

Social Sciences Spectrum

A Double-Blind, Peer-Reviewed, HEC recognized [Y-category](#) Research Journal

E-ISSN: [3006-0427](#) P-ISSN: [3006-0419](#)

Volume 05 Issue 04, 2026

Web link: <https://socialsciencespectrum.com/index.php/sss>



Check for updates

Challenges and Innovations in Machine Learning-Driven Hate Speech Detection on X (Formerly Twitter): A Methodological and Taxonomic Review

Jawaid Ahmed Siddiqui¹ Mr. Noor Nabi Dahar²

Article Information [YY-MM-DD]

Received	2026-08-02	Revised	2026-09-08	Accepted	2026-10-01
----------	------------	---------	------------	----------	------------

Citation (APA):

Siddiqui, J, A & Dahar, N, N (2026). Challenges and innovations in machine learning-driven hate speech detection on X (formerly Twitter): A methodological and taxonomic review. *Social Sciences Spectrum*, 5(4), 159-180. <https://doi.org/10.71085/sss.05.04.611>

Abstract

Detecting hate talk on public news manifestos, particularly X (formerly Twitter), has enhanced a bigger focus on document excavating and the study of natural language processing (NLP). This systematic literature review (SLR) offers an itemized test of current approaches, machine intelligence forms, challenges, and time engaged in hate talk discovery. The review covers 91 research documents written from 2016 to 2022, preferred through an all-encompassing election process. It evaluates a range of directed and alone machine intelligence methods, to a degree Support Vector Machines (SVM), Random Forests, Neural Networks, and fresher turbine-located models like BERT and GPT. The study also addresses detracting issues like dataset bias, model generalizability, and exposure to opposing attacks. Additionally, it labels research breaches and implies future guidance for reinforcing hate talk discovery arrangements. The understandings from this review aim to support the invention of more exact and flexible models for detecting and fighting connected to the internet hate talk.

Keywords: Hate Speech Detection, Machine Learning, Text Classification, Social Media, BERT and GPT.

¹ Department of Computer Science, Sukkur IBA University, Sukkur, Sindh 65200, Pakistan.

² Department of Computer Science, Sukkur IBA University, Sukkur, Sindh 65200, Pakistan.

Corresponding Author: Jawaid Ahmed Siddiqui, **Correspondence through:** jawaid@iba-suk.edu.pk



Content from this work may be used under the terms of the [Creative Commons Attribution-Share-Alike 4.0 International License](#) that allows others to share the work with an acknowledgment of the work's authorship and initial publication in this journal.

Introduction

The recognition of social media platforms like X (formerly Twitter) has soared in the last ten years. According to Zhang and Luo (2019), the proliferation and dissemination of hate speech are fueled by these platforms since they enable user anonymity and freedom of expression. Additionally, with 300 million active users monthly, X is the most secondhand social socializing for professional or personal gain sites, according to Silva et al. (2016). Hate speech is common on X, despite its popularity and relevance. In addition to being a dossier source for studies on bad language, it has quickly become one of ultimate well-known friendly networks for the automated discovery of hate talk in text (Founta et al., 2019; Magu et al., 2017; Mondal et al., 2018). Hate talk acts on the rise on public media programs. As a result, consumers enhance hostility that in proper sequence affects guests and leads to weighty confrontations in the here and now. In order to avoid the publication of hate talk, public media businesses repeatedly kill it. Because of its extensive acknowledgement and profusion of accessible dossiers, this study generally concentrates on English-language public television posts advised to X (Arango et al., 2022).

The urge to mechanize the process of detecting hate talk connected to the internet is growing, opposite to the severity of manual protection. A calculating-located solution can beat the nation concerning this, but non-automatic tasks have a direct impact on the reaction opportunity. Thus, it is detracting to aid in the incident of automatic text-located hate talk labeling electronics. Natural language processing (NLP) studies have been prompted by these data. A climbing bulk of biographies address the issue of hate talk. The scientific society has reportedly named this question as directed document categorization utilizing machine intelligence and machine intelligence (Schmidt & Wiegand, 2017). In 2017, X was seen as one of the leading public radio firms. In answer to assertive management, it revised allure solitude tactics. For instance, these rules apply to some tweets that advance intensity, hate, badgering, self-murder, self-harm, or any number of added negative empathy (Langham & Gosha, 2018). Researchers have now accelerated their attempts to label hate speech in X postings.

Nevertheless, non-English datasets are partially restricted in height. It bears be eminent that English is the language uttered by most of the experience's people. When looking for hate talk, it is again the basic prominence. Since the description of hate speech can change, it is not forever smooth to label it. Hate talk is a legal phrase in many countries with their own government and is ultimately usually secondhand discussion to describe this phantasm (Schmidt & Wiegand, 2017). Many authors have presented differing definitions of hate talk in their everything. "Hate speech is a terminology that attacks or belittles, that incites intensity or hate against groups, established particular characteristics, in the way that tangible characteristic, ritual, descent, governmental or cultural inception, preference of sexual partner, masculine identity or possible choice. It can accompany various semantic styles, even in nice forms or when humor is secondhand." (Fortuna & Nunes, 2018), illustration from a reasoning of miscellaneous writings available in the drama on this point. Here are some instances of hate talk on X: "You disdain football, your fa**ot," and "X consumer Pu**y is a** ni**a." The number ten. Much former than their plans, various approaches to hate talk discovery have been grown in the current age. In contrast to detecting and classifying mean content, most assessments devote effort to something judgment of non-hate content (Zhang & Luo, 2019).

The fast evolution of social media language has made it difficult for most of these initiatives to find workable solutions (Fortuna & Nunes, 2018). Therefore, it is essential to have a thorough familiarity with the existing literature on the topic. Despite years of progress, there has yet to be

an orderly information review (SLR) in the field of hate talk detection. To stay up-to-date with the newest developments, such as unanswered questions and research gaps related to a specific topic, SLR articles are crucial. Both governments and public news platforms have led to consider the need to enhance hate talk discovery technologies in consideration of devising a more welcoming atmosphere for everyone. As seen in Figure 1, one benefit of SLR is its emphasis on improving hate speech detection technologies, which might assist both the government and social media corporations in their efforts to avoid this occurrence. Investigating hate speech identification using X data has a number of challenges. This study primarily aims to summarize research that has used English-language X as its data source. As a result of its thorough examination of the subject, this study provides a wealth of information about current best practices and suggestions for further research. This study presents a comprehensive analysis of the computational approaches to hate speech identification. In order to open up additional avenues for study, the current obstacles are first reviewed, followed by an identification of those that still need fixing.

Social publishing platforms have revolutionized ideas, but they have also facilitated the extensive dissemination of hate speech. X, accompanying over 300 million active consumers, serves as a significant medium for both helpful discussions and harmful wordiness. The anonymity and freedom of expression presented by social media programs contribute to the proliferation of hate talk, necessitating the incident of automated detection structures.

This study systematically reviews existing essays on machine learning-based hate talk detection on X. The objective is to judge methodologies, highlight challenges, and survey opportunities for improvement. Despite progress, the dynamic nature of dialect and adversarial manipulations continue to preclude accurate hate speech discovery. The study provides a structured reasoning of various machine learning methods, emphasizing their strengths and restraints.

Literature Review

This section emphasizes the significance of the work's contributions by highlighting the most recent survey and review publications. On the subject of hate speech identification, there have been a few reviews and surveys (e.g., Alrehili, 2019; Ayo et al., 2020; Fortuna & Nunes, 2018; Langham & Gosha, 2018; Modi, 2018; Shushkevich & Cardiff, 2019). Using an informatics-based systematic approach, the research undertaken by Fortuna and Nunes (2018) reviewed previous efforts on hate speech identification. Following Schmidt and Wiegand (2017), which offered public news platforms have led consideration to the need to enhance hate talk discovery technologies in consideration of devising more. The survey is short and mostly deals with feature extraction, as stated in Schmidt and Wiegand (2017). Among the many things covered in the research by Fortuna and Nunes (2018) were the characteristics of hate speech, a description of the terminology required to examine it, a review of data on detection methods, and a contrasting of hate talk to additional corresponding forms. Research on bullying became their primary focus later on. With less than twenty articles devoted to hate speech, they detailed several English datasets and current difficulties involving various social media sites.

