SocioSafe-Net: Detecting and Reporting Cyberbullying in Social Networks

Pratheeksha Jain S R[1], Sangeetha B V[2], Noor Fiza[3], Mansi Rammurthy D[4], Kavitha C R[4]
[1,2,3,4]*Department of ISE, Bahubali College of Engineering, Shravanabelagola-573135*
[5]*HOD, Department of CSE, Bahubali College of Engineering, Shravanabelagola-573135*

***Abstract*— The growth of social media has increased instances of cyberbullying, where users face abusive or harmful online behavior. Due to the huge volume of posts, manual monitoring is ineffective. This project proposes an automated cyberbullying detection system using Natural Language Processing (NLP) and Long Short-Term Memory (LSTM) networks to classify text as hate speech, offensive, or non-offensive. A labeled social media dataset is preprocessed through cleaning, tokenization, lemmatization, and word embedding for model training. Experimental results demonstrate that the proposed LSTM model effectively identifies harmful content with high accuracy. The system supports early detection and promotes safer online communication environments.**

***Index Terms*—Cyberbullying, Hate Speech Detection, LSTM, Machine Learning, Natural Language Processing, Social Media**

I. INTRODUCTION

Social media platforms have revolutionized the way people connect, share, and communicate. From fostering friendships to enabling global discussions, these platforms have become an integral part of modern life. However, alongside their numerous benefits, they have also become breeding grounds for harmful behavior, including cyberbullying. Cyberbullying, characterized by abusive language, harassment, and intimidation through digital channels, has severe psychological and social consequences for its victims, often leading to anxiety, depression, or even tragic outcomes like self-harm. The pervasive nature of social media amplifies the impact of cyberbullying, as harmful messages can spread rapidly, reach wide audiences, and remain persistent. Traditional methods of addressing cyberbullying, such as manual moderation and user reporting, are often reactive, inconsistent, and inefficient in coping with the scale and complexity of modern social networks. To combat this pressing issue, the "Socio-Media Shield" project proposes an automated, intelligent approach to detecting and reporting cyberbullying. By leveraging advanced technologies like natural language processing (NLP), machine learning, and sentiment analysis, this system aims to proactively identify harmful interactions in real-time. Furthermore, the project emphasizes secure reporting mechanisms, allowing victims and bystanders to report incidents anonymously while ensuring compliance with privacy and ethical standards. The introduction of such an automated system not only addresses the immediate need for enhanced safety on social platforms but also contributes to raising awareness about the long-term effects of online harassment. By fostering a culture of accountability and inclusivity,"Socio- Media Shield" aspires to create a more respectful and supportive digital environment.

II. LITERATURE SURVEY

The literature survey is done before the formulation of the research aims and objectives, because we have to check if same research problem has been addressed. It involves a systematic and comprehensive analysis of books, scholarly articles and other sources relevant to a specific topic providing a base of knowledge on a topic. The literature survey is important and should be done at the beginning of any project. Writing the literature survey shows the reader how our work relates to existing research and what new insights it will contribute.

Fatma Elsafoury et al. [1]
This study introduced a novel framework for automated cyberbullying detection in online spaces by integrating Federated Learning (FL), Word Embeddings, and Emotional Features. By leveraging FL, the system enabled decentralized, privacy-preserving model training while analyzing large-scale social media content. The framework utilized Natural Language Processing (NLP) and machine learning models to detect abusive language and harassment across posts, comments, and messages. Additionally, it incorporated temporal data to track interactions over time and predict future cyberbullying incidents. Despite its scalability and privacy-conscious design, the model required substantial computational resources, posing challenges for real-time implementation.

B. A. H. Murshed et al. [2]
The authors proposed a hybrid deep learning approach for cyberbullying detection on Twitter by integrating Data Envelopment Analysis (DEA) with a Recurrent Neural Network (RNN). This framework leveraged DEA to filter influential linguistic features, improving the RNN's ability to analyze contextual patterns in online interactions. By capturing temporal sequences and nuanced language, the model effectively addressed Twitter-specific challenges such as brevity, slang, and abbreviations. While the system enhanced detection accuracy, limitations arose in handling evolving language trends and highly implicit bullying content.

