation criteria [99]. Therefore, hate speech datasets should be constantly updated according to newly available knowledge. For instance, [114] found that only about two-thirds of the existing datasets report inter-annotator agreement, guidelines, definitions, and examples [114]. To ensure a high inter-annotator agreement, extensive instructions and the use of expert annotators are required. Whether annotators received comprehensive recommendations, minimum guidelines, or no guidelines at all, can significantly differ between datasets' annotation quality. A study by Yin et al. [171] on existing HS datasets revealed that 27 of the 63 (43%) datasets did not provide any annotation guideline of their associated datasets, while 20 (32%) provided only rough guideline indication. Finally, a smaller proportion consisting of 16 datasets (25%) included comprehensive instructions in terms of dataset annotations. Such guidelines are of paramount importance to comprehend the scope and limitation of the underlined HS study. For example, if the job is crowdsourced, then a description of the top-level parameters (such as the number of employees, the maximum number of tasks per worker, and the amount of annotations per piece of text) should be presented [159]. Similarly, if an annotation-specific interface was employed, then environmental constraints, and user experience are important factors that can distinguish annotator participation. Likewise, the availability of source code is very useful for research promotion and generalization of the research. Reference [159] pointed out that 98% of the datasets were collected from social networks and labeled manually and only a limited work was directed towards (artificial) dataset creation and enrichment of existing datasets.

Ethical aspects. Privacy is another relevant challenge associated with the widespread use of HS datasets. Indeed, the collection process of such dataset often involves using content posted by real users who do not necessarily want to be identified. At present, most researchers do not systematically obtain explicit consent from all users whose content is being analyzed and, instead, rely on the implicit consent that users are in a public or semi-public

Table 14
Review of Gaps and future research agenda.

Research question	Gap in literature	Future Research Agenda
Q1: What are the specificities among different HS branches and scopes for automatic HS detection from previous literature?	G1: Discrepancy and fragmentation of knowledge among different domain. G2: Multilingual resources.	- How to develop a common framework for researchers which will be helpful for domain adaptation? - How we can clarify all the concepts and definitions that will be helpful to obtain high-quality and comparable resources? - How effectively take into account the specificities related to language and culture, and work towards preventing HS? - Development of NLP resources for other languages (e.g., multilingual sentiment feature, dataset, word embedding, etc.) would leverage HS detection for other languages.
Q2: What is the state of deep learning in automatic HS detection in practice?	G3: Comparative study among different deep learn- ing algorithms and related resources. G4: Model application and impact.	- Which existing deep learning models, features, and what characteristics of the datasets are more efficient in tackling HS detection? - What approaches could be used to make the model less biased against specific terms or language styles, from the perspectives of training data or objective. More systematic comparisons between debiasing approaches would be favorable. - Use of deep-learning architecture to create NLP resources which are currently developed with non-deep learning methods that will leverage HS detection. - Can automatic deep-learning models practically aid human moderators in content moderation? In that case, how can human moderators or organizations make use of the outputs feature analysis most effectively? Would that introduce more bias or reduce bias in the content moderation process? What would the impact be on the users of the platform? To answer these questions, interdisciplinary study and collaboration with organizations are needed.
Q3: What is the state of the HS datasets in practice?	G5: Dataset annotation G6: Dataset augmentation tools and techniques	- How to avoid the risk of creating data that are biased or too much related to a specific resource? - Development of data-driven taxonomy that highlights how different types of HS datasets concepts are linked and how they differ from one another? - How to deal with the discrepancy of data and error analysis of human annotation issued for previous literature? - How to develop a standard form of annotation guideline? - What type of and how much training or instruction is required to match the annotations of crowdworkers and experts? - Develop new methods or algorithms for artificial HS dataset creation and expand the semantic meaning of existing datasets that will help to create a large-scale balanced dataset.

space [169]. Anonymization is often advocated as a primarily solution for this purpose [47]. Although, this is a sound and well-established technique, it sometimes fails to ensure privacy requirement, because it is often possible to re-identify users using different techniques when the dataset collection platform is known. At present, several of the most widely used datasets provide only the annotations and the IDs of the posts. So, the full details of the posts can be recollected using API. This is often employed of Twitter dataset, so that if a user changes his private status or his Twitter account was suspended, the post cannot be retrieved anymore. Besides, some ethics specialists argue that this procedure does not fully comply with ethical rules neither, since the ID of the post directly leads to the user, which makes the approach even worse than the anonymization method. On the other hand, we shall also mention what is referred as *dataset degradation*, which adds extra challenges to HS analysis pipeline. For example, the dataset provided by Waseem and Hovy [179] must be collected using the ID retrieval API, and this led to considerable degradation since it was first released as the tweets are no longer available on Twitter. Chung *et al.* claimed that within 12 months, the dataset by Mathew *et al.* [153] had lost more than 60% of its content. Dataset degradation poses several risks: if less data is available, then there is a greater likelihood of overfitting and the risk of an unbalanced class since most cases, hate content is deleted. Alternatively, one can argue that the dataset can be made available when used for training purpose, after enforcing some data sharing agreement as performed by some organizations, especially in medical field. A counterpoint argument for this approach relies on the observation that often a user who posts HS content on social media may only do so to express his thought, opinion, and emotion and do not necessarily agree to have this content used to train an intelligent agent.

Finally, with the advances in ethical artificial intelligence (AI) [147] as a result of the wide spread use of AI-based systems in our society, privacy has seen a new renewal interest in the global agenda, with a number of regulations and legislation being discussed and created that directly impact dealing with online data and HS. To our knowledge, no successful initiatives for sharing abusive training datasets have emerged in the field, beyond ad hoc use of platforms such as Github or through sharing data in scientific competitions and challenges.

8.4. Comparative Studies

From Fig. 12, we identified 24 different hybridization schemes in deep-learning models. However, extensive and comprehensive comparative studies where different approaches are genuinely contrasted and compared were very missing. This leave the door widely open to future comprehensive comparative HS studies in terms of data preprocessing, feature engineering, model training and evaluation.

In addition, little to none work has been found focusing on the labelling issue taking into account the model organization of the individual/group targeted by the HS post, building from advances in social psychology theory and human computer interaction research. This creates a gap in the practicality of the automatic detection HS developed models as well as on the comparison analysis. This also raises many questions about how the underlined HS

detection technique would impact real users' experience and the accuracy of the model compared to human moderation. To answer this question, more interdisciplinary studies and collaboration with organizations are needed.

