Journal of Population Therapeutics
& Clinical Pharmacology

RESEARCH ARTICLE
DOI: 10.53555/jptcp.v31i4.5540

EXPLORING TEXTUAL HATE SPEECH IDENTIFICATION APPROACHES AND DATASETS: A SYSTEMATIC LITERATURE REVIEW AND META-ANALYSIS

Husnain Saleem[1*], Muhammad Javed[2], Muhammad Zubair Asghar[3], Muhammad Ahmad Jan[4], Maria Zuraiz[5], Aftab Ali[6], Asad Ullah[7]

[1*]Ph.D. Scholar, Gomal University, Dera Ismail Khan, Pakistan, Email: husnain@gu.edu.pk
[2]Assistant Professor, Gomal University, Dera Ismail Khan, Pakistan,
Email: javed_gomal@gu.edu.pk
[3]Associate Professor, Gomal University, Dera Ismail Khan, Pakistan,
Email: mzubairgu@gmail.com
[4] Assistant Professor, Gomal University, Dera Ismail Khan, Pakistan,
Email: muhammad.ahmadjan@gu.edu.pk
[5] Ph.D. Scholar, Gomal University, Dera Ismail Khan, Pakistan,
Email: maria.zuraiz@aack.au.edu.pk
[6] Ph.D. Scholar, Gomal University, Dera Ismail Khan, Pakistan, Email: aftab.ali01@nbp.com.pk
[7] Ph.D. Scholar, Gomal University, Dera Ismail Khan, Pakistan, Email: asadullahpushia@gmail.com

***Corresponding Author:** Husnain Saleem
*Ph.D. Scholar, Gomal University, Dera Ismail Khan, Pakistan, Email: husnain@gu.edu.pk

Abstract

There have been growing concerns about the influence of hate speech on social discourse and its ability to instigate violence and prejudice as it has spread widely across internet platforms. Researchers and service providers must now prioritize identifying and regulating hate speech. In this survey, we look at studies published between 2018 and 2023 that explore various aspects of hate speech identification. This review begins by pointing out the alarming growth of hate speech on the internet and its adverse effects, underscoring the importance of developing reliable identification mechanisms. Based on the papers' principal focuses, we classify them into one of five broad themes: dataset construction, algorithm development, bias analysis, multilingual and multimodal techniques, and ethical considerations. This systematic literature review and meta-analysis highlights the need for standardized evaluation metrics, more extensive datasets, and robust algorithms to deal with the ever-evolving nature of hate speech while pointing out the shortcomings of currently available hate speech identification methods and datasets. To effectively counteract online hate speech, researchers, legislators, and technology businesses will find this comprehensive assessment an invaluable resource, an in-depth overview of the hate speech identification landscape. Future research initiatives on this crucial topic can build upon the insights and problems given here.

Keywords: Bias Analysis, Deep Learning, Hate Speech Identification, Hate Speech Datasets, Machine Learning, Multilingual, Multimodal.

Introduction

The internet and social media have enabled people to communicate across great distances and in real time. This new era of digital connectivity has opened up previously unimaginable communication and information-sharing channels. However, this revolutionary shift in sharing information has left us vulnerable to a growing danger: hate speech. Derogatory, provocative, or discriminatory statements directed at persons or groups based on race, ethnicity, religion, sexual orientation, gender, or other legally protected characteristics constitute hate speech. It is an ugly phenomenon that threatens the foundations of democracy everywhere: fairness, tolerance, and respect for all. Hate speech has real-world repercussions, not just online. They materialize in ways that incite violence, spread harmful stereotypes, and drive wedges amongst communities. As well as hurting people, hate speech may damage communities by weakening trust and unity. Scholars, academics, and politicians have focused on creating tools and approaches for automatic hate speech identification due to the seriousness of the problem. Natural language processing and machine learning drive these technologies, which aim to identify and limit the spread of hate speech in cyberspace. The area of hate speech is dynamic and complicated, posing substantial obstacles to its identification and reduction as the rapid expansion of technology and the internet continues to transform the dynamics of online communication. Therefore, it is crucial for continuous efforts to tackle this critical issue to have a complete comprehension of the methodology, obstacles, and prospects in automatic hate speech identification.

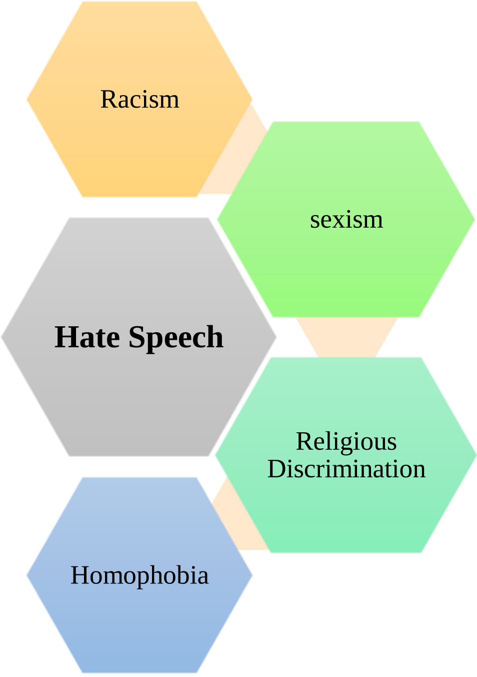

Figure 1 Types of Hate speech

While many studies have delved into this complex problem, an exhaustive and structured evaluation of the diverse methodologies and their associated implications must be more conspicuously present in the existing literature reviews (MacAvaney et al., 2019). This systematic literature review aims to rectify this glaring gap by conducting a comprehensive and well-structured investigation into hate speech identification research. Beyond the mere enumeration of various methods, our study assesses their efficacy, applicability, and limitations (Sap et al., 2019). Its overarching objective is to provide a comprehensive panorama of the myriad approaches to automatic hate speech identification. The foremost contribution of this review lies in its meticulous categorization of hate speech identification methods into several distinct categories, as facilitated by a custom-built taxonomy. From lexicon-based strategies to machine learning-based approaches, and everything in between (Al-Hassan & Al-Dossari, 2019; Gomez et al., 2020), there is a wide range of methods that can be classified under these umbrella terms. In our analysis, we focus on the specific benefits, drawbacks, and applications of each method. The selection of datasets is intrinsically linked to any inquiry in this field. Using the Islamophobic hate speech dataset (Sap et al., 2019) and the dataset (Sap et al., 2019) examples, we examine the fundamental features of datasets such as their size, thematic categorizations, and potential biases. This all-encompassing method guarantees an in-depth familiarity with the fundamentals of hate speech identification strategies. This systematic review (Davidson et al., 2019) examines not just the technological aspects of automatic hate speech identification research, but also its ethical and societal implications. It addresses issues of bias, such as sexism and racism that go beyond the technicalities of algorithms. This highlights the need for designing algorithms that are sensitive to cultural differences and subtleties in language (Sap et al., 2019).

Research Questions

1. What methods have been used, what obstacles have been encountered, and where do you see automatic hate speech identification going in the future?
2. Where do you see automatic hate speech identification datasets going from here?

The purpose of this review is to provide a thorough analysis of the techniques currently in use to identify hate speech. We hope to shed light on the current state, challenges, and possible future paths of automatic hate speech recognition to better serve the community.

Methodology

Our findings have been proven to be understandable, accurate, and applicable because we used an organized and comprehensive scientific approach in conducting this systematic literature review. Our initial search of reputable literature databases produced a respectable 500 scholarly papers. From an initial pool of thousands of papers, we were able to narrow it down to a manageable sample of 65 by using rigorous inclusion and exclusion criteria. We carefully drafted well-defined study questions and established specific inclusion and exclusion criteria to maintain methodological rigour. Through this iterative procedure, we were able to fine-tune the accuracy and utility of our pick. We conducted a comprehensive search of academic databases for our systematic review. This included the ACM Digital Library, IEEE Xplore, Google Scholar, and PubMed, among others. We searched for academic publications focused on hate speech identification in text, with a publication timeframe ranging from 2018 to 2023. We employed an array of pertinent search terms to ensure comprehensive data collection. These search terms encompassed concepts like "hate speech identification," "offensive language identification," "cyberbullying identification," "Sentiment Analysis (SA)," "text," "Deep Learning (DL)," "Machine Learning (ML),""NLP." And "natural language processing," Our methodology is intentionally centered on collecting data about automatic hate speech identification. We aimed to encompass studies in diverse languages and mediums, thereby highlighting the global nature of this issue. This language and medium neutrality is pivotal for comprehensively capturing the various strategies to combat hate speech across diverse contexts. Our approach prioritizes precision and relevance. Consequently, only studies aligning with the stringent selection criteria are considered for inclusion in our study. This refinement accelerates the research process and augments the practicality and significance of our findings.

Inclusion Criteria

- Studies focusing on automatic hate speech identification.
- Academic publications related to hate speech identification in text.
- Research papers published between 2018 and 2023.
- Studies encompass various languages and mediums to reflect the international scope of the issue.

Exclusion Criteria

- Studies not directly related to hate speech identification.
- Non-academic publications.

- Research papers published outside the defined timeframe (before 2018 or after 2023).
- Studies primarily focusing on languages or mediums do not contribute to exploring hate speech identification.

Our methodology incorporates a thematic analysis strategy akin to established best practices (Tranfield et al., 2003). This systematic approach allows us to identify patterns in hate speech identification methodologies systematically. We meticulously read each selected document, applied relevant codes to pertinent sections, and subsequently categorized these codes into thematic clusters for systematic insight synthesis. An integral outcome of our methodology is the identification of key challenges and obstacles encountered in the field of hate speech identification. This facet holds significance as it sheds light on researchers' and practitioners' difficulties, providing valuable insights for the broader discourse. Our methodology combines a well-defined research question framework, stringent inclusion and exclusion criteria, thematic analysis, and the identification of challenges to present a comprehensive overview of hate speech identification methods and their implications.

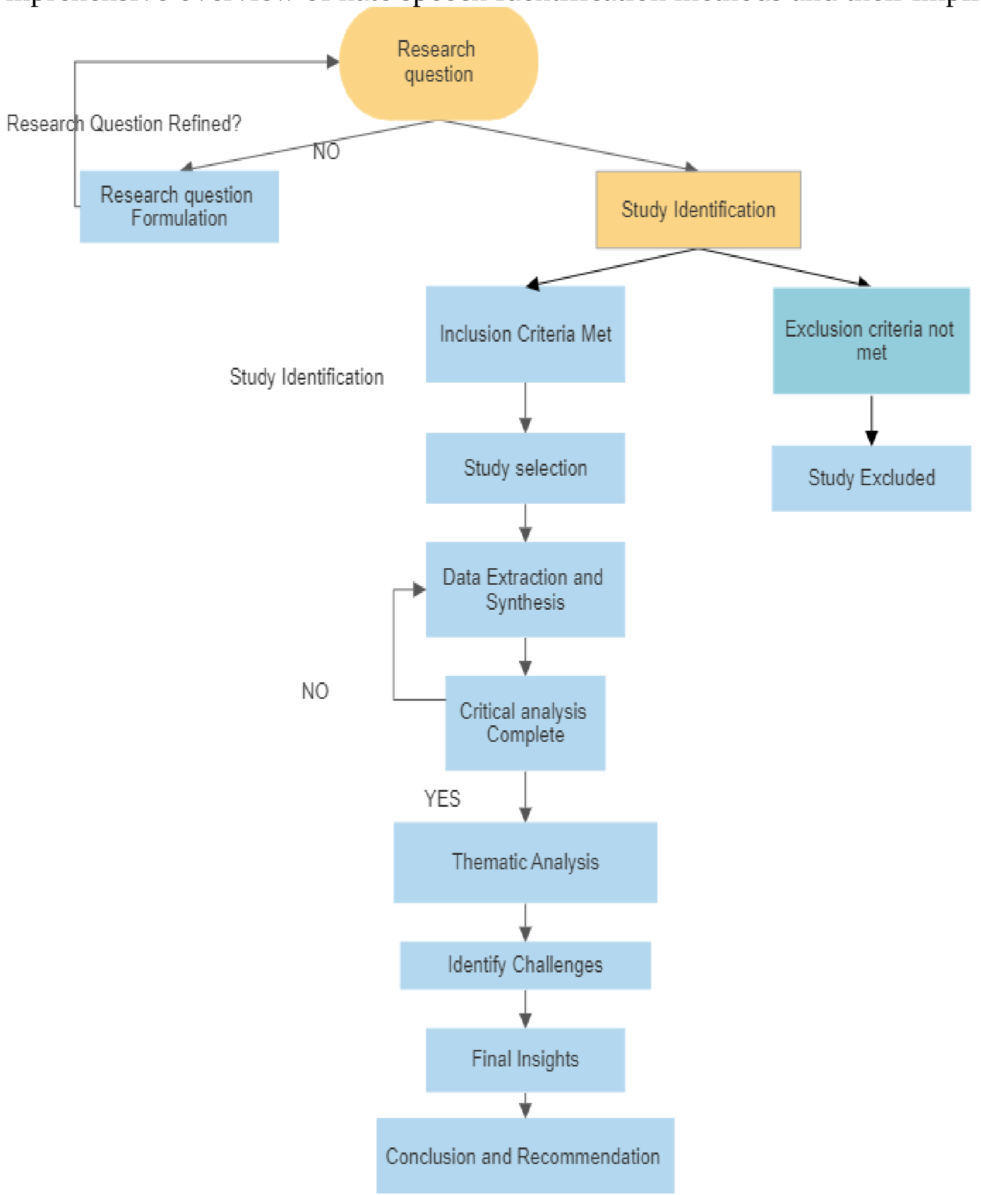
Research question
Research Question Refined?
NO
Research question Formulation
Study Identification
Inclusion Criteria Met
Exclusion criteria not met
Study Identification
Study selection
Study Excluded
Data Extraction and Synthesis
NO
Critical analysis Complete
YES
Thematic Analysis
Identify Challenges
Final Insights
Conclusion and Recommendation

