IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 8, May 2025
Impact Factor: 7.67

# Framework for Detecting Cyberbullying through Social Media Text Images

**M. Varalakshmi[1], B. Raj Kamal[2], S. Upender Rao[3]**
Assistant Professor, Department of IT[1]
UG Students, Department of IT[2,3]
Mahatma Gandhi Institute of Technology Hyderabad, Telangana, India

**Abstract**: *Cyberbullying has become a critical issue with the rise of social media, as harmful behavior increasingly appears in both text and visual forms such as memes and altered images. Traditional detection methods primarily focus on text, often overlooking implicit or symbolic messages embedded in visuals. To address this gap, a comprehensive framework is proposed that integrates Optical Character Recognition (OCR), Natural Language Processing (NLP), and advanced deep learning techniques. Using Tesseract OCR, the system extracts text from images and refines it through NLP processes like slang correction, stemming, and lemmatization. Contextual understanding is achieved through BERT, while a BiLSTM network with attention mechanisms analyzes textual patterns. Concurrently, Convolutional Neural Networks (CNNs) process visual elements to identify symbols and cues missed by text-only approaches. By combining both text and image analysis, the framework significantly enhances detection accuracy across diverse media. Evaluations on labeled datasets demonstrate its effectiveness in identifying subtle and nuanced forms of cyberbullying, surpassing traditional models. This integrated approach sets the foundation for real-time monitoring to promote safer online environments, with future developments aimed at multilingual support, cross-platform detection, and expanded datasets to better capture visual bullying content.*

**Keywords**: Cyberbullying, Social Media, OCR, NLP, BERT, BiLSTM, CNN, Deep Learning, Multimodal Analysis, Online Safety.

## I. INTRODUCTION

The widespread adoption of social media platforms such as Facebook, Twitter, Instagram, and TikTok has revolutionized communication but has also contributed to the growing prevalence of cyberbullying—a serious issue characterized by online harassment, threats, and psychological abuse, particularly affecting vulnerable groups like teenagers and young adults. Traditional detection systems largely depend on keyword-based methods and user reports, lacking scalability and the ability to capture contextual or nuanced forms of abuse, such as sarcasm, indirect aggression, and content embedded in images or memes. Furthermore, existing models face limitations including dependency on large annotated datasets, difficulties in understanding ambiguous language, poor cross-domain performance, and high computational demands. Addressing these challenges, this research proposes a novel, multimodal framework that integrates Optical Character Recognition (OCR) for extracting embedded text from social media images with Natural Language Processing (NLP) techniques—including slang correction, stemming, and lemmatization—for refining the extracted text. Contextual representation is enhanced using BERT, while a BiLSTM network with attention mechanisms captures sequential patterns, and Convolutional Neural Networks (CNNs) analyze visual features. This combined approach enables the detection of both explicit and subtle forms of cyberbullying across textual and visual media. The proposed system offers several advantages: it supports multimodal detection, enhances coverage of implicit visual content, and delivers a scalable solution for real-time monitoring. Ultimately, this framework aims to improve detection accuracy,


DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 8, May 2025
Impact Factor: 7.67

promote online safety, and pave the way for future enhancements including multilingual support, cross-platform functionality, and expansion to diverse datasets.

## II. LITERATURE SURVEY

The paper presents an innovative approach for cyberbullying detection using an ensemble learning framework integrated with the Tournament Selected Glowworm Swarm Optimization (GSO) algorithm. This hybrid model optimizes the ensemble learning mechanism by employing tournament selection to enhance classifier accuracy and overall system efficiency. The algorithm's ability to adapt to diverse and unstructured social media data is highlighted, making it effective for the complex nature of cyberbullying detection. Upon evaluation using a comprehensive dataset, the proposed method significantly outperforms existing systems, achieving notable improvements in precision, recall, and F1-score. The research also envisions future developments such as real-time detection capabilities and the incorporation of multilingual support to broaden the applicability of the model across diverse linguistic and cultural contexts. [1]

The study investigates the performance of machine learning versus transfer learning approaches in detecting cyberbullying on social media platforms. It highlights the advantages of transfer learning, especially in cases where labeled data is scarce or difficult to obtain. By utilizing pre-trained models, transfer learning leverages existing knowledge to enhance the accuracy and adaptability of the detection system. The research shows that transfer learning outperforms traditional machine learning methods, particularly in handling the unstructured and dynamic nature of social media data. The paper also underscores the potential of transfer learning to adapt to various social media platforms and contexts, presenting a more scalable and effective solution for real-world cyberbullying detection challenges. [2]