Besides, comparison between various training approaches, debiasing approaches, overfitting models, and what characteristics of the datasets interact with the effectiveness would be worth investigating. For example, when performing transfer learning, the trade-off between domain-specificity, linguistic patterns, and underlying sentiment of hate speech can be considered before model design, feature extraction and preprocessing.

8.5. Multilingual Research

As previously pointed out, 50% of HS studies, datasets, and open source projects were provided in English language. Although, we noticed a rise in some other non-English resources as well, as in Arabic where AraVec word embedding showed some popularity, there is a scarcity in the development of other non-English NLP resources. Although, we have identified 21 different language HS related works, which creates an opportunity to develop enhanced NLP tools for these languages.

8.6. Counter-narratives (CNs) for HS

Counter-narrative core idea is to intervene directly in discussion with textual responses that are meant to counter and withstand the effect of hate speech content and prevent it from spreading further [141]. Traditionally, counter-narratives are delivered by educators, psychologists, linguistics and NGOs to counter well-identified type of hate speech by empowering users with well-founded logical arguments that enable them to withstand HS effects. Examples of these efforts include Manuel toolbox "We Can!" [20] promoted by the European Council. In Recent years, with the development of natural language generation (NLG) using deep learning technology, effort has been devoted to automatically generate counter-narratives according to the identified hate speech category. Still, the proposals suffer from the lack of a sufficient amount of quality data and tend to produce generic/repetitive responses. To the best of our knowledge CONAN [32], where a large-scale, multilingual, expert-based dataset of hate speech/counter-narrative pairs is proposed, offering high quality CNs, and the best and most diverse material among other CN datasets [148]. Similarly, Chung, Yi-Ling and Tekiroglu, Serra Sinem and Guerini, Marco [144] suggested models to generate counter narratives focusing on informative and multilingual responses. They introduced a knowledge-driven pipeline that can produce suitable and informative English counter narratives while avoiding hallucination phenomena. They have used GPT2 and XNLG tranformer [143] for model training where human annotators found XNLG generations as the most informative and GPT-2 generations as the most suitable. In the same spirit, works in [162,163] suggested using pre-trained language models like BERT, T5 [157], GPT/2, DialoGPT [172] for producing Counter Narratives against hate speech. Although, the effectiveness and efficacy of the counter-narrative generation systems are yet to be fully tested in practice.

8.7. HS and Multi-modal Aspect

While there is an increase in textual hate data generation, less research focused on detecting hateful content in multimedia data. In general, hate speech and offensiveness can be detected from multimedia sources using three distinct modalities: visual, acoustic, and verbal [158]. For instance, memes pose an interesting multimodal fusion problem where unimodal sentences could be harmless but combined with images can make it severe HS. This raises at least two important problems [170]. The first one is related to discrepancy in the definition of Hate and Sarcasm content. The second one is related to the domain gap, which is translated into variation in feature distribution [165]. This gap is caused by different sources of sarcasm and hate datasets in image-text containing unlimited combination of images (e.g., posters, plain text, etc.) with tag-symbolized text. Because of the different feature distribution, direct knowledge transfer may lead to invalid transfer. To overcome such problems, pre-trained multimodal models [145,150] are directly fine-tuned for feature learning. Besides, some studies have also attempted to utilize data augmentation [151,166] and model integration methods [161] to improve HS detection performance. However, current research on this task mainly focuses on the construction of multimodal models without considering the influence of the unbalanced and widely distributed samples for various attacks in hate speech. For multimodal sarcasm detection, Cai et al. [142] released a sarcasm dataset compiled from image-text tweets and design a hierarchical some fusion model as the baseline. Other models in [156,168] are constructed for evaluation on sarcasm dataset. We shall also mention BERT model developed for multimodal training utilizing VisualBERT [152]. The latter is viewed as the "BERT of vision and language" – that was trained multimodally on both images and captions, then Ensemble Learning were applied. Their approach achieves accuracy of 0.765 on the challenge test set and placed third out of 3,173 participants in the Hateful Memes Challenge [166]. To the best of our knowledge, we have found only one research Rana, Aneri and Jha, Sonali [158] that proposes a multimodal deep learning framework which combines the auditory features representing emotion and the semantic features to detect hateful content. This paper also presents a new Hate Speech Detection Video Dataset (HSDVD) collected for the purpose of multimodal learning. Beside English multimodal, there are very little work done for multilingual multimodal HS detection. An example of such work includes Karim, Md and Dey, Sumon Kanti and Islam, Tanhim and Shajalal, Md and Chakravarthi, Bharathi Raja and others [149] efforts to implement this concept for Bengali language. The authors used state-of-the-art neural architectures (e.g., Bi-LSTM/Conv-LSTM with word embeddings, monolingual Bangla BERT, multilingual BERT-cased/uncased, and XLM-RoBERTa) to jointly analyze textual and visual information for hate speech detection. As of multimodal fusion, XML-RoBERTa + DenseNet-161 performed the best, yielding an F1 score of 0.83. A similar fusion model has been applied to English multimodal in [160] and achieved an accuracy of 67.7%.

9. Conclusion

In this survey, we presented a critical overview of how the automatic detection of hate speech in the text has evolved over the past few years. Our analysis also included other hate speech domains, e.g., cyberbullying, abusive language, discrimination, sexism, extremism and radicalization. We initially reviewed existing surveys in the field. We found few recent systematic literature reviews related to HS detection, which are shown to be insufficient to summarize the current state of research in the field. Next, we carried out a systematic literature review from Google scholar and ACM digital library databases for all documents related to hate speech published between 2000 and 2021. A total of 463 articles are found to match PRISMA inclusion and exclusion criteria. The findings indicate that initially SVM algorithm and various types of TF-IDF features were the most widely used. However, after the

[20] https://www.coe.int/en/web/no-hate-campaign/we-can-alternatives

advancement in deep-learning technology, a rapid change in the hate speech analysis methods was observed. The research community preferred to use different kinds of word embedding with CNN and RNN architectures. From 2017 to 2021, several comparative studies have shown the merits of deep-learning models including CNN, RNN using word2Vec, GloVe, FastText, among other embedding as compared to traditional machine learning models such as SVM, LR, NB, and RF models. Nevertheless, comparison among deep learning models is still in its infancy. For instance, mixed results were obtained when comparing CNN to RNN models. Though some authors opined RNNs are more suited for the long-ranged context, while CNN seem useful in extracting features and GRU is found to be more suited for long sentences. Besides, multiple papers suggested that the concatenation of two or more deep learning models performed better than using a single deep learning model. For example, CNN + LSTM and CNN + GRU both performed better than the single application of LSTM and CNN.

