Using Hate Sonar for Cloud-Based Hate Speech Detection in Social Media Platforms

S Sujatha
Department of Biomedical Engineering, Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai,
Tamil Nadu, India

Abstract
Hate Sonar for cloud-based hate speech identification in social media platforms improves real-time content recognition. HateSonar can reliably identify poisonous language using powerful machine learning techniques, improving online community monitoring. We want to automate hate speech detection on social media to make it safer and more inclusive. Technology analyzes and classifies text using natural language processing to reveal dangerous information. Automating the detection process, minimizing human moderators, and responding faster to potentially hazardous online conduct improves content moderation. This strategy may help create a more responsible digital world, promote pleasant connections, and safeguard people from online abuse. HateSpeech_Detection dataset for 5 campaigns Five parameters is employed. Hateful Comments: 124–694, Offensive Comments: 340–1302, Neutral Comments: 506–2943, Total Comments: 2509–4855, Detection Accuracy (%): 90.93–97.47. From Category_Breakdown dataset for 5 platforms Five parameters are employed. Racial Hate (%): 20.91–49.87, Political Hate (10.76–22.93), Gender-Based Hate (19.68–28.71), Other Categories (8.03–18.58), Total Hate Speech (%): 44.52–72.96. Data from Temporal_Trends for 5 platforms Five parameters are employed. Daily frequency is 133–452, weekly frequency is 995–2817, and monthly frequency is 4715–8437. Monday through Saturday are peak activity days, and 6-13 is peak hour.

Keywords
Hate Sonar, Hate speech detection, Social media moderation, Cloud-based technology, Content analysis

1. Introduction

Social internet has facilitated global communication but also spread hate speech. Hate speech harms people and communities and requires strong detection and mitigation methods. Manual, resource-intensive, and inconsistent content moderation methods are common. HateSonar detects hate speech easily and automatically using machine learning. HateSonar can efficiently and reliably detect and handle offensive material by processing massive volumes of social media data in real time in the cloud. HateSonar in cloud computing frameworks improves social media hate speech detection accuracy, scalability, and efficiency. The goal is to employ machine learning algorithms to categorize and detect hate speech for quick interventions. To make the internet safer, detect and remove dangerous information while protecting user privacy and free speech. HateSonar cloud integration allows scalability for massive datasets, real-time analysis, and safe user data storage. HateSonar can detect hate speech from innocent sentiments by identifying nuanced and complicated textual patterns using powerful natural language processing. The system's linguistic and context flexibility makes it useful across social media ecosystems.

Section 2 discusses hate speech on social media, its social effects, and moderating constraints. To solve these issues, scalable, accurate, and context-sensitive detection technologies are needed. Section 3 examines HateSonar's cloud-based technological implementation. HateSonar's machine learning model architecture, cloud integration for scalability and storage, and APIs for social media integration are discussed. The section also discusses sophisticated language processing and contextual analysis methods.

Section 4 assesses HateSonar's hate speech detection ability using precision, recall, and F1 scores. Case studies demonstrate the system's

DOI: https://doi.org/10.5281/zenodo.19315956

capacity to handle varied datasets, identify nuanced hate speech, and scale to meet social media platform needs. The fifth section finishes with HateSonar's involvement in improving social media content control. Multilingual support, real-time user input to improve identification algorithms, and multimedia information including images and videos are future goals.

2. Literature Survey

Online Hate Speech Identification Using Machine Learning. Research has focused on using machine learning to detect online hate speech. The system processes and categorises user-generated information to correctly identify hate speech across digital channels. Machine learning algorithms increase identification accuracy, improving content filtering and internet safety. This strategy emphasises the importance of modern technology in reducing online hate speech as part of a larger effort. [1]. Harm Metrics for Multimedia Online Hate Evaluation. Harm measures for hate speech assessment are lacking in multimedia communication platforms. This study emphasises the need for such measurements, especially for video and audio material. Harm metrics improve regulatory frameworks and internet user safety by assessing hate speech's negative impacts [2]. Multilingual Offensive Language Detection: A Comprehensive Survey. Given linguistic variety, offensive language identification across languages is a priority. The study examines multilingual offensive content detection technologies, which are necessary for inclusive moderation. Multilingual detection techniques make global sites safer and enforce community norms across languages [3]. Impact of Partisan Messaging on COVID-19 Social Distancing Adherence. The COVID-19 pandemic's social distancing rules showed that partisan social media messaging may influence public behaviour. This investigation shows how political statements affected health adherence. Digital communication affects public health responses, emphasising the necessity for ethical communications during health emergencies. [4].

