International Research Journal of Mathematics, Engineering and IT
ISSN: (2349-0322)
Association of Academic Researchers and Faculties (AARF)
Impact Factor- 8.132 Volume 12, Issue 04, April 2025
Website- www.aarf.asia, **Email** : editoraarf@gmail.com

Analyzing Online Youth Behavior: Machine Learning Approaches for Toxic Content Detection, Regional Trends, and Future Directions

Renu Kumari[1*] and Vijay Kumar[1]

[1,2]College of Commerce Arts and Science, Patliputra University,, Patna, 800020, Bihar, India.

Abstract

Social computing, fueled by the widespread use of social media plat- forms, has significantly reshaped the ways youth interact, express, and form opinions. In India, where digital adoption among young users is rising rapidly, the influence of online interactions has both empowering and concerning implications. This review paper presents a comparative study on the effect of social computing on youth, with a regional empha- sis on Bihar—one of the most populous yet under-researched states in India. The paper explores how machine learning techniques are being leveraged to analyze youth behavior, predict social media trends, and detect toxic content such as hate speech. We examine existing literature on social computing's psychological and social impact, especially among youth populations, and survey machine learning-based approaches to predictive analytics and hate speech detection on platforms like Twit- ter, Facebook, and Instagram. Through comparative insights between global, national, and regional studies, this paper highlights the cultural and linguistic nuances that challenge traditional machine learning (ML) models. The review also outlines current limitations in data availability, regional language processing, and algorithmic fairness. Finally, we iden- tify emerging trends and recommend future research directions focused on culturally contextualized artificial Intelligence (AI) systems that can aid in safer and more inclusive digital environments for youth. This

study aims to bridge the gap between technical innovation and socio- cultural understanding in youth-centered social computing research.

1 Introduction

1.1 Background and Motivation

In the last decade, social computing has emerged as a powerful domain that combines social behavior with computational intelligence. Platforms like Face- book, Twitter, and Instagram have not only revolutionized communication but also shaped public discourse, especially among the youth. In India, where over 66% of the population is below the age of 35, the influence of these platforms is particularly pronounced. As digital penetration deepens in Tier-2 and Tier-3 regions, understanding how youth engage with social media becomes essential from both sociological and technological standpoints.

While several global studies have examined the effects of social media on young users, there remains a scarcity of research that considers regional and cultural specificities—especially in states like Bihar. Youth in Bihar are rapidly adopting digital platforms, yet their patterns of interaction, exposure to mis- information, and vulnerabilities to toxic content remain underexplored. This gap motivates a deeper inquiry into how social computing affects this demo- graphic, and how machine learning (ML) can help uncover behavioral patterns and mitigate harmful content.

Hate speech on social media platforms like Twitter has become a growing concern, making it harder to ensure online safety and encourage healthy con- versations. To reduce its harmful impact on individuals and communities, it is important to detect and monitor such content effectively.In hate speech detec- tion, two broad categories of machine learning techniques are widely used: traditional ML (logistic regression, SVM, random forest) and deep learning approaches (RNN, LSTM, CNN, Transformers). Traditional methods typically rely on feature engineering, such as n-grams, TF-IDF, POS tags, and lexicon- based features. For example, support vector machines (SVM) and logistic regression often outperform others in tasks like OffensEval and HASOC using n-gram and embedding features, while random forest, especially when coupled with count vectorizers, has shown top accuracy in comparative studies[1, 2].