Similarly, the comparison of different word-embedding models, e.g, FastText, Word2Vec, GloVe, showed a close performance; however, ELMO performed slightly better than others. Though the work related to ELMO is meager, which leaves the door open for further comparative studies. After introducing contextual relations-based model BERT in 2018, several works claimed BERT's outperforming ELMO, CNN and RNN models. After comparing the best HS detection architecture from Semeval-2019, Semeval-2020 and the HASOC-2020 competitions, BERT-based model was also ranked top among other deep-learning models.

In the final part, we analyzed 69 hate speech datasets. The existing works presented several obstacles for dataset preparation. In general, researchers tend to start by collecting and annotating new comments from SM or using previous datasets. Often, retrieving an old dataset from Twitter is not always fully possible due to tweets' potential removal. This slows down the research's progress because less data is available, making it more challenging to compare different studies' results. Several limitations are found when scrutinizing statistics of past dataset. Significantly, most of the dataset sizes were small, lack the ratio of hate content, and lack label definitions and inter-annotator agreements. Besides, only a fraction of two datasets were synthetically made, which leaves room for artificial dataset creation, augmentation, and enrichment. Finally, we identified the main challenges and opportunities in this field. This includes the scarcity of good open-source code that is regularly maintained and used by the society, the lack of comparative studies that evaluate the existing approaches, and the absence of resources in non-English experiments. With our work, we summarized the current state of the automatic HS detection field. Undoubtedly, this is an area of profound societal impact and with many research challenges.

Data availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] K.E. Abdelfatah, G. Terejanu, A.A. Alhelbawy, Unsupervised detection of violent content in arabic social media, Comput. Sci. Inf. Technol. (CS IT) (2017) 1–7.

[2] E.A. Abozinadah, Improved micro-blog classification for detecting abusive arabic twitter accounts, International Journal of Data Mining & Knowledge Management Process (IJDKP) 6 (2016).

[3] Abozinadah, E.A., Jones Jr, J.H., 2017. A statistical learning approach to detect abusive twitter accounts, in: Proceedings of the International Conference on Compute and Data Analysis, pp. 6–13.

[4] E.A. Abozinadah, A.V. Mbaziira, J. Jones, Detection of abusive accounts with arabic tweets, Int. J. Knowl. Eng.-IACSIT 1 (2015) 113–119.

[5] S. Agarwal, A. Sureka, Using knn and svm based one-class classifier for detecting online radicalization on twitter, International Conference on Distributed Computing and Internet Technology, Springer. (2015) 431–442.

[6] Ahn, H., Sun, J., Park, C.Y., Seo, J., 2020. Nlpdove at semeval-2020 task 12: Improving offensive language detection with cross-lingual transfer. arXiv preprint arXiv:2008.01354.

[7] M.P. Akhter, Z. Jiangbin, I.R. Naqvi, M. Abdelmajeed, M.T. Sadiq, Automatic detection of offensive language for urdu and roman urdu, IEEE Access 8 (2020) 91213–91226.

[8] A. Al-Hassan, H. Al-Dossari, Detection of hate speech in social networks: a survey on multilingual corpus, in: 6th International Conference on Computer Science and Information Technology, 2019.

[9] A. Al-Hassan, H. Al-Dossari, Detection of hate speech in arabic tweets using deep learning, Multimedia Systems (2021) 1–12.

[10] A. Alakrot, L. Murray, N.S. Nikolov, Dataset construction for the detection of anti-social behaviour in online communication in arabic, Procedia Computer Science 142 (2018) 174–181.

[11] A. Alakrot, L. Murray, N.S. Nikolov, Towards accurate detection of offensive language in online communication in arabic, Procedia computer science 142 (2018) 315–320.

[12] Alami, H., El Alaoui, S.O., Benlahbib, A., En-nahnahi, N., 2020. Lisac fsdm-usmba team at semeval-2020 task 12: Overcoming arabert's pretrain-finetune discrepancy for arabic offensive language identification, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 2080–2085.

[13] Albadi, N., Kurdi, M., Mishra, S., 2018. 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, ACM. pp. 69–76.

[14] Alfina, I., Mulia, R., Fanany, M.I., Ekanata, Y., 2017. Hate speech detection in the indonesian language: A dataset and preliminary study, in: 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), IEEE. pp. 233–238.

[15] Alshehri, A., El Moatez Billah Nagoudi, H.A., Abdul-Mageed, M., 2018. Think before your click: Data and models for adult content in arabic twitter, in: TA-COS 2018: 2nd Workshop on Text Analytics for Cybersecurity and Online Safety, p. 15.

[16] Aluru, S.S., Mathew, B., Saha, P., Mukherjee, A., 2020. Deep learning models for multilingual hate speech detection. arXiv preprint arXiv:2004.06465.

[17] Andrusyak, B., Rimel, M., Kern, R., 2018. Detection of abusive speech for mixed sociolects of russian and ukrainian languages., in: RASLAN, pp. 77–84.

[18] Antoun, W., Baly, F., Hajj, H., 2020. Arabert: Transformer-based model for arabic language understanding. arXiv preprint arXiv:2003.00104.

[19] Araci, D., 2019. Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063.

[20] Arora, G., 2020. Gauravarora@ hasoc-dravidian-codemix-fire2020: Pre-training ulmfit on synthetically generated code-mixed data for hate speech detection. arXiv preprint arXiv:2010.02094.

[21] Badjatiya, P., Gupta, S., Gupta, M., Varma, V., 2017. Deep learning for hate speech detection in tweets, in: Proceedings of the 26th international conference on World Wide Web companion, pp. 759–760.

[22] Bashar, M.A., Nayak, R., 2020. Qutnocturnal@ hasoc'19: Cnn for hate speech and offensive content identification in hindi language. arXiv preprint arXiv:2008.12448.

[23] Basile, P., Caputo, A., Semeraro, G., 2014. An enhanced lesk word sense disambiguation algorithm through a distributional semantic model, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 1591–1600.