Common Crawl Corpus Geospatial Data Analysis. The Common Crawl Corpus offers unique geographical data analysis. This study sheds light on worldwide tendencies, particularly in social media, using large-scale web-scraped data. Spatial trend analysis may be used to many sectors, improving knowledge of geospatial data's function in digital communication [5]. Online content hate speech detection using machine learning. Machine learning methods have helped identify online hate speech. This work highlights the usefulness of supervised learning models in spotting objectionable language, helping create content quality algorithms. Hate speech identification improves online safety, supporting worldwide digital harassment prevention [6]. An analysis of COVID-19-related hate speech using social media memes. Memes are spreading COVID-19-related hate speech unexpectedly. Analysis of Twitter meme-sharing habits shows how visual media promotes hate speech. Social media companies must identify these tendencies to combat harmful non-text material, reflecting the changing nature of online communication difficulties. [7]. COVID-19 Sinophobia and Radicalisation Indicators. Sinophobia increased on Twitter and Reddit throughout the outbreak. This research shows how platform-based interventions might reduce xenophobia by detecting radicalisation. Understanding online mood fluctuations is crucial to reducing health crisis-fueled discrimination on popular sites [8].

Assessing COVID-19 Hate Speech Tweet Emotions and Trends. A tweet-emotion classification method has helped researchers grasp COVID-19 hate speech. This research examines the emotional undertones of hostile or insulting tweets to assess regional public attitude. Emotion categorisation helps detect negative and good public discourse changes, improving moderation and action [9]. Algorithmic and Qualitative Analysis of Twitter Anti-Gypsy Hate Speech. Combining algorithmic and qualitative research may reveal hate speech targeting particular populations, such as anti-Gypsy sentiment. Researchers use Twitter data to study the challenges of identifying and categorising ethnic minority hate speech. Advanced, context-sensitive solutions for hate speech are needed after integrating algorithmic and human-centric methodologies [10]. Explore social media annotator perspectives on misogynoir. The research examines misogynoir and annotators' viewpoints on gender and racial hate speech detection. By evaluating content annotation biases, it shows the need for diverse, well-trained annotator teams to accurately identify

DOI: https://doi.org/10.5281/zenodo.19315956

intersectional hate speech. This study is crucial to inclusive and equitable social media regulation [11]. Studying Far-Right Hate Speech on Gab: Methodological Challenges. Researching far-right rhetoric on Gab presents methodological, epistemological, and legal issues. This study examines content moderation discrepancies and privacy issues researchers confront. Understanding hate speech dynamics across social media platforms, even those with inadequate moderation, requires analysing hate speech on decentralised or fringe sites [12].

Age-Themed Cinema Dialogue Emotional Mapping. Exploring how age-themed film conversations reflect emotional undertones reveals social views on ageing. Ageism discourses are informed by this study's analysis of cinema storylines and terminology about older characters. Emotional mapping shows biases and preconceptions that shape public views and culture [13]. Hate Speech Detection in Published Content using Machine Learning.

This work refines methods for automatic moderation by using machine learning to recognise hate speech in internet material. Enhanced hate speech recognition methods reduce online abuse, making the internet safer. Offence language recognition on numerous platforms is becoming more difficult, although content moderation machine learning algorithms are improving [14]. Improved Hate Speech Detection using Cross-Platform Data. A more diversified dataset from numerous social media sites may enhance hate speech detection systems. This technique studies how cross-platform analysis improves language pattern and sentiment recognition across user populations. These findings help create more comprehensive hate speech detection technologies [15]. Twitter User Hate Speech Index Improves Hate Speech Detection. A Twitter user hate speech index enhances detection accuracy. This index-based technique lets computers account for user-specific language use, improving real-time hate speech detection. Personalised metrics enhance this method's comprehension of social media abusive language patterns [16].

