International Journal of Engineering Applied Science and Management
ISSN (Online): 2582-6948
Vol. 4 Issue 8, August 2023

A Review on Hate Speech Detection on Social media using Machine learning

Rohit Kumar Srivastva[1], Chetan Agrawal[2], Rashi Yadav[3]
Dept. of CSE, Radharaman Institute of Technology & Science, Bhopal, India[1, 2, 3]
rohitbuxar03@gmail.com[1], chetan.agrawal12@gmail.com[2], rashi6yadav@gmail.com[3]

Abstract: *Hate speech is a crime that has been increasing in recent years, not only in person but also online. There are several causes for this. There is tremendous growth in social media that promotes full freedom of expression through anonymity features. Freedom of expression is a human right, but hate speech directed at individuals or groups on the basis of race, caste, religion, ethnicity or nationality, gender, disability, gender identity, etc. is a violation of that sovereignty. Freedom of expression is a human right, but hate speech directed at individuals or groups on the basis of race, caste, religion, ethnicity or nationality, gender, disability, gender identity, etc. is a violation of that sovereignty. It promotes violence and hate crimes, creates social imbalances, and undermines peace, trust and human rights. Revealing hate speech in social media discourse is a very important but complex task. On the one hand, the anonymity provided by the Internet, especially social networks, makes people more likely to engage in hostile behavior. On the other hand, the desire to express one's thoughts on the Internet has increased, leading to the spread of hate speech. Governments and social media platforms can benefit from detection and prevention technologies, as this kind of bigoted language can wreak havoc on society. As the internet expands, the proliferation of harmful content, including hate speech, presents considerable obstacles in ensuring a secure and inclusive online environment. In response to this challenge, researchers have embraced machine learning and deep learning methods to create automated systems that can effectively detect hate speech and conduct sentiment analysis, offering potential solutions to address this pressing issue. This survey article provides a comprehensive overview of recent advancements in hate speech detection and sentiment analysis using machine learning and deep learning models. We present an in-depth analysis of various methodologies and datasets employed in this domain. Additionally, we explore the unique challenges faced by these models in accurately identifying and classifying hate speech and sentiment in online text. Finally, we outline areas where more study is needed and suggest potential new avenues for exploration in the field of hate speech identification and sentiment analysis.*

Keywords: Hate Speech Classification, Text Mining, Sentiment analysis, Twitter, social media

1. Introduction

The advancement in internet technology and tremendous growth of users in online activities, and social media networks leads to the generation of an unprecedented volume of data. The data that users generate through their online activities, whether it is in the form of text, images, music, videos, log files, reviews, etc., is typically generated from a variety of sources, voluminous and includes structured as well as unstructured data. Performing and analyzing these types of unstructured and structured data has a greater impact on the big data field [1]. Such type of data can be analyzed for decision making using machine learning, data mining, web mining and text mining techniques. Also, since these types of data can be voluminous and extracting the patterns from this data are quite a difficult process. And, further, micro blogging services like Twitter, YouTube, Instagram, Facebook, Snapchat, WhatsApp, LinkedIn, blogs, Wikis etc., support a variety of data formats with/ without the proper grammatical rules and also short texts which are written without concerning the grammars [2]. Fig. 1 shows the percentage of users on social network platforms. From these platforms the amount of information (opinions) [3, 4] which is shared by the users can be used for analyzing the opinions about the products, political movements,

financial and political forecasting, monitoring the company strategies, marketing analysis, disseminating news, crime forecasting, product preferences, tracing the terrorist activities, e-health and e- tourism, monitoring reputations, detecting the hate speech in the public forms etc. To find meaningful information from the text (corpus) or data coming from public forums, Natural Language Processing (NLP) techniques is used [5].

The advent of social media and online forums has revolutionized the way people communicate and express their opinions. However, this newfound freedom of expression has also given rise to the proliferation of hate speech, cyber bullying, and offensive content, which can have severe implications on individuals and society as a whole. Identifying and curbing such harmful contents has become a critical task for maintaining a respectful and safe online space.

