Review   IJMCT

Online Hate Speech on Social Media Platforms

Anushka Sharma[1], Ghanshyam Prasad Dubey[2,*], Avishi Shrivastava[1], Mohini Sharma[1], Ananya Likhar[1], Lokendra Sharma[3]

Abstract
Social media is an online interactive digital platform which allows large number of group of people to connect with each other virtually, to create content and share their thoughts, information, ideas, and much more in the form of text, pictures, and videos. But sometime people share their negative view, and their message is turned into hate speech. Hate speech is of two types: online and offline. Online hate speech spreads faster than offline hate speech, spreading on platforms like Instagram, Facebook, X (formerly Twitter), etc. because when people share something, then it gets shared by their followers or friends and it reach out to millions of people and it is difficult to monitor it. This leads to rise of conflict among them based on the gender, caste, religion, culture, and country. This paper will discuss the spread of hate speech on social media platform and its identification.

Keywords: Social media, hate speech, online platform, identification, cyberbullying, recurrent neural network (RNN)

INTRODUCTION

Through social media platforms, the internet has changed the way we communicate and interact with one another. It has enabled or given us freedom of speech to rapidly express and share our thoughts, ideas, and opinions with everyone. Social media has allowed more people to speak out, but some people use it to hurt them. This freedom of speech, however, has given rise to a darker side of internet interaction, namely hate communication. Discriminatory speech, hate speech, incitement to hatred, incitement to terrorism, and incitement to genocide are all examples of hate communications [1]. Hate speech is defined as any language or content that promotes hatred, toward a certain group of people based on their race, religion, gender, ethnicity, or other protected characteristics [2]. It is spreading on every social media platform very rapidly. It takes place on comment section or during live chats of any platform [3]. Different social media platforms have different policies against hate speech [4].

According to the research done in the United Kingdom, the United States, Finland, and Germany, roughly 50% of 15- to 30-year-olds are exposed to hostile material online, particularly on social networking sites [5]. Hate websites frequently employ a variety of techniques to attract and retain their audience, they make fantasies which seem real [6]. They use abusive language to spread negativity and hate that they want to spread among people [6].

TYPES OF ONLINE HATE
Terrorism

Because of the cruel and hurtful things that individuals post online, terrorism is on the rise. It is related to public reactions [7, 8]. For the purpose of propagandizing and glorification of violent crimes,

terrorist groups produce and distribute propaganda movies, pictures, and statements. They occasionally use social media for this purpose. Some people share these negative things on social media sites like Facebook, X (formerly Twitter), and many websites to get more people to join them. Social media has been used by terrorists to spread hate speech. Due to this people resort to crime or violence. By eliminating extremist information, keeping an eye out for suspicious activity, and working with government organizations, social media corporations and law enforcement organizations are thwarting terrorist actions. To stop extreme ideology and possible violence, it is imperative to strike a balance between security, privacy, and freedom of speech [7, 8].

Religion

People who are ignorant of a specific religion or who have unfavorable opinions about it may grow hostile against those who follow it. People targeting other religions, insulting them, using negative and abusive words and making negative comments based on their religious beliefs is a religious hate speech which spreads rapidly on social media [9, 10].

The most targeted religion on social media is Islam (Muslim) [11]. The way that Muslims are portrayed in the media, historical developments, and ongoing conflicts often result Muslims being frequently the subject of excessive attacks [11]. According to official government statistics, Muslims and Jews made up most of the victims of religious hate crimes in England and Wales in the year ending March 2022 [12, 13].

According to a data released by the Home Office, there were 155,841 hate crimes overall in 2022, a 26% increase from 2021 [13, 14].

Politics

Social media expressions of extreme hatred, threats, or misinformation directed towards people or organizations that have different political opinions are referred to as political hate. Some people may make threats, send offensive comments, or launch internet harassment campaigns in an attempt to harm those who hold different political opinions. Politicians who are the targets of hate speech are criticized for their political beliefs, deeds, or other characteristics using cruel words or demeaning remarks. When Donald Trump was president, for instance, there were instances of hate speech when people used derogatory words to oppose his actions or policies [14].