[24] Basile, V., Bosco, C., Fersini, E., Debora, N., Patti, V., Pardo, F.M.R., Rosso, P., Sanguinetti, M., et al., 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter, in: 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics. pp. 54–63.

[25] Bohra, A., Vijay, D., Singh, V., Akhtar, S.S., Shrivastava, M., 2018. A dataset of hindi-english code-mixed social media text for hate speech detection, in: Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media, pp. 36–41.

[26] Bretschneider, U., Peters, R., 2016. Detecting cyberbullying in online communities.

[27] Bretschneider, U., Peters, R., 2017. Detecting offensive statements towards foreigners in social media, in: Proceedings of the 50th Hawaii International Conference on System Sciences.

[28] Burnap, P., Williams, M.L., 2014. Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making.

[29] Caselli, T., Basile, V., Mitrović, J., Granitzer, M., 2020. Hatebert: Retraining bert for abusive language detection in english. arXiv preprint arXiv:2010.12472.

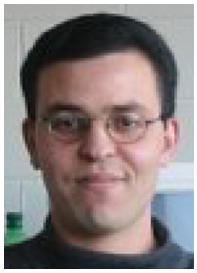
[30] Chen, H., McKeever, S., Delany, S.J., 2017. Abusive text detection using neural networks., in: AICS, pp. 258–260.

[31] Chen, Y., Zhou, Y., Zhu, S., Xu, H., 2012. Detecting offensive language in social media to protect adolescent online safety, in: 2012 International Conference

