Comprehensive Survey on Towards a Multi–Modal system for the detection of Cyberbullying and Fake Accounts in Social Networks

Mr. S. Sivaraman[1], Nandhitha S[2], Dr. N. Mohanapriya[3]

[1]Assistant Professor Department of Computer Science and Engineering, Vivekanandha College of Engineering for Women, Tiruchengode, Tamil Nadu, India.

[2]PG Scholar, Department of Computer Science and Engineering, Vivekanandha College of Engineering for Women, Tiruchengode, Tamil Nadu, India.

[3]Associate Professor Department of Computer Science and Engineering, Vivekanandha College of Engineering for Women, Tiruchengode, Tamil Nadu, India.

Abstract: **As social networks increasingly dominate everyday life, cyberbullying has become a significant concern, disproportionately affecting younger individuals. Cyberbullying causes severe emotional harm, leading to distress, anxiety and long-term psychological trauma. Teens propose permanent bans and criminal charges to combat cyberbullying. The current cyberbullying detection system relies too heavily on text analysis, limiting its effectiveness. It struggles to incorporate non-textual data like images, videos and comments, hindering detection of complex cases. The existing cyberbullying detection system has major flaws, including limited scope and outdated technology. It struggles to detect complex cases, neglecting non-textual data and leaving vulnerable users exposed. A comprehensive, advanced system is urgently needed to address these shortcomings and effectively protect users. CyberGuard, an advanced system, detects cyberbullying through multi-modal analysis and fake account identification. It integrates text, image/video and behavioral analysis for accurate detection and real-time monitoring. CyberGuard enhances user safety, reduces false reports, and paves the way for future enhancements in social media security.**

Index Terms: **Cyberbullying, Multi-modal analysis, Fake account identification, Behavioural analysis, Real-time monitoring, social media security**

I. INTRODUCTION

As social networks become increasingly embedded in modern communication, issues like cyberbullying and the proliferation of fake accounts have become pressing concerns. Cyberbullying, characterized by online harassment, threats, and defamation, affects millions globally, especially young users who are vulnerable to emotional and psychological harm. Fake accounts, on the other hand, are often created to deceive or impersonate others, facilitating harmful behaviours like spreading misinformation, perpetrating scams, and engaging in cyberbullying anonymously. Despite advancements in AI and machine learning, existing detection systems are limited in their accuracy and ability to adapt to new tactics used by malicious actors. Furthermore, while some platforms employ algorithms to identify and filter harmful content or suspicious profiles, they often fail to address the nuanced language and evolving methods of deception that define cyberbullying and fake accounts. Machine learning, data analysis and Natural Language Processing (NLP) are powerful tools for opposing cyberbullying and detecting fake accounts. Machine learning leverages techniques like NLP to analyse text, sentiment and behavioural patterns, quickly identifying harmful content and suspicious users. Data analysis further enhances this process by examining user interactions, activity patterns and account metadata to detect abusive behaviour and anomalies in real-time. NLP specifically identifies abusive language, hate speech and spam, while also flagging fake profiles with generic bios or unnatural language. These automated systems improve platform security and user trust by reducing reliance on manual moderation. However, they must be carefully designed to avoid bias and respect user privacy, fostering a safer and more reliable online environment.

II. LITERATURE REVIEW

[1] Umita Deepak Joshi, several studies focus on detecting fake profiles, particularly on platforms like