Figure 2 Proposed Structure of Hate Speech Identification

Results and Analysis

Our comprehensive literature study of hate speech identification, based on a meticulous analysis of 65 selected research papers, there are 20 papers on the lexicon-based approach, 15 papers on the TF-

IDF machine learning approach, 18 papers on the deep learning approach, and 12 papers on hybrid approach has yielded significant insights into the prevailing trends and methodologies in this domain. The identified papers offer a rich landscape of approaches and techniques for automatic hate speech identification. The most relevant papers are shown in Appendix A.

A notable trend across the reviewed papers is the prominent use of lexicon-based approaches for hate speech identification. Several studies, including those by (Ousidhoum et al., 2019) and (Mozafari et al., 2020), focus on constructing hate speech vocabularies or lexicons. These lexicons are foundational resources to identify and flag abusive language within texts. The prevalence of such lexicon-based methods underscores their practical utility in addressing hate speech identification challenges. Our analysis highlights the integration of machine learning strategies, particularly TF-IDF, in various studies. The works by (Velankar et al., 2022) and (Roy et al., 2020) explore the efficacy of TF-IDF as a numerical representation of word importance in distinguishing hate speech from non-hateful content. Combining TF-IDF with diverse machine learning techniques contributes to the methodological diversity in hate speech identification research. A compelling finding is the pivotal role of deep learning models in advancing the state-of-the-art in hate speech identification. The works of (Malik et al., 2022) exemplify this trend. Deep learning models, such as deep neural networks and Convolutional Neural Networks (CNNs), are lauded for their ability to capture nuanced linguistic nuances. This ability translates to exceptional accuracy in recognizing hate speech patterns within textual data. Our analysis reveals the increasing adoption of hybrid methodologies that amalgamate multiple identification strategies. Studies by (Gomez et al., 2020) and (Fortuna et al., 2019) stand out in this context. These researchers aim to enhance hate speech identification's effectiveness across diverse communication channels by fusing textual analysis with image processing or leveraging complementary techniques.

Table 1 No of paper

YEAR	No of papers
2018	4
2019	11
2020	16
2021	7
2022	17
2023	5

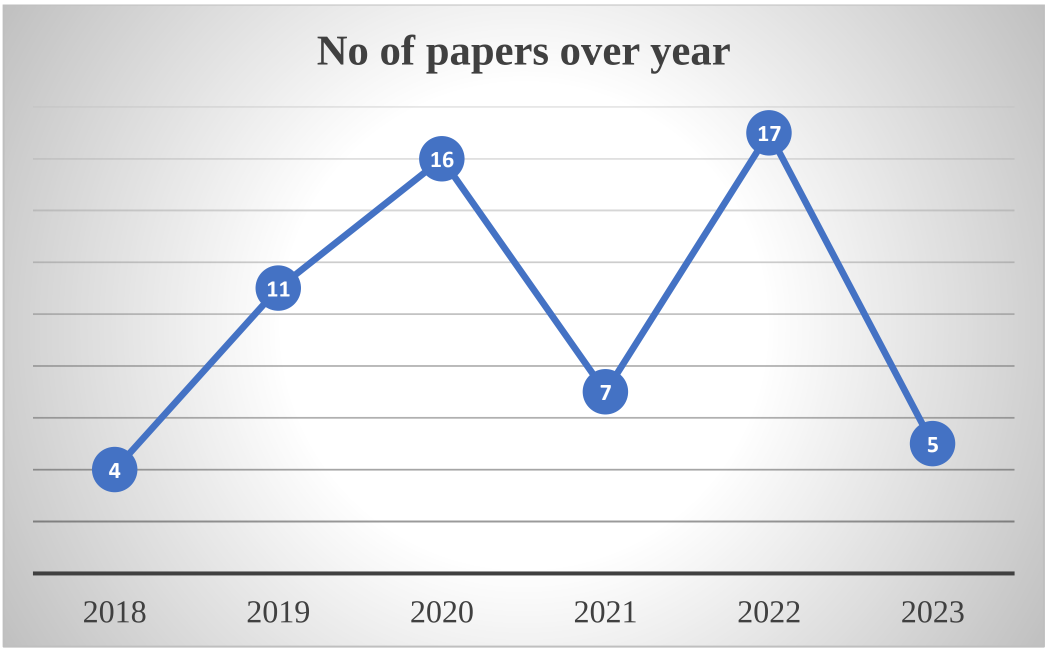

Figure 3 No of Papers over the year

Among the 65 selected papers, the distribution of methodologies is as follows.

Table 2 Machine learning methods used

Approach	Number of Papers	Percentage
Lexicon-Based Approaches	20	30.77%
TF-IDF and ML	15	23.08%
Deep Learning Models	18	27.69%
Hybrid Approaches	12	18.46%
Total	65	100%

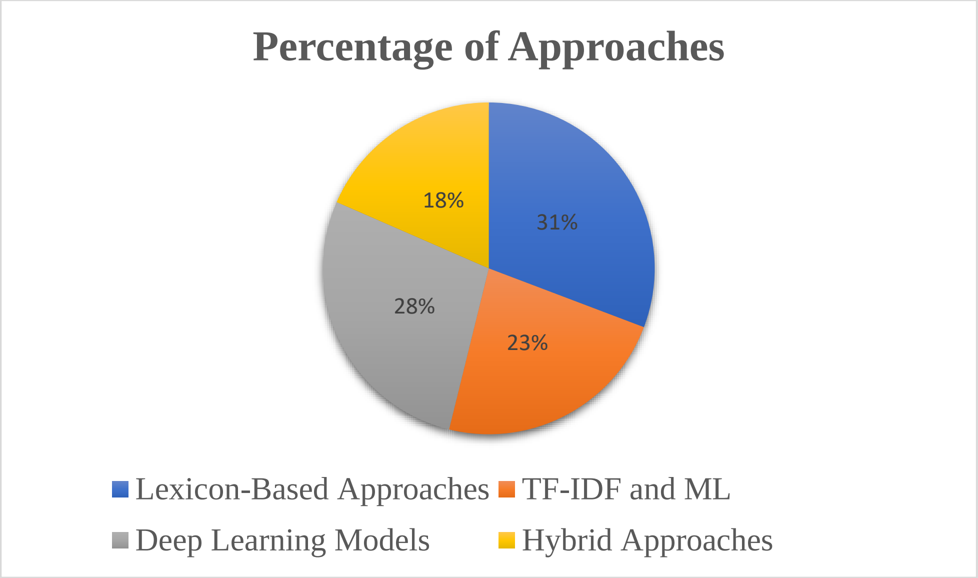

Figure 4 Percentage of Approaches

To identify hate speech, lexical approaches use predefined word and phrase lists or dictionaries. These strategies rely on overt linguistic indicators and patterns linked to xenophobic discourse. About 31% of the publications used lexicon-based approaches to hate speech identification, demonstrating the continued importance of rule-based strategies. One standard method combines TF-IDF and machine learning algorithms. Words are given relevance scores based on how often and how significantly they appear in the text. 23% of the publications used TF-IDF and machine learning algorithms. Deep learning techniques, often involving neural networks with multiple layers, have gained traction for their ability to capture complex patterns in text data. Approximately 28% of the papers in the study embraced deep learning methods, indicating their effectiveness in identifying nuanced hate speech patterns. Hybrid approaches combine different methodologies to improve the accuracy of hate speech identification. Methods such as the TF-IDF and machine learning may be included in these strategies. About 18 publications used hybrid methods, demonstrating the power of integrating different approaches.

This systematic review categorises the diverse methodologies, extracts thematic insights, and identifies challenges encountered in hate speech identification. The methodological diversity observed in the literature underscores the complexity of the hate speech identification problem and the ongoing efforts to address it through multifaceted strategies. This review's comprehensive coverage and analysis contribute to a holistic understanding of the field's methodological landscape and implications.

Automatic Hate Speech Identification Models

The reviewed literature shows various approaches used to identify hate speech. We group these approaches into thematic clusters, each representing a unique set of methods and procedures, to help readers better understand this field.

Lexicon-Based Methods for Hate Speech Identification

Using dictionaries with lists of words, phrases, and patterns known to be indicative of hate speech content is called a "lexicon-based method" for identifying such content. These lists are hand-crafted to include words and phrases typically found in hate speech, enabling algorithms to identify and categorize hate speech based on exact matches. This strategy uses linguistic analysis to look for certain phrases or terms within a text to establish whether or not it contains hate speech. There are several benefits to using a lexicon-based approach. They are simple to build since they employ a dictionary-like strategy in which the presence of particular keywords causes a determination of classification.

To better understand the difficulties and methods used in evaluating hate speech identification systems, (Bosco et al., 2018) describe the 2018 hate speech identification task. (Tehet al., 2018) create a list of objectionable keywords by identifying and classifying profane words in hate speech, providing a fundamental component for lexicon-based techniques. (Dewaniet al., 2021) Demonstrates using lexicon-based methodologies outside of the English language by focusing on developing computational linguistic resources for the automated identification of cyberbullying risks in Roman Urdu. To evaluate the efficacy of lexicon-based categorization in identifying poisonous, hateful, offensive, and abusive content, (Fortunaet al., 2020) conduct an empirical analysis of hate speech datasets, yielding useful insights into the drawbacks of lexicon-based techniques. To prove the value of lexicon-based strategies in sentiment-related tasks, (Mehmood et al., 2020) suggest an unsupervised lexical normalization approach to Roman Hindi and Urdu sentiment analysis. The importance of lexicon-based techniques in identifying and treating hate speech across languages is highlighted in a survey of multilingual corpora for identifying hate speech in social media platforms (Al-Hassan & Al-Dossari, 2019).An organized review of methods for combating hate speech (Arango et al., 2019) details the successes and failures of various lexicon-based strategies and their prospective effects. Introducing a hierarchically labelled hate speech dataset, Portuguese-specific lexicon-based models for hate speech identification may now be developed and evaluated (Fortuna et al., 2019). (Pitenis et al., 2020) This is an excellent example of using lexicon-based approaches in language-specific hate speech identification tasks because it focuses on foul language recognition in Greek.

A new study (Röttger et al., 2022) presents functional testing for multilingual hate speech identification models to emphasize the significance of evaluating the efficacy of lexicon-based techniques across many languages. Exploring the difficulties of obtaining transferable hate speech identification (Ruwandika & Weerasinghe, 2018) offers insights into the obstacles and potential solutions, focusing on the importance of lexicon-based approaches in comprehending bias and limitations.