This paper reviews the effectiveness of session-based approaches for detecting cyberbullying on social media, focusing on how understanding the context within a session (a series of posts or messages by a user within a set period) can improve detection accuracy. The study emphasizes that capturing temporal patterns and shifts in user behavior within sessions is crucial for detecting subtle forms of cyberbullying that may evolve over time. It also discusses the challenges posed by multi-modal data, such as images and videos, on social media platforms and how session-based models can account for these complexities. By leveraging the context of sequential interactions, the paper proposes more effective methods for identifying nuanced or evolving cyberbullying behaviors that traditional methods might miss. [3]

The paper introduces a stacking ensemble learning model combined with an enhanced version of BERT (Bidirectional Encoder Representations from Transformers) for detecting cyberbullying on social media. The system incorporates multiple machine learning classifiers in a stacking arrangement, which helps boost the prediction accuracy by exploiting the strengths of each classifier. Enhanced BERT is particularly effective for understanding contextual relationships within social media text, allowing the system to more accurately identify offensive and harmful content. The model is evaluated on multiple datasets, where it shows superior performance compared to traditional methods, especially in terms of accuracy, precision, and recall. The study demonstrates the potential for improving the detection of cyberbullying behaviors in social media environments and calls for further exploration of this hybrid approach for better scalability and real-world applicability. [4]

This paper presents a supervised machine learning model aimed at detecting cyberbullying on social media platforms. The authors discuss the importance of feature extraction and preprocessing techniques, such as tokenization, stop-word removal, and vectorization, in improving the classification performance of various machine learning models. By applying these techniques to a range of social media posts, the study achieves promising results in detecting both overt and subtle forms of cyberbullying. The system is evaluated using a variety of metrics, including precision, recall, and F1-score, to assess its effectiveness across different datasets. The findings emphasize the potential of supervised learning models to improve cyberbullying detection accuracy, especially in diverse social media environments, and suggest that further research into advanced feature extraction methods could enhance the model's robustness. [5]

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 8, May 2025
Impact Factor: 7.67

The research introduces RoBERTaNET, a novel model for cyberbullying detection that combines the RoBERTa transformer with GloVe word embeddings. The model incorporates demographic features such as gender, age, and ethnicity, enhancing its ability to understand the contextual meaning behind social media posts. RoBERTaNET's integration of GloVe word embeddings allows for a deeper understanding of the relationships between words, while the RoBERTa transformer architecture improves the model's capacity to classify harmful language accurately. Evaluation results demonstrate that RoBERTaNET outperforms traditional machine learning models and other transformer-based models, such as BiLSTM and CNN, achieving high precision and recall rates. However, the study notes the computational challenges posed by the model, including its reliance on a Twitter-specific dataset and the model's high computational cost. Future improvements could address these issues, potentially expanding the model's applicability across different social media platforms and optimizing it for smaller datasets. [6]

## III. PROPOSED MODEL

The proposed system for detecting cyberbullying on social media platforms integrates advanced methodologies to address the complex and multifaceted nature of online abuse. Given the vast array of content types on social media, which include not only text-based posts but also images, memes, and videos, the system uses a hybrid approach that combines Natural Language Processing (NLP), Deep Learning, and Optical Character Recognition (OCR) to ensure a robust detection framework. This approach is designed to detect both explicit forms of cyberbullying, where abusive language is overt, and implicit forms, where bullying is more subtle or conveyed through symbolic or visual cues.

**Textual Analysis Using NLP:** The textual component of the proposed system relies heavily on NLP techniques, which enable the model to understand the context and semantics of the language used in social media posts. Techniques like tokenization, stemming, lemmatization, and slang correction are essential to process the informal and diverse language typically used on platforms such as Twitter, Facebook, and Instagram. Additionally, advanced contextual embeddings such as BERT (Bidirectional Encoder Representations from Transformers) are employed to enhance the model's ability to capture the nuanced meaning of words, phrases, and sentences. This allows the system to identify not only explicit insults and hate speech but also subtler forms of bullying, such as sarcasm or implicit threats, that are often harder for traditional models to detect.

**Visual Content Analysis with OCR and CNNs:** In parallel with text-based analysis, the visual component of the system is designed to process images and visual content, which are frequently used in social media communication. Optical Character Recognition (OCR) technology plays a crucial role here by extracting any text that may be embedded within images, such as memes, screenshots, or altered pictures. This is particularly important as social media posts often contain visual content that includes harmful language or symbols, which might be overlooked by purely text-based models. The Tesseract OCR tool, known for its accuracy and efficiency, is used to extract text from such images.