1.2 Effect of Social Media on Youth: Benefits & risks (mental health, identity formation, engagement)

This section explores the effect of social media on youth by examining its benefits and risks—including impacts on mental health, identity formation, and engagement—drawing insights from global studies, the broader Indian con- text, and region-specific observations from Bihar. Social media has significantly reshaped youth behavior worldwide, acting both as a catalyst for connection and a source of psychological distress. Numerous global studies highlight its dual impact—enhancing peer communication and identity exploration while also contributing to mental health concerns such as anxiety, depression, lone- liness, and disrupted sleep patterns. A comprehensive review by Keles et al. found a strong correlation between social media usage and depressive symptoms in adolescents, attributing this to factors like cyberbullying, social comparison, and validation-seeking behavior. The displaced behavior theory explains that excessive time spent on digital platforms often replaces in-person social interactions that are crucial for emotional resilience, leading to isolation and diminished well-being. In parallel, identity formation—a core developmen- tal task in adolescence—is deeply influenced by curated online personas, as adolescents equate digital validation (likes, comments) with self-worth[3]. A study conducted in Germany by Michikyan et al. supported this, noting that teens increasingly blur the lines between their online and offline selves, which can complicate self-concept and social relationships[4]. Additionally, longitu- dinal studies from the UK and the US report that nighttime social media use negatively impacts sleep quality and academic focus. Despite these challenges, social platforms also offer avenues for support, community engagement, and access to mental health resources when used responsibly. Hence, global evi- dence paints a nuanced picture: while social media can empower youth through connectivity and self-expression, it simultaneously poses risks to psychological and developmental health, emphasizing the need for context-aware interven- tions and digital literacy education tailored to adolescent users.

In the Indian context, this section focuses on digital access among youth, high- lighting studies that examine the amplified impact of social media in rural and semi-urban areas. In India, the rise of social media among adolescents has intensified the dual forces of connectivity and vulnerability, with research showing significant effects on mental health. A qualitative survey of 204 Indian youths aged 14–23 found that excessive use of platforms like Instagram and Facebook is strongly correlated with elevated stress, anxiety, depression, sleep disturbances, compulsive night-time scrolling, and exposure to cyberbullying [5]. Supporting these findings, a report in The Indian Express highlights how young Indians (especially those aged 18–24) spend an average of 2.4 hours daily on social media, where curated online personas and 'internet validation' via likes often exacerbate feelings of inferiority, body image issues, and eat- ing disorders—particularly among teenage girls, with 32% reporting worsened self-perception due to Instagram 1. Additionally, these platforms have been implicated in spreading mental health misinformation, which can mislead vul- nerable users2. Local news reports further underscore the profound emotional toll: cases of cyberbullying leading to depression and even self-harm among school and college students have become increasingly common across cities like Lucknow3.Overall, the Indian context mirrors global concerns but is mag- nified by regional issues like gender-sensitive body image pressures, a high prevalence of smartphone addiction among youth, and the rapid shift toward image-centric platforms, calling for culturally tailored digital literacy educa- tion and mental health interventions

This issue is becoming more serious for society, individuals, researchers, and policymakers. Although automated tools have been developed to detect and monitor hate speech, their performance remains poor, indicating the need for more research in this area [28].This paper aims to review and compare existing literature on:

- The psychological and behavioral impact of social computing on youth,
- Machine learning techniques for predictive analytics and hate speech detection,
- Regional and cultural considerations in building context-aware ML models.

[1] https://indianexpress.com/article/lifestyle/health/mental-health-in-india-impact-of-social-media-on-young-indians-facebook-instagram-youtube-twitter-7778499/?utm source=chatgpt.com

[2] https://timesofindia.indiatimes.com/life-style/health-fitness/health-news/mental-health-myths-go-viral-on-social-media-experts-warn-of-potential-risks/articleshow/121545029.cms?utm source=chatgpt.comIn Bihar, the discourse surrounding youth engagement with social media remains nascent but revealing: according to the UDAYA (Understanding the Lives of Adolescents and Young Adults) survey, approximately 29.2% of ado- lescent girls in Bihar—higher than in neighbouring Uttar Pradesh—reported exposure to social media by 2018–19, with education level and family income strongly influencing digital access. Researchers have observed that increased exposure among better-off, urban, and educated youths is linked to both positive outcomes—such as improved awareness of sexual and reproductive health—but also risks, including heightened susceptibility to online harass- ment and misinformation [6]. Qualitative insights from neighboring regions illustrated that cyberbullying is emerging as a serious concern: a study in Luc- know noted that caste- or appearance-based trolling on platforms like Facebook contributed to depression and anxiety in teenagers[4]. While Bihar-specific mental health data associated with social media use remain limited, this regional evidence hints at a growing interface between socio-economic disparities and digital vulnerabilities. These include skewed access favoring wealthy urban youth and potential exposure to harmful content, underscoring the urgent need for localized research into psychological effects, digital literacy, and culturally sensitive interventions. Few studies highlighted that the rise of social media has made it easier for people to form online communities and stay anonymous, which has led to growing concerns about hate speech. [3]https://timesofindia.indiatimes.com/city/lucknow/cyberbullying-taking-toll-on-youngsters-mental health/articleshow/121171034.cms?utm source=chatgpt.com