Social Media Prediction of Targeted Violence. Social media analysis to forecast targeted violence is becoming popular. Through linguistic patterns and content, this research emphasises the need of early detection systems in spotting dangers. Predictive modelling using social media data helps address online behaviours that may lead to offline damage [17]. Twitter Political Discourse Changed by Machine Learning. Machine learning is changing Twitter politics, reflecting digital conversation. The research examines how algorithms affect user engagement and content reach, especially in political circumstances. Machine learning may improve political discourse online by regulating poisonous or provocative words [18]. Analytical Approach to Tweet User Engagement and Toxicity. The association between user engagement and tweet toxicity shows how inflammatory material influences user behaviour. Higher involvement with poisonous information shows moderating techniques to hide bad messages. Engagement patterns and toxicity measurements can improve social media regulations and interactions [19]. DataLab Platform for Trend Analysis and Intervention. DataLab is a complete platform for investigating data patterns and performing solutions, especially in social media. It helps researchers and analysts find linguistic trends and intervention areas by offering sophisticated data management capabilities. DataLab helps enhance content quality and user experience using data-driven insights [20].

3. Proposed Methodology

The suggested system utilizes HateSonar, a Python-based program for hate speech detection, integrated inside a cloud infrastructure. Integration with current systems is facilitated via adaptable API connectors, allowing for simple implementation. Figure 1 shows how HateSonar gathers and preprocesses social media data for hate speech identification. Social media networks like Twitter, Facebook, and Reddit broadcast comments, posts, and hashtags. APIs provide data into a secure cloud repository. Nuisance elimination, tokenisation, and data structuring for analysis are part of preprocessing.

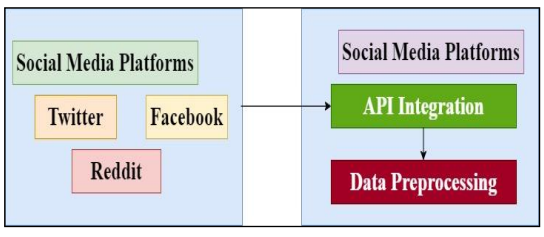

Figure.1. Block Diagram of Data Collection and Preprocessing

DOI: https://doi.org/10.5281/zenodo.19315956

A. Cloud Infrastructure and Data Governance

A resilient cloud infrastructure supports the system, providing scalable resources, secure storage, and real-time processing capabilities. Figure 2 shows HateSonar's hate speech detection methodology. Pre-processed data is sent to hate speech-classification machine learning algorithms. These models classify material as hostile, offensive, or neutral. Cloud storage stores results for analysis and visualisation. The workflow uses powerful NLP to understand text context. These algorithms let HateSonar accurately recognise and categorise hate speech in real time. This scalable framework detects dangerous information across numerous social media platforms and allows real-time moderation.

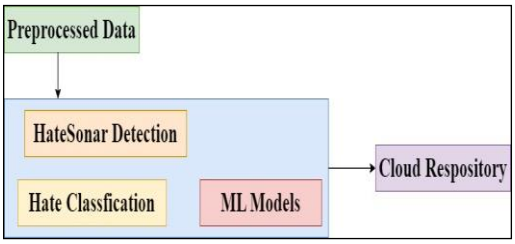

Figure.2. Block Diagram of Hate Speech Detection Workflow

B. Deployment of Machine Learning Models

HateSonar utilizes pre-trained machine learning models to detect hate speech in many languages and circumstances. The models are deployed in the cloud, facilitating real-time inference and updates. A bespoke pipeline is established to preprocess data via text tokenization, elimination of stop words, and transformation into numerical vectors. The classification algorithm uses these vectors to identify objectionable information, labeling it as hate speech, offensive language, or neutral. The system facilitates the incorporation of newer, more sophisticated models as they emerge, ensuring precision and pertinence. Figure 3 shows HateSonar's reporting architecture block diagram. Moderators and analysts use interactive dashboards to view detection findings. The dashboards provide hate speech trends, category breakdowns, and flagged material notifications.

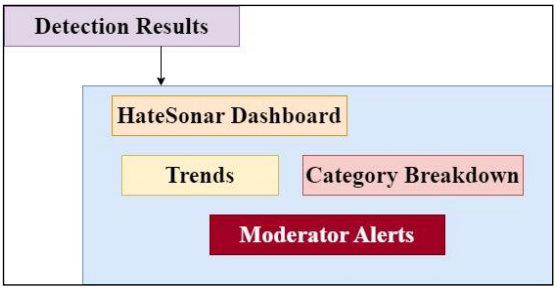

Figure.3. Block Diagram of Dashboard for Monitoring and Reporting

C. Immediate Surveillance and Notifications

The solution includes a real-time monitoring dashboard that offers insights into current hate speech patterns across observed platforms. Alerts are sent upon the detection of hate speech material, facilitating prompt replies. Alerts may be tailored to inform moderators, legal teams, or law enforcement authorities according to severity levels. The dashboard further facilitates visualization capabilities to exhibit trends, user engagement, and flagged material, assisting stakeholders in making educated judgments. Figure 4 shows the HateSonar-based hate speech detection system's data flow. APIs capture Twitter, Facebook, and Reddit comments and postings. Data for analysis is cleaned, tokenised, and formatted from these raw sources.