To demonstrate the usefulness of lexicon-based models in tackling a variety of speech-related problems, (Chakravarthi Muralidaran, 2021) presents results from a collaborative task on hope speech recognition for equality, diversity, and inclusion. The study of hate speech identification by NLP (Parihar et al., 2021) explores the uses and difficulties of lexicon-based techniques in combating hate speech. With an emphasis on the importance of lexicon-based approaches, (Lingiardi et al., 2020) examine the state of the art in multimodal and multilingual automatic hate speech identification.

The work of (Mandl et al., 2020) presents an overview of the HASCO track, elucidating the significance of lexicon-based strategies in identifying hate speech and offensive content across languages. To demonstrate the value of lexicon-based approaches in evaluating hate speech across languages, (Corazza et al., 2020) perform a multilingual evaluation of online hate speech identification. A recent study (Davidson et al., 2019) highlights the importance of lexicon-based techniques in tackling bias and fairness by examining racial prejudice in hate speech and abusive language identification datasets. The function of social media platforms in disseminating hate speech is explored in (Pereira-Kohatsu et al., 2019), which focuses on Twitter and investigates the identification and tracking of hate speech using lexicon-based methods. Offence identification in Greek is discussed (Pitenis et al., 2020), demonstrating the transferability of lexicon-based methods across languages and the significance of cultural context. Functional evaluations for multilingual hate speech identification models are presented in (Röttger et al., 2022). These tests show how important lexicon-based methods are for measuring how well hate speech recognition systems perform across languages.

TF-IDF & Machine Learning Methods for Hate Speech Identification

In natural language processing, TF-IDF is a popular method for quantifying the relative importance of individual terms in a document and a larger corpus. When determining a term's significance, it considers how often it appears within the document and throughout the entire corpus. Words with high TF-IDF values are helpful in classifying texts into their respective categories. In contrast, machine learning techniques cover a broad spectrum of algorithms that help computers infer meaning from data and perform tasks like prediction and categorization. Machine learning models can be educated with examples of hate speech and then used to determine whether an unlabelled text sample should be classified as hate speech.

A work by (Abro et al., 2020) compares different automatic algorithms for identifying hate speech using machine learning techniques, providing important insight into the efficacy of these systems. Particularly pertinent for monitoring social media platforms is the research conducted by (Ruwandika Weerasinghe, 2018) on identifying hate speech in social media through machine learning algorithms. This study (Sandaruwan et al., 2019) uses text mining and machine learning to identify hate speech in Sinhala social media, proving the utility of ML methods outside of the English language.

Automatic offensive language identification for Urdu and Roman Urdu is discussed by (Akhter et al., 2020), demonstrating the use of machine learning to identify objectionable text in these languages. The use of machine learning to identify vulnerable communities and shield them from online harm is investigated in "Hate Speech Identification on Social Media" (Mossie &Wang, 2020). (Ibrohim & Budi, 2019) Highlighting machine learning techniques in a multilingual setting by addressing multi-label hate speech and abusive language identification on Indonesian Twitter. Using machine learning methods, the authors (Mathew et al., 2019) look into the dissemination of hate speech on social media platforms and its effects.

(Pereira-Kohatsu et al., 2019) Highlights the relevance of classical machine learning methods in comprehending online hate speech dynamics by concentrating on identifying and monitoring hate speech on Twitter using machine learning. An analysis of the potential for racial prejudice in hate speech identification (Zhang &Luo, 2019) demonstrates the need to account for bias in machine learning models developed for this purpose.

The difficulties of hate speech identification tasks and the importance of machine learning methods are explored in depth in a new study (MacAvaney et al., 2019). This work supports the conclusion drawn from (Davidson et al., 2019), Evaluating racial bias in hate speech and abusive language identification datasets, that fairness and prejudice should be considered when developing machine learning models for hate speech identification. Using Greek as an example, (Pitenis et al., 2020) demonstrate how machine-learning approaches can identify offensive content in a given language. Problems with hate speech recognition systems are discussed, and machine learning-based remedies are proposed (Pawar et al., 2022) (Florio et al., 2020). This highlights the need for reliable machine-learning models. (Sap et al., 2019) investigate the potential for racial bias in hate speech identification, highlighting the difficulties in developing fair and unbiased machine learning models. In the view of (Aljarah et al., 2021), which emphasizes the interpretability and explainability of machine learning techniques for hate speech identification, we provide a post-hoc explanation of the environment in which hate speech classifiers operate.

Deep Learning Methods for Hate Speech Identification

A subset of machine learning techniques, deep learning employs artificial neural networks to discover and model data patterns automatically. These methods are becoming more and more popular since they can find complex and subtle patterns in large datasets. Input data is processed and transformed using neural networks, a deep learning model consisting of numerous layers of interconnected nodes or neurons. Neural networks develop meaningful representations from input data by iteratively adjusting internal parameters to minimize the difference between expected and actual outputs through forward and backward propagation. The literature review reveals that deep learning techniques, particularly neural networks, are widely used for hate speech identification. These approaches can identify intricate linguistic patterns of hate speech because of their capacity to identify delicate and nonlinear linkages within textual content.

(Gröndahl et al., 2018) Discusses how to avoid being identified by hate speech identifiers by utilizing adversarial tactics, drawing attention to the difficulties caused by deep learning models and adversarial instances. To demonstrate the use of deep learning models for this purpose, (Haq et al., 2020) introduce USAD, a smart system for slang and abusive text recognition in PERSO-Arabic-scripted Urdu. This study (Akram et al., 2023) highlights the efficacy of neural networks in identifying objectionable content by focusing on the automatic identification of offensive language for Urdu and Roman Urdu with deep learning models. (Saeed et al., 2021) deals with classifying harmful comments in Roman Urdu using deep learning models, proving the usefulness of neural networks for this task. The benefits of transfer learning approaches for enhancing hate speech recognition are explored (Ali et al., 2022) through the lens of Twitter hate speech identification using deep learning models. Focusing on a case study in hate speech identification, (Zimmerman et al., 2018) examine the need for sufficient explanatory text classifiers and place special emphasis on the interpretability of deep learning models. Using deep learning models to accommodate many modalities of hate speech (Chakravarthi & Muralidaran, 2022) tackles the problem of multimodal hate speech identification from Bengali memes and writings.

Insights into the efficacy of various neural network designs are provided through a comparison investigation of deep learning methods for hate speech identification (Malik et al., 2022). A characteristic extraction-based approach to hate speech identification is proposed by (Mohtaj et al., 2022), drawing attention to the importance of deep learning in the representation of features for hate speech identification. The usefulness of CNNs in identifying hate speech material is demonstrated in a new framework for hate speech identification (Roy et al., 2020). Using deep learning models, (Sharma et al., 2022) describe a method for identifying hate speech in Hindi-English code-switched language. The scalability of deep learning models for dealing with huge datasets is discussed (Toraman et al., 2022). The authors concentrate on using cross-domain transfer to identify hate speech on a big scale. Automatic hate speech identification systems are investigated (William et al., 2022), focusing on the importance of deep learning in obtaining high-performance hate speech identification. To improve identification accuracy, (Kapil Ekbal, 2020) offers a deep neural network-based multitask learning technique for hate speech identification.

Research on the efficacy of deep learning in languages other than English is presented in a paper by (Akram et al., 2023). The paper focuses on automatically identifying offensive language in Urdu using deep learning models. Automatic hate speech identification is discussed in (Zhou et al., 2020), focusing on using deep learning models to deal with several modalities and languages. (Sutejo & Lestari, 2018) looks into online hate speech, illuminating the difficulties and potential solutions associated with identifying hate speech utilizing deep learning models. One study that compares the effectiveness of monolingual and multilingual BERT for hate speech identification and text categorization is forthcoming (Velankar et al., 2022).

Hybrid Methods for Hate Speech Identification

When discussing strategies for identifying hate speech, "hybrid" refers to using multiple methods to strengthen the results' reliability and precision. These methods combine the benefits of several approaches to work around the weaknesses of any of them, such as lexicon-based methods, machine learning, or deep learning. Hybrid approaches strive to improve the overall performance of hate speech identification models by integrating different methodologies.

Research on automatic hate speech identification on social media is presented in (Alrehili, 2019). The study focuses on a hybrid approach to hate speech identification that utilizes machine learning and lexicon-based methodologies. Identifying hate speech in Asian languages such as Malayalam, Tamil, and Hindi is discussed (Dhanya &Balakrishnan, 2021). They provide an overview of methods, including hybrid ones, for doing so. With an emphasis on a hybrid method integrating lexical resources and machine learning techniques, (Wang et al., 2022) examines the problem of political hate speech identification and lexicon construction in Taiwan. Convolutional and Bi-directional Gated Recurrent Unit (Bi-GRU) networks are combined with a Capsule network in the new hate speech identification system HCovBi-caps (Khan et al., 2022). Automatic slur identification in Urdu and Roman Urdu using a hybrid approach that combines rule-based and machine-learning techniques to increase accuracy is reported (Akhter et al., 2020). Hope speech identification for equality, diversity, and inclusion using a hybrid approach is investigated by (Lingiardi et al., 2020), highlighting the significance of identifying positive material in the struggle against hate speech.The hybrid approach's efficacy in feature representation for hate speech identification is highlighted by proposing a feature extraction-based model for hate speech identification (Mohtaj et al., 2022). (Mehta et al., 2022) presents a hybrid strategy that combines machine learning and XAI techniques for interpretability and covers social media hate speech identification using XAI.To better demonstrate the advantages of multitasking in enhancing identification accuracy, (Kapil et al., 2020) propose a deep neural network-based multitask learning solution to hate speech identification. A hybrid approach to identifying abusive content is demonstrated by USAD, an intelligent system for slang and abusive text recognition in PERSO-Arabic-scripted Urdu (Haq et al., 2020). Natural language processing is discussed by (Parihar et al., 2021) and highlights the importance of a hybrid approach to identify hate speech accurately. The work by (Sharma et al., 2022) presents a hybrid method for identifying hate speech in the form of Hindi-English code-switching.

Datasets for Hate Speech Identification

The accessibility and suitability of datasets are crucial to investigating hate speech identification systems, as shown in Appendix B. These datasets are crucial for training and assessing hate speech identification models, as they are often extensively curated and annotated. We explore some of the most important datasets that have helped push this study area forward.

The UHSD dataset (Akhter et al., 2020) used a study that included hate speech and foul language written in standard and Roman Urdu. As part of their research, they are looking for ways to identify and categorize hate speech in various languages. Cyberbullying in Roman Urdu was studied by (Dewani et al., 2021). They used the Roman Urdu Cyberbullying Dataset to do so. Cyberbullying information in this language may be one focus of their investigation. In order to identify religious, sectarian, and racial bigotry in Urdu, the ISE-Hate dataset was used (Akram et al., 2023). Their studies may focus on identifying and countering religious and ethnic bigotry. Toxic remark classification in Roman Urdu was investigated by (Saeed et al., 2021). They used the Roman Urdu Toxic Remark Classification dataset. Toxic comments in this language setting will likely be the focus of their studies. Offensive Roman Urdu language in hate speech was studied using the P-Urdu-Offensive dataset (Parihar et al., 2021). Finding and analysing hate speech in this language may be part of their research. Offensive language in Urdu hate speech was studied by (Hussain et al., 2022) using the Offensive Language in Urdu Dataset. Finding and evaluating instances of hate speech in written Urdu may be part of their investigation. To investigate emotion identification and sentiment analysis in Urdu text, (Ullah et al., 2022) used the Urdu Emotion and Sentiment Analysis Dataset. Emotion and sentiment analysis are expected to be central to their studies. Research on sentiment analysis in Roman Hindi and Roman Urdu was conducted using the Roman Hindi and Urdu Sentiment Analysis Dataset (Mehmood et al., 2020). They may need to perform sentiment analysis tasks in various languages as part of their research.