Cyberbullying

Cyberbullying, a form of harassment, occurs through social networking sites, live messaging, harmful posts, comments, online games, texts, and pictures sent to mobiles [15–17]. With the boom of social media cyberbullying is increasing rapidly. In order to ruin someone's reputation or well-being, it frequently entails sending cruel messages, spreading untrue information, or posting humiliating content. Extreme mental health conditions, such as depression, anxiety, and, in some cases, suicidal thoughts, can be exacerbated by the emotional and psychological misery brought on by ongoing online harassment, humiliation in front of others, and the sense of powerlessness [18].

ANONYMITY

Through the use of usernames or pseudonyms, social media anonymity enables individuals to participate in online debates and content sharing without disclosing their actual identities, offering privacy and safety [19]. Individuals can feel free to discuss their ideas, beliefs, and personal experiences when they are not concerned about being identified [20, 21]. For internet users, the ability to remain anonymous online is a vital and powerful right. For millions of people globally, it aids in preserving freedom of speech and expression as well as safety. The protection given by anonymity might give some people the confidence to participate in cyberbullying, trolling, and hate speech since they feel less responsible for their activities [22]. Users frequently have the choice to preserve some sort of anonymity

when participating in online activities on social media sites like X (formerly Twitter), Facebook, Instagram, and YouTube. The user has anonymity to speak which leads to hate speech sometimes. For instance, during the 2016 U.S. elections, the Russia-based Internet Research Agency created fictitious social media profiles and organizations to remark on contentious social issues like immigration and race.

But it is technical because when users want to take the advantage of platform, most of their accounts are linked with email account, mobile number, and address through which they can be tracked easily if they spread some non-violent content, hate speech, cyberbullying on social media platform. Anonymity requires high technical skill to hide identity [23] as shown in Table 1.

ONLINE HATE SPEECH EFFECTS (PSYCHOLOGICAL)

The mental health of those who are the targets of hate speech might suffer from severe anxiety and depression, alcohol consumption, smoking, negative thinking, low self-esteem, helplessness, and social isolation. It has a longer effect on their health like stress, anxiety, and depression [25]. In a study, victims went through a severe state of terror because they think that threats they have received online may manifest offline at any time; victims of this kind of abuse are always concerned [26]. In terms of psychology, online hate speech on social media affects the mental health of human beings and they may even commit suicide due to hate speech. The negative comments on targeted users make them helpless, useless, and create a loss of faith as shown in Table 2. With the help of this table, we can analyze the hate speech in different hate categories. Understanding and combating hate speech, prejudice, and discrimination in society depends on the classification of hate targets [24].

DISTINCT SOCIAL MEDIA POLICIES TO DEAL WITH HATE SPEECH
X (Formerly Twitter)

Freedom of speech, but not freedom of reach, is the X (formerly Twitter) policy of 2022. X will not generate any advertising or other revenue from negatives or hate tweets Similar to the rest of the internet, you cannot find the tweet unless you specifically search it. It is entirely prohibited to promote violence, sensitive topics, and hate speech. It will remove negative content and stop promoting hate speech [4].

Table 1. Percentage of anonymous Tweets with their categories through different anonymous accounts on behavior, race, physical, and other categories [24].

Categories of Anonymity	Anonymously Posted Tweets
Behavior	46%
Race	55%
Random tweets	40%
Physical	49%
Others	46%

Table 2. How hate targets are categorized into specific hate categories.

Hate Categories	Hate Targets
Class	Poor people, rich people
Religious	Hindu, Muslim, Sikh
Gender	Male, female, transgender
Color	Black people, white people
Physical	Short people, tall people, dumb people, beautiful people
Disability	Blind people, deaf people
Ethnicity	Indian, Japanese, Chinese
Other	Drunk, etc.

Meta

Meta has a policy to remove content including hate speech, harassment, violence threats, and other content that may silence or hurt others. Meta allows you to control comments on your post. In order to choose what to read or share, Meta puts warnings on postings that are flagged as false. [4]. Facebook, Instagram, ad Threads are now a Meta company and Facebook and Instagram have a policy against hate speech that promotes hatred based on factors such as race, religion, handicap, gender, age, veteran status, and sexual orientation as hate. Facebook has provided few avenues for reporting instances of policy abuse. There are many options to deal with the individual who posts hate speech, unfriend them, take them off our list of friends, prevent them from contacting you, and report them.