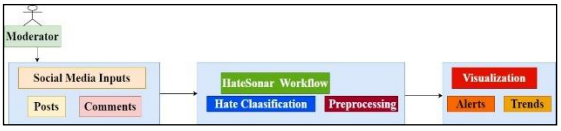

Figure.4. Data Flow Diagram: Hate Speech Detection Workflow

D. Multilingual Support

The algorithm is designed to identify hate speech across many languages used on social media. HateSonar is augmented by supplementary language models developed on datasets tailored to regional languages and dialects. A language detection module ascertains the language of the incoming text and directs it to the corresponding model. Cloud-based multilingual processing guarantees the system's scalability and adaptability to new language problems. Figure 5 shows the whole HateSonar ecosystem for social media hate speech detection and management. HateSonar's machine learning models preprocess and

DOI: https://doi.org/10.5281/zenodo.19315956

analyse data from numerous platforms. Moderators and analysts may see category breakdowns and trends on dashboards.

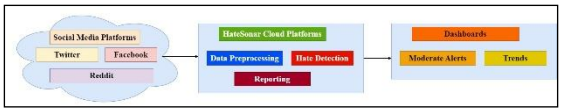

Figure.5. Overview Diagram of HateSonar Ecosystem for Hate Speech Detection

Table 1 of HateSonar is a robust cloud-based instrument designed to identify hate speech on social media networks. Notable characteristics include multilingual support, enabling the identification of dangerous information across many languages, thus ensuring worldwide applicability. Sentiment analysis ascertains the purpose behind postings, enabling sophisticated moderating.

Table 1. Features Of Hatesonar for Cloud-Based Hate Speech Detection

Feature	Description	Benefit	Application	Outcome
Multilingual Support	Detects hate speech in various languages	Widens applicability	Suitable for global platforms	Enhanced content moderation
Sentiment Analysis	Identifies the emotional tone of messages	Distinguished hate speech intent	Supports policy enforcement	Accurate moderation decisions
Scalability	Handles large-scale social media data	Accommodates growing user bases	Applicable to high-traffic sites	Real-time detection
Text Classification	Categorizes content into hate or neutral speech	Streamlines moderation processes	Automates filtering	Reduces manual workload
Integration Capability	Seamlessly integrates with social media APIs	Simplifies deployment	Easy adoption by platforms	Faster implementation

E. Sentiment and Context Analysis

The approach integrates sentiment and contextual analysis to improve the precision of hate speech identification. Sentiment analysis determines the emotional tone of information, while context analysis examines the surrounding language to discern purpose. These layers reduce false positives and negatives by enabling the algorithm to differentiate between hate speech and innocuous information. The cloud architecture executes these analyses concurrently, guaranteeing rapid processing times for extensive datasets.

F. Automation of Moderation Workflow

Automation is used to optimize content moderation processes. Content that has been flagged is prioritized according to its severity, with recommended actions for moderators including warnings, bans, or content removal. The automation lowers manual labor and guarantees uniform adherence to community norms. Additionally, the technology offers configurable moderation rules adapted to individual platform needs.

4. Results and Discussion

A. Model Training and Continuous Learning

Continuous learning processes are included to enhance model performance over time. User comments, moderator determinations, and evolving hate speech patterns are used to retrain and refine the HateSonar models. Cloud architecture enables the storage and processing of training datasets, enabling efficient deployment of model updates. This flexibility guarantees the system's efficacy against emerging trends of hate speech. The system includes educational materials to improve awareness and combat hate speech. Interactive modules, lessons, and campaigns are offered via a specific portion of the site. Cloud-hosted material guarantees accessibility and facilitates continuous updates of resources. The system further facilitates collaborations with entities dedicated to digital literacy and the prevention of hate speech. Figure 6 shows HateSonar hate speech detection data for five campaigns. The amount of hostile, offensive, and neutral comments, total comments, and detection accuracy percentages are metrics. Campaign 1 recognised 452 nasty comments with 92% accuracy.