The Saudi Twitter Hate Speech Dataset investigated hate speech and inflammatory language in Arabic tweets (Al-Hassan & Al-Dossari, 2019). As part of their study, they may look into the dynamics of hate speech on Twitter in the Saudi context. To investigate hate speech in English on Twitter, (Ali et al., 2022) used the Hate Speech on Twitter Dataset. They are probably analyzing the behavior and content of hate speech on Twitter as part of their research. Using the J-Hate dataset, (Aluru et al., 2020) dug into the prevalence of offensive language and hate speech in Japanese writing. One possible outcome of their study is a strategy for identifying and preventing hate speech in Japanese. The R-HSAB dataset, including profanity, abusive language, and hate speech written in Romanian (Xia et al., 2020), was used. Their studies may focus on methods for identifying and categorizing hate speech in Romanian. The Hate Speech in Persian (HSP) dataset was used to research hate speech and offensive language in Persian text (Corazza et al., 2020). One possible focus of their investigation is the analysis of Persian language hate speech. The HASOC dataset was utilized by (Davidson et al., 2019) and includes examples of hate speech and offensive language written in English, German, and Hindi. As part of their research, they are studying ways to identify and categorize hate speech in many languages. The HASOC 2019 dataset was used, which includes examples of abusive language and hate speech in English, German, and Hindi (Fortuna et al., 2019).

Identifying and analyzing instances of hate speech in online content may be part of their investigation. The HASOC 2020 dataset was used in the research (Fortuna et al., 2020). This dataset includes examples of hate speech and offensive language written in English, German, and Hindi. Their study likely focuses on methods of identifying and categorizing online expressions of hatred. (Ibrahim & Bud, 2019) Researched the identification of hate speech on Indonesian Twitter using the INACL-IMW 2019 dataset. In the context of Indonesian social media, their investigation may focus on identifying and analyzing hate speech. The A-HSAB dataset was used in the research by (Xia et al., 2020), and it includes examples of profanity, abuse, and hate speech written in Arabic. Potentially essential to their investigation is the examination of Arabic online hate speech.

The HSD 3.0 dataset was used, and it includes hate speech and abusive language in Hindi (Kapil & Ekbal, 2020). They are studying how to identify and analyze hate speech in the Hindi language for their research. The Bengali Hate Speech Dataset (BHSD) examined hate speech and objectionable language in Bengali (Karim et al., 2022). One possible focus of their investigation is the analysis of hate speech in Bengali. Using the HCovBi-caps dataset, (Khan et al., 2022) looked at hate speech classification in English text, with a particular emphasis on health-related deception during times of crisis. The MHSD dataset was used, which includes abusive language and hate speech in English, German, Hindi, and Italian (MacAvaney et al., 2019). One possible focus of their investigation is examining hate speech in many languages.Hate speech and offensive language in English, German, Hindi, and Konkani are the focus of the MultiHASOC dataset (Mathew et al., 2019). The analysis of hate speech in many languages may be a part of their research.

Discussion

Our in-depth analysis of hate speech identification approaches and data sets paves the way for various vital conversations illuminating this dynamic field's current landscape and promising future. The primary focus of the publications in this subfield is the creation of novel methods and machine-learning models for identifying instances of hate speech within text data. These works help drive the state of the art in hate speech identification forward by suggesting novel algorithms and methods. The focus of the papers in this section is on collections of offensive language online. Datasets are necessary for training and assessing hate speech identification algorithms must be created, curated, and analyzed as part of these processes. These data sets are essential for evaluating and enhancing the performance of hate speech identification algorithms in benchmarking situations. The worldwide nature of online communication has increased the significance of studies to identify hate speech in multilingual or cross-lingual environments. To prevent hate speech more broadly, these studies investigate the difficulties and potential solutions associated with identifying it in several languages. It is critical to have reliable evaluation metrics for hate speech identification systems. There are several ways that the efficacy of hate speech identification algorithms can be evaluated, and these papers cover a wide range of such methods. Researchers and practitioners can use these metrics to compare systems and monitor developments in the field. Internet hate speech is extremely widespread. It is vital to understand how hate speech takes form and spreads on social media platforms like Twitter and Facebook if we are to develop effective identification systems and remedies. However, there is not a perfect technique for identifying hate speech, and the level of success may vary widely across languages. Articles in this area focus on the challenge of inciting violence in languages other than English, specifically Urdu, Hindi, and Marathi. They achieve this by considering the specifics of each language. These collectives are representative of the wide range of research into identifying hate speech; they also highlight the need for improved models, standardized data sets, multilingual approaches, and stringent evaluation metrics. Due to the dynamic nature of the problem of hate speech in the digital era, multidisciplinary studies involving natural language processing, machine learning, and the social sciences are necessary to find practical solutions.

Our review has highlighted the complexity and importance of identifying hate speech in online communication. Collective efforts are required to develop ethical and practical hate speech identification systems due to crucial factors, including language diversity, model robustness, and real-world ramifications. The list of review datasets is shown in Appendix B.

Table 3 For the category of papers

Category of Papers	Meaning	Reference Papers
Hate Speech Identification Methods	Research focused on developing techniques and models for identifying hate speech.	(Abro et al., 2020), (Gröndahl et al., 2018), (Ruwandika & Weerasinghe, 2018)
Hate Speech Datasets	Papers discussing the creation, curation, or analysis of hate speech datasets.	(Bosco et al., 2018), (Alrehili, 2019), (Sandaruwan et al., 2019), (Florio et al., 2020), (Teh et al., 2018), (Dhanya & Balakrishnan, 2021), (Aziz et al., 2023), (Haq et al., 2020), (Akhter et al., 2020), (Akram et al., 2023), (Saeed et al., 2021), (Parihar et al., 2021), (Hussain et al., 2022), (Ullah et al., 2022), (Al-Hassan & Al-Dossari, 2019), (Ali et al., 2022), (Aluru et al., 2020), (Xia et al., 2020), (Corazza et al., 2020), (Davidson et al., 2019), (Khan et al., 2022), (MacAvaney et al., 2019), (Mathew et al., 2019), (Mossie & Wang, 2020),

		(Mubarak et al., 2020), (Ousidhoum et al., 2019), (Arango et al., 2019), (Roy et al., 2020), (Saleh et al., 2023), (Sap et al., 2019), (Satapara et al., 2022), (Sutejo & Lestari, 2018), (Toraman et al., 2022), (William et al., 2022), (Ruwandika & Weerasinghe, 2018), (Zhang & Luo, 2019)
Hate Speech Identification in Multilingual Settings	Research focusing on identifying hate speech in multilingual or cross-lingual contexts.	(Zimmerman et al., 2018), (Zhou et al., 2020), (Fortuna et al., 2020), (Malik et al., 2022), (Mandl et al., 2020), (Pawar et al., 2022), (Velankar et al., 2022)
Hate Speech Evaluation Metrics	Papers discussing metrics and evaluation techniques for assessing hate speech identification systems.	(Aljarah et al., 2021), (Malik et al., 2022), (Röttger et al., 2022)
Social Media Analysis	Research on analysing hate speech in the context of social media platforms.	(Zimmerman et al., 2018), (Chakravarthi & Muralidaran, 2021), (Zhou et al., 2020), (Fortuna et al., 2019), (Gomez et al., 2020), (Ibrohim & Budi, 2019), (Mozafari et al., 2020), (Pereira-Kohatsu et al., 2019).
Hate Speech in Specific Languages	Research focusing on hate speech identification in specific languages.	(Dewani et al., 2021), (Kapil & EkbaL, 2020), (Karim et al., 2022), (Mehmood et al., 2020), (Pitenis et al., 2020), (Sharma et al., 2022).

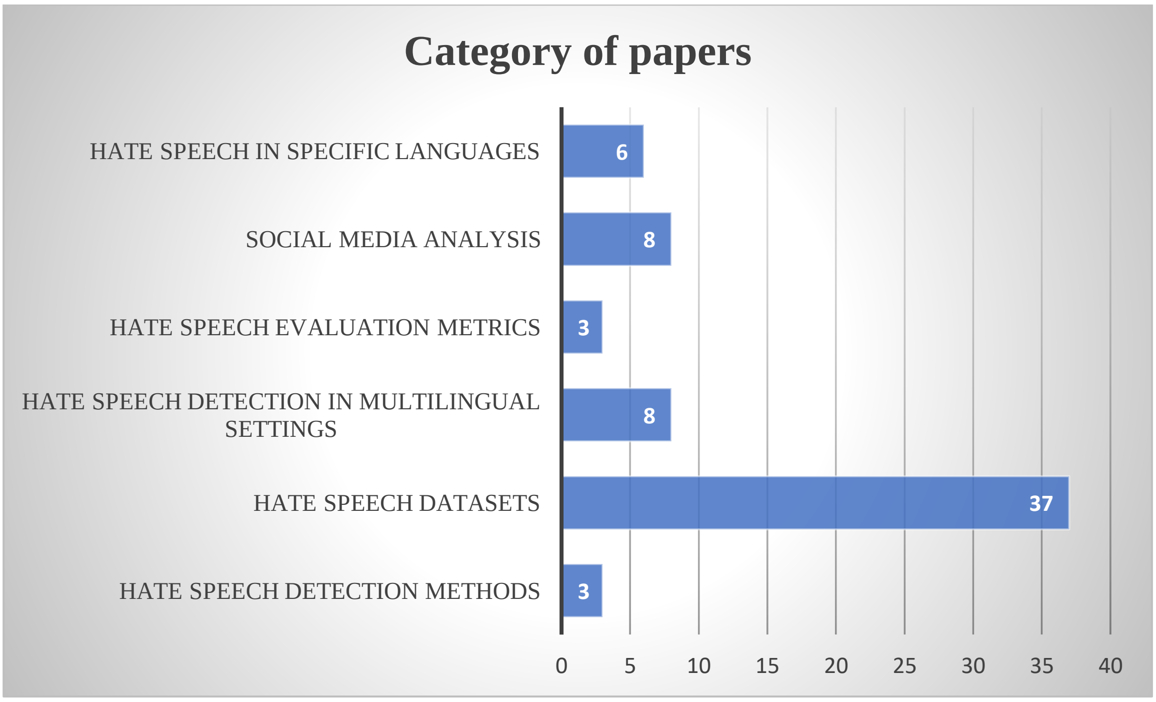

Figure 5 Category of paper in the study

Challenges of Machine Learning Models in Hate Speech Identification

A potential approach to reducing online toxicity, hate speech identification using machine learning algorithms faces significant obstacles. These difficulties stem from the multifaceted and nuanced nature of language in cyberspace. Managing ambiguity in contextual meaning is a significant obstacle.The meaning of a hateful statement might vary significantly from one situation to the next. It is important to remember that the definition of "hate speech" can vary depending on the setting (Abro et al., 2020). Machine learning models face a big hurdle when identifying sarcasm and irony, frequently employed in hate speech (Gröndahl et al., 2018). Models that can accurately identify hate speech in different languages are urgently needed due to the multilingual nature of the problem (Malik et al., 2022). Due to the lack of actual hate speech occurrences compared to non-hateful content, many existing models are biased (Ruwandika & Weerasinghe, 2018). As offenders find novel ways to circumvent hate speech identification algorithms, it becomes increasingly difficult to maintain accurate models (Gröndahl et al., 2018). A lack of annotated data for some languages and dialects prevents the construction of reliable hate speech identification models (Saeed et al., 2021). Semantically complicated constructs are common in hate speech, which makes it challenging for models to capture and understand meaning (Teh et al., 2018).

As (Dhanya & Balakrishnan, 2020) points out, models may fail to fare well when used in settings where hate speech is prevalent in a different culture. New slurs, slang, and symbols are constantly added to the hate speech vocabulary, making it difficult to maintain accurate models (Mohtaj et al., 2022). Inheriting biases from training data might cause models to make discriminatory judgements and even single out some populations for harm (Davidson et al., 2019). For legal and ethical considerations, hate speech identification programs must explain their findings (Aljarah et al., 2021). Multimodal identification algorithms must be created because hate speech frequently consists of text, images, and videos (Zhou et al., 2020). Complexity arises from different social channels having their own lingo and cultural standards, making identifying hate speech challenging (Gomez et al., 2020). A layer of difficulty is added to model development when real-time identification of hate speech is necessary to avert injury (Zimmerman a et al., 2018). For developers and platforms, striking a fair balance between protecting free expression and identifying hate speech presents ethical and legal issues (Mehmood et al., 2020). More work is needed to develop reliable metrics for evaluating hate speech identification systems (Pitenis et al., 2020). It can be especially difficult to acquire data and construct accurate models for less-resourced languages (Kapil & Ekbal, 2020). It is possible to use adversarial attacks to alter hate speech so that monitoring systems cannot identify it (Sap et al., 2019). For user confidence and responsibility, hate speech identification models must provide clear justifications for their findings (Aljarah et al., 2021). Models that do well in a lab setting might translate poorly to the real world, where data is often noisier and more varied (Malik et al., 2022). These obstacles underline the difficulty of creating trustworthy machine-learning models for hate speech identification and highlight the importance of maintaining investment in this field of study. Some deep learning architectures find it difficult to audit and explain their judgments because of their "black-box" nature, which makes it difficult to understand how and why a given prediction is generated. Machine learning algorithms show potential for hate speech identification, but they face several obstacles that must be overcome to improve in areas such as accuracy, fairness, and generalizability. To overcome these obstacles, we must take an interdisciplinary approach integrating linguistics, ethics, and technology to develop more accurate methods of identifying hate speech.