YouTube

YouTube has a user reporting system in place that allows viewers to submit reports of objectionable content in an anonymous manner. If someone is using hate content and posting or spreading it against the terms of conditions of YouTube, then YouTube has a right to cancel or terminate the account. Users on YouTube have limited alternatives for reporting videos that they believe violates their hate speech policy. With regard to specific content, there are two options: report abuse or flag the specific video [4].

IDENTIFYING HATE SPEECH ON SOCIAL MEDIA

An analysis of the increase of hate speech online, funded by UNESCO, shows that social media sites like Facebook and X (formerly Twitter) have mostly adopted defensive strategies as shown in Figure 1. To gain a more nuanced grasp of the problem, they may leverage their huge data collection and actual occurrences to better comprehend the dynamics of online hate speech [23].

This data has been taken by Google's website, which shows an annual increase in hate speech crime, emphasizing the importance of identifying hate speech to take action against individuals spreading hate on social media platforms.

Few research studies have attempted to analyze how hate speech spreads online, while several prediction models have attempted to forecast the likelihood that hate speech will be retweeted. These models, however, are trained on offensive material and do not offer any justifications for the distinction between offensive and neutral information [27, 28].

Identification or detection of hate speech is difficult because it is expressed in many forms like in languages, in sarcasm way, languages have different meaning at different places Due to this variability, it is challenging to create a universal algorithm that can correctly classify all instances of hate speech [27, 28].
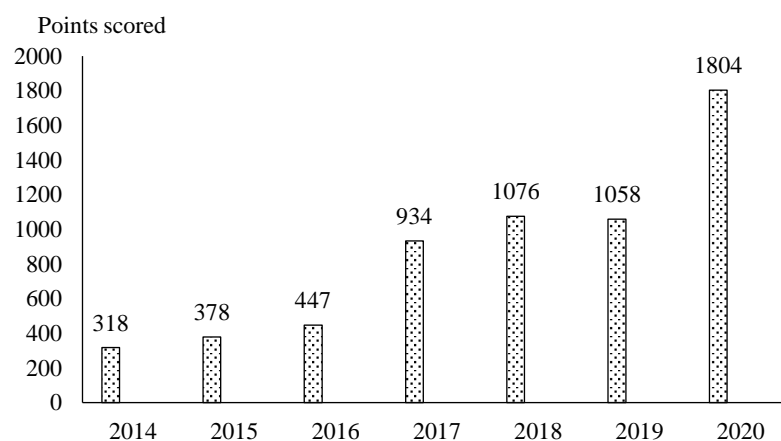

Figure 1. Cases of online hate speech in India.

Following are a few technical methods to detect hate speech content on social media platform.

Natural Language Processing

Natural language processing (NLP) is a powerful artificial intelligence tool that uses linguistic analysis and machine learning algorithms to identify hate speech on social media platforms. It makes it possible for computers to comprehend and analyze human language, making it a viable tool for detecting hate speech [29].

NLP Approaches to detect hate speech

Data collection and preprocessing of data

In this, first collect the dataset of hateful or non-hateful content from different dataset sources like Kaggle, Google, Amazon, Microsoft, etc. and preprocess this data before being fed to the classifiers (model) to remove unwanted noise, missing value, slang words etc. Data processing makes our data suitable for the task and improve the quality and accuracy of data to obtain a best result [30].

Tokenization and Feature extraction

In this we break sentences into words or tokens; the process is called tokenization.

Feature Extraction

It converts these words (tokenized text) into machine language or binary number because machine does not understand any language it only understands language in binary form. To capture the semantic meaning of words, methods like TF-IDF (term frequency–inverse document frequency) and word embeddings (such as Word2Vec, GloVe) are frequently utilized [30].

Model Building (Classifiers)

Classifiers

To build a machine learning model for detection of hate speech, first we have to choose classifiers like recurrent neural networks (RNNs), convolutional neural networks (CNNs), support vector machines (SVMs), decision tree, or many others to fed data and which will give best result [30].

Recurrent Neural Networks

RNN is an artificial neural network form and it is biased model. It works on sequential data and is used in speech and text analysis as shown in Figure 2. RNNs are also called feedback networks because they need previous data to predict the next one. They create a loop that aids in understanding the meaning of phrases depending on previous information. It analyzes the meaning of a paragraph word by word and stores it in a hidden layer. RNNs update the content of a hidden layer at each step as they examine text step by step in sequential order [31].