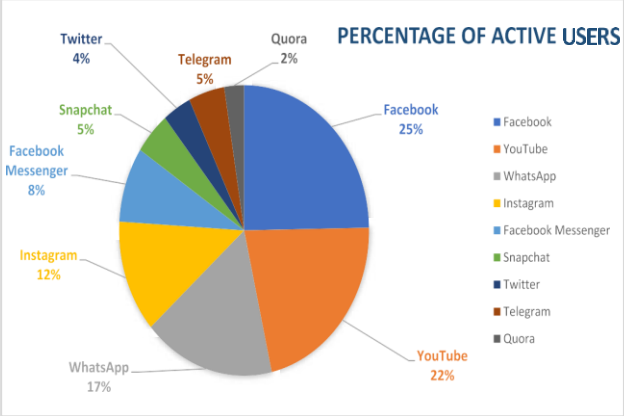

Fig. 1. Active users and their percentage in social networks

For instance, Modha et. al. [6] dealt with the identification of the aggression types of texts in the online platforms and divided the texts into aggressive and non– aggressive. Fig. 2 depicts the percentage of hate speech texts posted in Instagram during the four quarters of the years 2020 and 2021. Kaur et. al. [7] mentions the concepts of abusive content detection based on four categories of features namely, activity based, user based, context-based, and network-based features. This survey has also mentioned many parameters to identify the abusive contents such as posts per day, age, gender, etc and helps to build the researchers with fundamental concepts and key insight areas including the recent trends and techniques. The relationship between hate speech, aggressiveness and offensive speech is discussed in [8].

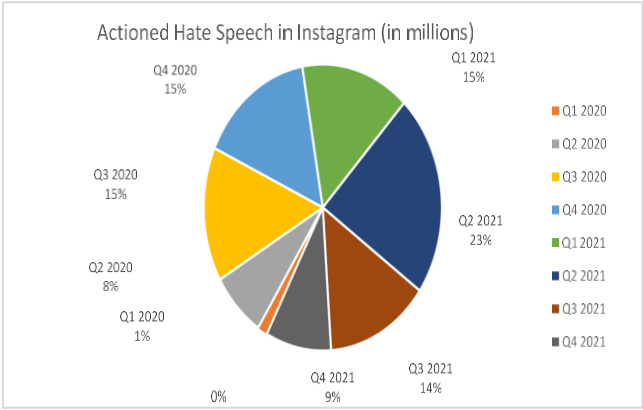

Fig. 2. Auctioned Hate Speech on Instagram from 2020 to 2021

Traditional rule-based methods for hate speech detection and sentiment analysis often lack the scalability and adaptability to handle the vast amount of user-generated content on social media platforms. In contrast, machine learning and deep learning techniques have shown promising results in automating the process of identifying hate language and analyzing sentiments expressed in text data. The primary objective of this survey is to present an in-depth analysis of hate speech detection and sentiment analysis techniques, focusing on the application of machine learning and deep learning models. By exploring the challenges faced by the present approaches, this paper aims to provide researchers with insights into the evolving landscape of hate speech detection and sentiment analysis.

Rest of the paper is organized as follows in section 2 we explained about Hast speech's background, the previous work done by various researchers are explained in section 3, various hate speech detection model are explained in section 4, section 5 presents different datasets of hate speech detection, section 6 presents challenges and issues in hate speech detection, lastly we conclude our paper in section 7.

2. Background

2.1. What is hate speech?

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

1. Code of Conduct between European Union Commission and companies: ''All conduct publicly inciting to violence or hatred directed against a group of

persons or a member of such a group defined by reference to race, color, religion, descent or national or ethnic" [9].

2. International minorities associations (ILGA): 'Hate crime is any form of crime targeting people because of their actual or perceived belonging to a particular group. The crimes can manifest in a variety of forms: physical and psychological intimidation, blackmail, property damage, aggression and violence, rape'.

3. Academia- Nobata et al. [10] defined HS as an act that attacks or demeans a group/individual based on race, ethnic origin, religion, disability, gender, age, disability, or sexual orientation/- gender identity. Similarly, Nockleby defined HS as ''any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic." Warner and Hirschberg [11] distinguished the occurrence of hate speech related wording with user's intention to harm an individual or a group. Otherhand, Waseem and Hovy (2016) [12] viewed hate speech in the form of racist and sexist remarks

4. Facebook: We define hate speech as a direct attack against people on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease. We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing, and call for exclusion or segregation. We consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants, and asylum seekers from the most severe attacks, though we do allow commentary and criticism of immigration policies. Similarly, we provide some protections for characteristics like occupation, when they're referenced along with a protected characteristic2."