Twitter. Fake profiles are often created by humans, bots, or cyborgs and spread harmful content like rumours or phishing attempts. Machine learning algorithms, including Neural Networks, Random Forest, XGBoost and LSTM, have been tested using datasets like the MIB dataset of Twitter profiles. XGBoost has been found to achieve high accuracy (96%), while Neural Networks also show strong performance (95%) in distinguishing real from fake profiles. Blocking or deleting fake profiles plays a crucial role in mitigating cybersecurity risks and enhancing online safety. [2] Zeinab Shahbazi, Fake news is a significant issue on social media and several efforts have been made to address it through the integration of blockchain and Natural Language Processing (NLP) with machine learning. A proposed system leverages blockchain's decentralized framework to ensure the authenticity of information, while smart contracts and the Proof-of-Authority protocol help maintain the trustworthiness of digital content. The solution aims to prevent the spread of fake news and improve social media security by validating users, ensuring accountability, and addressing misinformation. [3] Dr. Vijayakumar. V, With the rise of cyberbullying incidents on social media, automated detection methods have gained importance. A hybrid deep neural network model combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) has been proposed to detect cyberbullying in text and images, achieving accuracies of 86% and 85%, respectively. This highlights the potential of deep learning in improving cyberbullying detection across various media types. [4] Sandip Bankar, Research has also focused on detecting cyberbullying on Twitter using sentiment analysis combined with machine learning models like Support Vector Machines (SVM) and Term Frequency-Inverse Document Frequency (TF-IDF). This approach achieves 92% accuracy, providing a practical way for platform administrators to address harmful content quickly. The ultimate goal is to foster a safer, more respectful online environment by using automated tools to detect and prevent cyberbullying. [5] Nitika Kadam, another study aims to detect fake profiles early using machine learning models. By employing algorithms such as C4.5, Naive Bayes, SVM, Artificial Neural Networks (ANN) and K-Nearest Neighbours (KNN), the research explores preprocessing and dimensionality reduction techniques. The study concludes that using an 80-20% training-testing ratio reduces resource consumption, while a 70-30% ratio improves classification accuracy. [6] Naman Singh, while many solutions have been effective at detecting bot or cyborg accounts, distinguishing between real and fake accounts created by humans remains a challenge. Recent advancements in machine learning have led to breakthroughs in detecting human-created fake accounts. By training models on datasets of labelled fake and real accounts, machine learning models can accurately classify human-created fake accounts, providing a promising solution for social media platforms. [7] Mahmoud Ahmad al-Khasawneh, Cyberbullying detection has also extended beyond text analysis to include multi-modal approaches that combine data from photos, videos, comments and temporal information. Using hierarchical attention networks and bidirectional LSTM with attention, researchers have developed models that outperform existing ones by incorporating diverse media types. This multi-modal approach shows that analysing various forms of content is crucial for improving detection accuracy and addressing the complexity of cyberbullying. [8] Mohammed Hussein Obaida, A deep learning model focused on detecting cyberbullying among young users has been tested on datasets from Twitter, Instagram and Facebook. Using Long Short-Term Memory (LSTM) networks, the model achieved high accuracies: 96% for Twitter, 94% for Instagram and 91% for Facebook. This study highlights the adaptability of the model across different social media platforms and the importance of deep learning for feature extraction in cyberbullying detection. [9] Andrea Pereraa, In the context of cyberbullying detection, a supervised machine learning system has been proposed that employs techniques such as SVM, Logistic Regression (LR), sentiment analysis and N-gram analysis. This system detects multiple categories of cyberbullying, including race, physical appearance, sexuality and politics, using features like TF-IDF and profanity detection. The system's accuracy is achieved by training on a 70-30% data split, and future research aims to incorporate real-time detection and cross-platform monitoring. [10] Sarah Khaled, to address fake accounts, a novel SVM-NN algorithm combining Support Vector Machines (SVM) and Neural Networks (NN) has been proposed. This hybrid approach shows great promise for detecting fake Twitter accounts and bots by efficiently selecting features, reducing data dimensions, and using machine learning classification. The SVM-NN method outperforms standalone SVM and NN algorithms, achieving 98%

accuracy and offering a reliable solution for improving online security and trust.

Table 1: Performance Metrics

Title	Techniques	Platform/Dataset	Accuracy (%)	Concept
Fake Profile Detection	Neural Networks, XGBoost	MIB Twitter dataset	XGBoost: 96%, Neural Networks: 95%	High accuracy in identifying fake profiles
Integrated System (Blockchain and NLP)	Reinforcement Learning, Blockchain, Smart Contracts	Facebook	85%	Combats fake news, fake users, and posts
Cyberbullying Detection (Text and Image)	CNN and LSTM (Hybrid Deep Neural Network)	Instagram	Text:86%, Image: 86%	Detection of cyberbullying in text and images
Cyberbullying Detection (Twitter)	SVM and TF-IDF (Sentiment Analysis)	Twitter	92%	Practical detection for cyberbullying
Fake Profile Detection	C4.5, Bayes, SVM, ANN, KNN	Twitter dataset	80-20% training-testing ratio: 70% accuracy	Improves classification accuracy by adjusting ratio
Fake Profile Detection (Human-created)	Supervised ML (SVM, ANN)	Twitter, Instagram	90%	Targets human-created fake accounts
Multi-modal Cyberbullying Detection	LSTM and Hierarchical Attention Networks and MLP	Real-world datasets	Outperforms existing models	Uses photos, videos, comments, temporal data
Cyberbullying Detection (LSTM)	LSTM (Deep Learning)	Twitter, Instagram, Facebook	Twitter: 96%, Instagram: 94%, Facebook: 91%	High accuracies across multiple platforms
Cyberbullying Detection (SVM, LR, Sentiment Analysis)	SVM, Logistic Regression, N-gram, Sentiment Analysis	Twitter, Reddit	87%	Detects four cyberbullying categories
Fake Account Detection (SVM and NN)	SVM and NN (SVM-NN)	Twitter	94%	Outperforms standalone SVM and NN algorithms