DOI: https://doi.org/10.5281/zenodo.19315956

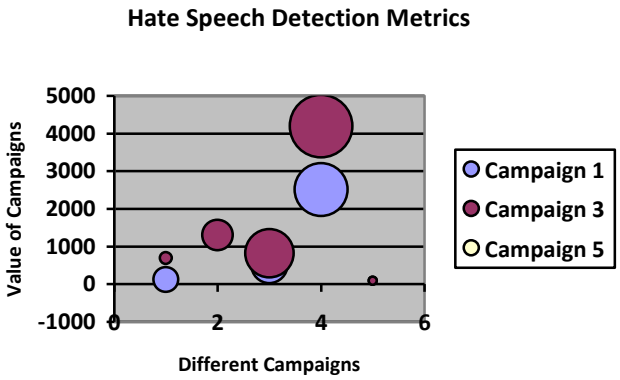

Figure. 6. Hate Speech Detection Metrics

Table 2 of HateSonar's process comprises five essential phases. Data gathering acquires user-generated material from platforms using APIs and web scraping techniques. The preprocessing phase purges raw text data of extraneous elements. Feature extraction discerns linguistic and contextual indicators of hate speech, transforming text into feature vectors. Detection uses machine learning techniques to categorise text as hateful or non-hateful.

Table 2. Workflow Of Hatesonar for Hate Speech Detection

Step	Description	Tools Used	Data Input	Output
Data Collection	Aggregates data from social media platforms	APIs and web scrapers	User-generated posts	Structured datasets
Preprocessing	Cleanses and formats text data	NLP tools	Raw text data	Filtered text
Feature Extraction	Extracts linguistic and contextual features	Text vectorization models	Preprocessed data	Feature vectors
Hate Speech Detection	Classifies text into hate or non-hate speech	Machine learning algorithms	Feature vectors	Labeled outputs
Reporting	Generates insights for moderation teams	Dashboards	Labeled outputs	Actionable analytics

B. Adherence to Policies and Regulatory Standards

Adherence to local and international rules is an essential element of the system. User data anonymization, secure log management, and audit trails are established to comply with regulatory requirements. Emerging technologies such as edge computing and federated learning may also be utilized to boost performance and privacy. Figure 7 categorises hate speech on five social media sites by race, politics, and gender. The proportion of hate speech recognised by each category is measured. For instance, Reddit has 40% racial hatred and 25% political hate. This report details hazardous material on each platform. Insights enable moderators and analysts target hate speech, making the internet safer.

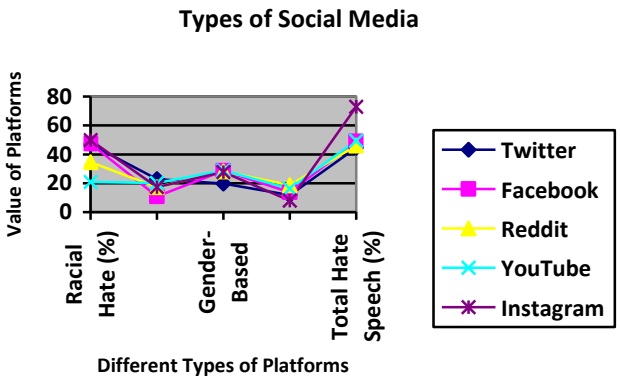

Figure.7. Category Breakdown Analysis

DOI: https://doi.org/10.5281/zenodo.19315956

C. Partnership with External Organizations

Collaborations with academic institutions, NGOs, and governmental organizations augment the system's efficacy. Collaborators provide access to specialized datasets, professional insights, and other resources for addressing hate speech. Cloud-based collaboration technologies promote safe and fast data exchange across stakeholders, encouraging a united strategy to hate speech prevention. Figure 8 shows hate speech detection temporal dispersion across five platforms. Data includes daily, weekly, and monthly frequencies and peak activity days and hours. Sundays are peak days on Reddit, with many angry comments in the evening. These data reveal user behaviour patterns, allowing preventive interventions during high-activity times. Social media networks may distribute moderation efforts and prevent hate speech surges using temporal insights.