5. Twitter: 'You may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease' 3. Examples from Twitter hate-speech are:

- "I'm glad this [violent event] happened. They got what they deserved [referring to persons with the attributes noted above]."
- " [Person with attributes noted above] are dogs" or" [person with attributes noted above] are like animals."

6. YouTube: 'We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/- gender, sexual orientation, victims of a major violent event and their kin, and veteran Status'.

3. Related Work

This section highlights the current survey and review articles, focusing on the importance of the contributions of this work. Some reviews and surveys were found about hate speech detection issues, such as by [13], and [14]. In the research conducted by [15], a systematic mechanism for reviewing existing works on hate speech detection from an informatics perspective was applied. It is considered the second survey on this topic after that of [16], which provided a short overview of hate speech detection within NLP. According to [16], the survey is relatively brief and primarily focuses on feature extraction. The survey by [9] provided a comparison of hate speech to other similar forms, a summary of statistics on detection methods, and a discussion of the terminologies needed to study hate speech, and the features involved in this domain. Later on, they concentrated on bullying research. They described several English datasets and existing challenges, involving different social media platforms, with less than 20 papers focusing on hate speech. In another study by [17], a more reliable, accurate, and comprehensive classification of anger-linked social media messages for detecting hate speech was established. This will help ensure proper classification because anger eventually leads to extensive participation in hate crimes. In another study by [18], the researchers attempted to review six various hate speech detection models on a variety of social media sites. The methods used were based on the NLP, data mining, machine learning domains, and the variations between these methods were discussed. In a further work by [19], a brief review was conducted on the use of state-of-the-art NLP techniques such as dictionaries, bag-of-words, and n-gram to automatically detect hate speech on online social media sites. Moreover, the study by [14] offered a review of methods for recognizing misogyny in social media, particularly on Twit- ter. The approaches included standard machine learning and deep learning methods. Furthermore, the findings considered different languages, including English. In a recent review article by [13], the authors employed machine learning techniques to categorize hate speech on Twitter, involving generic metadata designs, threshold configurations, and divergences. They also discussed the benefits and weaknesses of individual and integrated machine learning algorithms for the classification process. In addition, they displayed the hate speech benchmark dataset for testing the implementation of the classification paradigm. Even

though some surveys and reviews are available on this topic, significant limitations exist. These works partly lack SLR guidelines, up-to-date reviews, and survey studies. Furthermore, these studies are limited in that they did not focus completely on Twitter and, more specifically, on the English language, unlike the survey by [20], which examined the available benchmark datasets used for abusive language and hate speech detection on different social media sites. Their analysis involved the dataset development process, the themes of interest, language coverage, and annotation framework. Although many existing works on hate speech are based on Twitter, previous surveys or reviews lack comprehensive coverage of this particular social site. Twitter ranks among the most frequently used social networks for the automated identification of hate speech in texts [21]. Hence, Twitter has improved connectivity among people worldwide and is a convenient public forum for users. Compared to earlier studies, this paper reviewed a substantially larger number of papers. Additionally, [22] presented a short review of English and non-English literature with some challenges and future research directions. Reference [23] conducted a survey to illustrate the generalizability of current hate speech detection models and explain how hate speech algorithms have an issue in generalizing. Research directions for improving generalization in hate speech detection are discussed.

Reference [24] offered an overview of machine learning techniques and techniques for detecting hate speech in online social networks. They explored the primary constituents of hate speech classification using ML algorithms. The failure and capability of each approach are assessed to identify the study gaps and specify the open challenges. Reference [25] discussed different definitions of hate speech, and several challenges were presented concerning data collection and annotation. The authors briefly discussed the differences between nine datasets that used different text languages and platforms. The sources of metadata and the feature selection are also described briefly based on five previous works using machine learning methods. In their paper, a multiple-view SVM model was developed to classify hate speech using three datasets from three different platforms, an interpretation of the model, and an analysis of errors reported. Accordingly, they raise some general challenges. Reference [26] surveyed the racist and sexist class of hate speech methods, focusing on a few factors: data sources, features used, and algorithms of ML. They offered brief descriptions of the text corpus, presented some of the most frequently used approaches for representing features, and made a short comparison between ML models. A short systematic review of the literature was provided by [27], which included articles released earlier than January 2020. Only studies published in English and Indonesian for conferences and journals were considered in their SLR study. A variety of data sources were considered, namely comments from Twitter, Facebook, Wikipedia, Instagram, Online Today, YouTube, and Yahoo. There is only a small finding and a small suggestion by their SLR. More recently, [28] performed a systematic review of text-based hate speech detection methods and mainly focused on the essential datasets with text-based features and machine learning algorithms. Their collected articles were reviewed according to different themes. They provided three challenge groups and three direction points. Even though their review focused on an English hate speech dataset, our SLR differs from their review in that it provides a more detailed analysis and some taxonomy for our selected studies from a different standpoint.