Convolution Neural Network

CNN is a deep learning model used for text classification. CNN model is known as feed forward neural network as shown in Figure 3. A CNN has three layers: an input layer for text data, hidden layers including convolutional and pooling layers that learn detailed patterns and decrease dimensionality, and an output layer for probability predictions [32].

Support Vector Machine

SVM is a classification problem-solving supervised learning method. By locating the best hyperplane to separate data with the largest gap between classes, SVMs are able to determine the best decision limits. SVMs are effective in high-dimensional domains, can handle both linear and nonlinear data as shown in Figure 4. SVM can easily handle nonlinear data by using kernel functions which convert nonlinear data into higher dimensional spaces and data become linearly separable. They are useful for those models where accuracy and interpretability are crucial because they establish a distinct decision limit [33].
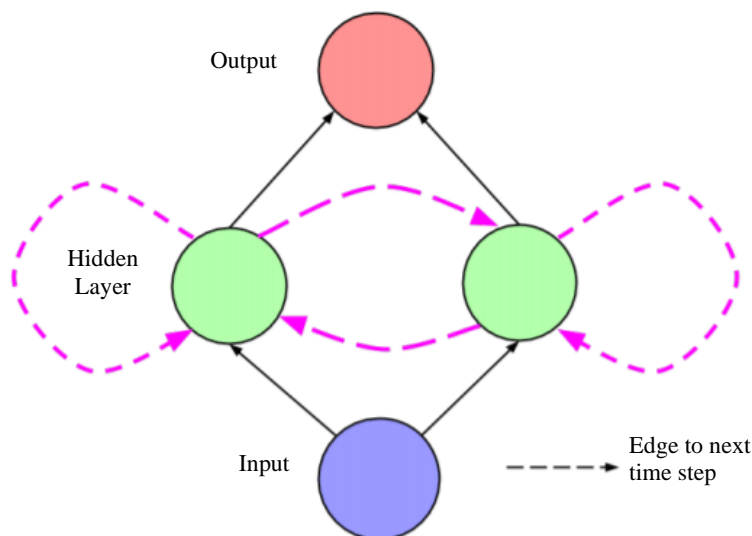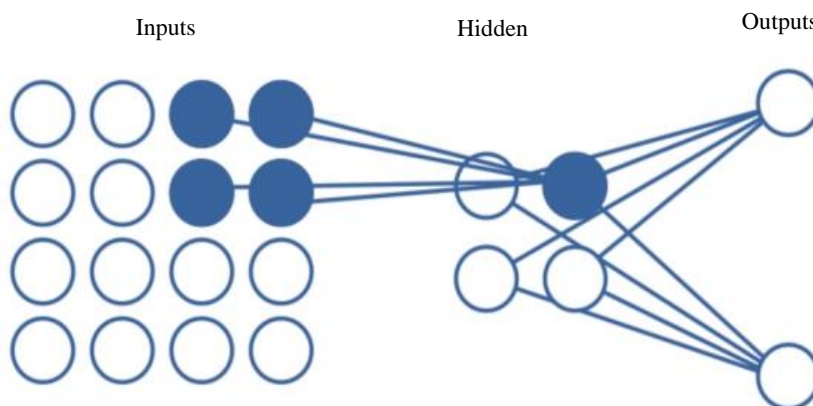
Figure 2. Recurrent neural network (RNN) [31].

Figure 3. Convolution neural network (CNN) [33].
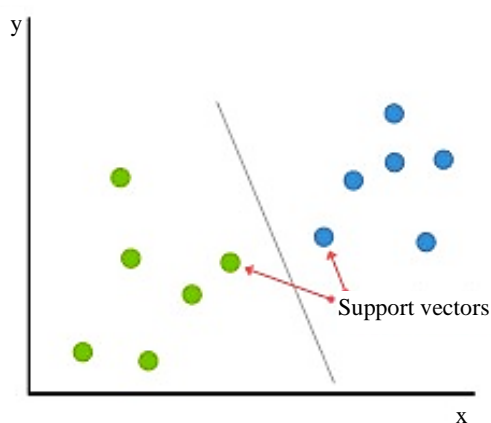

Figure 4. Support vector machine (SVM) [34].

A classification model is built using supervised learning techniques like SVM or neural networks and trained on a pre-processed dataset of instances of labelled hate speech and non-hate speech. The pre-processed data should be divided into training, validation, and test sets. For model training, use the training set. The training process teaches the model to spot patterns that separate hate speech from other types of discourse [34].