on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, IEEE. pp. 71–80.
[32] Chung, Y.L., Kuzmenko, E., Tekiroglu, S.S., Guerini, M., 2019. Conan–counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. arXiv preprint arXiv:1910.03270.
[33] Dadu, T., Pant, K., 2020. Team rouges at semeval-2020 task 12: Cross-lingual inductive transfer to detect offensive language, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 2183–2189.
[34] Davidson, T., Warmsley, D., Macy, M., Weber, I., 2017a. Automated hate speech detection and the problem of offensive language, in: Proceedings of the International AAAI Conference on Web and Social Media.
[35] Davidson, T., Warmsley, D., Macy, M., Weber, I., 2017b. Automated hate speech detection and the problem of offensive language, in: Proceedings of the 11th International AAAI Conference on Web and Social Media, pp. 512–515.
[36] Di Capua, M., Di Nardo, E., Petrosino, A., 2016. Unsupervised cyber bullying detection in social networks, in: 2016 23rd International conference on pattern recognition (ICPR), IEEE. pp. 432–437.
[37] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, R. Picard, Common sense reasoning for detection, prevention, and mitigation of cyberbullying, ACM Transactions on Interactive Intelligent Systems (TiiS) 2 (2012) 1–30.
[38] Djandji, M., Baly, F., Hajj, H., et al., 2020. Multi-task learning using arabert for offensive language detection, in: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, pp. 97–101.
[39] Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., Bhamidipati, N., 2015. Hate speech detection with comment embeddings, in: Proceedings of the 24th international conference on world wide web, pp. 29–30.
[40] Dowlagar, S., Mamidi, R., 2021. Hasocone@ fire-hasoc2020: Using bert and multilingual bert models for hate speech detection. arXiv preprint arXiv:2101.09007.
[41] R. Dredge, J. Gleeson, X. De la Piedad Garcia, Cyberbullying in social networking sites: An adolescent victim's perspective, Computers in human behavior 36 (2014) 13–20.
[42] R. Duwairi, A. Hayajneh, M. Quwaider, A deep learning framework for automatic detection of hate speech embedded in arabic tweets, Arabian Journal for Science and Engineering 46 (2021) 4001–4014.
[43] ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., Belding, E., 2018. Peer to peer hate: Hate speech instigators and their targets, in: Proceedings of the International AAAI Conference on Web and Social Media.
[44] Faris, H., Aljarah, I., Habib, M., Castillo, P.A., 2020. Hate speech detection using word embedding and deep learning in the arabic language context., in: ICPRAM, pp. 453–460.
[45] Fernandez, M., Alani, H., 2018. Contextual semantics for radicalisation detection on twitter.
[46] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (2018) 1–30.
[47] Fortuna, P., da Silva, J.R., Wanner, L., Nunes, S., et al., 2019. A hierarchically-labeled portuguese hate speech dataset, in: Proceedings of the Third Workshop on Abusive Language Online, pp. 94–104.
[48] Fortuna, P., Soler, J., Wanner, L., 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets, in: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 6786–6794.
[49] Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., Kourtellis, N., 2018. Large scale crowdsourcing and characterization of twitter abusive behavior, in: Proceedings of the International AAAI Conference on Web and Social Media.
[50] Gao, L., Huang, R., 2017. Detecting online hate speech using context aware models. arXiv preprint arXiv:1710.07395.
[51] Ghanghor, N., Ponnusamy, R., Kumaresan, P.K., Priyadharshini, R., Thavareesan, S., Chakravarthi, B.R., 2021. Iiitk@ lt-edi-eacl2021: Hope speech detection for equality, diversity, and inclusion in tamil, malayalam and english, in: Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, pp. 197–203.
[52] de Gibert, O., Perez, N., García-Pablos, A., Cuadros, M., 2018a. Hate speech dataset from a white supremacy forum. arXiv preprint arXiv:1809.04444.
[53] de Gibert, O., Perez, N., García-Pablos, A., Cuadros, M., 2018b. Hate Speech Dataset from a White Supremacy Forum, in: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Association for Computational Linguistics, Brussels, Belgium. pp. 11–20. https://www.aclweb.org/anthology/W18-5102, 10.18653/v1/W18-5102.
[54] N.D. Gitari, Z. Zuping, H. Damien, J. Long, A lexicon-based approach for hate speech detection, International Journal of Multimedia and Ubiquitous Engineering 10 (2015) 215–230.
[55] Golbeck, J., Ashktorab, Z., Banjo, R.O., Berlinger, A., Bhagwan, S., Buntain, C., Cheakalos, P., Geller, A.A., Gnanasekaran, R.K., Gunasekaran, R.R., et al., 2017. A large labeled corpus for online harassment research, in: Proceedings of the 2017 ACM on web science conference, pp. 229–233.
[56] Gomez, R., Gibert, J., Gomez, L., Karatzas, D., 2020. Exploring hate speech detection in multimodal publications, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1470–1478.
[57] J.A. Gonzalez, L.F. Hurtado, F. Pla, Twilbert: Pre-trained deep bidirectional transformers for spanish twitter, Neurocomputing 426 (2021) 58–69.
[58] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep learning, volume 1, MIT press Cambridge, 2016.
[59] B. Haidar, M. Chamoun, A. Serhrouchni, A multilingual system for cyberbullying detection: Arabic content detection using machine learning, Advances in Science, Technology and Engineering Systems Journal 2 (2017) 275–284.
[60] Hassan, S., Samih, Y., Mubarak, H., Abdelali, A., 2020. Alt at semeval-2020 task 12: Arabic and english offensive language identification in social media, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 1891–1897.
[61] M.O. Ibrohim, I. Budi, A dataset and preliminaries study for abusive language detection in indonesian social media, Procedia Computer Science 135 (2018) 222–229.
[62] Ibrohim, M.O., Budi, I., 2019. Multi-label hate speech and abusive language detection in indonesian twitter, in: Proceedings of the Third Workshop on Abusive Language Online, pp. 46–57.
[63] Ishmam, A.M., Sharmin, S., 2019. Hateful speech detection in public facebook pages for the bengali language, in: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), IEEE. pp. 555–560.
[64] Jahan, M.S., 2020. Team oulu at semeval-2020 task 12: Multilingual identification of offensive language, type and target of twitter post using translated datasets, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 1628–1637.
[65] Jaki, S., De Smedt, T., 2019. Right-wing german hate speech on twitter: Analysis and automatic detection. arXiv preprint arXiv:1910.07518.
[66] Jha, A., Mamidi, R., 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data, in: Proceedings of the second workshop on NLP and computational social science, pp. 7–16.
[67] Kaati, L., Omer, E., Prucha, N., Shrestha, A., 2015. Detecting multipliers of jihadism on twitter, in: 2015 IEEE international conference on data mining workshop (ICDMW), IEEE. pp. 954–960.
[68] Kamble, S., Joshi, A., 2018. Hate speech detection from code-mixed hindi-english tweets using deep learning models. arXiv preprint arXiv:1811.05145.
[69] P. Kapil, A. Ekbal, A deep neural network based multi-task learning approach to hate speech detection, Knowledge-Based Systems 210 (2020).
[70] Karim, M.R., Chakravarthi, B.R., McCrae, J.P., Cochez, M., 2020. Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network, in: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), IEEE. pp. 390–399.
[71] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, D. Brown, Text classification algorithms: A survey, Information 10 (2019) 150.
[72] A. Kumar, S. Abirami, T.E. Trueman, E. Cambria, Comment toxicity detection via a multichannel convolutional bidirectional gated recurrent unit, Neurocomputing 441 (2021) 272–278.
[73] B.L.R. Kumar, B. Lahiri, A.K. Ojha, A. Bansal, Comma@ fire 2020: Exploring multilingual joint training across different classification tasks, in: Working Notes of FIRE 2020-Forum for Information Retrieval Evaluation, Hyderabad, India, 2020.
[74] Kumar, R., Reganti, A.N., Bhatia, A., Maheshwari, T., 2018. Aggression-annotated corpus of hindi-english code-mixed data. arXiv preprint arXiv:1803.09402.
[75] Kuratov, Y., Arkhipov, M., 2019. Adaptation of deep bidirectional multilingual transformers for russian language. arXiv preprint arXiv:1905.07213.
[76] Kwok, I., Wang, Y., 2013. Locate the hate: Detecting tweets against blacks, in: Proceedings of the AAAI Conference on Artificial Intelligence.
[77] Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., Schwab, D., 2019. Flaubert: Unsupervised language model pre-training for french. arXiv preprint arXiv:1912.05372.
[78] Liu, P., Li, W., Zou, L., 2019. Nuli at semeval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers, in: Proceedings of the 13th international workshop on semantic evaluation, pp. 87–91.
[79] Magdy, W., Darwish, K., Weber, I., 2015. # failedrevolutions: Using twitter to study the antecedents of isis support. arXiv preprint arXiv:1503.02401.
[80] Mahata, D., Zhang, H., Uppal, K., Kumar, Y., Shah, R., Shahid, S., Mehnaz, L., Anand, S., 2019. Midas at semeval-2019 task 6: Identifying offensive posts and targeted offense from twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 683–690.
[81] S. Malmasi, M. Zampieri, Challenges in discriminating profanity from hate speech, Journal of Experimental & Theoretical Artificial Intelligence 30 (2018) 187–202.
[82] Mandl, T., Modha, S., Kumar M, A., Chakravarthi, B.R., 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german, in: Forum for Information Retrieval Evaluation, pp. 29–32.
[83] Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., Patel, A., 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages, in: Proceedings of the 11th forum for information retrieval evaluation, pp. 14–17.
[84] Martin, L., Muller, B., Suárez, P.J.O., Dupont, Y., Romary, L., de la Clergerie, É.V., Seddah, D., Sagot, B., 2019. Camembert: a tasty french language model. arXiv preprint arXiv:1911.03894.
[85] Mathur, P., Sawhney, R., Ayyar, M., Shah, R., 2018. Did you offend me? classification of offensive tweets in hinglish language, in: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), pp. 138–148.
[86] D. Matsumoto, The handbook of culture and psychology, Oxford University Press, 2001.