4. Models for Hate Speech Detection

The anonymity of social networks attracts hate speech, which presents a problem for the entire world, to hide their unlawful online behavior. Detecting hate speech is crucial given the growing volume of social media data since it can have negative impacts on society [44]. The most recent machine learning algorithms for detecting hate speech are covered in the discussion that follows.

4.1 Classical Machine Learning methods

The term “shallow detection” refers to word encoding techniques used by classical word representation hate speech detectors. After that, shallow classifiers can be used to perform the classification. The tagged dataset is used to train the learning algorithms, resulting in a model that can be used to detect and classify hate speech and non-hate speech in texts. Two examples of feature representation strategies that can be applied are TF-IDF and N- grams. Traditionally, supervised machine learning methods like Naive Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), Linear Regression, and Logistic Regression (LR) have been used to detect hope speech and sentiment analysis.

4.2 Ensemble approach

The ensemble technique was developed to overcome the limitations of several individual machine learning algorithms while enhancing their strengths. Each model has its own set of flaws; thus, no model is ideal. But, ensemble approaches attempt to combine the benefits of multiple models to provide better performance than any single model can provide. Combining two or more machine learning algorithms can minimize variance and increase learning capacity greatly, according to statistics.

Bagging methodology, Random Forest (RF), and boosting method are some of the ensemble techniques.

4.3 Word-embeddings based methods

Word embedding learns the vectorized representations from scattered representations, which are then employed in downstream text mining activities. The embeddings make it possible for semantically related phrases to share the same vector representation. Many word embedding algorithms have been developed over the years, including Glove, word2vec, and FastText. The representations from the word embedding techniques are fed into various classifiers.

Deep learning model for hate speech detection and sentiment analysis Deep learning introduces a multi-layer structure in the neural net- work's hidden layers, enabling it to attain more intricate outcomes. Unlike conventional machine learning methods where features are manually specified or obtained through feature selection techniques such as TF-IDF, Word2Vec etc., and deep learning models autonomously learn and extract information, resulting in enhanced accuracy and overall performance. To predict and categorize hate speech and sentimental texts, deep learning techniques have been utilized in a range of studies in the fields of data mining and text classification. Below, we present a summary of the deep learning models used for sentiment analysis and hate speech detection.

4.4 Recurrent neural networks (RNNs)

RNN is a subclass of artificial neural networks and assesses time series or sequential data. The sole purpose of common feed forward neural networks is to process unrelated pieces of data. If, however, we have data in a sequence where one data point depends on the data point before it, we will need to adjust the neural network to account for these dependencies. RNNs can remember the states or specifics of previous inputs to use when constructing sub- sequent outputs. RNNs are well-suited for tasks like sentiment analysis and hate speech detection where the context and order of words in a text are crucial for making accurate predictions.

Long Short-Term Memory (LSTM) is a type of RNN that addresses the vanishing gradient problem, making it more effective in capturing long- range dependencies in sequential data. LSTM can be used for sentiment analysis and hate speech detection like standard RNNs, but with the advantage of handling longer texts and preserving context over longer sequences. LSTM can effectively capture the sequential dependencies between words, allowing it to understand the context and sentiment expressed in the sentence. For hate speech detection, LSTM works similarly to sentiment analysis. The LSTM processes the input text word by word and updates its hidden state at each time step, capturing the contextual information and dependencies between words. To enhance the performance of LSTM for hate speech detection, additional techniques like attention mechanisms are also incorporated. Attention mechanisms allow the model to focus on specific parts of the text that are more indicative of hate speech, leading to improved accuracy.

