CNN Based Malicious Abuse Detection System

Avancha Sai Geetha Vani[1], Mr.P.Mareswara Rao[2], , Mrs.M.Revati[3]

[1]PG student ,Department Of Computer Science & Engineering, Mother Teresa Institute Of Science And Technology Autonomous, Sanketika Nagar, Kothuru (V), Sathupally - 507303, Khammam Dist., Telangana,India

[2,3,4]Assistant Professor , Department Of Computer Science & Engineering, Mother Teresa Institute Of Science And Technology Autonomous, Sanketika Nagar, Kothuru (V), Sathupally - 507303, Khammam Dist., Telangana,India

***Abstract*—**The wide spread adoption of internet technologies, social networking platforms, and digital communication channels has led to a significant rise in harmful online activities. Issues such as cyber bullying, phishing attacks, spam communication, offensive content, fake accounts, and other forms of malicious behavior pose serious challenges to both individuals and organizations. Conventional security and monitoring techniques often struggle to identify these threats effectively because malicious actors continuously alter their communication patterns, vocabulary, and behavioral characteristics. To overcome these limitations, this study presents a CNN-based Malicious Abuse Detection System that utilizes deep learning techniques for the automatic identification and classification of abusive online content.

Furthermore, the framework can support automated moderation and monitoring applications across social media networks, messaging platforms, online communities, and organizational communication systems. The model reduces the need for extensive manual intervention while maintaining reliable performance in the presence of noisy and incomplete data. Overall, the Research highlights the effectiveness of deep learning-based solutions in strengthening Cyber security measures and promoting safer digital communication. The proposed CNN-based malicious abuse detection framework provides an intelligent, efficient, and Scalable approach for identifying and mitigating modern online threats.

I. Introduction

The rapid expansion of internet connectivity and digital communication technologies has fundamentally changed the way individuals, organizations, and communities interact. Social networking websites, instant messaging applications, online discussion forums, email services, and collaborative digital platforms have become integral components of modern society. These technologies facilitate seamless communication, knowledge sharing, business transactions, education, and entertainment across geographical boundaries. While the digital revolution has created numerous opportunities for social and economic development, it has also introduced significant cyber security challenges. The increasing dependence on online communication systems has provided malicious actors with new avenues to conduct harmful and abusive activities on a large scale.

Malicious abuse in cyberspace has emerged as a serious concern affecting users worldwide. Such abuse can take many forms, including cyber bullying, hate speech, phishing attacks, identity

Doi:10.48047/ijiee.2026.16.2.46

impersonation, online harassment, spam campaigns, misinformation dissemination, fraudulent communications, and the distribution of harmful content. These malicious activities not only compromise the privacy and security of users but also cause psychological distress, financial losses, reputational damage, and reduced trust in digital platforms. The growing volume of user-generated content across online platforms further complicates the task of identifying and controlling abusive behavior effectively.

II. LITERATURE SURVEY

The rapid growth of social media platforms and online communication has led to an increase in abusive, offensive, and malicious content. Automated abuse detection systems have become essential for maintaining safe online environments. Researchers have explored various machine learning and deep learning techniques to identify and classify abusive content effectively.

Yoon Kim (2014) introduced Convolution Neural Networks (CNNs) for sentence classification and demonstrated that CNNs can achieve excellent performance in text categorization tasks. The study showed that CNNs automatically extract meaningful features from text without extensive manual feature engineering, making them suitable for abuse detection applications.

Davidson et al. (2017) developed a hate speech detection framework using machine learning techniques on Twitter data. The authors distinguished between hate speech, offensive language, and non-offensive content. Their work highlighted the challenges of contextual interpretation and dataset imbalance in abuse detection systems.

Zhang, Zhao, and LeCun (2015) proposed character-level CNN models for text classification. Their research demonstrated that CNN architectures can effectively learn textual patterns directly from raw text, reducing dependence on handcrafted linguistic features. The approach achieved competitive performance across multiple classification tasks.

Badjatiya et al. (2017) investigated deep learning techniques for detecting hate speech using word embeddings and neural networks. Their experimental results showed that deep learning models outperform traditional machine learning approaches in capturing semantic relationships among words and identifying abusive content.

Park and Fung (2017) introduced a CNN-based framework for cyberbullying detection on social media platforms. The study incorporated semantic and contextual features to improve classification accuracy and demonstrated the effectiveness of deep learning in identifying harmful online interactions.

Jigsaw Toxic Comment Classification Challenge (2018) provided a large-scale dataset for toxic comment detection and encouraged the development of advanced machine learning models. Several CNN-based approaches achieved high classification accuracy, establishing CNNs as a reliable method for abuse detection tasks.