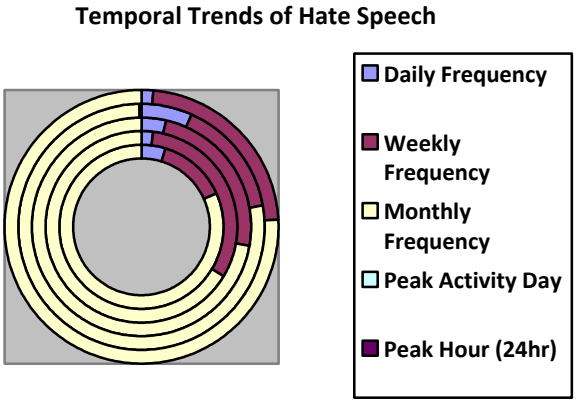

Figure.8. Temporal Trends of Hate Speech

D. User Reporting Interface

A user-centric reporting interface has been created to allow platform users to easily report hate speech. Reports are evaluated using HateSonar's categorization features, and legitimate instances are elevated for further action. The interface enables users to monitor the progress of their reports, therefore augmenting transparency and confidence. Cloud-based interfaces provide uniform performance across devices and locations. Sophisticated preprocessing systems adjust to the variety of data types present on social media networks. This phase employs Natural Language Processing (NLP) approaches to standardize text, addressing slang, acronyms, and emojis often seen in hate speech. Domain-specific preprocessing is added to cater for platform-specific norms and peculiarities. The cloud framework facilitates pipeline modification, guaranteeing alignment with platform APIs and data formats.

E. Analysis of Historical Data

Historical data analytics modules uncover long-term patterns in hate speech across social media platforms. These findings underpin policy modifications, content regulation tactics, and the formulation of public awareness initiatives. Cloud-based analytics tools easily analyze and show historical data, allowing comprehensive reporting and actionable insights. AI-driven classification allows the system to aggregate reported information into categories such as political hate speech, racial insults, or gender-based discrimination. This classification aids moderators in prioritizing and addressing certain forms of hate speech with more efficacy. Machine learning models developed on varied datasets guarantee precise categorization, bolstered by scalable cloud infrastructure.

5. Conclusion

HateSonar for cloud-based hate speech identification on social media platforms improves automated content moderation, but it has drawbacks. Language complexities and cultural differences make it hard to recognize all hate speech. Managing privacy when analyzing massive volumes of user-generated material is also important. HateSonar makes online settings safer by improving the speed and scalability of harmful material detection. No system is perfect at differentiating objectionable language from genuine debate, therefore false positives and negatives are a problem. The technology will be improved for contextual comprehension, applied to many languages and platforms, and updated with user input to identify hate speech more accurately.

DOI: https://doi.org/10.5281/zenodo.19315956

Five parameters are taken from HateSpeech_Detection dataset for five campaigns. Hateful Comments: 124–694, Offensive Comments: 340–1302, Neutral Comments: 506–2943, Total Comments: 2509–4855, Detection Accuracy (%): 90.93–97.47. Five parameters are taken from Category_Breakdown dataset for five platforms. Racial Hate (%): 20.91–49.87, Political Hate (10.76–22.93), Gender-Based Hate (19.68–28.71), Other Categories (8.03–18.58), Total Hate Speech (44.52–72.96). Data from Temporal_Trends for 5 platforms Five parameters are employed. Daily frequency is 133–452, weekly frequency is 995–2817, and monthly frequency is 4715–8437. Monday through Saturday are peak activity days, and 6-13 is peak hour.