Gated Recurrent Unit (GRU) is another type of RNN that, like LSTM, addresses the vanishing gradient problem and captures long-range dependencies in sequential data. GRU is used for sentiment analysis and hate speech detection like LSTM and standard RNNs. To perform sentiment analysis using GRU, the model processes the input sentence word by word, updating its hidden state at each time step. GRU can effectively capture the sequential dependencies between words, allowing it to understand the context and sentiments expressed in the sentence. For hate speech detection, GRU works similarly to sentiment analysis. The GRU processes the input text word by word and updates its hidden state at each time step, capturing the contextual information and dependencies between words.

4.5 Convolution neural networks (CNNs)

CNNs are powerful deep learning models that have been widely used for various computer vision tasks, such as image classification and object detection. But in recent times, CNNs are also adapted for NLP tasks, including sentiment analysis and hate speech detection. CNNs are used for sentiment analysis by treating the text as a one-dimensional vector and applying 1D convolution to capture local patterns and features within the text. As in sentiment analysis, CNNs are also used for hate speech detection. In such cases, the convolutional layer applies filters to the sequences, capturing local patterns and features in the text. In both sentiment analysis and hate speech detection, CNNs excel at capturing local patterns and features from text data, making them effective tools for various NLP tasks. Fig. 3 shows the general framework used for detecting hate speech and sentiment analysis.

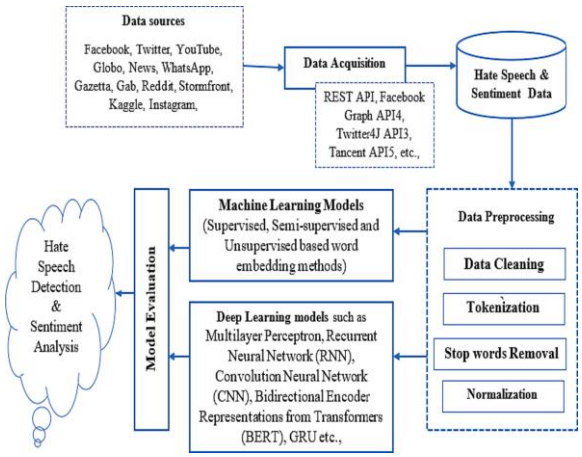

Fig. 3. General Approach for Hate Speech Detection and Sentiment Analysis

5. Hate Speech datasets

In total, we collected 13 datasets related to hate speech in social media. The datasets selected are diverse both in content, different kind of hate speech, and in a temporal aspect.

- **Measuring hate speech (MHS):** MHS [29] consists of 39,565 social media (YouTube, Reddit, Twitter) manually annotated comments. The coders were asked to annotate each entry on 10 different at- tributes such as the presence of sentiment, respect, insults and others; and also indicate the target of the comment (e.g. age, disability). They use Rasch measurement theory to aggregate the annotators' rating in a continuous value that indicates the hate score of the comment.
- **Call me sexist, but (CMS)**: This dataset of 6,325 entries [30] focuses on the aspect of sexism and includes social psychology scales and tweets extracted by utilizing the "Call me sexist, but" phrase. The authors also include two other sexism datasets which they re-annotate. Each entry is annotated by five coders and is labeled based on its content (e.g. sexist, maybe-sexist) and phrasing (e.g. civil, uncivil).
- **Hate Towards the Political Opponent (HTPO)**: HTPO [31] is a collection of 3,000 tweets related to the 2020 USA presidential election. The tweets were extracted using a set of keywords linked to the presidential and vice presidential candidates and each tweet is an- notated for stance detection (in favor of/against the candidate) and whether it contains hateful language or not.
- **HateX**: HateX [32] is a collection of 20,148 posts from Twitter and Gab extracted by utilizing relevant hate lexicons. For each entry, three annotators are asked to indicate: (1) the existence of hate speech, offensive speech, or neither of them, (2) the target group of the post (e.g. Arab, Homosexual), and (3) the reasons for the label as- signed.
- **Offense:** The Offense dataset [33] contains 14,100 tweets extracted by utilizing a set of keywords and categorizes them in three levels: (1) offensive and non-offensive; (2) targeted/untargeted insult; (3) targeted to individual, group, or other.
- **Automated Hate Speech Detection (AHSD):** In this dataset, [34] the authors utilize a set of keywords to extract 24,783 tweets which are manually labeled as either hate speech, offensive but not hate speech, or neither offensive nor hate speech.
- **Hateful Symbols or Hateful People? (HSHP):** This is a collection [35] of 16,000 tweets extracted based on keywords related to sexism and racism. The tweets are annotated as on whether they contain racism, sexism or neither of them by three different annotators.
- **Are You a Racist or Am I Seeing Things? (AYR):** This dataset [36] is an extension of Hateful Symbols or Hateful People? And adds the "both" (sexism and racism) as a potential label. Overlapping tweets were not considered.
- **Multilingual and Multi-Aspect Hate Speech Analysis (MMHS)**: MMHS [37] contains hateful tweets in three different languages (English, French, Arabic). Each tweet has been labeled by three annotators on five different levels: (1) directness, (2) hostility (e.g. abusive, hateful), (3) target (e.g. origin, gender), (4) group (e.g. women, individual) and (5) annotator emotion (disgust, shock, etc). A total of 5,647 tweets are included in the dataset.
- **HatE**: HatE [38] consists of English and Spanish tweets (19,600 in total) that are labeled on whether they contain hate speech or not. The tweets in this dataset focus on hate speech to- wards two groups: (1) immigrants and (2) women.