Mozafari et al. (2019) compared deep learning architectures including CNN, LSTM, and hybrid models for offensive language detection. Their findings indicated that CNN-based models offer faster training times while maintaining competitive performance, making them suitable for real-time content moderation systems.

From the reviewed literature, it is evident that CNN-based models provide effective feature extraction capabilities and achieve high accuracy in abuse detection tasks. However, challenges such as sarcasm detection, contextual understanding, multilingual content processing, and dataset bias remain active research areas. The proposed CNN-Based Malicious Abuse Detection System aims to leverage deep learning techniques to improve the identification of abusive and malicious content while maintaining efficient computational performance.

III. RESEARCH METHODOLOGY

Doi:10.48047/ijiee.2026.16.2.46

The proposed CNN-Based Malicious Abuse Detection System adopts a deep learning approach to identify and classify abusive, offensive, and malicious textual content from online platforms. The methodology begins with the collection of a labeled dataset containing abusive and non-abusive comments obtained from publicly available sources. The collected data undergoes several preprocessing steps to improve data quality and model performance. These steps include converting text to lowercase, removing punctuation, URLs, special characters, and stop words, followed by tokenization and sequence padding. The cleaned text is then transformed into numerical representations using word embedding techniques, which capture semantic relationships among words and provide meaningful input to the neural network.

Convolution Neural Network (CNN) architecture is employed to automatically extract significant textual features from the embedded input sequences. The convolution layers apply multiple filters to identify local patterns and contextual information associated with abusive language. These extracted features are further processed through max-pooling layers, which reduce dimensionality while preserving the most informative characteristics of the text. The resulting feature maps are flattened and passed to fully connected layers that perform the final classification task. Depending on the classification objective, either a sigmoid or soft max activation function is utilized in the output layer to categorize the text into abusive or non-abusive classes, or into multiple abuse-related categories.

The model is trained using the Adam optimization algorithm and an appropriate loss function to minimize classification errors and improve predictive performance. During training, the dataset is divided into training, validation, and testing subsets to ensure reliable evaluation and prevent over fitting. The effectiveness of the proposed system is assessed using standard performance metrics, including accuracy, precision, recall, F1-score, and confusion matrix analysis. By combining advanced text preprocessing techniques with CNN-based feature extraction, the proposed methodology provides an efficient and accurate solution for detecting malicious and abusive content in online communication environments.

IV. EXISTING SYSTEM

The existing malicious abuse detection systems are primarily based on conventional Machine Learning (ML) techniques that are designed to identify harmful online content and suspicious user activities. These systems are commonly used in social media platforms, online forums, email filtering applications, and cyber security monitoring tools to detect spam, phishing messages, hate speech, cyber bullying, and other forms of digital abuse. Popular machine learning algorithms employed in such systems include Support Vector Machines (SVM), Naïve Bayes, Decision Trees, and Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest classifiers.

In traditional abuse detection frameworks, the first stage involves collecting and preprocessing large amounts of textual or behavioral data. Data preprocessing activities include cleaning unwanted symbols, removing duplicate entries, eliminating stop words, and converting text into a standardized format. After preprocessing, feature extraction techniques are applied to convert textual information into numerical representations that machine learning algorithms can understand. Common feature extraction methods include Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), n-gram analysis, keyword frequency analysis, and statistical feature generation.

Disadvantages

1. Dependence on Manual Feature Engineering
2. Limited Contextual Understanding
3. Inability to Capture Semantic Relationships
4. Vulnerability to Evasion Techniques
5. High False Positive and False Negative Rates

Doi:10.48047/ijiee.2026.16.2.46

6. Poor Scalability for Large Datasets

7. Reduced Adaptability to Emerging Threats

8. Inefficient Handling of Unstructured Data

9. Limited Anomaly Detection Capability

10. Frequent Maintenance Requirements

V. PROPOSED SYSTEM

The proposed CNN-Based Malicious Abuse Detection System introduces a hybrid deep learning framework that combines Convolution Neural Networks (CNN) and Auto encoder algorithms to improve abuse detection performance. The proposed system is designed to automatically learn complex hidden patterns from online communication data without relying heavily on manual feature engineering.