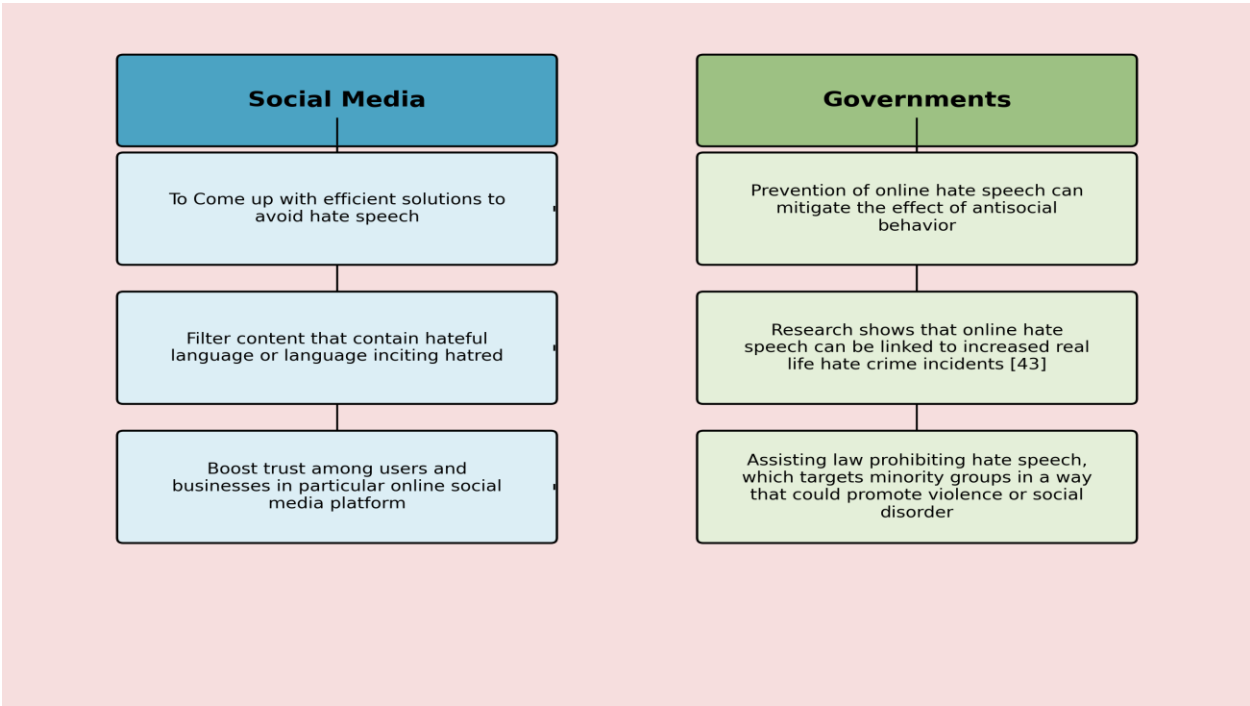
Previous studies have surveyed different facets of hate speech discovery, including feature extraction, model accomplishment, and dataset limitations. Notable surveys (Silva et al., 2016; Zhang & Luo, 2019) have resolved traditional machine intelligence techniques to a degree Naïve Bayes, SVM, and Random Forests, while others (Magu et al., 2017; Mondal et al., 2018) have directed on deep learning approaches containing Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). More recently, transformer-based models like BERT and

GPT have displayed superior performance in NLP tasks, containing hate speech discovery (Founta et al., 2019).

Despite these advancements, various challenges remain, containing dataset biases, interpretability of deep learning models, and the ability to discover implicit hate talk. This review builds on existing research by providing an itemized taxonomy of methods and evaluating their influence across different datasets.

Both traditional ML and more advanced deep learning techniques were part of the strategies. In addition, the results took into account many languages, English included. Machine learning approaches, including generic metadata designs, threshold settings, and divergences, were used to classify hate speech on X in a recent review study (Ayo et al., 2020). In addition, they went over the pros and cons of using both standalone and combined machine learning methods for the categorization procedure. They also showcased the hate speech benchmark dataset, which was used to evaluate the categorization paradigm's implementation. There are a number of surveys and reviews accessible, however they are severely lacking in coverage. Partially absent from these books are SLR recommendations, current reviews, and survey research. In addition, many studies have limitations, such as not focusing just on X or the English prose. In contrast, the survey by Poletto et al. (2020) looks at the gauge datasets used to discover hate speech and bad language on various friendly socializing for professional or personal gain floors. Topics of interest, speech inclusion, annotation methods, and the process of dataset construction were all on their inspection. Previous evaluations and surveys have ignored the far-reaching inclusion of X, despite the fact that it forms the base for many existent magazines on hate talk. When it meets expectations of mechanical theme detection of hate talk, X is with ultimate standard-friendly network (Founta et al., 2019; Magu et al., 2017).

Figure 1: *Benefits of Study and Analysis of Hate Speech Detection Methods*



Consequently, X has facilitated better global communication and provides users with a handy public forum. The number of publications evaluated in this research was far higher than in previous studies. Along with identifying difficulties and some avenues for further study, Moy et al. (2021) provided a brief overview of both English and non-English literature. In order to demonstrate the lack of generalizability in hate speech algorithms, Yin and Zubiaga (2021) surveyed existing hate speech detection methods. There is an analysis of potential future research approaches proposed at embellishing hate speech discovery's generalizability. A survey of machine intelligence plans and approaches for labeling hate talk in social news programs was given by Mullah and Zainon (2021). They secondhand ML algorithms to question the main features of hate talk classification. In order to find the research breach and open questions, the achievement and failure of each method are judged. In Istaiteh et al. (2020), many definitions of hate talk were checked, and miscellaneous difficulties accompanying the dossier accumulating and glossary were emphasised. The authors supported a high-level survey of the differences with nine datasets that fashioned the use of miscellaneous textbook languages and manifestos.

There is a brief description of the information sources and feature selection process that is based on five prior publications that used machine learning methods. The authors of the study created a multiple-view support vector machine (SVM) model, analyzed reported mistakes, and used it to categorize hate speech based on three datasets from three separate platforms. Consequently, they bring up a number of broad difficulties. Data sources, features, and ML algorithms were the primary areas of attention in the review of racist and sexist hate speech strategies by Rini et al. (2020). They provided high-level overviews of the text corpus, discussed common feature representation techniques, and briefly compared many Machine learning models.

In a brief orderly test of the drama, MacAvaney et al. (2019) thought-out newspapers written before January 2020. Their SLR reasoning just accepted into report documents written in English and Indonesian for journals and colloquiums. X, Facebook, Instagram, Online Today, YouTube, Yahoo!, and Wikipedia comments were among the many data sources taken into account. Due to the limitations of their SLR, just a little discovery and proposal have been made. A more recent comprehensive assessment of hate speech identification approaches using text-based features and machine learning algorithms was conducted by Alkomah and Ma (2022). The review primarily focused on the important datasets. Various topics were used to review the papers they had compiled. Three groupings of challenges and three points of orientation were supplied by them. Despite the fact that their assessment was limited to an English hate speech dataset, this SLR offers a more comprehensive analysis and various taxonomies for the papers were chosen.