[4] https://timesofindia.indiatimes.com/city/lucknow/cyberbullying-taking-toll-on-youngsters-mental-health/articleshow/121171034.cms?utm source=chatgpt.com

We emphasize a dual perspective: first, the social impact of digital interaction on youth; second, the role of ML as both a tool for understanding and inter- vention. Special focus is placed on Bihar to provide insights that may inform regional policy-making and the development of inclusive AI systems.

2 Related Work

2.1 Machine Learning in Social Computing & Hate Speech Detection

In the domain of hate speech detection on social media, traditional machine learning algorithms such as logistic regression, K-nearest neighbor (KNN), support vector machines (SVM), Naive Bayes, decision tree and random forests have demonstrated robust performance, particularly when paired with classical feature extraction methods like TF-IDF and n-grams.

For instance, Das et al. compared logistic regression, SVM, KNN na¨ıve Bayes, decision tree and random forest on a Twitter hate speech dataset, finding that SVM, Decision Tree and random forest achieved top accura- cies of around 95.5% 96.2% and 98.2%, respectively [7]. Complementing this, a research effort analyzing similar classifiers on Twitter reported logistic regression delivering around 80% AUC, while both SVM and random forest outperformed it in binary classification tasks; random forest, in particular, showed resilience to noise and superior feature importance estimation due to its ensemble nature [8, 9]. Standard pipelines typically involve pre-processing and vectorization—often TF-IDF—followed by classifier training, with random for- est being favored for its ensemble voting and overfitting mitigation, and SVM valued for its capacity to construct optimal hyperplane separators [10, 11].

These methods continue to serve as strong baselines in hate speech detection frameworks, offering high interpretability and efficiency, though their perfor- mance may decline when applied to code-mixed Indian datasets, underscoring the need for culturally and linguistically aware model adaptation.

Priyadarshini I. et al. highlighted that hate speech not only spreads violence and hatred but also demands significant computing resources and con- stant monitoring by both humans and algorithms. Although many AI-based approaches have been developed to address this issue, there is still a need for more efficient solutions that perform well and require less time to train, especially given the growing volume of online content. To tackle this, the use of transfer learning with pre-trained models has been proposed, allowing for faster and more reusable hate speech detection on social media platforms[12]. Detecting hate speech and telling it apart from offensive content remains a challenge for many existing machine learning models. This is mainly because most current approaches treat hate speech classification as a multi-class prob- lem, which limits their accuracy. To address this, Khan et al. introduced a new perspective by treating hate speech detection on social media as a multi- label problem. Their system, called ’Hate Classify,’ uses the HSD-DT model to categorize posts into three labels: hate speech, offensive, or not offensive [13]. In fact, Turki & Roy found random forest superior to bagging and AdaBoost models[14].

Greevy and Smeaton [15] present an early supervised learning approach to automatically classify racist web content using Support Vector Machines (SVMs). They explore different representations of text—namely bag-of-words

(unigrams), bigrams, and part-of-speech tags—to determine which features most effectively support SVM-based classification. Their work emphasizes the complexity of detecting racism, noting that the mere presence of certain words often fails to capture the nuanced nature of racist discourse. Their’s Key findings include: Feature engineering matters: The bigram representation outperforms bag-of-words in capturing contextual clues important in racist language, Part-of-speech tagging adds nuanced value: Including syntax-based features enhances the SVM’s ability to detect subtler forms of racism that may bypass simple keyword filters and High baseline accuracy: Their SVM models achieve strong classification performance, suggesting that such methods can serve as reliable tools for automated racism detection in textual data. This study offers a solid foundation for examining traditional ML methods in hate or racist speech detection. It highlights the importance of choosing the right textual features—even with simpler models like SVM—and sets a method- ological baseline to contrast against more advanced deep learning and hybrid models.