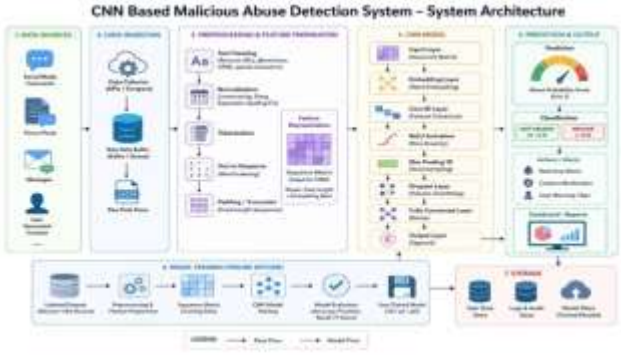
The first stage of the proposed system involves data collection and preprocessing. Data is gathered from social media platforms, online forums, chat systems, and cyber security datasets. Preprocessing operations such as text cleaning, normalization, stemming, stop-word removal, tokenization, and vectorization are applied to improve data quality. Word embedding techniques convert textual data into numerical representations suitable for deep learning models.

Advantages

The proposed CNN-Based Malicious Abuse Detection System provides numerous advantages over traditional machine learning approaches. One of the primary advantages is automatic feature extraction. Unlike conventional systems that depend on manual feature engineering, CNN models automatically learn relevant patterns and semantic relationships from raw input data. This improves detection accuracy and reduces human effort.

Another important advantage is improved contextual understanding. CNN models can analyze complex sentence structures, semantic meanings, and contextual relationships within online communication data. This enables the system to detect hidden abusive intent, offensive language variations, and evolving malicious patterns more effectively.

VI. SYSTEM ARCHITECTURE



language structures, slang words, or continuously evolving cyber threats. In contrast, the CNN model automatically learns meaningful representations from the dataset, enabling the system to detect abusive behavior more efficiently and accurately. The experimental results demonstrate that CNN-based abuse detection systems outperform many conventional machine learning approaches such as Naive Bayes, Decision Trees, and Support Vector Machines in terms of precision, recall, F1-score, and overall accuracy. The model performs effectively even when dealing with large datasets containing noisy and unstructured textual information. The system can successfully identify different forms of malicious content that may negatively affect individuals or organizations. Therefore, the proposed model contributes significantly to modern cybersecurity and content moderation solutions.

VII. RESULTS AND OUTCOMES

Doi:10.48047/ijiee.2026.16.2.46

Fig:7.1: Output 1 Home Page

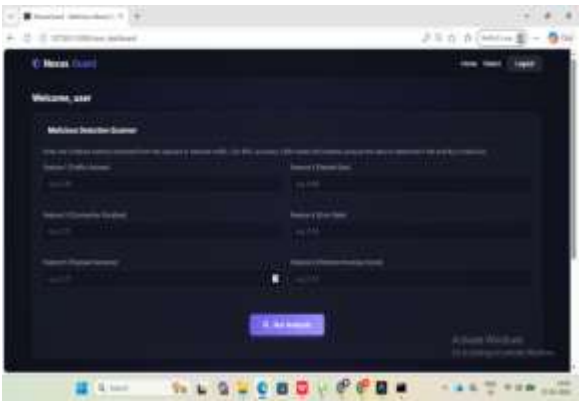


Fig:7.2: Output 2 Prediction Page

VIII. CONCLUSION

The CNN-based malicious abuse detection system provides an intelligent and automated approach for identifying harmful, abusive, and malicious activities in digital communication platforms. The system uses Convolutional Neural Networks (CNNs) to analyze textual patterns, extract hidden features, and classify messages or content into abusive and non-abusive categories with high accuracy. Traditional machine learning techniques often depend heavily on manual feature extraction and predefined rules, which may fail when handling complex.

IX. REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learning Representations (ICLR)*, San Diego, CA, USA, 2015.

[4] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 1–9.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778.

[6] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[7] J. Kim, J. Kim, H. L. T. Thu, and H. Kim, "Long short term memory recurrent neural network classifier for intrusion detection," in *Proc. Int. Conf. Platform Technology and Service (PlatCon)*, Jeju, South Korea, 2016, pp. 1–5.

[8] W. Wang, M. Zhu, J. Wang, X. Zeng, and Z. Yang, "End-to-end encrypted traffic classification with one-dimensional convolution neural networks," in *Proc. IEEE Int. Conf. Intelligence and Security Informatics (ISI)*, Beijing, China, 2017, pp. 43–48.

[9] R. Vinayakumar, K. P. Soman, and P. Poornachandran, "Applying convolutional neural network for network intrusion detection," in *Proc. Int. Conf. Advances in Computing, Communications and Informatics (ICACCI)*, Udupi, India, 2017, pp. 1222–1228.

Doi:10.48047/ijiee.2026.16.2.46

[10] N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, "A deep learning approach to network intrusion detection," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 1, pp. 41–50, Feb. 2018.

Doi:10.48047/ijiee.2026.16.2.46