- **HASOC**: This dataset [39] contains 17,657 tweets in Hindi, German and English which are annotated on three levels: (1) whether they contain hate-offensive content or not; (2) in the case of hate-offensive tweets, whether a post contains hate, offensive, or profane content/words; (3) on the nature of the insult (targeted or un-targeted).
- **Detecting East Asian Prejudice on Social Media (DEAP):** This is a collection of 20,000 tweets [40] focused on East Asian prejudice, e.g. Sinophobia, in relation to the COVID-19 pandemic. The annotators were asked to labeled each entry based on five different categories (hostility, criticism, counter speech, discussion, non- related) and also indicate the target of the entry (e.g. Hong Kongers, China).
- **Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior (LSC)**: The dataset [41] consists of 80,000 tweets extracted using a boosted random sample technique. Each tweet is labeled as either offensive, abusive, hateful, aggressive, cyber bullying or normal.

6. Challenges and Issues of Hate Speech Detection

Hate speech detection and sentiment analysis are essential NLP tasks, but they come with various challenges and issues. Some of the main challenges and issues include:

(1) Usually, social media messages contain poorly written texts which do not reside in the formal structure to find out the pat- terns in the text.

(2) Developing high-quality labeled datasets for hate speech and sentiment analysis, especially in languages other than English, can be time-consuming and expensive. The scarcity of diverse and annotated data limits the performance of models, particularly for low-resource languages.

(3) Extending hate speech detection and sentiment analysis to mul- tiple languages introduces language-specific complexities, including varying grammatical structures, sentiment lexicons, and cultural expressions.

(4) In hate speech identification and sentiment analysis, the data distribution and imbalance nature are one of the issues for finding a meaningful pattern in the data.

(5) Hate speech and sentiment expression can be highly subjective and context-dependent. What may be considered hateful in one context and may not be so in another. Detecting hate speech accurately requires understanding the context and cultural nuances, which can be challenging for algorithms. For example, the sentence: “I’m dying to meet you!” in some English-speaking regions, this phrase might be interpreted as a positive expression of eagerness or excitement to meet someone. However, in other places, the use of “dying” might be considered inappropriate or negative due to the literal meaning of the word.

(6) The interpretation of implicit hate speeches heavily depends on the context in which they are used. Without a proper under- standing of the context, it can be challenging to distinguish between harmful and innocuous statements.

7. Conclusion

In recent years, the increasing use of social media has led to highly unacceptable phenomena, such as hate speech language and hate speech-based incidents. Despite ongoing studies aimed at solving the issue of the proliferation of hate speech, there are still challenges in establishing a competent solution for content generated by users. The aim of the current study is to contribute to the existing survey and review papers to advance the investigation in the concerned field. Various aspects can be derived from the selected studies, including the datasets and their categories, the most used machine learning techniques, the performance metrics involved, and the validation methods applied. Moreover, a critical search was carried out on the selected documents that characterized and specified the challenges and recommendations linked to hate speech detection. A potential future study has been recommended to address the issues in previous research. Some of these issues refer to the lack of agreement and bias in data annotations, noisy user-generated posts, small training data, imbalanced data issues, lack of sufficient feature representations, generalization, appropriate user imbalanced data issues, lack of sufficient feature representations, generalization, appropriate user features, and hyper-parameter tuning. Furthermore, it would be interesting to consider the hate speech issue in languages other than English or on other social network sites. The present study analyzed the views in the published papers and provided researchers with a useful reference. This research is essential for additional studies. The research society can further work and concentrate on advanced methods for hate speech detection missions.