Kwok and Wang [16] addressed the escalating concern of anti-Black hate speech on Twitter—a platform where Black users are disproportionately rep- resented (25% of users) compared to their U.S. population share (13.5%). Recognizing this overrepresentation and the presence of overtly racist content, the authors set out to build a lightweight supervised classifier to identify tweets targeting Black individuals or communities.

They curated a dataset of tweets from diverse Twitter accounts, manually labeled as ”racist” or ”nonracist.” This binary annotation process was done with minimal resource investment, enabling them to scale the data collection with minimal cost. They employed a traditional machine learning pipeline: Feature Extraction:Bag-of-words and simple lexical features, Model Training:** A straightforward Na¨ıve Bayes clas- sifier and Evaluation:10-fold cross-validation across the corpus. The Na¨ıve Bayes model achieved an average accuracy of 76% on single-tweet classification. However, they noted a tendency for the model to mislabel tweets containing derogatory slurs against Black people as hate speech even if context clarified the meaning (e.g. quoting a slur in a neutral or condemnatory way). They demonstrated feasibility of flagging anti-Black tweets using low-resource ML approaches. There are some limitations: such as Lack of contextual sensitivity leads to over-flagging, absence of nuanced labels (e.g., degrees of hate, targeted groups beyond race).

Gitari et al.[17] introduced a structured, lexicon-based supervised clas- sifier designed to detect hate speech in online discourses such as forums and blogs, focusing specifically on three domains: race, nationality, and reli- gion. Authors begin by applying subjectivity analysis to filter out neutral or objective sentences—retaining only subjective content likely to contain hate- ful intent. They then develop a hate speech lexicon, comprised of semantic terms and expressions relevant to the targeted domains.

Using this lexicon, they extract features from individual sentences and aggregate them to represent documents. These features include sentiment and subjectivity scores,presence of lexicon terms, and other semantic markers. A supervised learn- ing model is then trained on these labeled corpora. Evaluation on a standard hate speech dataset demonstrates that combining subjectivity filtering with lexicon-based semantic analysis significantly improves detection precision and recall compared to simpler lexicon-only methods. The authors emphasize that their lexicon-enhanced classifier is particularly effective in real-world web dis- course, where hate speech is often nuanced and context-dependent. They also highlight the approach's scalability: as more annotated data becomes available, more advanced machine learning techniques (like SVM or maximum entropy models) can be layered on top of this lexicon-based framework to further boost performance. Their's key Contributions are: Integration of subjectivity detection to filter irrelevant content, Construction of a targeted lexicon for race, nationality, and religion domains and a classifier combining lexical and semantic features that outperforms basic lexicon-only baselines.

Weiqiang Jin and colleagues proposed a method called PMTL-DisCo (Prompting Multi-Task Learning guided by Dissemination Consistency) to detect fake news, especially when limited examples are available.

Their approach goes beyond traditional use of pre-trained language models by com- bining advanced techniques like masked language models, multi-task learning, and prompt-based tuning. PMTL-DisCo introduces three key innovations: (1) it enhances feature learning through a process called news distributed represen- tation optimization, which uses the consistency of how similar news is shared;

(2) it applies an improved prompt-tuning method using high-quality, expanded label words through adaptive multi-label verbalization; and (3) it boosts pre- diction accuracy with a multi-neighbor reasoning technique that leverages the trustworthiness of socially connected news items [18].

While hate speech can be expressed both online and offline, its frequency and impact have significantly increased with the growth of social media. This study focuses on identifying and analyzing unstructured data from social media posts that aim to spread hate through their comments. To address this, all social media platforms should recognize the extent of hate speech, using a framework called SA developed by Rodriguez A. et al., which combines data analysis with natural language processing [19].

