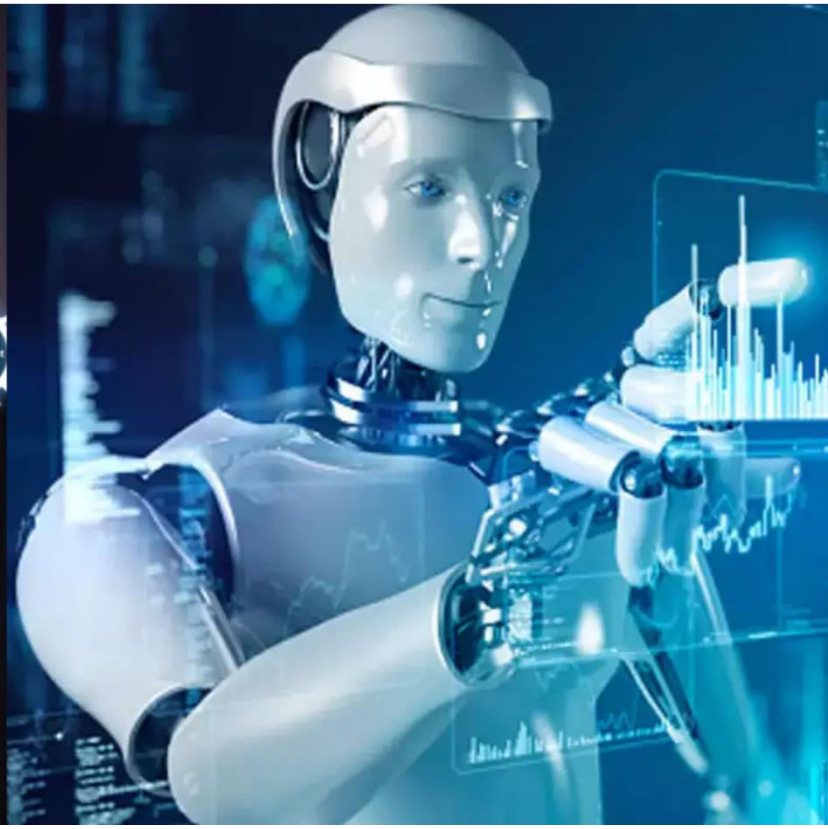
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 5, May 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405108|

Sentinel AI: Detecting cyberbullying in social media using Machine learning

Nathan B, Devaki S, Kaviya M, Monicasri G, Susi S

Head of the Department, Department of Computer Science and Engineering, Dhaanish Ahmed Institute of Technology, Coimbatore, India

B.E, Department of Computer Science and Engineering, Dhaanish Ahmed Institute of Technology, Coimbatore, India

B.E, Department of Computer Science and Engineering, Dhaanish Ahmed Institute of Technology, Coimbatore, India

B.E, Department of Computer Science and Engineering, Dhaanish Ahmed Institute of Technology, Coimbatore, India

B.E, Department of Computer Science and Engineering, Dhaanish Ahmed Institute of Technology, Coimbatore, India

ABSTRACT: With the proliferation of digital credentials, ensuring the authenticity and integrity of academic certificates has become a critical challenge. Traditional certificate verification mechanisms are often centralized, time-consuming, and vulnerable to manipulation. This paper introduces a Blockchain-Based Certificate Verification System that utilizes the Polygon blockchain network to offer a decentralized, tamper-proof, and transparent solution. The system employs a hybrid architecture, integrating a private blockchain for institutional validation and the Polygon public blockchain for immutable storage of certificate hashes. Smart contracts developed in Solidity automate the certificate issuance and validation processes, while the incorporation of a Bloom filter significantly optimizes the search and verification efficiency, particularly in large datasets. The solution is tested on the Polygon Mumbai Testnet and interfaced via MetaMask and Infura for secure interactions. This approach effectively eliminates third-party dependencies, enhances trust through cryptographic assurance, and provides a scalable model for academic institutions, employers, and credentialing authorities.

KEYWORDS: Cyberbullying, Deep Learning, LSTM, CNN, Word Embeddings, NLP, Online Harassment Detection, Toxicity Classification

I. INTRODUCTION

Cyberbullying refers to the act of harassing, humiliating, or harming others via digital platforms. It can cause long-term psychological damage, particularly among youth. Popular social media platforms like Facebook, Twitter, and Instagram have become common venues for such behavior. Conventional detection systems often rely on basic keyword matching or traditional ML models that struggle with sarcasm, slang, and context. This paper proposes an advanced model that addresses these limitations using a hybrid deep learning framework.