Challenges of Hate Speech Identification Datasets

Datasets play a critical role when training and testing hate speech identification models. However, the quality and efficacy of machine learning algorithms built for this purpose may be compromised by the difficulties in generating and curating hate speech datasets.There is typically a significant underrepresentation of hate speech in hate speech datasets compared to non-hate speech datasets (Abro et al., 2020). Poor hate speech identification has been linked to model bias towards the majority class (Gröndahl et al., 2018). There are likely to be discrepancies in labelling because of the subjective nature of hate speech annotation, which might be affected by the annotators' prejudices (Davidson et al., 2019). Dataset quality can be impacted by the potential for annotators to arrive at various conclusions on hate speech (Arango et al., 2019). Since hate speech can be expressed in many different ways, it is difficult to collect representative data (Zhou et al., 2020) across various languages. Annotation and identification methods for multimodal hate speech, which includes text, images, and videos, must be more sophisticated (Gomez et al., 2020). Because of the importance of context, sarcasm, and cultural references in hate speech, it can be challenging to discern just on text alone.

Identification is made more difficult by using slang, regional dialects, and euphemisms (Aluru et al., 2020). As time passes, hate speech changes such that it might evade monitoring (MacAvaney et al., 2019). It is difficult to keep data sets current with changing trends in hate speech (Sap et al., 2019). Ethical issues about exposing annotators to objectionable information are raised when creating hate speech datasets because building these datasets can include curating harmful content (Arango et al., 2019). Annotators' safety is paramount (Xia et al., 2020). Davidson et al. (2019) found that hate speech identification datasets may reflect the underlying biases of the platforms where they were collected. These prejudices can propagate into inaccurate hate speech identification algorithms (Sap et al., 2019). It is difficult to collect enough labeled data for hate speech identification in low-resource languages (Dhanya & Balakrishnan, 2021). Development in such languages is hampered by a dearth of resources (Ullah et al., 2022). It is difficult to understand the reasoning behind the conclusions made by many hate speech identification methods, especially deep learning models (Aljarah et al., 2021). This is crucial for honesty and transparency (Mehta & Passi, 2022). Generalization issues have been identified in hate speech models trained on a single language or culture (Zhang & Luo, 2019). Adapting across languages and cultures can be difficult. Traditional criteria, such as accuracy, may not reflect real-world performance, making it difficult to choose appropriate evaluation metrics for hate speech identification (Mathew et al., 2019). In the context of potential harm, metrics should account for false positives and negatives. (Chakravarthi & Muralidaran, 2021). To avoid being caught, bad actors may try to trick hate speech identification systems with manipulated input (Gröndahl et al., 2018). Being resistant to these kinds of attacks is essential. There are several obstacles in the way of hate speech identification datasets, including poor data quality, prejudice, context, the ever-changing nature of hate speech, and ethical concerns. Researchers and practitioners in this area must solve these obstacles to create robust and equitable hate speech identification systems. Many different and difficult obstacles exist when working with hate speech identification datasets. In order to overcome these obstacles, researchers must seriously consider the methods used to create datasets, how bias might be reduced, and the criteria used to evaluate results. These challenges must be overcome to develop reliable and effective hate speech identification models that work in a range of linguistic and cultural contexts.

Challenges of Hate Speech Identification Feature Sets

Successful hate speech identification models require careful consideration in feature selection and careful engineering. However, some obstacles can hinder the models' performance and generalizability during this process. The difficulty stems partly from the features' limited ability to express data. (Davidson et al., 2019) Many hate speech identification models use training data that may not adequately represent the variety of languages, cultures, and circumstances in which hate speech occurs. Because of this, underrepresented groups may receive inferior service from biased models. According to research (Sharma et al., 2022), online users frequently transition between languages when communicating. It is challenging to spot hate speech in works that use a variety of languages and dialects. Hate speech identification models need to consider the context in which specific words or phrases are used, as researched by (Aljarah et al., 2021). It is possible that a word that's insulting in one setting is not so in another. For reliable identification, context is essential. (Gomez et al., 2020) Note that hate speech is often disguised as irony or sarcasm, making it difficult for models to distinguish between real hate speech and parody.

According to (Gröndahl et al., 2018), attackers can circumvent hate speech identification systems through deliberate text manipulation. To evade discovery, adversaries may use misspellings, character substitutions, or other methods. Hate speech identification models may inherit biases existing in the training data, according to (Sap et al., 2019), which can result in unjust or erroneous judgments, particularly regarding underrepresented groups. According to (Pawar et al. 2022), managing vast amounts of user-generated information on social media in real time is difficult. We need models that can scale and efficient feature sets. There needs to be a happy medium between hate speech identification and user privacy, according to (MacAvaney et al., 2019). Users may have privacy concerns if their data is collected and analyzed for identifyingreasons. According to (Saeed et al., 2021), more fine-grained feature sets and models are necessary to identify specific forms of hate speech, such as targeting specific communities or based on various traits. Hate speech identification methods, according to (Chakravarthi & Muralidaran, 2021), should justify their actions, especially when content is identified or users are punished. Reference: This intricacy and the necessity for continual research to enhance the reliability and fairness of identification methods, as well as to address the ever-changing character of online hate speech, are brought into sharp focus by these difficulties. Researchers keep digging into new feature sets, methods, and datasets to address these issues and promote more welcoming and safe digital communities. To keep up with the ever-increasing amount of stuff on the web, the difficulties of using static, inflexible feature sets for hate speech identification underscore the necessity for culturally aware, contextually aware feature engineering approaches. To create models that can accurately capture the nuanced linguistic patterns present in hate speech in various languages, cultural settings, and contexts, it is essential to find solutions to these problems.

Future Research Directions in Hate Speech Identification

Improvements in natural language processing and a deeper appreciation of the difficulties inherent in automatic hate speech identification are driving this change. Several promising fresh areas of inquiry have emerged as scientists work to improve hate speech identification models' accuracy and sturdiness. Recent research has highlighted the difficulty and significance of hate speech identification in languages other than English (Abro et al., 2020; Ruwandika & Weerasinghe, 2018; Akhter et al., 2020). Hate speech identification models for various languages, emphasizing low-resource languages, should be developed and improved in future studies. The importance of context and multimodal content in hate speech identification has been emphasized by several recent studies, including those by (Gröndahl et al., 2018) and (Gomez et al., 2020). For greater precision, hate speech recognition programs should investigate ways to incorporate contextual and multimodal information efficiently. (Sap et al., 2019) and (Davidson et al., 2019) talk about how to address the problem of bias in hate speech identification algorithms and datasets. Future research should focus on developing and analyzing bias reduction strategies to maintain fairness in hate speech identification and avoid potential harm and misclassification. Model explainability is emphasized by (Karim et al., 2022; Aljarah et al., 2021) for hate speech identification. Improved trust and accountability can be achieved by creating interpretable models and explainability methods. The difficulties in identifying hate speech that uses figurative language, slang, or coded terms are discussed by (Mehmood et al., 2020) and (Akram et al., 2023). Methods for efficient identification and categorization of such material can be the subject of future studies. Both (Malik et al., 2022) and (Karim et al., 2022) recognize the challenges that zero-shot and few-shot circumstances present for hate speech identification methods. Methods to improve model performance when there is insufficient labelled data should be investigated in future studies. (Florio et al., 2020) note the importance of investigating domain adaptation strategies to fulfill the demand for hate speech identification algorithms that can operate across several domains and platforms. It is acknowledged by (Xia et al., 2020) that standardized evaluation measures are needed for hate speech identification. Metrics that accurately reflect the practical significance of hate speech identification can be proposed and validated in future studies. It is crucial to factor in user feedback and consider how hate speech identification might affect people. Evaluation and risk reduction measures from the end user's perspective can be explored in subsequent research (Mehta & Passi, 2022). Privacy concerns, free speech, and regulatory consequences are ethical and legal issues that need to be investigated in the context of hate speech identification (Sutejo & Lestari, 2018).

The future of this field of study lies in determining the lasting impacts of hate speech and creating efficient intervention measures. Creating scalable, real-time hate speech identification algorithms for online communities and social media is ongoing. Future studies should focus on practical applications (Kapil & Ekbal, 2020). The identification of hate speech in Sinhala, a low-resource language, was discussed by (Sandaruwan et al., 2019). Creating tools and datasets for hate speech identification in languages with sparse data should be a priority for future studies. To identify religious, sectarian, and racial bigotry in Urdu, (Akram et al., 2023) presented a benchmark corpus. Similar standards for other languages and cultures to combat other forms of hate speech can be investigated in future research. (Saeed et al., 2021) studied the problem of labelling hate speech in Classical Urdu. This system could be applied to additional languages and other forms of harmful substance identification. A technique for emotion recognition and sentiment analysis in resource-constrained Urdu was proposed by (Ullah et al., 2022). Extending this work to other low-resource languages and investigating potential uses beyond hate speech identification are possible next steps. Unsupervised lexical normalization for sentiment analysis in Roman Hindi and Urdu was described by (Mehmood et al., 2020). Lexical normalization techniques can be studied further to better sentiment analysis and, by extension, hate speech identification. Using semantic and embedding models, (Hussain et al., 2022) found hate speech in Urdu. Improving the identification accuracy of these models and investigating further linguistic nuance can be the focus of future research. Roman Urdu hate speech prediction was mapped geographically (Aziz et al., 2023). This method can be expanded in future studies to examine the global dissemination of hate speech in languages and areas other than English. (Zhou et al., 2020) reviewed research on automatic hate speech identification across multiple media types and languages. In the future, researchers may improve hate speech identification by using more complex multimodal models. (Badjatiya et al., 2017) investigated whether or not an explanation of text classifiers is necessary or sufficient for hate speech identification. Comprehensive approaches for explaining the judgments of text classifiers, especially in complicated circumstances, can be developed with more research. A collaborative assignment on hope speech identification was undertaken by (Chakravarthi & Muralidaran, 2021). In subsequent research, more robust models and datasets for recognizing 'hope speech' in many languages can be investigated. (Mossie & Wang's, 2020) research used hate speech identification to identify marginalized groups. In the future, researchers can learn more about the specific problems that marginalized groups confront and use that information to create effective responses. (Toraman et al., 2022) investigated the use of cross-domain transfer for identifying hate speech at scale. Possible next steps in hate speech identification research include investigating cutting-edge cross-domain transfer methods. Contextual comprehension, linguistic diversity, bias reduction, transparency, and real-time capabilities are all areas where hate speech identification research could benefit from further development. Researchers can aid in the creation of more reliable, objective, and efficient hate speech identification models by focusing on these issues.