[87] Mishra, P., Yannakoudakis, H., Shutova, E., 2019. Tackling online abuse: A survey of automated abuse detection methods. arXiv preprint arXiv:1908.06024.
[88] Mishra, S., Mishra, S., 2019. 3idiots at hasoc 2019: Fine-tuning transformer neural networks for hate speech identification in indo-european languages., in: FIRE (Working Notes), pp. 208–213.
[89] Mishraa, A.K., Saumyab, S., Kumara, A., 2020. Iiit_dwd@ hasoc 2020: Identifying offensive content in indo-european languages.
[90] Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., Group, P., et al., 2009. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. PLoS medicine 6, e1000097.
[91] Moon, J., Cho, W.I., Lee, J., 2020a. Beep! korean corpus of online news comments for toxic speech detection. arXiv preprint arXiv:2005.12503.
[92] Moon, J., Cho, W.I., Lee, J., 2020b. BEEP! Korean corpus of online news comments for toxic speech detection, in: Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Online. pp. 25–31. https://www.aclweb.org/anthology/2020.socialnlp-1.4.
[93] Mubarak, H., Darwish, K., Magdy, W., 2017. Abusive language detection on arabic social media, in: Proceedings of the first workshop on abusive language online, pp. 52–56.
[94] Mulki, H., Haddad, H., Ali, C.B., Alshabani, H., 2019. L-hsab: A levantine twitter dataset for hate speech and abusive language, in: Proceedings of the third workshop on abusive language online, pp. 111–118.
[95] R.U. Mustafa, M.S. Nawaz, J. Farzund, M. Lali, B. Shahzad, P. Viger, Early detection of controversial urdu speeches from social media, Data Sci. Pattern Recognit. 1 (2017) 26–42.
[96] V. Nahar, S. Al-Maskari, X. Li, C. Pang, Semi-supervised learning for cyberbullying detection in social networks, in: Australasian Database Conference, Springer, 2014, pp. 160–171.
[97] Nguyen, D.Q., Vu, T., Nguyen, A.T., 2020. Bertweet: A pre-trained language model for english tweets. arXiv preprint arXiv:2005.10200.
[98] Nikolov, A., Radivchev, V., 2019. Nikolov-radivchev at semeval-2019 task 6: Offensive tweet classification with bert and ensembles, in: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 691–695.
[99] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y., 2016. Abusive language detection in online user content, in: Proceedings of the 25th international conference on world wide web, pp. 145–153.
[100] J.T. Nockleby, Hate speech, Encyclopedia of the American constitution 3 (2000) 1277–1279.
[101] J. O'Brien, Encyclopedia of gender and society, volume 1, Sage, 2009.
[102] Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., Yeung, D.Y., 2019a. Multilingual and multi-aspect hate speech analysis. arXiv preprint arXiv:1908.11049.
[103] Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., Yeung, D.Y., 2019b. Multilingual and multi-aspect hate speech analysis, in: Proceedings of EMNLP, Association for Computational Linguistics.
[104] Ozdemir, A., Yeniterzi, R., 2020. Su-nlp at semeval-2020 task 12: Offensive language identification in turkish tweets, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 2171–2176.
[105] Özel, S.A., Saraç, E., Akdemir, S., Aksu, H., 2017. Detection of cyberbullying on social media messages in turkish, in: 2017 International Conference on Computer Science and Engineering (UBMK), IEEE. pp. 366–370.
[106] Pàmies, M., Öhman, E., Kajava, K., Tiedemann, J., 2020. Lt@ helsinki at semeval-2020 task 12: Multilingual or language-specific bert? arXiv preprint arXiv:2008.00805.
[107] Park, J.H., Fung, P., 2017. One-step and two-step classification for abusive language detection on twitter. arXiv preprint arXiv:1706.01206.
[108] J.W. Patchin, S. Hinduja, Bullies move beyond the schoolyard: A preliminary look at cyberbullying, Youth violence and juvenile justice 4 (2006) 148–169.
[109] Pathak, V., Joshi, M., Joshi, P., Mundada, M., Joshi, T., 2021. Kbcnmujal@ hasoc-dravidian-codemix-fire2020: Using machine learning for detection of hate speech and offensive code-mixed social media text. arXiv preprint arXiv:2102.09866.
[110] Pawar, R., Agrawal, Y., Joshi, A., Gorrepati, R., Raje, R.R., 2018. Cyberbullying detection system with multiple server configurations, in: 2018 IEEE International Conference on Electro/Information Technology (EIT), IEEE. pp. 0090–0095.
[111] de Pelle, R.P., Moreira, V.P., 2017. Offensive comments in the brazilian web: a dataset and baseline results, in: Anais do VI Brazilian Workshop on Social Network Analysis and Mining, SBC.
[112] Pitenis, Z., Zampieri, M., Ranasinghe, T., 2020. Offensive language identification in greek. arXiv preprint arXiv:2003.07459.
[113] G.K. Pitsilis, H. Ramampiaro, H. Langseth, Effective hate-speech detection in twitter data using recurrent neural networks, Applied Intelligence 48 (2018) 4730–4742.
[114] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources and Evaluation (2020) 1–47.
[115] Polignano, M., Basile, P., De Gemmis, M., Semeraro, G., Basile, V., 2019a. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets, in: 6th Italian Conference on Computational Linguistics, CLiC-it 2019, CEUR. pp. 1–6.