2.2 Deep Learning in Social Computing & Hate Speech Detection

Traditional approaches remain strong baselines, particularly when combined with transformer-derived embeddings to boost performance beyond standalone deep models. On the other hand, deep learning methods directly learn rep- resentations from text: convolutional neural networks (CNN) capture local phrase patterns, recurrent models like RNN or LSTM model sequential con-text, and pre-trained transformers (e.g., BERT) enable fine-tuned, context-rich understanding.Convolutional Neural Networks (CNNs) are effective at recog- nizing local features and layered patterns, making them useful for extracting spatial details from spectrograms, images, and even text through word embed- dings. However, when it comes to understanding long-term relationships in data, Recurrent Neural Networks (RNNs)—especially LSTMs and GRUs—are more suitable. RNNs are designed to handle sequences, allowing them to cap- ture context and patterns over time, which is essential for processing speech or long pieces of text. Anand & Eswari reported LSTM and CNN with GloVe achieved over 97% accuracy on hate speech datasets[20], while transformer models fine-tuned for HS detection consistently outperform CNN/LSTM when sufficient multilingual and code-mixed data are available. Hybrid systems that layer transformer embeddings onto traditional classifiers (e.g., logistic regres- sion or MLP) have even matched or exceeded pure deep models, showcasing the continued importance of both paradigms in designing effective hate speech detection systems[1].

As the number of internet users rapidly increased, problems like cyberbul- lying and hate speech also began to rise. This article focuses on the issue of hate speech on Twitter. Hate speech often spreads false information to pro- voke hatred. It usually targets specific groups based on gender, religion, race, or disability, as noted by Roy, P. K. et al. When such speech makes individuals or communities feel attacked or discouraged, it can sometimes lead to unexpected criminal behavior [21].

This study, based on the work by Toktarova et al., presents a detailed approach that combines deep learning and traditional machine learning tech- niques for hate speech detection on Twitter. It also compares the performance of different methods. To ensure reliable results, the models are trained on a well-constructed dataset, created using data from multiple sources and labeled by experts [22]. Ruqaya Abdulhasan Abed and her team proposed a modified Convolu- tional Neural Network-based Intrusion Detection System (CNN-IDS). They used the UNSW-NB15 dataset to train and test their models. For selecting the most relevant features, they applied Principal Component Analysis (PCA) and Singular Value Decomposition (SVD). The selected features were then classified using three techniques: Ridge Regression (RR), Stochastic Gradient Descent (SGD), and CNN. The model supports both binary and multi-class classification. The findings highlight that

PCA and SVD significantly improve model accuracy. Notably, the Ridge Regression model achieved high accuracy in binary classification, increasing from 98.13% to 99.85% [23].

By automatically detecting hate speech in web content, models based on natural language processing and machine learning provide a means to make online platforms safer. The biggest challenge, however, is getting enough exam- ples annotated to train these models. This work constructs a unified hate speech representation by combining two separate datasets using a transfer learning approach.

To project and compare various datasets, develop a two- dimensional visualization tool for the built hate speech representation by Yuan,L. et al.[24]. Weiqiang Jin and colleagues proposed a method called Detection Yet See Few (Detect YSF) for detecting fake news. This approach is designed to work well even with limited labeled data by combining two techniques: adversarial semi-supervised learning and contrastive self-supervised learning. Detect YSF is built on Transformer-based pre-trained language models (like BERT and RoBERTa), and it fine-tunes them using a special method based on masked language modeling and pseudo-prompts. The training process includes two key improvements: (1) a simple contrastive self-supervised learning method is used to improve the quality of sentence-level semantic representations; and (2) a Generative Adversarial Network (GAN) is developed using random noise and fake news examples to generate challenging embeddings. This GAN uses Multi- Layer Perceptrons (MLPs) and a separate PLM encoder. Finally, the model applies semi-supervised adversarial learning to refine the embeddings further during prompt tuning, improving DetectYSF's performance [25].

Mehta, H. et al. explored how artificial intelligence can be effectively used to detect hate speech. Their research focused on understanding how advanced AI models make decisions and how those decisions can be explained. They used two datasets to demonstrate the ability of AI to identify hate speech.

During data preprocessing, they removed inconsistencies, cleaned the tweet text, performed tokenization and lemmatization, and simplified categorical variables to prepare a clean dataset for training the models[26].