Conclusion

As online platforms struggle to deal with the rise of unpleasant and dangerous information, hate speech identification has emerged as a critical issue in the modern era. By analyzing the findings of 65 carefully chosen papers, this systematic literature review has offered a thorough overview of the present research in hate speech identification. The review of these studies has uncovered numerous important themes and potential avenues for future study on this pressing topic. First and foremost, there needs to be a way to adequately stress the significance of multilingual and cross-lingual hate speech identification. Although many investigations have been conducted on English language materials, it has been stressed how important it is to establish reliable hate speech identification algorithms in languages that lack enough resources. Because online hate speech affects people worldwide, researchers need to focus on creating algorithms and datasets that can reliably recognize it in various languages. Another major trend in hate speech identification is contextual and multimodal analysis. Adding contextual and multimodal data, such as pictures and videos, can greatly increase the accuracy of hate speech identification algorithms. Future studies should look for creative ways to incorporate these many data sources into frameworks for identifying hate speech. Issues of bias and impartiality hamper the identification of hate speech. Unintentional harm might result from using biased datasets and models. To avoid unfairness in hate speech identification and misclassification, researchers need to work on creating and assessing bias mitigation approaches in the future. For models to earn users' trust and shoulder responsibility, they must be easy to explain and interpret. To make these systems more understandable and approachable, especially in complicated settings, researchers should put effort into creating interpretable models and explainability methodologies. Using slang and coded language in hate speech has its difficulties. To make sure that hate speech identification models are flexible enough to account for shifting linguistic norms, future studies should look into methods for efficient identification and classification of such content. Pay close attention to situations with zero or few learning opportunities. In these cases, where there is a need for labelled data, novel approaches are required to improve model performance. In the future, researchers should look for ways to improve hate speech identification models' performance in low-data settings.

It has been emphasized that domain adaptability and generalization are crucial for developing flexible hate speech identification systems. Researchers should investigate domain adaptation approaches to ensure that models can function well across various online platforms and communities. There is consensus that we need more consistent metrics for assessment. Metrics that go beyond typical performance indicators and represent the real-world impact of hate speech identification can be proposed and validated in future studies. The importance of user-cantered strategies in hate speech identification cannot be overstated. Researchers should focus on user-centric evaluation and mitigation measures to reduce adverse effects and improve services. Privacy issues, the right to free expression, and the ramifications of government regulation should all be at the forefront of future studies. Finding a middle ground between identifying hate speech and protecting fundamental rights is difficult but essential. Future studies should focus on identifying the long-term consequences of hate speech and creating efficient intervention measures. Finding solutions to reduce the harm that hate speech causes is essential to making the internet more welcoming for everyone. Hate speech identification solutions that work in real-time and can be scaled up are in great demand. Future studies should focus on developing deployable applications for online communities and social media platforms. In conclusion, identifying hate speech is an evolving and complex field. By underlining the importance of multilingual techniques, multimodal analysis, fairness issues, interpretability, and more, this systematic literature review has established a road map for future research. Scholars can help create safer and more welcoming online communities by tackling these issues and pursuing these lines of inquiry, which will lead to the creation of more advanced hate speech identification systems. Researchers must think outside the box to effectively counteract the

Appendix A: Methods and Approaches for Hate Speech Identification

Paper	Title	Dataset	Approach	Results	Limitations
(Abro et al., 2020)	Automatic hate speech detection using machine learning: A comparative study	Twitter Hate Speech Dataset	Machine Learning Classifiers	Achieved 79% off overall accuracy by using the bigram feature with the Support vector machine algorithm.	Limited to English, Small Dataset
(Gröndahl et al., 2018)	All you need is "love" evading hate speech identification	Not Specified	Evasion Techniques Analysis	Identified Weaknesses in Identification Models	Focus on Evasion Techniques only
(Bosco et al., 2018)	Overview of the Evalita 2018 hate speech identification task	Evalita 2018 Hate Speech Identification Task Dataset	Task Overview	Provided Overview of the Task	Limited to Italian language only, Task-Specific Paper
(Alrehili, 2019)	Automatic hate speech identification on social media: A brief survey	Various Hate Speech Datasets	Survey	Summarized Hate Speech Identification Methods	Survey Paper
(Sandaruwan et al., 2019)	Sinhala hate speech identification in social media using text mining and machine learning	Sinhala Hate Speech Dataset	Text Mining and Machine Learning	Achieved recall value as 0.84 with 92.33% accuracy with the use of character trigram with Multinomial Naïve Bayes	Limited to Sinhala

(Wang et al., 2022)	Political Hate Speech Detection and Lexicon Building: A Study in Taiwan	Taiwan Hate Speech Dataset	Deep learning & lexicon based approach	BERT achieved 73.2% F1- score and lexicon based approach achieved 57.1% F1-score.	Focused on only Political Hate Speech, Limited to Chinese language
(The et al., 2018)	Identifying and categorizing profane words in hate speech	Not-specified	Lexicon based approach	Identified Profane Words in Hate Speech	Limited to Profanity Identification
(Dhanya & Balakrishnan, 2021)	Hate speech identification in Asian languages: A survey	Various Asian Language Datasets	Survey	Summarized Hate Speech Identification in Asian Languages	Survey Paper
(Aziz et al., 2023)	Geo-Spatial Mapping of Hate Speech Prediction in Roman Urdu	Roman Urdu Hate Speech Dataset	Geo-Spatial Mapping	93% accuracy was attained utilizing the suggested feed-forward neural network and random forest with fastText word embedding.	Limited to Roman Urdu
(Haq et al., 2020)	USAD: An intelligent system for slang and abusive text detection in PERSO-Arabic-scripted Urdu	PERSO-Arabic-scripted Urdu Dataset	Slang and Abusive Text Detection using Lexicon based approach	The model identifies 72.6% correctly as abusive or non-abusive Tweet.	Limited to PERSO-Arabic-scripted Urdu
(Akhter et al., 2020)	Automatic detection of offensive language for urdu and roman urdu	Urdu and Roman Urdu Offensive Language Dataset	Machine Learning Classifiers	On Roman Urdu and Urdu datasets, respectively, LogitBoost and SimpleLogistic beat the other models, achieving 99.2% and 95.9% values of F-measure.	Limited to Urdu and Roman Urdu only
(Dewani et al., 2021)	Development of computational linguistic resources for automated identification of textual cyberbullying threats in the Roman Urdu language	Roman Urdu Cyberbullying Dataset	Computational Linguistic Resource Development	Developed Resources for Roman Urdu Cyberbullying Identification	Limited to Roman Urdu
(Akram et al., 2023)	ISE-Hate: A benchmark corpus for inter-faith, sectarian, and ethnic hatred detection on social media in Urdu	ISE-Hate Corpus	Machine learning and deep learning techniques	The best method for identifying hateful material in Urdu tweets turned out to be BERT.	Focused on only Urdu language, Limited to inter-faith, sectarian, and ethnic hatred detection on social media
(Saeed et al., 2021)	Roman Urdu toxic comment classification	Roman Urdu Toxic Comment Dataset	Toxic Comment Classification	Classified Toxic Comments in Roman Urdu	Limited to Roman Urdu
(Parihar et al., 2021)	Hate speech identification using natural language processing: Applications and challenges	Various Hate Speech Datasets	NLP-Based Identification	Discussed NLP-Based Approaches and Challenges	Survey Paper

(Hussain et al., 2022)	Identification of offensive language in Urdu using semantic and embedding models	Urdu Offensive Language Dataset	Semantic and Embedding Models	Utilized Semantic and Embedding Models for Identification	Limited to Urdu
(Ullah et al., 2022)	A novel approach for emotion identification and sentiment analysis for low resource Urdu language based on CNN-LSTM	Low Resource Urdu Emotion and Sentiment Analysis Dataset	CNN-LSTM Approach	Developed Approach for Emotion and Sentiment Analysis in Urdu	Limited to Urdu and Low Resource
(Mehmood et al., 2020)	An unsupervised lexical normalization for Roman Hindi and Urdu sentiment analysis	Roman Hindi and Urdu Sentiment Analysis Dataset	Lexical Normalization	Developed Lexical Normalization Approach for Sentiment Analysis	Limited to Roman Hindi and Urdu
(Al-Hassan & Al-Dossari, 2019)	Identification of hate speech in social networks: A survey on multilingual corpus	Various Multilingual Hate Speech Datasets	Survey	Surveyed Hate Speech Identification in Multilingual Contexts	Survey Paper
(Ali et al., 2022)	Hate speech identification on Twitter using transfer learning	Twitter Hate Speech Dataset	Transfer Learning	Utilized Transfer Learning for Hate Speech Identification	Focused on Twitter Data
(Aluru et al., 2020)	Deep learning models for multilingual hate speech identification	Various Multilingual Hate Speech Datasets	Deep Learning Models	Developed Deep Learning Models for Multilingual Hate Speech Identification	Multilingual Focus
(Chakravarthi & Muralidaran, 2021)	Findings of the shared task on hope speech identification for equality, diversity, and inclusion	Shared Task Datasets	Hope Speech Identification	Shared Task Findings for Hope Speech Identification	Focused on Hope Speech Identification
(Corazza et al., 2020)	A multilingual evaluation for online hate speech identification	Various Multilingual Hate Speech Datasets	Multilingual Evaluation	Conducted Multilingual Evaluation of Hate Speech Identification	Multilingual Focus
(Davidson et al. 2019)	Racial bias in hate speech and abusive language identification datasets	Various Hate Speech Datasets	Bias Analysis	Analyzed Racial Bias in Hate Speech Datasets	Bias Analysis Focus
(Fortuna et al., 2019)	A hierarchically-labeled portuguese hate speech dataset	Portuguese Hate Speech Dataset	Dataset Development	Developed Portuguese Hate Speech Dataset with Hierarchy	Dataset Development Focus
(Fortuna et al., 2020)	Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of hate speech datasets	Various Hate Speech Datasets	Dataset Analysis	Analyzed and Compared Hate Speech Dataset Classifications	Dataset Analysis Focus
(Gomez et al., 2020)	Exploring hate speech identification in multimodal publications	Multimodal Hate Speech Dataset	Multimodal Identification	Explored Hate Speech Identification in Multimodal Content	Multimodal Focus
(Ibrohim & Budi, 2019)	Multi-label hate speech and abusive language	Indonesian Hate Speech Dataset	Multi-label Identification	Achieved Multi-label Identification in Indonesian Twitter	Limited to Indonesian

	identification in Indonesian Twitter				
(Kapil & Ekbal, 2020)	A deep neural network based multi-task learning approach to hate speech identification	Various Hate Speech Datasets	Multi-task Learning	Developed Multi-task Learning Approach for Hate Speech Identification	Multi-task Learning Focus
(Karim et al., 2022)	Multimodal hate speech identification from Bengali memes and texts	Bengali Hate Speech Dataset	Multimodal Identification	Developed Multimodal Hate Speech Identification from Memes and Texts	Multimodal Focus
(Khan et al., 2022)	HCovBi-caps: Hate speech identification using convolutional and Bi-directional gated recurrent unit with Capsule network	HCovBi-caps Dataset	Capsule Network	Utilized Capsule Network for Hate Speech Identification	Focused on HCovBi-caps
(MacAvaney et al., 2019)	Hate speech identification: Challenges and solutions	Various Hate Speech Datasets	Challenges and Solutions	Discussed Challenges and Proposed Solutions in Hate Speech Identification	Challenge Discussion
(Malik et al., 2022)	Deep learning for hate speech identification: A comparative study	Various Hate Speech Datasets	Deep Learning Models	Compared Deep Learning Models for Hate Speech Identification	Comparative Study
(Mandl et al., 2020)	Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in Tamil, Malayalam, Hindi, English, and German	HASOC 2020 Datasets	Task Overview	Provided Overview of the HASOC 2020 Task	Task-Specific Paper
(Mehta & Passi, 2022)	Social Media Hate Speech Identification Using Explainable Artificial Intelligence (XAI)	Various Hate Speech Datasets	XAI-Based Identification	Utilized XAI for Hate Speech Identification	XAI Focus
(Mohtaj et al., 2022)	A Feature Extraction based Model for Hate Speech Identification	Various Hate Speech Datasets	Feature Extraction	Developed Feature Extraction Model for Hate Speech Identification	Feature Extraction Focus
(Mozafari et al., 2020)	A BERT-based transfer learning approach for hate speech identification in online social media	Various Hate Speech Datasets	Transfer Learning with BERT	Utilized BERT for Hate Speech Identification	Transfer Learning with BERT
(Mubarak et al., 2020)	Arabic offensive language on Twitter: Analysis and experiments	Arabic Offensive Language Dataset	Offensive Language Analysis	Analyzed Offensive Language on Twitter	Limited to Arabic
(Ousidhoum et al., 2019)	Multilingual and multi-aspect hate speech analysis	Various Multilingual Hate Speech Datasets	Multilingual Analysis	Conducted Multilingual and Multi-aspect Hate Speech Analysis	Multilingual Focus