Decision Tree Algorithm

It is a supervised approach that is used for classification and regression tasks. It is a tree-like structure which contains leaf node and decision node. While leaf nodes represent the result of decisions and do not have any more branches, decision nodes are used to make any kind of decision and have numerous branches. It starts with root node from where decision tree starts, internal nodes stand in for its characteristics or features of dataset, branch represent the decision rules, and last leaf nodes represents the results and do not contain any more branches as shown in Figure 5. With the decision tree technique, a model is built to forecast a target variable's value by picking up basic choice rules deduced from the properties of the data. It is best for decision making to classify the text as shown in Table 3 [35].

Semantic Analysis

Understanding the meaning behind the words is made easier by NLP techniques like sentiment analysis and semantic analysis. It is useful to search the true meaning behind the words. By analyzing the semantic structure, sentiment, and relationships between words, phrases, and concepts, the program can distinguish between hateful content and non-hateful literature. Transformers (like BERT and Roberta) and other complex models must be able to capture contextual information. By leveraging annotated datasets of hate speech, these algorithms are trained and fine-tuned, resulting in their increased ability to accurately label and attenuate hate speech while addressing problems with sarcasm, inferred meaning, and context.
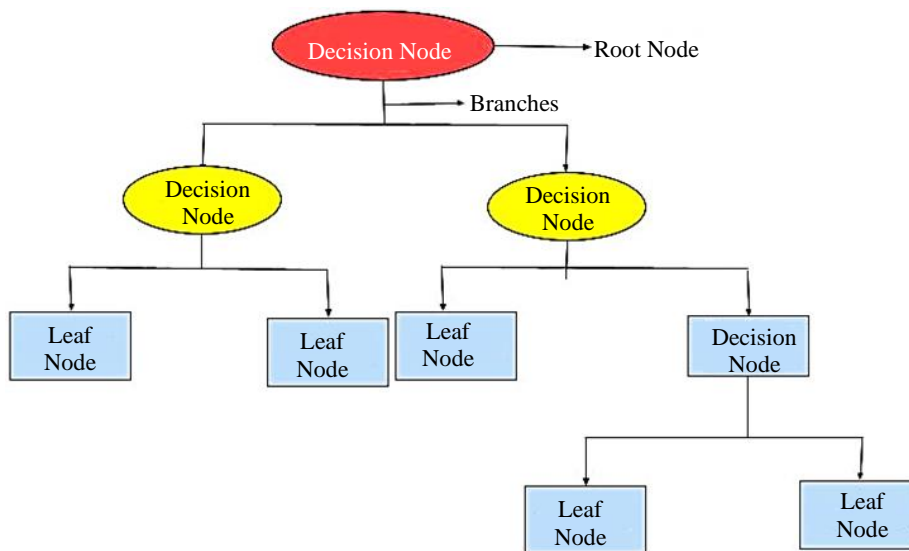

Figure 5. Decision tree [35].

Table 3. Comparison table.

Classifiers	Type of Model	Uses	Advantages	Disadvantages
Support vector machine (SVM)	Supervised learning model	Classification and regression	It can easily deal with nonlinear data using kernel fucntion.	Not effective for large dataset
Recurrent neural network (RNN)	Deep learning model	Prediction, text classification	Best for time series prediction as it has a previous data information	Training a model is difficult.
Convolutional neural network (CNN)	Deep learning model	Image recognition, classification	Efficient image processing	More training data is required
Decision tree	Supervised learning model	Classification and regression	It can easily deal with categorical and numerical data.	A tiny alteration in the data may result in a bigger alteration.

Testing of Model

Utilize the trained model to process social media material in real time. The technology analyzes the content and predicts if it includes hate speech when fresh posts and comments are made. Create a reporting system so that users may alert you to material that the model might have missed. To guarantee precise and nuanced identification, combine automatic NLP detection with human moderation. Testing makes sure the hate speech recognition model works effectively in practical situations and achieves its objectives of reducing harmful material while preventing needless content removal. To adjust to changing language and online behavior, regular monitoring and testing are crucial.