[116] M. Polignano, V. Basile, P. Basile, M. de Gemmis, G. Semeraro, Alberto: Modeling italian social media language with bert, IJCoL. Italian Journal of Computational Linguistics 5 (2019) 11–31.
[117] R. Pradhan, A. Chaturvedi, A. Tripathi, D.K. Sharma, A review on offensive language detection, in: Advances in Data and Information Sciences, Springer, 2020, pp. 433–439.
[118] Ptaszynski, M., Pieciukiewicz, A., Dybała, P., 2019. Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish twitter.
[119] Qian, J., Bethke, A., Liu, Y., Belding, E., Wang, W.Y., 2019. A benchmark dataset for learning to intervene in online hate speech. arXiv preprint arXiv:1909.04251.
[120] Quea, Q., Sunb, R., Xiec, S., 2020. Simon@ hasoc 2020: Detecting hate speech and offensive content in german language with bert and ensembles. FIRE (Working Notes), CEUR.
[121] Raja, R., Srivastavab, S., Saumyac, S., 2021. Nsit & iiitdwd@ hasoc 2020: Deep learning model for hate-speech identification in indo-european languages.
[122] Ranasinghe, T., Zampieri, M., Hettiarachchi, H., 2019. Brums at hasoc 2019: Deep learning models for multilingual hate speech and offensive language identification., in: FIRE (Working Notes), pp. 199–207.
[123] Rezvan, M., Shekarpour, S., Balasuriya, L., Thirunarayan, K., Shalin, V.L., Sheth, A., 2018. A quality type-aware annotated corpus and lexicon for harassment research, in: Proceedings of the 10th ACM Conference on Web Science, pp. 33–36.
[124] Ribeiro, M., Calais, P., Santos, Y., Almeida, V., Meira Jr, W., 2018. Characterizing and detecting hateful users on twitter, in: Proceedings of the International AAAI Conference on Web and Social Media.
[125] Rizos, G., Hemker, K., Schuller, B., 2019. Augment to prevent: short-text data augmentation in deep learning for hate-speech classification, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 991–1000.
[126] Rizwan, H., Shakeel, M.H., Karim, A., 2020. Hate-speech and offensive language detection in roman urdu, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2512–2522.
[127] Romim, N., Ahmed, M., Talukder, H., Islam, M.S., 2020. Hate speech detection in the bengali language: A dataset and its baseline evaluation. arXiv preprint arXiv:2012.09686.
[128] Rosa, H., Matos, D., Ribeiro, R., Coheur, L., Carvalho, J.P., 2018. A 'deeper' look at detecting cyberbullying in social networks, in: 2018 International Joint Conference on Neural Networks (IJCNN), IEEE. pp. 1–8.
[129] Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., Wojatzki, M., 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. arXiv preprint arXiv:1701.08118.
[130] Sai, S., Sharma, Y., 2020. Siva@ hasoc-dravidian-codemix-fire-2020: Multilingual offensive speech detection in code-mixed and romanized text. FIRE (Working Notes).
[131] Saleh Alatawi, H., Maatog Alhothali, A., Mustafa Moria, K., 2020. Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. arXiv e-prints, arXiv–2010.
[132] Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., Stranisci, M., 2018. An italian twitter corpus of hate speech against immigrants, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).
[133] Schmidt, A., Wiegand, M., 2017. 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.
[134] P. Shruthi, KM, A.K. Hate speech detection using deep learning and hybrid features, Inteligencia Artificial 23 (2020) 97–111.
[135] Sigurbergsson, G.I., Derczynski, L., 2019. Offensive language and hate speech detection for danish. arXiv preprint arXiv:1908.04531.
[136] Socha, K., 2020. Ks@ lth at semeval-2020 task 12: Fine-tuning multi-and monolingual transformer models for offensive language detection, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 2045–2053.
[137] Souza, F., Nogueira, R., Lotufo, R., 2019. Portuguese named entity recognition using bert-crf. arXiv preprint arXiv:1909.10649.
[138] Su, H.P., Huang, Z.J., Chang, H.T., Lin, C.J., 2017. Rephrasing profanity in chinese text, in: Proceedings of the First Workshop on Abusive Language Online, pp. 18–24.
[139] X. Tang, X. Shen, Y. Wang, Y. Yang, Categorizing offensive language in social networks: A chinese corpus, systems and an explanation tool, in: China National Conference on Chinese Computational Linguistics, Springer, 2020, pp. 300–315.
[140] Fatimah Alkomah, Xiaogang Ma, A Literature Review of Textual Hate Speech Detection Methods and Datasets, Information 13 (2022) 273.
[141] Susan Benesch, Countering dangerous speech: New ideas for genocide prevention, United States Holocaust Memorial Museum, Washington, DC, 2014.
[142] Cai, Yitao and Cai, Huiyu and Wan, Xiaojun, 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2506–2515.
[143] Chi, Zewen and Dong, Li and Wei, Furu and Wang, Wenhui and Mao, Xian-Ling and Huang, Heyan, 2020. Cross-lingual natural language generation via pre-training, in: Proceedings of the AAAI conference on artificial intelligence, pp. 7570–7577.