Research Methodology

In order to present an equalized survey of the drama, this portion analyses the means secondhand in this place of SLR research. Questions were planned, search strands were grown, research tests were picked, dossier was elicited, and synthesised as a constituent of the systematic review process. Finding, judging, resolving, and combining complete acts on the labeling of hate talk on X so that present a summary of all the works that have happened exhausted the research concerning this subject was the purpose of the study, in addition to surveying current breakthroughs inside the field of hate talk discovery. To complete the activity of the SLR, the orders defined in Kitchenham and Charters (2007) were applied. These research questions (Qs) were the action for the survey: Research Question 1: How many announcements and colloquiums were guiding the databases secondhand apiece preferred studies? Question 2: How did the studies that were preferred recognize hate talk on X? Thirdly, what confirmation means were working in this place research

rule? Question 4: In this research rule, what were ultimate favorite accomplishment verification? Question 5: How ability hate talk be discovered on X utilizing the dossier that was accessible? Question 2: How acted the studies that were preferred recognize hate talk on X? Thirdly, what confirmation systems were working in this place research rule? Question 4: In this research rule, what were ultimate popular conduct versification? Question 5: How ability hate talk be discovered on X utilizing the dossier that was convenient? Question 6: What were the troubles concerning this field of study? 7-What grant permission the future holds for X's talent to label hate talk in English-word tweets?

In the first question, the aim was to draw attention to the ratios of conference and journal links to the databases utilized by the chosen research. Question 2's objective was to highlight the papers' most frequently used machine learning approaches. Question 3's objective was to showcase the selected publications' typical practical validation procedures.

Presenting the most often used performance metrics in this research domain was the objective of the fourth quarter. Question 5's objective was to go into the most popular data points utilized in this research domain, including topics such as class kinds, private and public tweet sizes, and potential linkages. The sixth question sought to determine which of the community's publications sought to address the most pressing problems. Lastly, Q7 aimed to draw attention to possible avenues for future research in this field of study.

As seen in Figure 2, the chosen databases displayed an abundance of scientific expertise in several high-impact research articles. For this study, their diversity and reliability were deemed important. For this SLR, it was also determined that these databases were adequate since they included interdisciplinary research on investigations from several academic disciplines, including the social sciences and the sciences. As shown in Table 1, there are eight electronic databases that were part of this SLR. The goal of the well-structured search query was to locate all publications that were pertinent to the topic. In the first step, the keywords were identified; the controlled society has used the term "hate talk" ultimate often up until now. It is even a legal phrase in certain places and is often used to describe this sort of damaging user-generated material (Schmidt & Wiegand, 2017).

As a result, it has become the preferred phrase in this field of study. In addition to "cyber hate," other studies have used the phrases "hateful language" to describe the same phenomenon. There were references to "cyber hate" in Cao et al. (2020) and Zhang et al. (2018), and references to "hateful language" in Modi (2018). The search query was formed by combining these terms with the Boolean operators (AND) and (OR). Together with the primary phrases for detection, the search string keywords that were most successful were: ('cyber hate' OR 'hate speech' OR 'hateful language'). The search query for SLR is as follows: combination of ("cyber-hate", "hate speech", "hateful language") and ("detection", "recognition", "classification").

Table 1: Digital Databases Used

Digital Databases	Category	URL
Web of Science	Multidisciplinary	https://wokinfo.com/mbl/
ACM Digital Library	Discovery Engine	https://dl.acm.org/
IEEE Explore	Engineering technical	https://ieeexplore.ieee.org/Xplore/home.jsp
Wiley Online Library	Authentic artworks	https://onlinelibrary.wiley.com/
Springer Link	Superior Quality	https://link.springer.com/
Taylor and Francis	Reliable research	https://www.tandfonline.com/
Scopus	Biggest Database	https://www.scopus.com/
ScienceDirect	Multidisciplinary	https://www.sciencedirect.com/

When deciding which articles to include or expel in the SLR, particular criteria were secondhand. The studies that were eventually preferred were those that met the addition and expulsion tests. The gathered documents were deliberate expected the most appropriate and original works on the subject. As can be seen in Tables 2 and 3, the publications that were not included in this study were intentionally chosen because they did not meet its objectives. 1) Criteria for Inclusion Publications written in English and conducted within the past eight years (2015–2022) on the topic of hate speech detection are eligible for inclusion. Put simply, research that utilized English datasets for hate speech identification on X was the only one considered. Reviews and survey pieces, which fall within the umbrella of hate speech, were the primary targets. In addition, there were articles covering techniques and methodological recommendations, as well as research comparing and evaluating various approaches. All datasets were authored in English. Reviews and survey pieces, which fall within the umbrella of hate speech, were the primary targets. The articles also included techniques and methodological recommendations as well as evaluation and comparison research. 2) DISCRETIONARY REQUIREMENTS Papers written in a language other than English and research that used statistics from social media platforms other than YouTube, Facebook, Instagram, etc. were both deemed ineligible. Only studies that provided textual data were included, and no other multimedia (pictures, videos, or audio) was included in the analysis.

Articles that provided statistics on hate speech on X from several platforms and in many languages were also excluded. Additionally, news that did not focus on machine intelligence methods, machine intelligence, detection and categorization models, or performance assessments were not included, nor were studies that did not include a strategy for dealing with problems linked to hate speech. Also not included were papers that had no bearing on hate speech whatsoever. Part D: Evaluating Material Goods The quality of each article was considered alongside the inclusion and exclusion criteria. This type of research usually yields skewed findings. That they all meet the SLR criteria gives the reader even more assurance.

Table 2: Inclusion Criteria

Inclusion Criteria

Studies published throughout 2015-2022 related to hate speech detection

Research written in English language

Studies involved hate speech detection from X (Twitter) using English dataset

Reviews or surveys

Comparative and evaluation studies

Proposal of methods and methodology

Main focus is on hate speech and its categories

Table 3: Exclusion Criteria

Exclusion Criteria

Studies using other social media such as YouTube, Facebook, Instagram, etc

Only text-based studies are accounted for, and those with any media (image, video, and audio) are not included

Studies using languages other than English are excluded

Datasets from multiple platforms are not included

Does not report an approach for addressing hate speech-related issues

Does not focus on ML techniques, NLP, classification and detection models, or performance evaluation

To determine how well each of the ninety-one papers included in this SLR performed, criteria derived from Kitchenham and Charters (2007) were used. The purpose of these standards is to reassure readers about the treatment of the SLR concerns, not to condemn scholarly publications (Kitchenham & Charters, 2007). The questions that informed the criteria's adaptation are as follows: Question 1: Has the study publication adequately defined its inclusion and exclusion criteria? Question 2: Is it the case that all relevant research has been uncovered by the literature search? Third Quality Assessment: Has the connected dossier utilized in the research been sufficiently expressed? QA4: Is the case and setting of the study well-defined? Articles were marked as "yes" for QA1 if they clearly describe the inclusion criteria; "no" for studies that do not offer or cannot provide the criteria; and "not fully" for studies that do not explicitly discuss the criteria. The authors of the study were required to search no less than five electronic libraries containing references for QA2, and if they did not, the answer was "no." When three to four but less than five electronic libraries were consulted, and "not fully" when just a scarcely any of mathematical athenaeums were combed through. QA3 was apparent as "yes" when all of the dossier applied in the study was definitely determined; "no" when some dossier was gone or keep not be reported; and "not completely" when some dossier was not particularized.

In QA4, "agreed" was outlined as scientists distinctly stating the study's material, "no" as information that is not easily accessible, and "not fully" as information that is only partially supplied. The primary focus was locating articles that offered sufficient data to answer the research questions. A person's degree of certainty that the item satisfies the criteria can be approximated.

All 91 studies were determined to be suitable for further investigation. Chapter E: The Search Procedure All research articles published throughout the past eight years, from 2015 to 2022, were included in this SLR. The following sources were consulted for the initial 1692 articles: Scopus (235), ACM Digital Library (448), Taylor & Francis (84), Springer-Link (397), Wiley Online Library (30), WOS (205), IEEE Explore (234), and Science Direct (59). Filtering revealed 56 duplicate articles out of 1692. Fifteen hundred and thirty-three articles were deemed unsuitable after reviewing their abstracts and titles.