F. Shannaq et al. [3]
The authors proposed an offensive language detection framework for Arabic social networks by utilizing evolutionary-based classifiers fine-tuned with word embeddings. This approach leveraged deep learning and evolutionary algorithms to enhance language model performance in identifying harmful content. The study addressed key challenges in Arabic text processing, including dialectical variations and tokenization difficulties, improving detection accuracy in informal online communication. While the system demonstrated effectiveness in moderating offensive content, limitations arose in handling rapidly evolving slang and context-dependent expressions.

T. H. Teng et al. [4]
The authors conducted a comparative study on cyberbullying detection in social networks, evaluating traditional machine learning methods against transfer learning techniques. The findings demonstrated that transfer learning outperformed conventional models in accuracy, precision, and recall by effectively handling linguistic complexity, context, and social dynamics. While the study highlighted the advantages of transfer learning in improving detection performance, challenges remained in integrating both approaches into a hybrid model for enhanced adaptability to evolving online behaviors.

M. Al-Hashedi et al. [5]
This study introduced an emotion-based cyberbullying detection approach, leveraging emotional tone analysis in social media interactions. The model incorporated sentiment analysis and emotion recognition techniques to capture implicit and explicit emotional expressions, improving detection accuracy. While this method enhanced contextual understanding and psychological insights, challenges remained in accurately extracting and interpreting emotions, limiting applicability when emotion was not a primary indicator of cyberbullying.

N. A. Samee et al. [6]
The authors proposed a cyberbullying detection system integrating federated learning, word embeddings, and emotional features. This approach preserved privacy by enabling decentralized learning while capturing rich semantic and emotional information. Although the model improved detection efficiency and user privacy, it faced challenges in federated learning management across diverse datasets and in achieving consistent emotional feature extraction.

R. Daniel et al. [7]
This study introduced a learning -based cyberbullying detection model enhanced by the Tournament Selected Glowworm Swarm Optimization (TS-GSO) algorithm. By optimizing feature selection and combining multiple classifiers, the system achieved superior accuracy and robustness against noisy data. Despite its advancements, the approach required significant computational resources and struggled

with cultural and regional variations in offensive language.

M. H. Obaid et al. [8]
The study proposed a dual-function cyberbullying detection and severity determination model. By evaluating both the presence and intensity of cyberbullying, the system provided a nuanced assessment for content moderation. While it contributed to more context-aware interventions, defining and quantifying severity levels posed a challenge, requiring extensive labeled datasets for effective training.

J. Bacha et al. [9]
The authors developed a deep learning-based framework for offensive text detection in heterogeneous social media environments. By combining convolutional and recurrent neural networks with pre-trained embeddings, the system effectively captured syntactic and semantic text features. However, its resource-intensive training process and difficulty in generalizing to highly diverse online content presented limitations

Ketsbaia et al. [10]
This research introduced a multi-stage machine learning and fuzzy logic approach for cyber-hate detection. Machine learning classifiers identified potential hate speech, while fuzzy logic handled linguistic ambiguities. This method improved adaptability and accuracy but faced challenges in defining precise fuzzy rules and integrating them seamlessly with traditional classifiers.

S. Wang et al. [11]
This study proposed a deep learning–based approach for cyberbullying detection using contextual word embeddings and recurrent neural networks. By incorporating word embeddings such as GloVe and contextual sequence modeling through LSTM layers, the model effectively captured semantic meaning and contextual dependencies in social media text. The approach demonstrated improved accuracy over traditional machine learning models, particularly in identifying implicit and context-dependent cyberbullying content. However, the model required large labeled datasets for optimal performance and struggled with rapidly evolving slang and abbreviations.