(Pereira-Kohatsu et al., 2019)	Identifying and monitoring hate speech in Twitter	Twitter Hate Speech Dataset	Twitter Hate Speech Identification	Identifyed and Monitored Hate Speech on Twitter	Twitter-Focused Study
(Pitenis et al., 2020)	Offensive language identification in Greek	Greek Offensive Language Dataset	Offensive Language Identification	Identifyed Offensive Language in Greek	Limited to Greek
(Röttger et al., 2022)	MULTILINGUAL HATECHECK: Functional Tests for Multilingual Hate Speech Identification Models	Various Multilingual Hate Speech Datasets	Functional Testing	Conducted Functional Tests for Multilingual Hate Speech Identification Models	Multilingual Focus
(Roy et al., 2020)	A framework for hate speech identification using deep convolutional neural network	Various Hate Speech Datasets	Deep CNN	Created Deep Learning Framework to Identify Hate Speech	Deep Learning Focus
(Saleh et al., 2023)	Identification of hate speech using BERT and hate speech word embedding with deep model	BERT-Based Hate Speech Identification	BERT and Deep Model	Utilized BERT and Word Embeddings for Hate Speech Identification	Model-Centric Study
(Sap et al., 2019)	The risk of racial bias in hate speech identification	Various Hate Speech Datasets	Bias Analysis	Analyzed Risk of Racial Bias in Hate Speech Identification	Bias Analysis Focus
(Satapara et al., 2022)	Overview of the hasoc subtrack at fire 2022: Hate speech and offensive content identification in English and Indo-Aryan languages	HASOC 2022 Subtrack Datasets	Task Overview	Provided Overview of the HASOC 2022 Subtrack	Task-Specific Paper
(Sharma et al., 2022)	Ceasing hate with moh: Hate speech identification in Hindi–English code-switched language	Hindi–English Code-Switched Hate Speech Dataset	Code-Switched Hate Speech Identification	Identifyed Hate Speech in Code-Switched Language	Code-Switched Language Focus
(Toraman et al., 2022)	Large-scale hate speech identification with cross-domain transfer	Various Hate Speech Datasets	Cross-Domain Transfer	Applied Cross-Domain Transfer for Large-Scale Hate Speech Identification	Cross-Domain Transfer Focus
(Velankar et al., 2022)	Mono vs. Multilingual Approaches in Code-Mixed Hate Speech Identification	Hindi-English Code-Mixed Hate Speech Dataset	Code-Mixed Hate Speech Identification	Compared Mono and Multilingual Approaches for Code-Mixed Hate Speech	Code-Mixed Language Focus
(Zhang & Luo, 2021)	Overview of the shared task on hope speech identification for equality, diversity, and inclusion	Shared Task Datasets	Hope Speech Identification	Provided Overview of the Hope Speech Identification Shared Task	Task-Specific Paper
(Mathew et al., 2019)	Spread of hate in online social media	341K user dataset with 21 million posts	The dynamics of post dispersion on Gab between hateful and non-hateful users	The fact that the hateful users are far more closely connected is a significant result.	Focuses on propagation study, not a specific dataset or approach.

(William et al., 2022)	Machine Learning based Automatic Hate Speech Recognition System	standard publicly available dataset	Machine learning	The testing findings showed that bigram features performed best with 79 percent overall accuracy when utilized with the bigram feature set.	Identifying automated hate speech messages can be made easier with the findings of our investigation. It will also be used as a benchmark for future research into existing automatic text classification algorithms, based on the results of the various comparisons.
(Pawar et al., 2022)	Challenges for Hate Speech Recognition System: Approach based on Solution	Language complexity, differing views on what constitutes hate speech	Machine learning	the art SVM results that are easier to comprehend than neural approaches	The challenges that this endeavour faces on a technological and practical level
(Mossie & Wang, 2020)	Vulnerable community identification using hate speech identification on social media	dataset for Amharic texts, we crawled Facebook pages to prepare the corpus	classical and deep learning-based classification algorithm	A strategy for identifying hate speech on social media that targets marginalized communities at risk	This can also encourage the way towards the development of policies, strategies, and tools to empower and protect vulnerable communities.
(Ruwandika & Weerasinghe,2018)	Identification of hate speech in social media	local English text dataset	Machine learning techniques	Tf-idf features performed best with an F-score of 0.719	-
(Lingiardi,et al., 2020)	Mapping Twitter hate speech towards social and sexual minorities: A lexicon-based approach to semantic content analysis	Tweets	Lexicon-based approach	Give a current picture of community attitudes and behaviours towards social, racial, sexual, and gender minorities. This information can be used to guide national and local programs aimed at preventing intolerance.	-
(Zhou,et al., 2020)	Deep learning based fusion approach for hate speech identification	data sets of the SemEval 2019 Task 5	Deep learning-based fusion approach	The results show that the accuracy and F1-score of the classification are significantly improved.	The degree of integration is not deep enough. to achieve the practical significance of performance at a little extra cost.
(Sutejo & Lestari,2018)	Indonesia hate speech identification using deep learning	We utilized both textual and acoustic features	Deep learning	The best model using textual feature obtained Fl-score 87.98% which is higher than the model of using acoustic feature only (Fl-score 82.5%), and the model of using acoustic and	-

				lexical features (Fl-score 86.98%).	
(Alshalan & Al-Khalifa,2020)	A deep learning approach for automatic hate speech identification in the Saudi Twittersphere	a public dataset of 9316 tweets labeled as hateful,	Deep learning approach	The dataset indicated that the CNN model performed the best, with (AUROC) of 0.89 and an F1-score of 0.79.	Other features can be also incorporated with word embeddings such as the user's gender, age, and location. Moreover, to alleviate the lack of the context problem
(Aljarahet al., 2021)	Intelligent identification of hate speech in Arabic social network: A machine learning approach	Hate speech based on Arabic context on the Twitter network.	Machine learning approach	The processed dataset is tested using a Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT), and Random Forest (RF), with the best results obtained by RF employing the feature set of (TF-IDF) and profile-related attributes.	-
(Elzayadyet al., 2023)	A hybrid approach based on personality traits for hate speech identification in Arabic social media	personality trait features from Arabic text	Hybrid approach based on personality traits	the proposed approach is superior in terms of the macro-F1 score by achieving 82.3%	extend our proposed framework to include multi-personality trait features rather than binary.
(Xia et al., 2020)	Demoting Racial Bias in Hate Speech Identification	African American English data set	Machine learning	Our approach can significantly lower the false positive rate for AAE text while having a negligible impact on the classification accuracy of hate speech.	-
(Arango et al., 2019)	Hate Speech Identification is Not as Easy as You May Think: A Closer Look at Model Validation	Large scale social platforms	Machine learning	The outcomes of cutting-edge systems show that supervised methods perform nearly flawlessly, but only on particular datasets.	Data overfitting and sampling issues and less in accuracy.
(Zimmerman et al., 2018)	Improving Hate Speech Identification with Deep Learning Ensembles	Twitter	Deep learning	Using a publicly available hate speech evaluation dataset, this method outperforms the original study by roughly 5 points in terms of F-measure.	Difficulties experienced with reproducibility of DL methods and comparison of findings from other work.
(Florio et al., 2020)	Time of Your Hate: The Challenge of Time in Hate Speech Identification on social media	Contro l'odio" dataset	Machine learning	The outcomes demonstrate how sensitive AlBERTo is to the fine-tuning set's temporal distance.	An event-heavy dataset is encountered by a supervised classification model.

Appendix B: Datasets for Hate Speech Identification

Dataset Name	Size	Categories of the Dataset	Languages	Reference
HateEval	17,000	Hate Speech, Offensive Language, Others	African-American English	(Abro et al., 2020)
L-HSAB	5,000	Hate Speech, Abusive Language, Profanity	English	(Gröndahl et al., 2018)
EVALITA 2018	Varies	Hate Speech, Offensive Language	Italian, English, and Others	(Bosco et al., 2018)
Hate Speech Identification Dataset (HSDD)	-	Hate Speech	Arabic	(Alrehili, 2019)
SL-HSD	2,300	Hate Speech, Offensive Language	Sinhala	(Sandaruwan et al., 2019)
Profanity Identification Dataset	40,000	Profane Words in Hate Speech	English	(Teh et al., 2018)
Offensive Language Identification Dataset	1,600	Offensive Language in Hate Speech	Malay	(Dhanya & Balakrishnan, 2021)
Roman Urdu Hate Speech Dataset	9,000	Hate Speech in Roman Urdu	Roman Urdu	(Aziz et al., 2023)
USAD	20,000	Slang and Abusive Text Identification	PERSO-Arabic-scripted Urdu	(Haq et al., 2020)
Urdu Hate Speech Dataset (UHSD)	8,991	Hate Speech, Offensive Language	Urdu, Roman Urdu	(Akhter et al., 2020)
Roman Urdu Cyberbullying Dataset	3,000	Cyberbullying Threats	Roman Urdu	(Dewani et al., 2021)
ISE-Hate	10,000	Inter-faith, Sectarian, and Ethnic Hatred Identification	Urdu	(Akram et al., 2023)
Roman Urdu Toxic Comment Classification	11,964	Toxic Comment Classification	Roman Urdu	(Saeed et al., 2021)
P-Urdu-Offensive	5,000	Offensive Language in Hate Speech	Roman Urdu	(Parihar et al., 2021)
Offensive Language in Urdu Dataset	5,000	Offensive Language in Hate Speech	Urdu	(Hussain et al., 2022)
Urdu Emotion and Sentiment Analysis Dataset	25,000	Emotion Identification and Sentiment Analysis	Urdu	(Ullah et al., 2022)
Roman Hindi and Urdu Sentiment Analysis Dataset	25,000	Sentiment Analysis in Roman Hindi and Urdu	Roman Hindi, Roman Urdu	(Mehmood et al., 2020)
Saudi Twitter Hate Speech Dataset	20,000	Hate Speech, Offensive Language	Arabic	(Al-Hassan & Al-Dossari, 2019)
Hate Speech on Twitter Dataset	16,500	Hate Speech on Twitter	English	(Ali et al., 2022)
J-Hate	4,355	Hate Speech, Offensive Language	Japanese	(Aluru et al., 2020)
Hate Speech in Persian (HSP)	15,800	Hate Speech, Offensive Language	Persian	(Corazza et al., 2020)
Hate Speech Dataset (HASOC)	5,000	Hate Speech, Offensive Language	English, German, Hindi	(Davidson et al., 2019)
HASOC 2019	3,500	Hate Speech, Offensive Language	English, German, Hindi	(Fortuna et al., 2019)
HASOC 2020	7,778	Hate Speech, Offensive Language	English, German, Hindi	(Fortuna et al., 2020)
INACL-IMW 2019	1,500	Hate Speech Identification in Indonesian Twitter	Indonesian	(Ibrohim & Budi 2019)
HSD 3.0	3,000	Hate Speech, Offensive Language	Hindi	(Kapil & Ekbal, 2020)

Bengali Hate Speech Dataset (BHSD)	3,000	Hate Speech, Offensive Language	Bengali	(Karim et al., 2022)
HCovBi-caps	1,000,000	Hate Speech Classification	English	(Khan et al., 2022)
Multilingual Hate Speech Dataset (MHSD)	5,000	Hate Speech, Offensive Language	English, German, Hindi, Italian	(MacAvaney et al., 2019)
(Multi HASOC) A Multilingual Dataset for Hate Speech Identification	8,000	Hate Speech, Offensive Language	English, German, Hindi, Konkani	(Mathew et al., 2019)

Disclosure
This research is part of PhD. program enrolled by the primary author at the Institute of Computing & Information Technology (ICIT), Gomal University, Dera Ismail Khan, Pakistan.

Conflicts of Interest
The authors declare that they have no conflicts of interest.