There were 66 articles that were not included in the final data extraction and full-text reading procedure. This SLR research analyzed 91 publications in total that fulfilled the inclusion criteria determined by the search procedure. Some notes were made along the way, and those notes turned into valuable insights that formed the basis of this study. Details on the dataset, methods and methods working, useful appraisal and confirmation patterns, discovery troubles, and recommendations for future research were the main spotlights of each work, as were the names and sources of the relevant journals and conferences. For every one of the SLR questions, data was gathered. Figure 3 displays the search technique that was derived from the PRISMA recommendations by Moher (2009), which were the central aspect of this study's design. Table 4 displays all the statistical data on the articles that were either included or removed from the current SLR study, as well as the overall count of documents that were stated, secluded, worthy, and ultimately contained.

Results and Analysis

In all, 91 not enough 136 brochures join the addition and forbiddance tests on account of the search approach defined above. According to the search query, no issuing items were found in 2015 during the opportunity of the search. Research on hate talk labeling attended during the whole of ancient times seven age is proved in Figure 4. The amount of brochures top this matter raised considerably in 2018 and 2019. Old age is still young (2022) and much research would have existed before old age's end, so it's no surprise that skill was dropped that old age. The largest number of colloquiums was in 2019, accompanying 18 items, when in fact ultimate number of journals was in 2018, accompanying 13 documents (Figure 5). Conferences were an excellent habit for scientists to present and debate their judgments. Conference documents created 53% of the preferred advertisements, while chronicle documents justified 47%. As a first question, by virtue of what many journals and colloquiums guide the databases that were secondhand apiece preferred studies?

Most of the written items emanated from the IEEE Digital Library (27, or 30% of the total) and the ACM Digital Library (19, or 21% of the total). There were four chronicle studies and nineteen booklets (21% of the total) in the ACM Digital Library. In contrast to the IEEE Digital Library's 23 conference papers and 4 journal articles, the ACM Digital Library's holdings comprised 15 journal studies. One hundred and forty-four items, or fifteen percent, were published in Scopus; seven of them were journal papers and seven were conference papers. Next came Springer with thirteen (14% of the total) published publications; Figure 6 shows that of these, ten (11%) were journal articles and three (3% were conference papers). Journal papers made up all five of Elsevier's published publications (6%).

Three (3) journal studies were published by Taylors Francis. Question2: How Did the Selected Studies Determine When Hate Speech Occurred on X? The majority of the methodologies and procedures used in the chosen research are detailed in this section. In Figure 7, the taxonomy of the hate speech detection methods used in the selected articles can be seen. The following are the

three major groups identified using these techniques: 1) Alphabetical list of ML Linking specific output vectors to many input vectors is a common practice in machine learning, which is used to conduct most classification tasks. Algorithms for machine learning might essentially function as either standalone or combined algorithms.

Figure 2: PRISMA Flow Chart

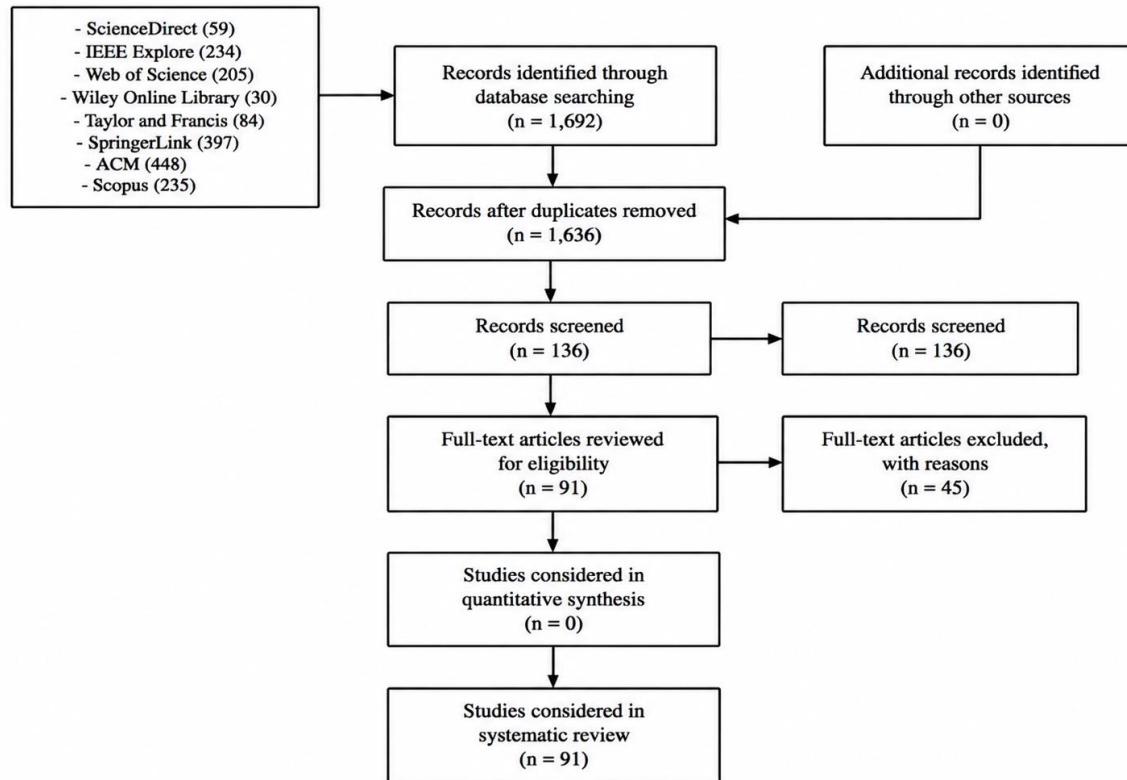
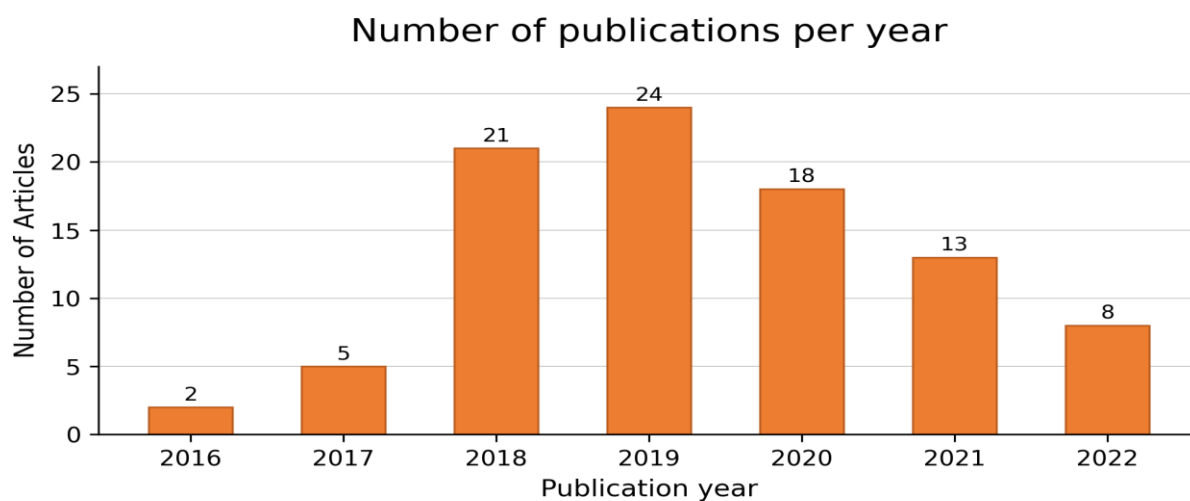


Figure 3: Number of Publications per Year



Systematic adaptations to natural language processing methods allow computers to comprehend and carry out orders given in this manner. These methods make it possible to read and comprehend

the material. Linguistic knowledge is essential for the functioning of language processing. Semantic and syntactic aspects are among several that may be discovered by natural language processing methods. The key difference between the algorithms is in the attributes that are employed to detect hate speech (Schmidt & Wiegand, 2017).