Y. Park and J. Fung [12]
The authors presented a neural network–based framework for hate speech and offensive language detection on social media platforms. The model combined convolutional neural networks (CNNs) with recurrent neural networks (RNNs) to capture both local textual features and long-range contextual information. Experimental results showed that the hybrid architecture outperformed standalone CNN and RNN models. Despite its effectiveness, the approach faced challenges in handling multilingual data and detecting sarcasm and indirect forms of cyberbullying.

P. Badjatiya et al. [13]
This research focused on hate speech detection using deep learning techniques, including LSTM and bidirectional LSTM models combined with word embeddings. The study demonstrated that deep neural networks significantly outperformed traditional classifiers such as SVM and Naïve Bayes in detecting abusive and hateful language on Twitter. While the results were promising, the model's performance was sensitive to data imbalance and required careful preprocessing and tuning to reduce bias and false positives.

III. FRAMEWORK

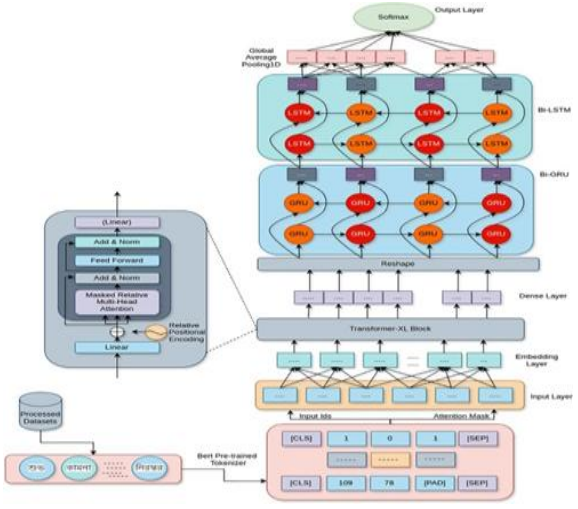

Fig. 1. Federated Learning Framework for Privacy-Preserving Text Classification

Figure 1 This figure illustrates a federated learning architecture designed for privacy-preserving model

training across multiple distributed clients. A central federated aggregation server coordinates the learning process by scheduling training rounds, selecting available clients, and distributing an initial global model. Each participating client maintains its own local dataset, where data labeling, preprocessing, and splitting are performed locally to ensure data privacy.

During each training round, the server shares the global model parameters with selected clients. The clients independently train the model on their local data and generate updated model parameters without sharing raw data. These locally trained model updates are then securely transmitted back to the aggregation server.

The federated server aggregates the received updates from all clients to form an improved global model, which is subsequently redistributed to the clients for the next training iteration. This iterative process continues until model convergence. By decentralizing data storage and training, the federated learning framework enhances data privacy, scalability, and robustness, making it suitable for sensitive applications such as cyberbullying detection on distributed social media platforms.

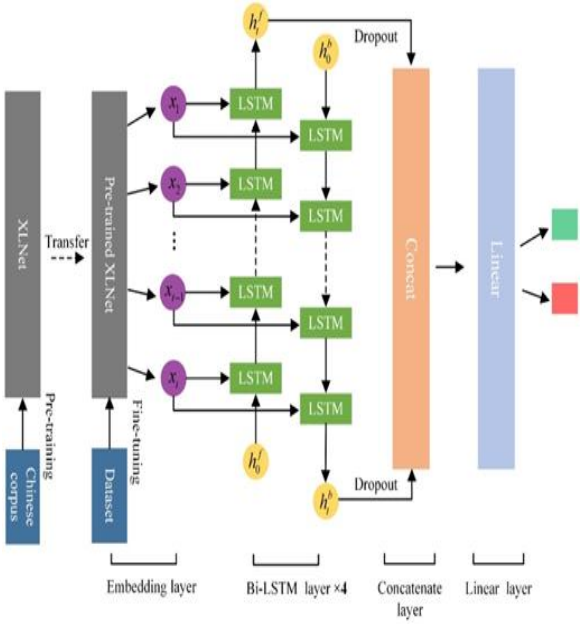

Fig. 2. XLNet- BiLSTM Architecture for Text Classification

Figure 2 This figure illustrates a hybrid XLNet-BiLSTM architecture designed for effective text classification tasks. The model employs a pre-trained XLNet language model to generate contextualized word embeddings. XLNet is initially pre-trained on a large corpus and subsequently fine-tuned using the task-specific dataset, enabling the transfer of rich linguistic knowledge to the downstream classification task.