References

1. Abro, S., Shaikh, S., Khand, Z. H., Zafar, A., Khan, S., & Mujtaba, G. (2020). Automatic hate speech identification using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications*, *11*(8).
2. Akhter, M. P., Jiangbin, Z., Naqvi, I. R., Abdelmajeed, M., & Sadiq, M. T. (2020). Automatic identification of offensive language for urdu and roman urdu. *IEEE Access*, *8*, 91213-91226.
3. Akram, M. H., Shahzad, K., & Bashir, M. (2023). ISE-Hate: A benchmark corpus for inter-faith, sectarian, and ethnic hatred identification on social media in Urdu. *Information Processing & Management*, *60*(3), 103270.
4. Al-Hassan, A., & Al-Dossari, H. (2019, February). Identification of hate speech in social networks: a survey on multilingual corpus. In *6th international conference on computer science and information technology* (Vol. 10, pp. 10-5121).
5. Ali, R., Farooq, U., Arshad, U., Shahzad, W., & Beg, M. O. (2022). Hate speech identification on Twitter using transfer learning. *Computer Speech & Language*, *74*, 101365.
6. Aljarah, I., Habib, M., Hijazi, N., Faris, H., Qaddoura, R., Hammo, B., ... & Alfawareh, M. (2021). Intelligent identification of hate speech in Arabic social network: A machine learning approach. *Journal of Information Science, 47*(4), 483-501.
7. Alrehili, A. (2019, November). Automatic hate speech identification on social media: A brief survey. In *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)* (pp. 1-6). IEEE.
8. Alshalan, R., & Al-Khalifa, H. (2020). A deep learning approach for automatic hate speech identification in the saudi twittersphere. *Applied Sciences, 10*(23), 8614.
9. Aluru, S. S., Mathew, B., Saha, P., & Mukherjee, A. (2020). Deep learning models for multilingual hate speech identification. *arXiv preprint arXiv:2004.06465*.
10. Arango, A., Pérez, J., & Poblete, B. (2019, July). Hate speech identification is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval* (pp. 45-54).
11. Aziz, S., Sarfraz, M. S., Usman, M., Aftab, M. U., & Rauf, H. T. (2023). Geo-Spatial Mapping of Hate Speech Prediction in Roman Urdu. *Mathematics*, *11*(4), 969.
12. Bosco, C., Felice, D. O., Poletto, F., Sanguinetti, M., & Maurizio, T. (2018). Overview of the evalita 2018 hate speech identification task. In *Ceur workshop proceedings* (Vol. 2263, pp. 1-9). CEUR.
13. Chakravarthi, B. R., &Muralidaran, V. (2021, April). Findings of the shared task on hope speech identification for equality, diversity, and inclusion. In *Proceedings of the first workshop on language technology for equality, diversity and inclusion* (pp. 61-72).

14. Corazza, M., Menini, S., Cabrio, E., Tonelli, S., & Villata, S. (2020). A multilingual evaluation for online hate speech identification. *ACM Transactions on Internet Technology (TOIT)*, *20*(2), 1-22
15. Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial bias in hate speech and abusive language identification datasets. *arXiv preprint arXiv:1905.12516*.
16. Dewani, A., Memon, M. A., & Bhatti, S. (2021). Development of computational linguistic resources for automated identification of textual cyberbullying threats in Roman Urdu language. *3 c TIC: cuadernos de desarrollo aplicados a las TIC*, *10*(2), 101-121.
17. Dhanya, L. K., & Balakrishnan, K. (2021, June). Hate speech identification in Asian languages: a survey. In *2021 international conference on communication, control and information sciences (ICCISc)* (Vol. 1, pp. 1-5). IEEE.
18. Elzayady, H., Mohamed, M. S., Badran, K. M., & Salama, G. I. (2023). A hybrid approach based on personality traits for hate speech identification in Arabic social media. *International Journal of Electrical and Computer Engineering*, *13*(2), 1979.
19. Florio, K., Basile, V., Polignano, M., Basile, P., & Patti, V. (2020). Time of your hate: The challenge of time in hate speech identification on social media. *Applied Sciences*, *10*(12), 4180.
20. Fortuna, P., da Silva, J. R., Wanner, L., & Nunes, S. (2019, August). A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online* (pp. 94-104).
21. Fortuna, P., Soler, J., & Wanner, L. (2020, May). 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).
22. Gomez, R., Gibert, J., Gomez, L., & Karatzas, D. (2020). Exploring hate speech identification in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 1470-1478).
23. Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018, January). All you need is" love" evading hate speech identification. In *Proceedings of the 11th ACM workshop on artificial intelligence and security* (pp. 2-12).
24. Haq, N. U., Ullah, M., Khan, R., Ahmad, A., Almogren, A., Hayat, B., & Shafi, B. (2020). USAD: an intelligent system for slang and abusive text identification in PERSO-Arabic-scripted Urdu. *Complexity*, *2020*, 1-7.
25. Hussain, S., Malik, M. S. I., & Masood, N. (2022). Identification of offensive language in Urdu using semantic and embedding models. *PeerJ Computer Science*, *8*, e1169.
26. Ibrohim, M. O., & Budi, I. (2019, August). Multi-label hate speech and abusive language identification in Indonesian Twitter. In *Proceedings of the third workshop on abusive language online* (pp. 46-57).
27. Kapil, P., & Ekbal, A. (2020). A deep neural network based multi-task learning approach to hate speech identification. *Knowledge-Based Systems*, *210*, 106458.
28. Karim, M. R., Dey, S. K., Islam, T., Shajalal, M., & Chakravarthi, B. R. (2022, November). Multimodal hate speech identification from bengali memes and texts. In *International Conference on Speech and Language Technologies for Low-resource Languages* (pp. 293-308). Cham: Springer International Publishing.
29. Khan, S., Kamal, A., Fazil, M., Alshara, M. A., Sejwal, V. K., Alotaibi, R. M., ... & Alqahtani, S. (2022). HCovBi-caps: hate speech identification using convolutional and Bi-directional gated recurrent unit with Capsule network. *IEEE Access*, *10*, 7881-7894.
30. Lingiardi, V., Carone, N., Semeraro, G., Musto, C., D'Amico, M., & Brena, S. (2020). Mapping Twitter hate speech towards social and sexual minorities: A lexicon-based approach to semantic content analysis. *Behaviour & Information Technology*, *39*(7), 711-721.
31. MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech identification: Challenges and solutions. *PloS one*, *14*(8), e0221152.

32. Malik, J. S., Pang, G., & Hengel, A. V. D. (2022). Deep learning for hate speech identification: a comparative study. *arXiv preprint arXiv:2202.09517*.
33. Mandl, T., Modha, S., Kumar M, A., & Chakravarthi, B. R. (2020, December). Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation* (pp. 29-32).
34. Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019, June). Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science* (pp. 173-182).
35. Mehmood, K., Essam, D., Shafi, K., & Malik, M. K. (2020). An unsupervised lexical normalization for Roman Hindi and Urdu sentiment analysis. *Information Processing & Management*, *57*(6), 102368.
36. Mehta, H., & Passi, K. (2022). Social Media Hate Speech Identification Using Explainable Artificial Intelligence (XAI). *Algorithms*, *15*(8), 291.
37. Mohtaj, S., Schmitt, V., & Möller, S. (2022). A Feature Extraction based Model for Hate Speech Identification. *arXiv preprint arXiv:2201.04227*.
38. Mossie, Z., & Wang, J. H. (2020). Vulnerable community identification using hate speech identification on social media. *Information Processing & Management*, *57*(3), 102087.
39. Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). A BERT-based transfer learning approach for hate speech identification in online social media. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8* (pp. 928-940). Springer International Publishing.
40. Mubarak, H., Rashed, A., Darwish, K., Samih, Y., & Abdelali, A. (2020). Arabic offensive language on twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*.
41. Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., & Yeung, D. Y. (2019). Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049*.
42. Parihar, A. S., Thapa, S., & Mishra, S. (2021, June). Hate speech identification using natural language processing: Applications and challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 1302-1308). IEEE.
43. Pawar, A. B., Gawali, P., Gite, M., Jawale, M. A., & William, P. (2022, April). Challenges for hate speech recognition system: approach based on solution. In *2022 International conference on sustainable computing and data communication systems (ICSCDS)* (pp. 699-704). IEEE.
44. Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Identifying and monitoring hate speech in Twitter. *Sensors*, *19*(21), 4654.
45. Pitenis, Z., Zampieri, M., & Ranasinghe, T. (2020). Offensive language identification in Greek. *arXiv preprint arXiv:2003.07459*.
46. Röttger, P., Seelawi, H., Nozza, D., Talat, Z., & Vidgen, B. (2022). MULTILINGUAL HATECHECK: Functional Tests for Multilingual Hate Speech Identification Models. *arXiv preprint arXiv:2206.09917*.
47. Roy, P. K., Tripathy, A. K., Das, T. K., & Gao, X. Z. (2020). A framework for hate speech identification using deep convolutional neural network. *IEEE Access*, *8*, 204951-204962.
48. Ruwandika, N. D. T., & Weerasinghe, A. R. (2018, September). Identification of hate speech in social media. In *2018 18th international conference on advances in ICT for emerging regions (ICTer)* (pp. 273-278). IEEE.
49. Saeed, H. H., Ashraf, M. H., Kamiran, F., Karim, A., & Calders, T. (2021). Roman Urdu toxic comment classification. *Language Resources and Evaluation*, 1-26.
50. Saleh, H., Alhothali, A., & Moria, K. (2023). Identification of hate speech using BERT and hate speech word embedding with deep model. *Applied Artificial Intelligence*, *37*(1), 2166719.
51. Sandaruwan, H. M. S. T., Lorensuhewa, S. A. S., & Kalyani, M. A. L. (2019, September). Sinhala hate speech identification in social media using text mining and machine learning. In *2019 19th*

International Conference on Advances in ICT for Emerging Regions (ICTer) (Vol. 250, pp. 1-8). IEEE.

52. Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019, July). The risk of racial bias in hate speech identification. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1668-1678).
53. Satapara, S., Majumder, P., Mandl, T., Modha, S., Madhu, H., Ranasinghe, T., ... & Premasiri, D. (2022, December). Overview of the hasoc subtrack at fire 2022: Hate speech and offensive content identification in english and indo-aryan languages. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation* (pp. 4-7).
54. Sharma, A., Kabra, A., & Jain, M. (2022). Ceasing hate with moh: Hate speech identification in hindi–english code-switched language. *Information Processing & Management*, *59*(1), 102760.
55. Sutejo, T. L., & Lestari, D. P. (2018, November). Indonesia hate speech identification using deep learning. In *2018 International Conference on Asian Language Processing (IALP)* (pp. 39-43). IEEE.
56. Teh, P. L., Cheng, C. B., & Chee, W. M. (2018, March). Identifying and categorising profane words in hate speech. In *Proceedings of the 2nd International Conference on Compute and Data Analysis* (pp. 65-69).
57. Toraman, C., Şahinuç, F., & Yilmaz, E. H. (2022). Large-scale hate speech identification with cross-domain transfer. *arXiv preprint arXiv:2203.01111*.
58. Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British journal of management*, *14*(3), 207-222.
59. Ullah, F., Chen, X., Shah, S. B. H., Mahfoudh, S., Hassan, M. A., & Saeed, N. (2022). A novel approach for emotion identification and sentiment analysis for low resource Urdu language based on CNN-LSTM. *Electronics*, *11*(24), 4096.
60. Velankar, A., Patil, H., & Joshi, R. (2022, November). Mono vs multilingual bert for hate speech identification and text classification: A case study in marathi. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition* (pp. 121-128). Cham: Springer International Publishing.
61. Wang, C. C., Day, M. Y., & Wu, C. L. (2022). Political Hate Speech Identification and Lexicon Building: A Study in Taiwan. *IEEE Access*, *10*, 44337-44346.
62. William, P., Gade, R., esh Chaudhari, R., Pawar, A. B., & Jawale, M. A. (2022, April). Machine learning based automatic hate speech recognition system. In *2022 International conference on sustainable computing and data communication systems (ICSCDS)* (pp. 315-318). IEEE.
63. Xia, M., Field, A., & Tsvetkov, Y. (2020). Demoting racial bias in hate speech identification. *arXiv preprint arXiv:2005.12246*.
64. Zhang, Z., &Luo, L. (2019). Hate speech identification: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, *10*(5), 925-945.
65. Zhou, Y., Yang, Y., Liu, H., Liu, X., & Savage, N. (2020). Deep learning based fusion approach for hate speech identification. *IEEE Access*, *8*, 128923-128929.
66. Zimmerman, S., Kruschwitz, U., & Fox, C. (2018, May). Improving hate speech identification with deep learning ensembles. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018).*