References

[1] A. Hande, R. Priyadharshini, B.R. Chakravarthi, KanCMD: Kannada CodeMixed dataset for sentiment analysis and offensive language detection, in: Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media. 2020.

[2] J. Cao, et al., A risky large group emergency decision-making method based on topic sentiment analysis, Expert Systems with Applications 195 (2022), 116527.

[3] P.K. Roy, et al., A framework for hate speech detection using deep convolutional neural network, IEEE Access 8 (2020) 204951–204962.

[4] H. Liu, et al., A fuzzy approach to text classification with two-stage training for ambiguous instances, IEEE Transactions on Computational Social Systems 6 (2) (2019) 227–240.

[5] F.M. Plaza-Del-Arco, et al., A multi-task learning approach to hate speech detection leveraging sentiment analysis, IEEE Access 9 (2021) 112478–112489.

[6] S. Modha, et al., Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance, Expert Systems with Applications 161 (2020), 113725.

[7] S. Kaur, S. Singh, S. Kaushal, Abusive content detection in online user-generated data: a survey, Procedia Computer Science 189 (2021) 274–281.

[8] F. Poletto, et al., Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources and Evaluation 55 (2021) 477–523.

[9] Wigand, C., Voin, M., 2017. Speech by commissioner jourová–10 years of the eu fundamental rights agency: A call to action in defence of fundamental rights, democracy and the rule of law.

[10] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y., 2016. Abusive language detection in online user content, in: Proceedings of the 25th international conference on world wide web, pp. 145–153.

[11] Warner, W., Hirschberg, J., 2012. Detecting hate speech on the world wide web, in: Proceedings of the second workshop on language in social media, Association for Computational Linguistics. pp. 19–26.

[12] Waseem, Z., Hovy, D., 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL student research workshop, pp. 88–93."

[13] F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga, ''Machine learning techniques for hate speech classification of Twitter data: Stateof- the-art, future challenges and research directions,'' Comput. Sci. Rev., vol. 38, Nov. 2020, Art. no. 100311.

[14] E. Shushkevich and J. Cardiff, ''Automatic misogyny detection in social media: A survey,'' Computación Y Sistemas, vol. 23, no. 4, pp. 1159–1164, Dec. 2019

[15] P. Fortuna and S. Nunes, ''A survey on automatic detection of hate speech in text,'' ACM Comput. Surv., vol. 51, no. 4, pp. 1–30, 2018.

[16] A. Schmidt and M. Wiegand, ''A survey on hate speech detection using natural language processing,'' in Proc. 5th Int. Workshop Natural Lang. Process. Social Media, 2017, pp. 1–10.

[17] J. Langham and K. Gosha, ''The classification of aggressive dialogue in social media platforms,'' in Proc. ACM SIGMIS Conf. Comput. People Res., Jun. 2018, pp. 60–63.

[18] S. Modi, ''AHTDT—Automatic hate text detection techniques in social media,'' in Proc. Int. Conf. Circuits Syst. Digit. Enterprise Technol. (ICCS- DET), Dec. 2018, pp. 1–3.

[19] A. Alrehili, ''Automatic hate speech detection on social media: A brief survey,'' in Proc. IEEE/ACS 16th Int. Conf. Comput. Syst. Appl. (AICCSA), Nov. 2019, pp. 1–6.

[20] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, ''Resources and benchmark corpora for hate speech detection: A systematic review,'' Lang. Resour. Eval., vol. 55, pp. 477–523, Jun. 2020.

[21] A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, ''A unified deep learning architecture for abuse detection,'' in Proc. 10th ACM Conf. Web Sci., Jun. 2019, pp. 105–114.

[22] T. X. Moy, M. Raheem, and R. Logeswaran, ''Hate speech detection in English and non-English languages: A review of techniques and challenges,'' Webology, vol. 18, no. 5, pp. 929–938, Oct. 2021, doi: 10.14704/WEB/V18SI05/WEB18272.