Weiqiang Jin and colleagues introduced a method called CAPE-FND for detecting fake news using context-aware prompt engineering. This method improves the performance of large language models (LLMs) by using a self- adjusting strategy that refines prompts through random search. It also includes context-specific rules, background information, and analogy-based reasoning to reduce errors or hallucinations from the model. CAPE-FND continually improves the original prompts to get better results. Experiments using GPT- 3.5-turbo on several public datasets showed that CAPE-FND can sometimes perform better than both GPT-4.0 and human evaluators in zero-shot and few-shot scenarios [23].

Weiqiang Jin and colleagues proposed a framework named Det2Ver that combines rumor detection and fact verification into a single system. Det2Ver uses information from the rumor detection process to improve fact-checking by creating customized prompt templates and using prompt-tuned large language models (LLMs) like T5. Their results highlight the effectiveness of this approach. In particular, Det2Ver boosts fact verification perfor- mance—especially the macro-F1 score—through cross-task knowledge sharing, outperforming other prompt-tuning methods in few-shot and zero-shot exper- iments across three widely used datasets [27].

3 Discussion and Future trends

The detection of hate speech has garnered significant attention from researchers, leading to the development of various techniques aimed at address- ing the issue. Both online and offline manifestations of hate, including hate crimes and hate speech, can be mitigated by effectively identifying and reg- ulating such content on social media platforms. Despite considerable efforts, Natural Language Processing (NLP)-based approaches for hate speech detec- tion remain limited in their effectiveness. The rapid evolution of language on social media continues to outpace existing models, making it difficult to imple- ment robust and generalizable solutions. Additionally, many current methods exhibit inconsistencies and face unresolved challenges, prompting scholars to propose future research directions to address these limitations. One growing concern is the increasing prevalence of misogynistic or gender-targeted lan- guage, which underscores the urgency of further investigation. Importantly, the presence of potentially hateful words in a text does not always imply hate speech, while subtler expressions lacking explicit slurs often pose greater challenges for accurate classification.

Deep learning has emerged as a prominent approach for text classification and is increasingly being applied to tackle the growing volume of hate speech across social media platforms. However, as noted by some researchers, this task also demands an understanding of social and cultural contexts, which machines often lack. The notable inconsistency between human annotations in hate speech classification further indicates that automated systems may strug- gle with this complex task. Utilizing features such as unigram dictionaries and broader hate speech patterns could enhance the detection of offensive con- tent, claims, or objectionable messages. Additionally, the influence of humor in hate speech detection remains an underexplored area and presents a valu- able direction for future research. Despite rapid advancements in this field, progress is hindered by a lack of multilingual datasets, with most studies focus- ing solely on English-language content. Consequently, there is a pressing need for research in other languages and for developing cross-lingual models, as generalization across languages remains underdeveloped. Current methods also suffer from inconsistency and unresolved challenges, underscoring the need for further exploration. Based on the reviewed literature, several emerging trends and future research directions are proposed to address these ongoing issues.

We reviewed several widely used hate speech datasets, examining their key features, classification schemes, intended purposes, and data formats. Most of these datasets are text-based, while only a few focus on visual content, such as hostile memes—examples include MultiOFF and

MMHS150K. To date, there appears to be no publicly available dataset dedicated to hate speech in video form, making the creation of new image and video datasets a notable chal- lenge for future research. Additionally, the limited availability of open-access datasets remains a significant obstacle, as many researchers do not release the datasets they develop. This lack of accessibility hinders the ability to replicate studies and perform meaningful comparisons across different research efforts.

In recent years, researchers have focused on detecting hate speech across multiple languages by developing diverse datasets. However, there remains a limited availability of labeled datasets in non-English languages. Despite this scarcity, existing labeled data in these languages can still be utilized with various benchmark models for evaluation.

Emojis have become a common means of expressing emotions and sen- timents online, and they also play a significant role in the dissemination of hateful or offensive content on social media. As a result, handling text that includes emojis can be considered a specialized area within pre-processing, aimed at improving the detection of hostile language. Additionally, some meta-heuristic optimization methods have been proposed to address hate speech detection. In the future, parameter optimization strategies—alongside exist- ing techniques such as the Firefly Metaheuristic Optimization (FMO) and Ant Lion Optimization (ALO)—could further enhance the effectiveness of detecting hateful content.