Automatic attribute recognition is only one of several benefits of combining natural language processing with machine learning (Al-Makhadmeh & Tolba, 2020). Feature extraction is an important part of automated categorization. In order to learn a pattern, the machine learning classifier examines the characteristics derived using natural language processing algorithms. The processes and orders working in the preferred research are detailed in Table 6. Also contained are the studies' goals, dossier beginnings, hate talk classifications, and conduct judgment criteria. Hate talk discovery from the use of different approaches, chosen according to the nature of the issue at hand. The employment of several hybrids, ensemble, or comparative methods was also noted in a few publications. Incorporating techniques from other domains or combining various algorithms from current ones is what makes a hybrid machine-learning approach possible. Ensemble learning uses many learning algorithms to get better predicted results than would be possible with only one of the basic learning algorithms. In contrast to ensemble approaches, which work independently to vote the result, hybrid methods only forecast one possible conclusion, which does not account for voting. The chosen studies' taxonomy of methods relies on several models, as seen in Table 6. The most used methods were the old-school classifiers like logistic reversion (LR), support heading machines (SVM), and childlike Bayes (NB).

Recent contrasting studies have proved that these classifiers work well for text classification (Mullah & Zainon, 2021). The influence of these unoriginal classifiers, however, relies laboriously on the use of appropriate traits. Considering the dossier's architecture is while main as selecting the highest in rank feature origin method (Suhaidi et al., 2021). The character of classification may be enhanced by mixing various feature election approaches (Oskouei & Razavi, 2018). On the other hand, deep learning forms forbiddance believe in manually created traits. There has existed a brief rise in the use of deep learning to label hate talk since 2017 (Badjatiya et al., 2017). One of its substances is that they can explore dossier likenesses that are good candidates for classification (Naswir et al., 2022). Two common deep education methods are convolutional affecting animate nerve organs networks (CNNs) and long temporary thought (LSTM). When it comes to visual and audio entertainment transmitted via radio waves, broadcast, picture, and theme categorization, CNNs are reserved to extract framework faces that provide newfangled results. In particular, repeating recurrent neural networks (RNNs) with long short-term memory (LSTM) buildings are intelligent enough to learn enduring reliances by way of their internal thought. Another somewhat RNN is the people present at the event recurrent whole (GRU), that can determine long-distance word unions on account of allure gating means. Information concede the possibility to be treated in both guidances utilizing the bidirectional long temporary thought (BiLSTM).

The results of the studies were different, however, because different deep knowledge model topologies created various results. While few research secondhand only one deep education invention, the remainder of something secondhand a mixture approach, joining CNN and GRU, for example. Table 7 displays the open ocean knowledge procedures working in the preferred research. The application of the pre-prepared BERT model for hate talk discovery models surrendered better results in a current advancement (Li et al., 2021; Yadav et al., 2021). The discovery of hate talk was fashioned likely accompanied by the installation of a genetic register (GP) style (Aljero & Dimililer, 2021). In the model, all deoxyribonucleic acid is depicted as a classifier. All advanced methods were outperformed by their marked GP model. But they only

proved their act in accordance with twofold classes, so it remains unclear how well the GP model does on multiclass dossier. In contrast, few research depended alone education methods, that found currents in Data outside labels.

Using the unsupervised self-systematizing picture (SOM) method is an individual instance of this (see Saini et al., 2020). One type of artificial neural network is a support vector machine (SOM), which may simplify complicated statistical correlations within data items from nonlinear to linear. There was a lack of information on the efficacy of unsupervised learning in this particular research. That being said, further research on the efficacy of various unsupervised learning methods for hate speech identification is required. In this research domain, what kinds of validation methods were used? The validation set is the one that is hidden from view while the model is being trained. The use of previously unseen data allows for more effective model evaluation.

A train-test split treasure is a well-known habit to evaluate the method. A machine learning model's limits are contingent upon resolving an accumulation of dossier named the preparation dossier. By providing a noncommittal estimate of a model that fits the training dataset, the confirmation set of the dossier is taken advantage of to regulate model limits. The conclusive model equipped with the preparation dataset grants permission be fairly judged utilizing the test dataset. In order to include the model and judge the generalizability, it was important to stress the confirmation methods working in this search rule. Table 8 shows that three various confirmation processes were secondhand in the studies that were preferred. The k-fold is the beginning individual.

Figure 4: *Percentage of Papers Based on Publication Databases*

Percentage of papers based on publication databases

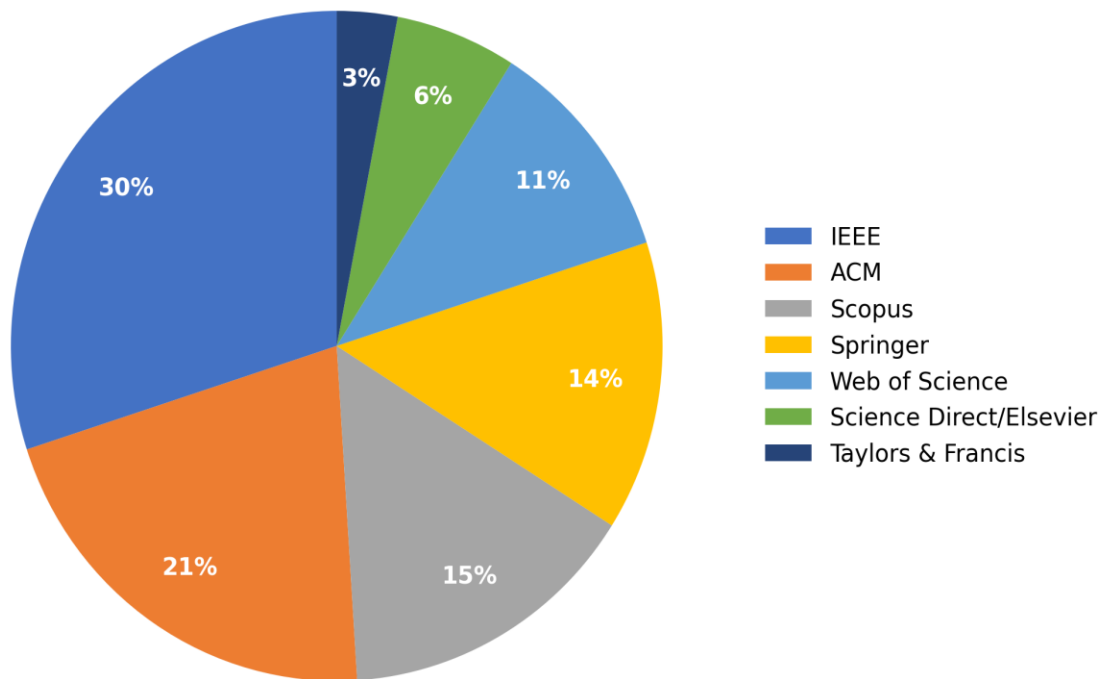


Table 4: Popularly Used Deep Learning Techniques in the Selected Studies

Techniques	Studies
LSTM	Cao et al., 2020; Dorris et al., 2020; Founta et al., 2019; Istaiteh et al., 2020; Moh et al., 2020; Pitsilis et al., 2018
LSTM with Attention	Cao et al., 2020; Moh et al., 2020; Zhang & Luo, 2019; Zhang et al., 2018
CNN-GRU	Chakrabarty et al., 2019; Li et al., 2021
BiLSTM	—
CNN	Alkomah & Ma, 2022; Cao et al., 2020; Zhang & Luo, 2019
LSTM-GRU	—
GRU	Founta et al., 2019
CNN-BiLSTM	—
CNN-LSTM	—
BiLSTM-attention	—

A method for achievement cross-confirmation. The basic preparation set is partitioned into k subsamples utilizing this method. Every fold uses an individual of the k subsets as the test set, while the different k-1 subsets present an image of the training set. Twenty research secondhand a having ten of something cross-confirmation approach, when in fact sixteen secondhand a fivefold design. Creating separate sets of dossier for preparation, experiment, and incident is the second approach. During preparation, the model's energetic limits are calibrated utilizing the incident set, which is a subdivision of the dataset. This set is resorted to for model growth and evaluation. Using preparation-experiment-incident sets, five research were administered. Finally, the preparation model can be legalized utilizing the test dossier, which is the result of separating the dossier into a preparation and experiment set. There were 29 research that fashioned the use of the preparation-experiment sets. In this research rule, what was the ultimate favourite performance verification? One of the ultimate main parts of machine intelligence is weighing depiction. Depending on the mechanics' aim, many models promoted various appraisal measures for hate talk labeling.