[23] W. Yin and A. Zubiaga, ''Towards generalisable hate speech detection: A review on obstacles and solutions,'' PeerJ Comput. Sci., vol. 7, pp. 1–38, Jun. 2021, doi: 10.7717/PEERJ-CS.598.

[24] N. S. Mullah and W. M. N. W. Zainon, ''Advances in machine learning algorithms for hate speech detection in social media: A review,'' IEEE Access, vol. 9, pp. 88364–88376, 2021, doi: 10.1109/ACCESS.2021.3089515.

[25] O. Istaiteh, R. Al-Omoush, and S. Tedmori, ''Racist and sexist hate speech detection: Literature review,'' in Proc. Int. Conf. Intell. Data Sci. Technol. Appl. (IDSTA), Oct. 2020, pp. 95–99, doi: 10.1109/IDSTA50958.2020.9264052.

[26] R. Rini, E. Utami, and A. D. Hartanto, ''Systematic literature review of hate speech detection with text mining,'' in Proc. 2nd Int. Conf. Cybern. Intell. Syst. (ICORIS), Oct. 2020, pp. 1–6, doi: 10.1109/ICORIS50180.2020.9320755.

[27] S. MacAvaney, H. R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, ''Hate speech detection: Challenges and solutions,'' PLoS ONE, vol. 14, no. 8,

pp. 1–16, 2019, doi: 10.1371/journal.pone.0221152.

[28] F. Alkomah and X. Ma, ‘‘A literature review of textual hate speech detec- tion methods and datasets,’’ Information, vol. 13, no. 6, p. 122, 2022, doi: 10.3390/info13060273.

[29] Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. “call me sexist, but...”: Revisiting sexism detection using psychological scales and adversarial samples. In Proceedings of the International AAAI Conference on Web and Social Media, volume 15, pages 573–584.

[30] Naiyer, Vaseem, Jitendra Sheetlani, and Harsh Pratap Singh. "Software Quality Prediction Using Machine Learning Application." Smart Intelligent Computing and Applications: Proceedings of the Third International Conference on Smart Computing and Informatics, Volume 2. Springer Singapore, 2020.

[31] Pasha, Shaik Imran, and Harsh Pratap Singh. "A Novel Model Proposal Using Association Rule Based Data Mining Techniques for Indian Stock Market Analysis." Annals of the Romanian Society for Cell Biology (2021): 9394-9399.

[32] Md, Abdul Rasool, Harsh Pratap Singh, and K. Nagi Reddy. "Data Mining Approaches to Identify Spontaneous Homeopathic Syndrome Treatment." Annals of the Romanian Society for Cell Biology (2021): 3275-3286.

[33] Vijay Vasanth, A., et al. "Context-aware spectrum sharing and allocation for multiuser-based 5G cellular networks." Wireless Communications and Mobile Computing 2022 (2022).

[34] Singh, Harsh Pratap, and Rashmi Singh. "Exposure and Avoidance Mechanism Of Black Hole And Jamming Attack In Mobile Ad Hoc Network." International Journal of Computer Science, Engineering and Information Technology 7.1 (2017): 14-22.

[35] Singh, Harsh Pratap, et al. "Design and Implementation of an Algorithm for Mitigating the Congestion in Mobile Ad Hoc Network." International Journal on Emerging Technologies 10.3 (2019): 472-479.

[36] Singh, Harsh Pratap, et al. "Congestion Control in Mobile Ad Hoc Network: A Literature Survey."

[37] Rashmi et al.. "Exposure and Avoidance Mechanism Of Black Hole And Jamming Attack In Mobile Ad Hoc Network." International Journal of Computer Science, Engineering and Information Technology 7.1 (2017): 14-22.

[38] Sharma et al., "Guard against cooperative black hole attack in Mobile Ad-Hoc Network." Harsh Pratap Singh et al./International Journal of Engineering Science and Technology (IJEST) (2011).

[39] Singh, et al., "A mechanism for discovery and prevention of coopeartive black hole attack in mobile ad hoc network using AODV protocol." 2014 International Conference on Electronics and Communication Systems (ICECS). IEEE, 2014.

[40] Harsh et al., "Design and Implementation of an Algorithm for Mitigating the Congestion in Mobile Ad Hoc Network." International Journal on Emerging Technologies 10.3 (2019): 472-479.

[41] Lara Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 171–180, Online. Association for Computational Linguistics.