4 Conclusion

Studies provided a detailed summary of different hate speech detection meth- ods used on social media platforms like Twitter, highlighting both their strengths and weaknesses. These methods include traditional techniques such as Support Vector Machines (SVMs) and decision trees, as well as advanced approaches like deep learning, recurrent neural networks (RNNs), and transfer learning. The study also explores the role of sentiment analysis and natural language processing (NLP) in improving detection accuracy.

The comparison of different models and datasets helps understand how well each approach works in tackling hate speech online. While there has been progress, most existing systems focus only on text or audio data, which can miss important context. For example, text-based models often struggle with sarcasm, indi- rect hate speech, or mixed-language content. Similarly, audio-based models are affected by background noise and differences in speakers' voices. Video-based hate speech detection remains under-researched due to high computational demands and the challenge of combining visual information effectively. More- over, many models lack transparency, making it difficult to understand how they arrive at their decisions. These limitations highlight the need for a multi- modal hate speech detection framework (MHSDF) that combines text, audio, and visual data to improve accuracy, robustness, and interpretability.

Credit authorship contribution statement

All authors have contributed equally.

Declarations

Funding and/or Conflicts of interests/Competing interests

The authors have no conflict of interest/competing interest to declare that are relevant to the content of this article.

References

[1] G. Ramos, F. Batista, R. Ribeiro, P. Fialho, S. Moro, A. Fonseca, R. Guerra, P. Carvalho, C. Marques, and C. Silva, "A comprehensive review on automatic hate speech detection in the age of the transformer," *Social Network Analysis and Mining*, vol. 14, no. 1, p. 204, 2024.

[2] F. Alkomah and X. Ma, "A literature review of textual hate speech detection methods and datasets," *Information*, vol. 13, no. 6, p. 273, 2022.

[3] B. Keles, N. McCrae, and A. Grealish, "A systematic review: the influence of social media on depression, anxiety and psychological distress in ado- lescents," *International journal of adolescence and youth*, vol. 25, no. 1, pp. 79–93, 2020.

[4] M. Michikyan, K. Subrahmanyam, and J. Dennis, "Facebook use and academic performance among college students: A mixed-methods study with a multi-ethnic sample," *Computers in Human Behavior*, vol. 45, pp. 265–272, 2015.

[5] V. V. Taddi, R. K. Kohli, and P. Puri, "Perception, use of social media, and its impact on the mental health of indian adolescents: A qualitative study," *World Journal of Clinical Pediatrics*, vol. 13, no. 3, p. 97501, 2024.

[6] R. Saha, P. Paul, S. Yaya, and A. Banke-Thomas, "Association between exposure to social media and knowledge of sexual and reproductive health among adolescent girls: evidence from the udaya survey in bihar and uttar pradesh, india," *Reproductive health*, vol. 19, no. 1, p. 178, 2022.

[7] S. Das, K. Bhattacharyya, and S. Sarkar, "Performance analysis of logistic regression, naive bayes, knn, decision tree, random forest and svm on hate speech detection from twitter," *International Research Journal of Innovations in Engineering and Technology*, vol. 7, no. 3, p. 24, 2023.

[8] S. K. Mohapatra, S. Prasad, D. K. Bebarta, T. K. Das, K. Srinivasan, and Y.-C. Hu, "Automatic hate speech detection in english-odia code mixed social media data using machine learning techniques," *Applied Sciences*, vol. 11, no. 18, p. 8575, 2021.

[9] D. S. Wankhede, A. Manikjade, N. Meher, T. Atkale, A. Ghule, D. Gujar *et al.*, "Analyzing the performance of naive bayes, logistic regression, svm and random forest for identifying hate speech from twitter social media," in *2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*. IEEE, 2023, pp. 1–6.

[10] F. T. Boishakhi, P. C. Shill, and M. G. R. Alam, "Multi-modal hate speech detection using machine learning," in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 2021, pp. 4496–4499.

[11] R. Reghunathan and A. Asha, “Hate speech detection in conventional lan- guage on social media by using machine learning,” *International Journal of Engineering Research*, vol. 11, no. 06, 2022.