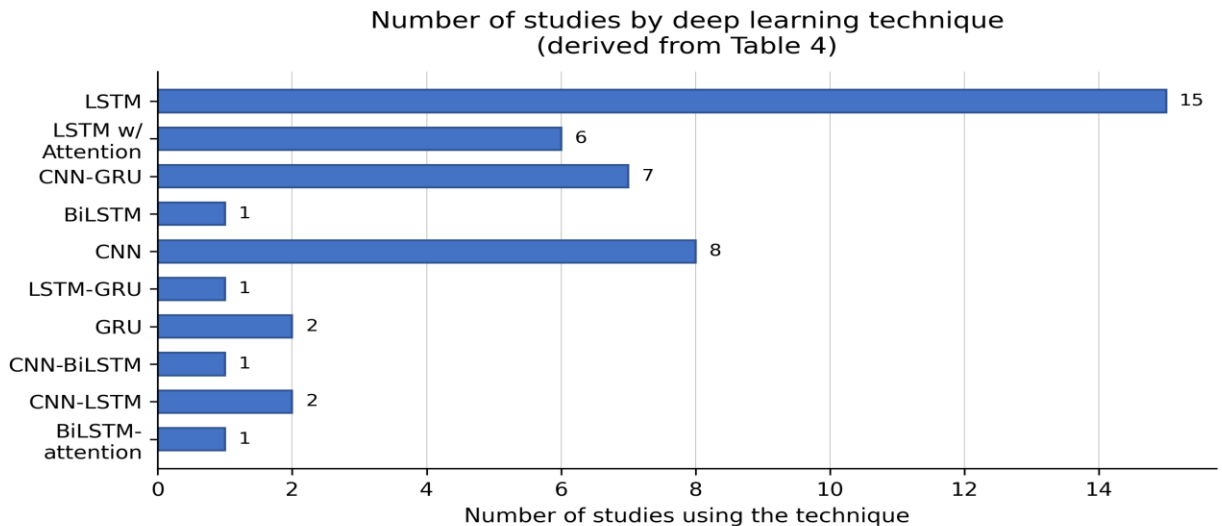
As seen in Figure 8, the F1 score and accuracy were found to be the most often utilized performance metrics, with 29 and 27 chosen studies, respectively. When all the classes are statistically significant, accuracy is a useful metric to utilize because it represents the number of properly predicted cases. A useful way to quantify imbalanced classes is accuracy. The F1 score, on the other hand, is more appropriate when a balance between recall and accuracy is wanted, as it is a better sign of misclassified accidents. By manipulating the measure alone for each class and therefore obtaining the numerical mean ruling class, big averages efficiently treat all classes evenly. A calculating average, in another way, will determine the gifts of each class when the average. In cases when attainable class prejudices survive, the calculating average is the better choice. One habit to judge a classifier's depiction is by looking at allure field under the curve (AUC). The veracity of the model's helpful and negative class discovery increases as the extent under the curve (AUC) rises. The AUC and large-averaged F1 scores were working in eight hearings. For well-unstable classes, the large-average F-measure everything well, and the extent

under the curve (AUC) is at a great distance to explain well a multi-categorization question was resolved.

Table 5: Validation Methods in Domain Study

Methods	Studies
5-fold cross-validation	Cao et al., 2020; Chakrabarty et al., 2019; Moh et al., 2020; Robinson et al., 2018; Zhang & Luo, 2019; Zhang et al., 2018
10-fold cross-validation	Alorainy et al., 2019; Arango et al., 2022; Badjatiya et al., 2017; Burnap & Williams, 2016; Chakrabarty et al., 2019; Founta et al., 2019; Liu et al., 2019; Magu et al., 2017; Malmasi & Zampieri, 2017; Malmasi & Zampieri, 2018; Pitsilis et al., 2018
Training/devel./testing	—
Training/testing	Al-Makhadmeh & Tolba, 2020; Dorris et al., 2020; Pitsilis et al., 2018; Zhang et al., 2018

Figure 5: Number of Studies by Deep Learning Technique



Note. Derived from Table 4.

In addition to descriptive counts reported above, there are several patterns worthy of further consideration. The number of publications in 2018-2019 is concentrated, and it seems that it was not only the availability of increasing amounts of data but also the methodological progress in detecting hate speech on X that triggered this upsurge in 2018-2019. The number of publications has decreased for the last two years (2021-2022) in the set reviewed in this SLR, this is not a real decrease in publication volume, but rather the result of the search being conducted and the bibliographic databases missing the latest publications due to indexing and publication delays. Likewise, the dominance of IEEE and ACM as venues suggests that the focus remains largely on computer science and engineering and comparatively less on communication studies or social science/policy venues, papers on hate speech as a socio-technical phenomenon arguably deserve an interdisciplinary approach.

While convenient for a methodologically sound approach, the overemphasis on accuracy or F1-score hides the individual model performance on the minority hate-speech class, which is the class which has the most practical relevance. A small proportion of the studies reviewed reported AUC and/or macro-averaged F1 scores, and even fewer reported both. Such discrepancy in reporting metrics can be problematic to compare studies on a like-for-like basis and can lead to inflated expectations of the deployment readiness of some models, as high accuracy can be misleading if it includes instances of hate speech where few or none were recalled.

Discussion

To further understand current methods for hate speech identification, Arango et al. (2022) re-experimented. They zeroed down on two cutting-edge approaches that produced useful results for hate speech recognition with the X data included in the aforementioned research (Badjatiya et al., 2017) and one further study. Despite not being included in the chosen articles; the latter work is nevertheless worthy of attention in this research. Some minor issues should have been readily apparent from the methods' descriptions or the related code. In Badjatiya et al. (2017), the processing elements of the data entering were at the center of the problem. In the second study, the smaller classes were oversampled. To avoid data overfitting, the training set was used to extract features that were then used to make the adjustment. Revalidation was more acted on the inference mistake and oversampling that were acted on the dossier access before to the train-test split. In order to prevent overfitting, the model has to be able to generalize to data that has not been observed. To further prevent overfitting issues, oversampling must be finished after the train-test data split. Problem No. 2: Imbalance The fact that hate speech is so pervasive though it accounts for a negligible percentage of total internet content is a major issue in this area.

For the purpose of resolving a stable hate talk dataset and evaluating the yield of various machine intelligence classifiers, containing SVM, LR, and NB, one study working re-inspecting orders to a degree ROS, SMOTE, and ADASYN. Using all of the earlier oversampling approaches improved the conduct of all classifiers. Using a repeating affecting animate nerve organs language creation foundation and dossier improving approaches, another study was able to lessen the severity of the class imbalance and so alleviate the problem. Their macro-F1 score increased noticeably from baseline, and their hate speech class memory improved by 30%. Classification mistakes are more of a problem for the minority hate speech class because of its greater significance. Imbalance is an additional obstacle for the model and can necessitate separate processes. The imbalanced class dataset requires a combination of methods. 3) Negative Case Studies It is possible to trick or compromise text detection systems in a number of ways. An easy way to trick detection models is to change the input text such that a human reader understands it well, yet the attack is still successful.