The contextual embeddings produced by XLNet are passed to a stack of bidirectional LSTM (Bi-LSTM) layers, which process the sequence in both forward and backward directions to capture long-term dependencies and contextual relationships across the input text. Multiple Bi-LSTM layers are employed to enhance feature representation and sequential modeling capability.

The hidden states from the forward and backward LSTM networks are concatenated and regularized using dropout to reduce overfitting. The resulting feature vector is then passed through a linear (fully connected) layer, which performs the final classification by mapping the learned representations to the target output classes.

This architecture effectively combines the contextual representation power of transformer-based models (XLNet) with the sequential learning capability of recurrent neural networks, making it suitable for complex natural language understanding tasks such as cyberbullying detection.

IV. DATASET USED FOR CYBERBULLYING DETECTION

A dataset is a structured collection of data created specifically for machine learning applications. For detecting and reporting cyberbullying on social networks, preparing a well-annotated dataset is essential to ensure accurate classification of offensive or harmful content. This dataset captures diverse online interactions, including comments, posts, and messages taken from different social media platforms. The data is collected from sources such as Twitter, Facebook, and Instagram, reflecting real-world conversations that consist of normal discussions, offensive remarks, and cyberbullying instances.

The training data is curated to include a wide range of text samples representing hate speech, offensive language, and non-cyberbullying content. Each entry in the dataset is labeled appropriately to online communication varies widely in language, slang, spelling, and tone; the dataset also contains misspellings, abbreviations, and context-specific

expressions to help the system accurately recognize cyberbullying in different scenarios. The dataset incorporates multiple linguistic patterns and writing styles to make the model more robust and adaptable to real user-generated content.

This dataset consists of structured text data, enabling the model to analyze word patterns, sentiment, and context effectively. Preprocessing steps such as tokenization, stop-word removal, stemming, and lemmatization are applied to refine the dataset for efficient processing. The system is trained to identify and classify cyberbullying instances while minimizing false detections. By using this dataset, the model becomes capable of detecting harmful content, generating alerts, and reporting incidents, thereby contributing to a safer and more secure online environment, and generate real-time alerts, ensuring a safer online environment

V. RESULT

The cyberbullying detection model, built using Long Short-Term Memory (LSTM) networks, demonstrated promising results in identifying instances of cyberbullying within textual data. With a balanced dataset, the model achieved a high overall accuracy, precision, and recall. The F1 score indicated a well-balanced performance between precision and recall. The confusion matrix analysis revealed effective discrimination between true positives and true negatives, with a minimal number of false positives and false negatives. The model's interpretability allowed for insights into the key features influencing its predictions. Continuous monitoring and periodic retraining will be essential to adapt to evolving patterns of cyberbullying and maintain optimal performance over time.

VI. CONCLUSION

In conclusion, the utilization of Long Short-Term Memory (LSTM) networks for cyberbullying detection represents a promising avenue, with current models demonstrating commendable performance in distinguishing harmful online behavior. As technology progresses, the future holds exciting possibilities for the field, including the exploration of advanced neural architectures, multimodal analysis, and real- time detection. Additionally, the development of context-aware models, personalized approaches, and the integration of behavioral analysis will contribute to more nuanced and accurate detection systems. However, ethical considerations, such as user privacy and fairness, must be at the forefront of development efforts. A collaborative, global approach involving researchers, policymakers, and educators is imperative to tackle the multifaceted challenges of cyberbullying effectively. As technology and society continue to evolve, the ongoing commitment to innovation, education, and ethical standards will be crucial in creating robust and responsible solutions for the detection and prevention of cyberbullying.