[12] I. Priyadarshini, S. Sahu, and R. Kumar, “A transfer learning approach for detecting offensive and hate speech on social media platforms,” *Multimedia Tools and Applications*, vol. 82, no. 18, pp. 27 473–27 499,
2023.

[13] M. U. Khan, A. Abbas, A. Rehman, and R. Nawaz, “Hateclassify: A service framework for hate speech identification on social media,” *IEEE Internet Computing*, vol. 25, no. 1, pp. 40–49, 2020.

[14] T. Turki and S. S. Roy, “Novel hate speech detection using word cloud visualization and ensemble learning coupled with count vectorizer,” *Applied Sciences*, vol. 12, no. 13, p. 6611, 2022.

[15] E. Greevy and A. F. Smeaton, “Classifying racist texts using a support vector machine,” in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004, pp. 468–469.

[16] I. Kwok and Y. Wang, “Locate the hate: Detecting tweets against blacks,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 27, no. 1, 2013, pp. 1621–1622.

[17] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, “A lexicon-based approach for hate speech detection,” *International Journal of Multimedia and Ubiquitous Engineering*, vol. 10, no. 4, pp. 215–230, 2015.

[18] W. Jin, N. Wang, T. Tao, M. Jiang, Y. Xing, B. Zhao, H. Wu, H. Duan, and G. Yang, “A prompting multi-task learning-based veracity dis- semination consistency reasoning augmentation for few-shot fake news detection,” *Engineering Applications of Artificial Intelligence*, vol. 144, p. 110122, 2025.

[19] A. Rodriguez, Y.-L. Chen, and C. Argueta, “Fadohs: framework for detec- tion and integration of unstructured data of hate speech on facebook using sentiment and emotion analysis,” *IEEE Access*, vol. 10, pp. 22 400–22 419,
2022.

[20] M. Anand and R. Eswari, “Classification of abusive comments in social media using deep learning,” in *2019 3rd international conference on com- puting methodologies and communication (ICCMC)*. IEEE, 2019, pp. 974–977.

[21] P. K. Roy, A. K. Tripathy, T. K. Das, and X.-Z. Gao, “A framework for hate speech detection using deep convolutional neural network,” *IEEE Access*, vol. 8, pp. 204 951–204 962, 2020.

[22] A. Toktarova, D. Syrlybay, B. Myrzakhmetova, G. Anuarbekova,
G. Rakhimbayeva, B. Zhylanbaeva, N. Suieuova, and M. Kerimbekov, “Hate speech detection in social networks using machine learning and deep learning methods,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 5, 2023.

[23] W. Jin, Y. Gao, T. Tao, X. Wang, N. Wang, B. Wu, and B. Zhao, “Veracity-oriented context-aware large language models–based prompting optimization for fake news detection,” *International Journal of Intelligent Systems*, vol. 2025, no. 1, p. 5920142, 2025.

[24] L. Yuan, T. Wang, G. Ferraro, H. Suominen, and M.-A. Rizoiu, “Trans- fer learning for hate speech detection in social media,” *Journal of Computational Social Science*, vol. 6, no. 2, pp. 1081–1101, 2023.

[25] W. Jin, N. Wang, T. Tao, B. Shi, H. Bi, B. Zhao, H. Wu, H. Duan, and G. Yang, “A

veracity dissemination consistency-based few-shot fake news detection framework by synergizing adversarial and contrastive self- supervised learning," *Scientific Reports*, vol. 14, no. 1, p. 19470, 2024.

[26] H. Mehta and K. Passi, "Social media hate speech detection using explainable artificial intelligence (xai)," *Algorithms*, vol. 15, no. 8, p. 291, 2022.

[27] W. Jin, M. Jiang, T. Tao, H. Zhou, X. Wang, B. Zhao, and G. Yang, "Can rumor detection enhance fact verification? unraveling cross-task synergies between rumor detection and fact verification," *IEEE Transactions on Big Data*, 2024.

[28] M. S. Jahan and M. Oussalah, "A systematic review of hate speech auto- matic detection using natural language processing," *Neurocomputing*, vol. 546, p. 126232, 2023.