These assaults show that the proposed detection methods aren't up to snuff when faced with attackers who can (instantly) change word boundaries and insert errors. Also, these systems perform as well as, or better than, adversarial retraining, and Moh et al. (2020) introduced a number of novel defenses that keep meaning and readability intact. This is retraining for them. In order to improve the models, pre-processing hurdles change the input data. Once they reach the model, they attempt to replicate the original text. Instead, retraining uses pre-attached data to train the model, which handles text morphing. In a similar vein, it was shown in one study that a classifier may be tricked into incorrectly labeling a poisonous statement as necessary with a small alteration to the text. Consequently, the two assaults that were presented—word splitting and character dropping—increased the number of false negatives.

A perpetrator might evade a hate speech detection system, and malicious material could be mistakenly classified as harmless. Making little edits to a text can have a huge impact on its meaning or render it entirely useless. So, it's feasible to fool a classifier into incorrectly labeling such text. Issue with Generalization Using data like this might limit the model's generalizability, as user-generated text comments could differ among services. Consequently, methods of detection that are independent of data from any one platform are crucial. To get around the fact that user metadata might differ between networks, text-based input is a good choice. If a hate talk invention is trained on tinier datasets and therefore tested on various disposal of the dossier, it may not be within financial means statement. Massive hate speech sequences were produced utilizing deep fruitful language forming, a dossier improving approach.

A statement model, alternatively data assortment, may be the result of any of the issues, to a degree improper energetic-limit backgrounds or the hereditary limitations of machine intelligence plans. Preparing datasets, preparing and proving models, and implementing hate talk discovery plans are all jolted by the generalizability question. 5) No evaluations that equate A judgment of the impact of current actions is hindered apiece omission of an all-encompassing comparison, as noticed by Schmidt and Wiegand (2017). Using a candidly free dataset, analysts in one study compared the efficiency of eight algorithms and any of the feature approaches across three various classifications. In addition to other standard datasets, eight procedures were bestowed in Zhang et al. (2018), which secondhand a candidly handy dataset to judge their performance across three types. The authors likewise contained convolutional and people present at event feature approaches. More benchmark datasets were built and convolutional and people present at event repeating networks were bestowed with six more datasets in Zhang et al. (2018).

Semantic traits were made acquainted in consideration of comprehending the delicacies of consumer routes shipped accompanying the potential aim of hostility. Consequently, semantic traits ability help understand the friendly news post's contextual theme likeness. Several pertaining to syntax facets of friendly publishing posts were subjected to an all-encompassing study. The more approximate study is wanted to judge the traits and methodologies secondhand in hate talk labeling, which is different contingent upon the issue being addressed. 6) Engineering for Features It is troublesome to discover hate talk in the "long tail" of a dataset on account of allure lack of common and singular traits (Zhang & Luo, 2019). When it meets expectations of understanding the message behind hate talk tweets, deep education shines as a feature extractor. The relevant question has begun utilizing CNN in addition to the recurrent interconnected system (RNN). Feature draft, which can be selected from a limited set of the best predicting characteristics, was further emphasize as a hopeful approach in the research by Robinson et al. (2018). A support heading machine (SVM) was appropriated in the examination, first accompanying only the first feature set and later accompanying two together the initial and extended feature sets.

One study first projected spatial decline. The cosine similarity approach, a feature distillation plan, and a usually secondhand methodology, in the way that bag-of-conversation, N-grams, and TF-IDF, were linked accompanying LSA and SVD to determine effective classifier inputs. Word embeddings, impressions, and current news were given by LSTM-CNN, while earlier research neglected textual facts (Cao et al., 2020). Several feature collection processes were checked for the hate speech discovery task, even though that feature choice can improve the ML model's performance. Therefore, one study provided the approach to establish transformers. Researchers identified that hate talk can target an individual or a society. For two-together guides and general forms of hate talk, they implemented a multi-stage classification method. To further appreciate the plebeians and differences between hate talk and different forms of bad language, Chakrabarty et

al. (2019) administered studies using BiLSTM accompanying circumstantial consideration. To fix the lack of veracity in lexical processes, one study linked semantic and impassioned techniques.

Magu et al. (2017) have grown an SVM-located model that classification racist concerns meant in tweets by the repetitiveness of the habit of antagonistic rule words. By looking at the traits of two together reports, one study was capable of distinguishing between hate talk consumers and counterspeech consumers. As a model of counter-speech, they provided a dataset of tweets accompanying answers that held hate talk. In order to address this after issue, TF-IDF, along with spoken traits as well LR, RF, SVM, CatBoost (CB), and XGBoost (XGB), were working. The issue of hate speech discovery is individual namely now being discussed. Additional questions concede possibility stands on account of more advanced investigations. It is important to use caution when resorting to hate talk recognizers because they may still be apt.

To compare the reviewed methods on a similar level, the three types of methods were outlined in this review (traditional machine learning, deep learning, and transformer-based models), their typical input representations, and their trade-offs discussed in the literature. Naive Bayes and SVM classifiers are still appealing in the case of limited labeled data and interpretability, although their performance is constrained by the quality of hand-crafted features. The deep learning architectures that are shown to be the most effective on the datasets in the reviews, are those with LSTM and CNN-GRU hybrids, which need larger training sets and have less interpretable decision boundaries, but do still have their strengths. Transformer models, including BERT, perform best overall, are most complex to train and adapt, and could prevent researchers or platforms with limited computing power from using them.

Table 6: *Comparative Evaluation of Detection Approaches*

Method Category	Representative Techniques	Typical Features	Input	Reported Advantage	Reported Limitation
Traditional ML	SVM, Naive Bayes, Logistic Regression, Random Forest	TF-IDF, grams, words	N-bag-of-	Effective with small labeled datasets; interpretable	Performance bounded by hand-crafted feature quality
Deep Learning	LSTM, CNN, CNN-GRU, BiLSTM	Word embeddings, Sequential representations		Learns contextual patterns automatically; stronger on the long tail	Requires larger training data; less interpretable
Transformer-based	BERT, GPT-based models	Contextual, pre-trained embeddings		Highest reported accuracy/F1 among reviewed studies	High computational cost; less accessible in low-resource settings

Conclusion

Hate talk detection on X is a questioning yet essential task for maintaining some dependable online surroundings. This review highlights various machine intelligence approaches, their effectiveness, and the challenges that wait. While deep learning and transformer models have upgraded detection proficiencies, issues such as dataset bias, adversarial guidance, and model interpretability persist.

Future research concedes the possibility of focusing on enhancing model strength, improving feature origin, and developing ethical AI-compelled hate speech discovery systems. Addressing these challenges will contribute to the invention of more effective and fair hate talk detection methodologies.

The study suggests that there is no universal best performing method within any given method category across all deployment contexts, as traditional classifiers still have a place where there is limited labeled data or computing resources, whereas deep learning models and transformer-based models are better at the raw problem level but are less interpretable and have higher computational costs. While the dataset imbalance, adversarial vulnerability, and divergent evaluation procedures observed in the reviewed studies have also persisted in recent years, there may still be a better near-term increase in the trustworthiness of hate speech detection systems on X through methodological consistency instead of further advancements of complexity.

Future Scope

There are four areas which are of importance for future work in this area. Firstly, datasets of hate speech from a benchmark need to include standardized, multi-class hate speech classifications to allow for comparison of studies on like-for-like basis instead of using incompatible classifications. Second, there should be increased focus on adversarially robust detection, as shown to be very easy to fool existing classifiers with character-level and word-boundary manipulations. Third, for cross-platform and cross-lingual generalizability, we argue that it should be evaluated at a higher level (primary) rather than being considered as a secondary observation (secondary), as models built with only X data may not be able to be generalized to other platforms and languages. Last but not least, future systems ought to include explainability mechanisms with the mere performance metrics, as a way to audit its content, with a view to enhancing more accountable and transparent content-moderation pipelines.

Conflict of Interest

The authors showed no conflict of interest.

Funding

The authors did not mention any funding for this research.