VII. FUTURE SCOPE

The future scope of cyberbullying detection using LSTM and related technologies holds promise for advancements in advanced neural architectures, multimodal analysis, real-time detection, personalized models, explainable AI, and global collaboration. As technology evolves, there is a growing emphasis on context-aware models, behavioral analysis, transfer learning, and adversarial robustness. The ethical considerations surrounding user privacy, bias, and fairness are crucial, necessitating the development of guidelines and standards for responsible deployment. Additionally, ongoing efforts in education and awareness are essential for fostering responsible online behavior. The interdisciplinary nature of addressing cyberbullying, coupled with advancements in technology and a global collaborative approach, will likely shape the future of effective and ethical cyberbullying.

REFERENCES

[1] F. Elsafoury, A. Abdelaziz, M. Ahmed, and A. E. Hassan, “Privacy-Preserving Cyberbullying Detection Using Federated Learning and Emotional Features,” Journal of Information Security and Applications, vol. 68, 2022.

[2] B. A. H. Murshed, M. S. Rahman, and M. A. Hossain, “A Hybrid Deep Learning Framework for Cyberbullying Detection on Twitter,” Social Network Analysis and Mining, vol. 11, no. 1, 2021.

[3] F. Shannaq, R. Al-Hmouz, and A. Al-Fayoumi, “Offensive Language Detection in Arabic Social Networks Using Evolutionary-Based

Classifiers," Applied Soft Computing, vol. 95, 2020.

[4] T. H. Teng, Y. C. Lim, and K. H. Ng, "A Comparative Study of Machine Learning and Transfer Learning Approaches for Cyberbullying Detection," IEEE Access, vol. 8, pp. 144430–144442, 2020.

[5] M. Al-Hashedi, M. M. Al-Shehri, and F. Saeed, "Emotion-Based Cyberbullying Detection on Social Media Using Sentiment Analysis," Computers & Security, vol. 92, 2020.

[6] N. A. Samee, M. Al-Sarem, and M. A. Al-Hashmi, "Federated Learning-Based Cyberbullying Detection with Word Embeddings and Emotional Features," Future Generation Computer Systems, vol. 118, 2021.

[7] R. Daniel, S. P. Raja, and K. Shankar, "Cyberbullying Detection Using Tournament Selected Glowworm Swarm Optimization," Journal of Ambient Intelligence and Humanized Computing, vol. 12, 2021.

[8] M. H. Obaid, A. A. Al-Shargabi, and S. Al-Mansoori, "Cyberbullying Detection and Severity Analysis Using Deep Learning," International Journal of Advanced Computer Science and Applications, vol. 11, no. 9, 2020.

[9] J. Bacha, A. Ghenai, and Y. A. Abdelrahman, "Deep Learning-Based Offensive Text Detection Across Heterogeneous Social Media Platforms," Expert Systems with Applications, vol. 168, 2021.

[10] N. Ketsbaia, G. Mikeladze, and T. Natsvlishvili, "A Multi-Stage Machine Learning and Fuzzy Logic Approach for Cyber-Hate Detection," Knowledge-Based Systems, vol. 215, 2021.

[11] S. Wang, Z. Liu, and F. Sun, "Contextual Word Embedding-Based Cyberbullying Detection Using LSTM Networks," Neural Computing and Applications, vol. 33, pp. 13567–13580, 2021.

[12] Y. Park and J. Fung, "Neural Network-Based Hate Speech and Offensive Language Detection on Social Media," Proceedings of the ACL Workshop on Abusive Language Online, 2017.

[13] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep Learning for Hate Speech Detection in Tweets," Proceedings of the 26th International World Wide Web Conference (WWW), 2017.