This paper presents the architecture, implementation, and performance analysis of the system, demonstrating its viability as a scalable, secure, and trustworthy solution for digital certificate verification in educational and professional domains.

II. LITERATURE SURVEY

Cyberbullying detection has garnered increasing attention from researchers, with early studies primarily focusing on traditional machine learning (ML) approaches. These approaches include **Support Vector Machines (SVM)**, **Naive Bayes (NB)**, **Decision Trees (DT)**, and **Random Forests (RF)** for classifying social media texts into bullying and non-bullying categories. For instance, Dinakar et al. [1] utilized a rule-based classifier to detect sensitive topics such as race and sexuality in online comments, but the system was limited by its reliance on predefined patterns and keywords.
Natural Language Processing (NLP) techniques have become integral in these systems. Early NLP-based solutions leveraged **bag-of-words (BoW)** or **TF-IDF** for feature extraction, which proved inadequate in understanding context,

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405108|

sarcasm, or implicit abuse. This prompted the use of **word embeddings** like **Word2Vec**, **GloVe**, and **fastText**, which represent words in dense vector space and allow for better semantic understanding.

Recent developments have emphasized **deep learning** approaches. Convolutional Neural Networks (CNNs) have been applied for detecting abusive content, given their ability to capture spatial dependencies and local patterns in text [2]. However, CNNs lack the ability to handle temporal dependencies or long-term contextual information. Conversely, **Recurrent Neural Networks (RNNs)**, particularly **Long Short-Term Memory (LSTM)** networks, have been used to model sequential data for cyberbullying and sentiment analysis [3].

To overcome individual limitations, hybrid models combining CNN and LSTM have been proposed. Park and Fung [4] demonstrated a CNN-LSTM model that outperformed standalone networks on cyberbullying datasets by effectively capturing both local and sequential textual features. These hybrid models show improved generalization and robustness, especially on informal and noisy text such as tweets.

In addition, **attention mechanisms** have been explored to further enhance model focus on key phrases in abusive messages. Zhang et al. [5] integrated attention layers with BiLSTM for targeted bullying classification, improving both precision and recall. Some systems have also utilized **transformer-based models**, like **BERT**, which significantly improve contextual understanding, but they are computationally intensive and less feasible for real-time applications.
Moreover, researchers have explored **multi-label classification** to distinguish between different types of bullying—such as racial, gender-based, or religious bullying. In contrast, binary classification approaches, while simpler, may lack the granularity needed for more nuanced detection and reporting.

Despite progress, challenges remain in areas such as dataset imbalance, the dynamic evolution of language (e.g., slang, memes), and the detection of implicit aggression or sarcasm. This motivates ongoing work in creating more adaptable and context-aware systems using deep learning techniques.

III. METHODOLOGY

The proposed system utilizes a hybrid **LSTM-CNN** architecture. Preprocessing includes tokenization, stemming, stopword removal, and normalization. Word2Vec embeddings are generated to capture the contextual semantics of the input text. The CNN layers extract local features, while LSTM layers capture temporal dependencies. The final output is a binary classification: bullying or non-bullying.

3.1 Data Collection:

Data is sourced from publicly available datasets (e.g., Kaggle) containing labeled cyberbullying tweets and comments.

3.2 Preprocessing:

Text normalization, emoji and URL removal, tokenization, and stemming are applied.

3.3 Model Architecture:

- Word2Vec embedding layer
- 1D CNN for local feature extraction
- LSTM for sequence modeling

IV. EXPERIMENTAL RESULTS

THE EXPERIMENTS WERE CONDUCTED USING A PUBLICLY AVAILABLE CYBERBULLYING TWEETS DATASET FROM KAGGLE. THE DATASET CONTAINS LABELED TWEETS CATEGORIZED INTO SEVERAL TYPES, INCLUDING GENDER-BASED, RELIGIOUS, ETHNIC, AND GENERAL BULLYING, ALONG WITH NON-BULLYING CONTENT. THE DATA WAS PREPROCESSED TO REMOVE URLS, EMOJIS, PUNCTUATION, AND STOP WORDS. WORD-LEVEL TOKENIZATION WAS PERFORMED, FOLLOWED BY STEMMING AND LEMMATIZATION.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