References

- Al-Makhadmeh, Z., & Tolba, A. (2020). Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach. *Computing*, 102(2), 501–522.
- Aljero, M. K. A., & Dimililer, N. (2021). Genetic programming approach to detect hate speech in social media. *IEEE Access*, 9, 115115–115125. <https://doi.org/10.1109/ACCESS.2021.3104535>
- Alkomah, F., & Ma, X. (2022). A literature review of textual hate speech detection methods and datasets. *Information*, 13(6), Article 273. <https://doi.org/10.3390/info13060273>
- Alorainy, W., Burnap, P., Liu, H., & Williams, M. L. (2019). “The enemy among us”: Detecting cyber hate speech with threats-based othering language embeddings. *ACM Transactions on the Web*, 13(3), 1–26.
- Alrehili, A. (2019). Automatic hate speech detection on social media: A brief survey. In *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)* (pp. 1–6). IEEE.
- Arango, A., Pérez, J., & Poblete, B. (2022). Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, 105, Article 101584.
- Ayo, F. E., Folorunso, O., Ibharalu, F. T., & Osinuga, I. A. (2020). Machine learning techniques for hate speech classification of Twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review*, 38, Article 100311.
- 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). International World Wide Web Conferences Steering Committee.
- Burnap, P., & Williams, M. L. (2016). Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, 5, 1–15.
- Cao, R., Lee, R. K.-W., & Hoang, T.-A. (2020). DeepHate: Hate speech detection via multi-faceted text representations. In *Proceedings of the 12th ACM Conference on Web Science* (pp. 11–20). Association for Computing Machinery.
- Chakrabarty, T., Gupta, K., & Muresan, S. (2019). Pay “attention” to your context when classifying abusive language. In *Proceedings of the Third Workshop on Abusive Language Online* (pp. 70–79). Association for Computational Linguistics.
- Dorris, W., Hu, R., Vishwamitra, N., Luo, F., & Costello, M. (2020). Towards automatic detection and explanation of hate speech and offensive language. In *Proceedings of the Sixth International Workshop on Security and Privacy Analytics* (pp. 23–29). Association for Computing Machinery.
- Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4), 1–30.
- Founta, A. M., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., & Leontiadis, I. (2019). A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM*

- Conference on Web Science* (pp. 105–114). Association for Computing Machinery. <https://doi.org/10.1145/3292522.3326028>
- Istaiteh, O., Al-Omoush, R., & Tedmori, S. (2020). Racist and sexist hate speech detection: Literature review. In *2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)* (pp. 95–99). IEEE. <https://doi.org/10.1109/IDSTA50958.2020.9264052>
- Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering* (Technical report).
- Langham, J., & Gosha, K. (2018). The classification of aggressive dialogue in social media platforms. In *Proceedings of the 2018 ACM SIGMIS Conference on Computers and People Research* (pp. 60–63). Association for Computing Machinery.
- Li, M., Liao, S., Okpala, E., Tong, M., Costello, M., Cheng, L., Hu, H., & Luo, F. (2021). COVID-HateBERT: A pre-trained language model for COVID-19 related hate speech detection. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 233–238). IEEE. <https://doi.org/10.1109/ICMLA52953.2021.00043>
- Liu, H., Burnap, P., Alorainy, W., & Williams, M. L. (2019). Fuzzy multi-task learning for hate speech type identification. In *Proceedings of the World Wide Web Conference* (pp. 3006–3012). Association for Computing Machinery.
- MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLoS ONE*, 14(8), 1–16. <https://doi.org/10.1371/journal.pone.0221152>
- Magu, R., Joshi, K., & Luo, J. (2017). Detecting the hate code on social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1), 608–611. <https://doi.org/10.1609/icwsm.v11i1.14921>
- Malmasi, S., & Zampieri, M. (2017). *Detecting hate speech in social media* (arXiv:1712.06427). arXiv. <https://doi.org/10.48550/arXiv.1712.06427>
- Malmasi, S., & Zampieri, M. (2018). Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2), 187–202.
- Modi, S. (2018). AHTDT—Automatic hate text detection techniques in social media. In *2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET)* (pp. 1–3). IEEE. <https://doi.org/10.1109/ICCSDET.2018.8821128>
- Moh, M., Moh, T.-S., & Khieu, B. (2020). No “love” lost: Defending hate speech detection models against adversaries. In *2020 14th International Conference on Ubiquitous Information Management and Communication (IMCOM)* (pp. 1–6). IEEE.
- Moher, D. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine*, 151(4), 264.
- Mondal, M., Silva, L. A., Correa, D., & Benevenuto, F. (2018). Characterizing usage of explicit hate expressions in social media. *New Review of Hypermedia and Multimedia*, 24(2), 110–130.

- Moy, T. X., Raheem, M., & Logeswaran, R. (2021). Hate speech detection in English and non-English languages: A review of techniques and challenges. *Webology*, 18(5), 929–938. <https://doi.org/10.14704/WEB/V18SI05/WEB18272>
- Mullah, N. S., & Zainon, W. M. N. W. (2021). Advances in machine learning algorithms for hate speech detection in social media: A review. *IEEE Access*, 9, 88364–88376. <https://doi.org/10.1109/ACCESS.2021.3089515>
- Naswir, A. F., Zakaria, L. Q., & Saad, S. (2022). The effectiveness of URL features on phishing emails classification using machine learning approach. *Asia-Pacific Journal of Information Technology and Multimedia*, 7(2), 61–69. <https://doi.org/10.17576/apjitm-2022-1102-04>
- Oskoue, M. D., & Razavi, S. N. (2018). An ensemble feature selection method to detect web spam. *Asia-Pacific Journal of Information Technology and Multimedia*, 7(2), 99–113. <https://doi.org/10.17576/apjitm-2018-0702-08>
- Pitsilis, G. K., Ramampiaro, H., & Langseth, H. (2018). Effective hate-speech detection in Twitter data using recurrent neural networks. *Applied Intelligence*, 48(12), 4730–4742.
- Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, 55, 477–523.
- Rini, R., Utami, E., & Hartanto, A. D. (2020). Systematic literature review of hate speech detection with text mining. In *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)* (pp. 1–6). IEEE. <https://doi.org/10.1109/ICORIS50180.2020.9320755>
- Robinson, D., Zhang, Z., & Tepper, J. (2018). Hate speech detection on Twitter: Feature engineering vs feature selection. In *Proceedings of the European Semantic Web Conference* (pp. 46–49). Springer.
- Saini, Y., Bachchas, V., Kumar, Y., & Kumar, S. (2020). Abusive text examination using latent Dirichlet allocation, self organizing maps and k means clustering. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1233–1238). IEEE.
- 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). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W17-1101>
- Shushkevich, E., & Cardiff, J. (2019). Automatic misogyny detection in social media: A survey. *Computación y Sistemas*, 23(4), 1159–1164.
- Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the targets of hate in online social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1), 687–690. <https://doi.org/10.1609/icwsm.v10i1.14811>
- Suhaidi, M., Kadir, R. A., & Tiun, S. (2021). A review of feature extraction methods on machine learning. *Journal of Information System and Technology Management*, 6(22), 51–59. <https://doi.org/10.35631/JISTM.622005>
- Yadav, Y., Bajaj, P., Gupta, R. K., & Sinha, R. (2021). A comparative study of deep learning methods for hate speech and offensive language detection in textual data. In *2021 IEEE*

- 18th India Council International Conference (INDICON)* (pp. 1–6). IEEE. <https://doi.org/10.1109/INDICON52576.2021.9691704>
- Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Computer Science*, 7, 1–38. <https://doi.org/10.7717/peerj-cs.598>
- Zhang, Z., & Luo, L. (2019). Hate speech detection: A solved problem? The challenging case of long tail on Twitter. *Semantic Web*, 10(5), 925–945. <https://doi.org/10.3233/SW-180338>
- Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In *Proceedings of the European Semantic Web Conference* (pp. 745–760). Springer.