References

1. O. Tverdokhlib, V. Vysotska, P. Pukach and M. Vovk, “Information Technology for Identifying Hate Speech in Online Communication Based on Machine Learning,” in Data-Centric Business and Applications: Modern Trends in Financial and Innovation Data Processes, vol. 1, pp. 339-369, 2024.
2. J. Omidosu and M. Kyobe, “View No Evil and Listen to No Evil: The Need for Harm Metrics to Evaluate Online Hate on Multimedia Communication Channels,” IEEE International Conference on Advanced Video and Signal Based Surveillance , pp. 1-8, 2024.
3. S. R. Thumala, H. Madathala and V. M. Mane, "Azure Versus AWS: A Deep Dive into Cloud Innovation and Strategy," International Conference on Electronics and Renewable Systems (ICEARS), pp. 1047-1054, 2025.
4. R. Govind, N. Garg, and L. Carter, “Selling Hope Versus Hate: The Impact of Partisan Social Media Messaging on Social Distancing During the COVID-19 Pandemic,” European Journal of Marketing, vol. 58, no. 2, pp. 632-658, 2024.
5. H. Madathala, S. R. Thumala, and G. Yeturi, “Optimizing cloud migration: Designing robust architectures for seamless transition from on-premises to Azure for SAP and database systems,” International Journal of Engineering Technology Research & Management, vol. 9, no. 1, 2025.
6. C. Sinyangwe, D. Kunda and W. P. Abwino, “Detecting Hate Speech and Offensive Language Using Machine Learning in Published Online Content,” Zambia ICT Journal, vol. 7, no. 1, pp. 79-84, 2023.
7. N. Vishwamitra, K. Guo, S. Liao, J. Mu, Z. Ma, L. Cheng, Z. Zhao and H. Hu, “Understanding and Analyzing COVID-19-Related Online Hate Propagation Through Hateful Memes Shared on Twitter,” in Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, pp. 103-107, 2023.
8. [8] M. Costello, N. Vishwamitra, S. Liao, L. Cheng, F. Luo and H. Hu, “COVID-19 and Sinophobia: Detecting Warning Signs of Radicalization on Twitter and Reddit,” Cyberpsychology, Behavior, and Social Networking, vol. 26, no. 7, pp. 546-553, 2023.
9. S. R. Thumala, V. Mane, T. Patil, P. Tambe and C. Inamdar, "Full Stack Video Conferencing App using TypeScript and NextJS," 3rd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), pp. 1285-1291, 2025.
10. S. Pasta, “Hate Speech Research: Algorithmic and Qualitative Evaluations. A Case Study of Anti-Gypsy Hate on Twitter,” Research on Education and Media, vol. 15, no. 1, pp. 130-139, 2023.
11. V. Ramesh, “Performance benefits of reactive frameworks,” International Journal of Computer Applications, vol. 975, pp. 8887, 2025.
12. N. Ganesh, A. Sriram, S. N. Krishnan, and T. S. Rao, “Simultaneous enhancement and detection of brain tumors using GAN,” in Intelligent Computing—Proceedings of the Computing Conference. Cham, Switzerland: Springer Nature Switzerland, pp. 206–220, 2025.
13. D. E. Takan, S. Takan and K. O. Paker, “Yaşlılık Temalı Sinema Filmlerinde Diyalogların Duygu Haritası,” İletişim Kuram ve Araştırma Dergisi, vol. 66, pp. 1-20, 2024.
14. C. Sinyangwe, D. Kunda, and W. P. Abwino, “Detecting Hate Speech and Offensive Language Using Machine Learning in Published Online Content,” Zambia ICT Journal, vol. 7, no. 1, pp. 79-84, 2023.
15. S. R. Thumala, “Importance of business continuity and disaster recovery (BCDR) methodologies for organizations: A comparison study between AWS and Azure,” International Journal of Science and Research (IJSR), vol. 11, no. 12, pp. 1406–1415, 2022.
16. V. Ramesh, “Evaluating Apache Kafka performance and operational efficiency: A

DOI: https://doi.org/10.5281/zenodo.19315956

comparative study of ZooKeeper and KRaft architectures," International Journal of Computer Applications, vol. 187, no. 46, pp. 12–18, 2025.
17. L. Kaati, A. Shrestha and N. Akrami, "Predicting Targeted Violence from Social Media Communication," 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 383-390, 2022.
18. H. Madathala, S. R. Thumala, B. Barmavat, and K. K. S. Prakash, "Functional consideration in cloud migration," International Peer Reviewed/Refereed Multidisciplinary Journal (EIPRMJ), vol. 13, no. 2, 2024.
19. N. Salehabadi, A. Groggel, M. Singhal, S. S. Roy and S. Nilizadeh, "User Engagement and the Toxicity of Tweets," arXiv preprint arXiv:2211.03856, pp. 1-18, 2022.
20. Y. Xiao, J. Fu, W. Yuan, V. Viswanathan, Z. Liu, Y. Liu, G. Neubig and P. Liu, "Datalab: A Platform for Data Analysis and Intervention," arXiv preprint arXiv:2202.12875, pp. 1-14, 2022.

DOI: https://doi.org/10.5281/zenodo.19315956