CHALLENGES IN IDENTIFICATION

Artificial intelligence and machine learning techniques are used by social media apps like X (formerly Twitter), Facebook, Instagram, YouTube, Yahoo etc. to detect the hate speech but these techniques are not completely successful because machines cannot think like humans [36]. If the target term is misspelled or certain letters have been substituted with symbols, the artificial intelligence algorithms will not be able to identify a user's post or remark as hate speech or as violating community standards [36]. Because hate speech can take many different forms and because sarcasm and context are difficult for computers to comprehend, detecting it is difficult. Word usage varies between cultures, which makes it challenging for robots. It might be difficult to distinguish between actual hate speech and things that only seem offensive. Furthermore, the definition of hate speech is not always clear-cut. Some individuals attempt to deceive computers, which need a lot of instances to learn from [36].

CONCLUSION

This paper discussed how online hate speech propagate on social media platforms. It has an impact on people's mental health causings anxiety, depression etc., which, in turn, causes them to commit crimes and commit suicide. Haters target the groups of people, politicians, celebrities, etc. on social media and hate speech cases are growing rapidly in India every year. In order to quickly identify and eliminate hate speech, social media firms must invest in advanced algorithms and artificial intelligence tools to identify hate speech because it is clear that artificial intelligence algorithms are not completely capable to detect hate speech.

Social media platforms (X [formerly Twitter], YouTube, Facebook, Instagram) have a policy to detect and stop hate speech. They take actions against the haters and ban their account, but they are not completely efficient. We have discussed artificial intelligence and machine learning algorithms like SVM, decision tree algorithm, CNN, RNN, etc., which help to identify hate speech.

Hate speech must be removed in order to maintain user confidence and trust since it encourages users to interact with potentially harmful content. Although it will not be simple, it can be stopped by cooperation. Social media hate speech can be explicitly prohibited by law, and individuals can be taught proper online behavior by reporting and removing offensive content.

REFERENCES

1. Ghanea N. The concept of racist hate speech and its evolution over time. In: United Nations Committee on the Elimination of Racial Discrimination's day of thematic discussion on Racist Hate Speech. 81st Session, Geneva Switzerland, August 28, 2012. pp. 1–16.
2. Zhang Z, Luo L. Hate speech detection: a solved problem? The challenging case of long tail on Twitter. Semantic Web. 2019; 10 (5): 925–945.
3. Allan J. The harm in hate speech. Constitutional Commentary. 2013; 29 (1): 59–80.
4. Ruwandika NDT, Weerasinghe AR. Identification of hate speech in social media. In: 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, September 26–29, 2018. pp. 273–278.
5. Keipi T, Näsi M, Oksanen A, Räsänen P. Online Hate and Harmful Content: Cross-National Perspectives. London, UK: Taylor & Francis; 2016.

6. Hirvonen K. Sweden: when hate becomes the norm. Race Class. 2013; 55 (1): 78–86.

7. Williams ML, Burnap P, Sloan L. Towards an ethical framework for publishing Twitter data in social research: taking into account users' views, online context and algorithmic estimation. Sociology. 2017; 51 (6): 1149–1168.

8. Miró-Llinares F, Moneva A, Esteve M. Hate is in the air! But where? Introducing an algorithm to detect hate speech in digital microenvironments. Crime Sci. 2018; 7 (1): 1–2.

9. Răileanu RC. Religion-based user generated content in online newspapers covering the Colectiv Nightclub fire. Revista Română de Comunicare și Relații Publice. 2016; 18 (2): 55–65.

10. Pew Research Center. Global Restrictions on Religion Rise Modestly in 2015, Reversing Downward Trend. Pew Research Center Report. Washington, DC, USA: Pew Research Center; 2017.

11. Horsti K. Digital Islamophobia: the Swedish woman as a figure of pure and dangerous whiteness. New Media Soc. 2017; 19(9): 1440–1457.

12. Jazeera A. Most victims of religious hate crimes are Muslims and Jews: UK. [Online]. Al Jazeera. 2022. Available at https://www.aljazeera.com/news/2022/10/6/hate-crimes-rise-by-26-percent-in-england-and-wales

13. Williams ML, Burnap P, Javed A, Liu H, Ozalp S. Hate in the machine: anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. Brit J Criminol. 2020; 60 (1): 93–117.

14. Sainudiin R, Yogeeswaran K, Nash K, Sahioun R. Characterizing the Twitter network of prominent politicians and SPLC-defined hate groups in the 2016 US presidential election. Soc Netw Anal Mining. 2019; 9: 1–5.