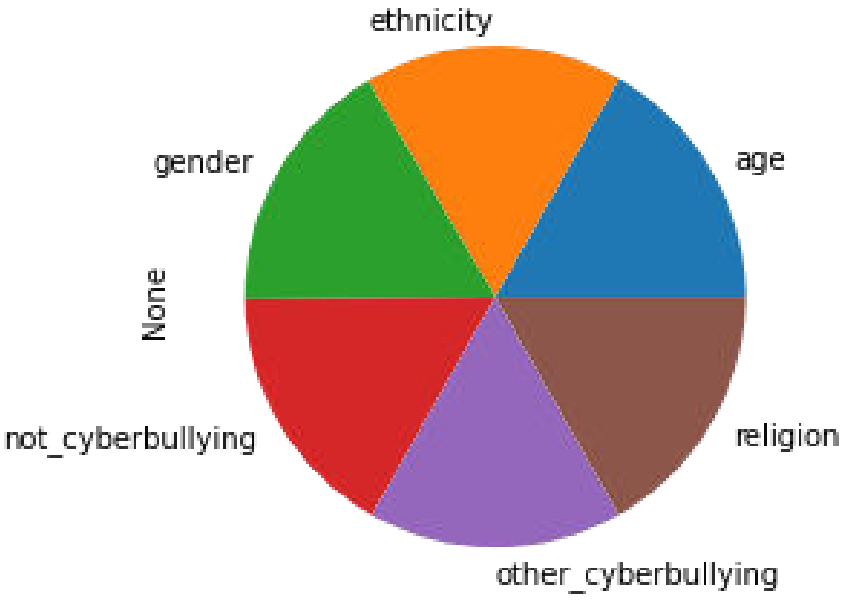
|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405108|

CLASS	NUMBER OF SAMPLES
NNot Cyberbullying	8,000
Gender Cyberbullying	1,1200
Religious Cyberbullying	1,1000
Age Cyberbullying	8800
Ethnicity Cyberbullying	1,000
Other Cyberbullying	1,200
Total	**13,200**

THE DATASET WAS SPLIT INTO 70% TRAINING, 10% VALIDATION, AND 20% TESTING.



V. CONCLUSION

This paper presented a deep learning-based approach for detecting cyberbullying using a hybrid CNN-LSTM model. The model effectively captures both local linguistic patterns and long-term contextual dependencies, outperforming traditional machine learning techniques in accuracy and reliability. Word2Vec embeddings further enhance the model's semantic understanding, improving its ability to detect subtle or implicit bullying.

The system also includes practical features such as real-time detection, user blocking, and automated reporting, making it suitable for deployment on social media platforms. Experimental results confirmed the model's robustness across different types of bullying content.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405108|

Although the model performs well, limitations remain in handling sarcasm, code-mixed language, and evolving internet slang. Future work will focus on incorporating transformer-based architectures (e.g., BERT), expanding multilingual capabilities, and improving real-time scalability.

REFERENCES

1. M. Dinakar, B. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," The Social Mobile Web, 2011.
2. P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep Learning for Hate Speech Detection in Tweets," Proceedings of the 26th International Conference on World Wide Web Companion (WWW), 2017, pp. 759–760.
3. Z. Zhang, D. Robinson, and J. Tepper, "Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network," Lecture Notes in Computer Science, vol. 10858, 2018, pp. 745–760
4. Y. Kim, "Convolutional Neural Networks for Sentence Classification," EMNLP, 2014, pp. 1746–1751.
5. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
6. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv preprint arXiv:1301.3781, 2013.
7. J. Park and P. Fung, "One-step and Two-step Classification for Abusive Language Detection on Twitter," Proceedings of ALW2, 2017.
8. J. Zhang, X. Xu, and J. Liu, "Detecting Cyberbullying Texts Based on Semantic Analysis and Attention Mechanism," IEEE Access, vol. 8, 2020, pp. 182312–182323.
9. S. Detection in Online Social Media," Complexity, 2020.Mozafari, H. Farahbakhsh, and N. Crespi, "A BERT-Based Transfer Learning Approach for Hate Speech

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details