III. ANALYSIS

Detecting cyberbullying and fake accounts on social media requires a combination of traditional machine learning (ML) and advanced deep learning (DL) techniques. Algorithms like SVM paired with feature extraction methods like TF-IDF or N-gram are effective for classifying abusive text. Neural networks, particularly LSTM, excel at detecting sequential patterns in conversations, while hybrid models like CNN-LSTM analyse both spatial and temporal aspects of content. For fake account detection, XGBoost identifies patterns in user behaviour and profile attributes and ANN captures complex relationships in data. Traditional methods like Bayes and C4.5 Decision Trees provide interpretable results, while modern advancements leverage Graph Neural Networks (GNNs) and Hierarchical Attention Networks (HAN). Emerging technologies like Reinforcement Learning, Blockchain, and Smart Contracts enhance detection by introducing dynamic adaptability and transparency. Integrating multi-modal data and explainable AI could further improve detection accuracy and robustness. To detect cyberbullying, LSTM and CNN-LSTM hybrids are highly valued for analysing sequential text and multi-modal content, while Hierarchical Attention Networks (HAN) enhance interpretability. In fake account detection, Graph Neural Networks (GNN) excel at analysing relational patterns and

XGBoost is effective for structured data. Reinforcement Learning (RL) adds adaptability, learning new behaviour dynamically. Hybrid approaches combining these algorithms are emerging as the most robust solutions.

IV. CONCLUSION

The rise of fake accounts, misinformation and cyberbullying on social media platforms poses significant challenges to online security and user trust. However, advances in machine learning and deep learning have led to the development of effective detection systems, ranging from identifying fake profiles to detecting harmful content. Techniques such as XGBoost, Neural Networks, LSTM, and hybrid models like CNN-LSTM for cyberbullying detection have demonstrated high accuracies, offering significant promise in mitigating these risks. Additionally, integrating technologies like blockchain for fake news detection and the combination of SVM-NN algorithms for fake account detection further enhance security and privacy on social media platforms. These solutions have the way for safer online environments by addressing both automated and human-created threats. As these models continue to evolve and integrate multi-modal data, future research should focus on refining their accuracy, expanding to real-time applications and ensuring their ability to handle emerging threats effectively.

REFERENCES

[1] Mahmoud Ahmad Al-Khasawneh, Muhammad Faheem, (member, ieee), Ala Abdulsalam Alarood, Safa Habibullah, and Eesa Alsolami Toward Multi-Modal Approach for Identification and Detection of Cyberbullying in Social Networks, 10.1109/ACCESS.2024.3420131 27 June2024. https://ieeexplore.ieee.org/document/10574823

[2] Mohammed Hussein Obaida, Saleh Mesbah Elkaffas, Shawkat Kamal Guirguis. Deep Learning Algorithms for Cyber-Bulling Detection in Social Media Platforms, 10.1109/ACCESS.2024.3406595, 28 May 2024. https://ieeexplore.ieee.org/document/10540101

[3] Cyberbullying Detection System on Social Media Using Supervised Machine Learning Author links open overlay panel Andrea Perera, Pumudu Fernando, Volume 239, 2024, Pages 506 - 516. https://www.sciencedirect.com/science/article/pii/S1877050924014431

[4] Sarah Khaled, Neamat El-Tazi, Hoda M. O. Mokhtar, Detecting Fake Accounts on Social Media, DOI: 10.1109/BigData.2018.8621913 December 2018, https://www.researchgate.net/publication/330629456_Detecting_Fake_Accounts_on_Social_Media

[5] Naman Singh, Tushar Sharma, Abha Thakral, Tanupriya Choudhury, Detection of Fake Profile in Online Social Networks Using Machine Learning, 10.1109/ICACCE.2018.8441713 23 August 2018, https://ieeexplore.ieee.org/document/8441713

[6] Nitika Kadam and Sanjeev Kumar Sharma Social Media Fake Profile Detection Using Data Mining Technique, doi: 10.12720/jait.13.5.518-523, March 1, 2022.

[7] Umita Deepak Joshi, Vanshika, Ajay Pratap Singh, Tushar Rajesh Pahuja, Smita Naval and Gaurav Singal, Fake Social Media Profile Detection, August 2021, DOI:10.1002/9781119769262.ch11, https://www.researchgate.net/publication/353892230_Fake_Social_Media_Profile_Detection

[8] Zeinab Shahbazi and Yung-Cheol Byun, Fake Media Detection Based on Natural Language Processing and Blockchain Approaches, DOI: 10.1109/ACCESS.2021.3112607, 14 September 2021, https://ieeexplore.ieee.org/abstract/document/9536745

[9] Multimodal Cyberbullying Detection using Hybrid Deep Learning Algorithms Dr. Vijayakumar V*1, Dr Hari Prasad D2, Adolf P3, https://dx.doi.org/10.37622/IJAER/16.7.2021.568-574

[10] Sandip Bankar, Surekha Janrao, Preeti Gupta, Rohini Patil, Cyberbullying Detection on Twitter using Sentiment Analysis, Vol. 20 No.3(2024), https://journal.esrgroups.org/jes/article/view/4992