15. Huang Q, Singh VK, Atrey PK. Cyber bullying detection using social and textual analysis. In: Proceedings of the 3rd International Workshop on Socially-Aware Multimedia, Orlando, FL, USA, November 7, 2014. pp. 3–6.

16. Agatston PW, Limber S, Kowalski RM. Cyberbullying: Bullying in the Digital Age. New York, NY, USA: Wiley-Blackwell; 2012.

17. Bayzick J, Kontostathis A, Edwards L. Detecting the presence of cyberbullying using computer software. In: WebSci '11: Proceedings of the 3rd International Web Science Conference, Koblen, Germany, June 15–17, 2011.

18. Rosenthal SR, Buka SL, Marshall BD, Carey KB, Clark MA. Negative experiences on Facebook and depressive symptoms among young adults. J Adolesc Health. 2016; 59 (5): 510–516.

19. Citron DK, Norton H. Intermediaries and hate speech: fostering digital citizenship for our information age. Boston Univ Law Rev. 2011; 91: 1435.

20. Graham G. The Internet: A Philosophical Examination. London, UK: Routledge; 1999.

21. Suler J. The online disinhibition effect. Cyberpsychol Behav. 2004; 7 (3): 321–326.

22. Citron DK. Hate Crimes in Cyberspace. Cambridge, MA, USA: Harvard University Press; 2014.

23. Gagliardone I, Gal D, Alves T, Martinez G. Countering Online Hate Speech. Paris, France: UNESCO Publishing; 2015.

24. Sellars A. Defining Hate Speech. Berkman Klein Center Research Publication No. 2016-20. Public Law Research Paper No. 16-48. Boston, MA, USA: Boston University School of Law; 2016.

25. Brown A. Hate Speech Law: A Philosophical Examination. New York, NY, USA: Taylor & Francis; 2015.

26. Awan I, Zempi I. The affinity between online and offline anti-Muslim hate crime: dynamics and impacts. Aggression and Violent Behav. 2016; 27: 1–8.

27. Cheng J, Adamic L, Dow PA, Kleinberg JM, Leskovec J. Can cascades be predicted? In: Proceedings of the 23rd International Conference on World Wide Web, Seoul, South Korea, April 7–11, 2014. pp. 925–936.

28. Lin KY, Lee RK, Gao W, Peng WC. Early prediction of hate speech propagation. In: 2021 International Conference on Data Mining Workshops (ICDMW), Auckland, New Zealand, December 7–10, 2021. pp. 967–974.

29. Mathur P, Shah R, Sawhney R, Mahata D. Detecting offensive Tweets in Hindi-English code-switched language. In: Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, Melbourne, Victoria, Australia, July 15–20, 2018. pp. 18–26.
30. Torregrosa J, Bello-Orgaz G, Martínez-Cámara E, Ser JD, Camacho D. A survey on extremism analysis using natural language processing: definitions, literature review, trends and challenges. J Ambient Intell Humanized Comput. 2023; 14 (8): 9869–9905.
31. Lipton ZC, Berkowitz J, Elkan C. A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:1506.00019. May 29, 2015. Available at https://arxiv.org/abs/1506.00019
32. Abril D. Hateful posts on Facebook and Instagram soar. [Online]. Fortune. 2020. Available at https://fortune.com/2020/08/11/facebook-instagram-hate-speech-violations-q2/
33. Gambäck B, Sikdar UK. Using convolutional neural networks to classify hate-speech. In: Proceedings of the First Workshop on Abusive Language Online, Vancouver, British Columbia, Canada, July 30–August 4, 2017. pp. 85–90.
34. Berwick R. An idiot's guide to support vector machines (SVMs). [Online]. Mit.edu. 2003. Available at https://web.mit.edu/6.034/wwwbob/svm.pdf
35. Javatpoint. Decision Tree Algorithm in Machine Learning. [Online]. Javatpoint. 2021. Available at https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm
36. Di Fátima B, Munoriyarwa A, Gilliland A, Msughter AE, Vizcaíno-Verdú A, Gökaliler E, Capoano E, Yu H, Alikılıç İ, González-Aguilar JM, Tsene L. Hate speech on social media: a global approach. Quito, Ecuador: Pontificia Universidad Católica del Ecuador; 2023.