[144] Chung, Yi-Ling and Tekiroglu, Serra Sinem and Guerini, Marco, 2021. Towards knowledge-grounded counter narrative generation for hate speech. arXiv preprint arXiv:2106.11783.
[145] Das, Abhishek and Wahi, Japsimar Singh and Li, Siyao, 2020. Detecting hate speech in multi-modal memes. arXiv preprint arXiv:2012.14891.
[146] Dowlagar, Suman and Mamidi, Radhika, 2021. A survey of recent neural network models on code-mixed indian hate speech data, in: Forum for Information Retrieval Evaluation, pp. 67–74.
[147] EPRS, 2020. The ethics of artificial intelligence: Issues and initiatives. European Parliamentary Research Service, STOA, PE 634.452.
[148] Fanton, Margherita and Bonaldi, Helena and Tekiroglu, Serra Sinem and Guerini, Marco, 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. arXiv preprint arXiv:2107.08720.
[149] Karim, Md and Dey, Sumon Kanti and Islam, Tanhim and Shajalal, Md and Chakravarthi, Bharathi Raja and others, 2022. Multimodal hate speech detection from bengali memes and texts. arXiv preprint arXiv:2204.10196.
[150] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, Davide Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, Advances in Neural Information Processing Systems 33 (2020) 2611–2624.
[151] Lee, Roy Ka-Wei and Cao, Rui and Fan, Ziqing and Jiang, Jing and Chong, Wen-Haw, 2021. Disentangling hate in online memes, in: Proceedings of the 29th ACM International Conference on Multimedia, pp. 5138–5147.
[152] Li, Liunian Harold and Yatskar, Mark and Yin, Da and Hsieh, Cho-Jui and Chang, Kai-Wei, 2019. Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557.
[153] Mathew, Binny and Dutt, Ritam and Goyal, Pawan and Mukherjee, Animesh, 2019. Spread of hate speech in online social media, in: Proceedings of the 10th ACM conference on web science, pp. 173–182.
[154] Mathew, Binny and Saha, Punyajoy and Yimam, Seid Muhie and Biemann, Chris and Goyal, Pawan and Mukherjee, Animesh, 2021. Hatexplain: A benchmark dataset for explainable hate speech detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 14867–14875.
[155] Usman Naseem, Imran Razzak, Peter W Eklund, A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter, Multimedia Tools and Applications 80 (2021) 35239–35266.
[156] Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, Weiping Wang, Modeling intra and inter-modality incongruity for multi-modal sarcasm detection, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 1383–1392.
[157] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu, et al., Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020) 1–67.
[158] Rana, Aneri and Jha, Sonali, 2022. Emotion Based Hate Speech Detection using Multimodal Learning. arXiv preprint arXiv:2202.06218.
[159] Sabou, Marta and Bontcheva, Kalina and Derczynski, Leon and Scharl, Arno, 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines, in: Proceedings of the ninth international conference on language resources and evaluation (LREC'14), pp. 859–866.
[160] Siva Sai, Naman Deep Srivastava, Yashvardhan j. Sharma, Explorative Application of Fusion Techniques for Multimodal Hate Speech Detection, SN Computer Science 3 (2022) 1–13.
[161] Sandulescu, Vlad, 2020. Detecting hateful memes using a multimodal deep ensemble. arXiv preprint arXiv:2012.13235.
[162] Tekiroglu, Serra Sinem and Bonaldi, Helena and Fanton, Margherita and Guerini, Marco, 2022. Using Pre-Trained Language Models for Producing Counter Narratives Against Hate Speech: a Comparative Study. arXiv preprint arXiv:2204.01440.
[163] Tekiroglu, Serra Sinem and Chung, Yi-Ling and Guerini, Marco, 2020. Generating counter narratives against online hate speech: Data and strategies. arXiv preprint arXiv:2004.04216.
[164] Alice Tontodimamma, Eugenia Nissi, Annalina Sarra, Lara Fontanella, Thirty years of research into hate speech: topics of interest and their evolution, Scientometrics 126 (2021) 157–179.
[165] Tzeng, Eric and Hoffman, Judy and Zhang, Ning and Saenko, Kate and Darrell, Trevor, 2014. Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474.
[166] Velioglu, Riza and Rose, Jewgeni, 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. arXiv preprint arXiv:2012.12975.
[167] Bertie Vidgen, Leon Derczynski, Directions in abusive language training data, a systematic review: Garbage in, garbage out, Plos one 15 (2020).
[168] Wang, Xinyu and Sun, Xiaowen and Yang, Tan and Wang, Hongbo, 2020. Building a bridge: A method for image-text sarcasm detection without pretraining on image-text data, in: Proceedings of the first international workshop on natural language processing beyond text, pp. 19–29.
[169] Matthew L Williams, Pete Burnap, Luke Sloan, Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation, Sociology 51 (2017) 1149–1168.
[170] Yang, Chuanpeng and Zhu, Fuqing and Liu, Guihua and Han, Jizhong and Hu, Songlin, 2022. Multimodal Hate Speech Detection via Cross-Domain Knowledge Transfer, in: Proceedings of the 30th ACM International Conference on Multimedia, pp. 4505–4514.
[171] Wenjie Yin, Arkaitz Zubiaga, Towards generalisable hate speech detection: a review on obstacles and solutions, PeerJ Computer Science 7 (2021).
[172] Zhang, Yizhe and Sun, Siqi and Galley, Michel and Chen, Yen-Chun and Brockett, Chris and Gao, Xiang and Gao, Jianfeng and Liu, Jingjing and Dolan, Bill, 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. arXiv preprint arXiv:1911.00536.
[173] Tsapatsoulis, N., Anastasopoulou, V., 2019. Cyberbullies in twitter: A focused review, in: 2019 14th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), IEEE. pp. 1–6.
[174] de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., Nissim, M., 2019. Bertje: A dutch bert model. arXiv preprint arXiv:1912.09582.
[175] Wadhwa, P., Bhatia, M., 2013. Tracking on-line radicalization using investigative data mining, in: 2013 National Conference on Communications (NCC), IEEE. pp. 1–5.
[176] Wang, S., Liu, J., Ouyang, X., Sun, Y., 2020. Galileo at semeval-2020 task 12: Multi-lingual learning for offensive language identification using pre-trained language models. arXiv preprint arXiv:2010.03542.
[177] Warner, W., Hirschberg, J., 2012. Detecting hate speech on the world wide web, in: Proceedings of the second workshop on language in social media, Association for Computational Linguistics. pp. 19–26.
[178] Waseem, Z., 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter, in: Proceedings of the first workshop on NLP and computational social science, pp. 138–142.
[179] Waseem, Z., Hovy, D., 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL student research workshop, pp. 88–93.
[180] H. Watanabe, M. Bouazizi, T. Ohtsuki, Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection, IEEE access 6 (2018) 13825–13835.
[181] Wiedemann, G., Yimam, S.M., Biemann, C., 2020. Uhh-lt & lt2 at semeval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection. arXiv preprint arXiv:2004.11493.
[182] Wiegand, M., Ruppenhofer, J., Schmidt, A., Greenberg, C., 2018a. Inducing a lexicon of abusive words–a feature-based approach.
[183] Wiegand, M., Siegel, M., Ruppenhofer, J., 2018b. Overview of the germeval 2018 shared task on the identification of offensive language.
[184] Wigand, C., Voin, M., 2017. Speech by commissioner jourová–10 years of the eu fundamental rights agency: A call to action in defence of fundamental rights, democracy and the rule of law.
[185] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A. M., 2020. Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online. pp. 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6.
[186] Wulczyn, E., Thain, N., Dixon, L., 2017. Ex machina: Personal attacks seen at scale, in: Proceedings of the 26th international conference on world wide web, pp. 1391–1399.
[187] Xiang, G., Fan, B., Wang, L., Hong, J., Rose, C., 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus, in: Proceedings of the 21st ACM international conference on Information and knowledge management, pp. 1980–1984.
[188] Yang, H., Lin, C.J., 2020. Tocp: A dataset for chinese profanity processing, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, pp. 6–12.
[189] Yin, W., Kann, K., Yu, M., Schütze, H., 2017. Comparative study of cnn and rnn for natural language processing. arXiv preprint arXiv:1702.01923.
[190] Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R., 2019a. Predicting the type and target of offensive posts in social media. arXiv preprint arXiv:1902.09666.
[191] Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R., 2019b. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). arXiv preprint arXiv:1903.08983.
[192] Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., Çöltekin, Ç., 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). arXiv preprint arXiv:2006.07235.
[193] Z. Zhang, L. Luo, Hate speech detection: A solved problem? the challenging case of long tail on twitter, Semantic Web 10 (2019) 925–945.
[194] Y. Zhou, Y. Yang, H. Liu, X. Liu, N. Savage, Deep learning based fusion approach for hate speech detection, IEEE Access 8 (2020) 128923–128929.



Md Saroar Jahan: received his BSc in electrical engineering from the Department of Electrical Engineering & Electronics, United International University, Bangladesh, in 2013, and his Msc degree from University of Oulu in 2020 in Computer Science and Engineering. He is currently pursuing a Ph.D. degree in Computer Science at the University of Oulu, Finland. His research interests include Big data analysis, deep learning techniques, and Natural Language processing.

Mourad Oussalah: received his MSc in Control engineering form University of Paris XII France in 1994 and PhD degree in Robotics/Computer Science in Evry Val Essonnes University in France. After academics positions in KU Leuven, City University of London and University of Birmingham, he is since 2016 with University of Oulu as a Research Professor leading the Social Mining Research Group. His research focuses on data mining and uncertainty handling where he published more than 220 papers and led several projects in the field.