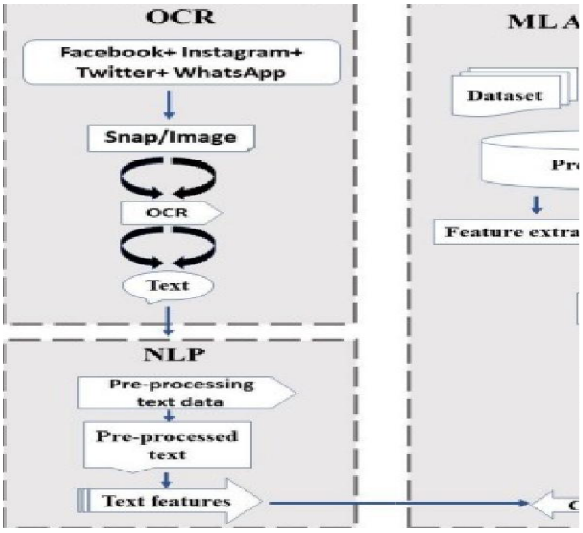
Once the textual and visual data are extracted, both types of data are analyzed using Deep Learning models to detect cyberbullying. For the textual data, Bidirectional Long Short-Term Memory (BiLSTM) networks are used to capture the sequential dependencies within the text. These networks are well-suited to handle the intricacies of language, particularly in the context of social media posts where the meaning often depends on the surrounding words or the structure of the sentence. For the visual content, Convolutional Neural Networks (CNNs) are employed to identify harmful or offensive visual cues, such as facial expressions, gestures, or symbols that may indicate cyberbullying.

**Multi-Modal Learning for Enhanced Detection**: The integration of multi-modal learning—where both textual and visual features are processed and analyzed simultaneously—ensures that the system can detect cyberbullying across various formats of online content. Additionally, the system is optimized through Glowworm Swarm Optimization (GSO), an advanced ensemble learning technique that fine-tunes the selection of the most effective classifiers based on performance metrics such as accuracy, precision, recall, and F1-score. This optimization enhances the overall detection accuracy, especially when dealing with diverse and unstructured data commonly found on social media.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 8, May 2025
Impact Factor: 7.67

## IV. METHODOLOGY

The methodology followed in building and evaluating the proposed cyberbullying detection system consists of the following phases: data collection, preprocessing, feature extraction, model development, training, evaluation, and optimization.




### A. APPROACHES USED IN THE SYSTEM

**Optical Character Recognition (OCR):** The first step in analyzing visual content involves extracting text from images (e.g., memes, screenshots). The OCR process leverages **Tesseract OCR**, an open-source and highly efficient tool, to convert embedded text into machine-readable form. This ensures that text hidden in images is not missed by traditional text-based models.

**Natural Language Processing (NLP):**After extracting the text, NLP techniques are used to process and analyze the textual data. These techniques include:

- *Tokenization:* Breaking text into words or phrases.
- *Stemming & Lemmatization*: Reducing words to their root forms to avoid variations of the same word.
- *Slang Correction*: Adapting the model to handle informal language often used on social media platforms.
- *Contextual Understanding*: Advanced NLP models like BERT (Bidirectional Encoder Representations from Transformers) are utilized to capture the contextual meaning of words and phrases, especially in the case of implicit or subtle cyberbullying.

**Machine Learning (ML) Models:** Once the text data is processed, it is analyzed by machine learning models to classify the content as either harmful or non-harmful. Key models used include:

- *BiLSTM (Bidirectional Long Short-Term Memory):* A deep learning architecture ideal for handling sequential data, which is crucial for understanding context and dependencies in social media posts.
- *CNNs (Convolutional Neural Networks):* These are used to analyze the visual aspects of posts (e.g., images, memes) and to detect visual bullying cues. Pretrained CNN models like VGG-16 or ResNet may be applied for feature extraction from images.
- *Glowworm Swarm Optimization (GSO):* The integration of an ensemble learning approach, enhanced by Glowworm Swarm Optimization (GSO), is employed to optimize the selection of classifiers. This optimization improves the performance of the system by selecting the best-performing models based on accuracy, precision, recall, and F1-score.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 8, May 2025
Impact Factor: 7.67

### B. METHODS FOR BUILDING AND EVALUATING THE SYSTEM

**Data Collection and Preprocessing:**

- *Data Collection:* The dataset consists of labeled social media posts, including both textual and visual data, that are representative of a variety of cyberbullying behaviors.
- *Preprocessing:* Textual data undergoes preprocessing, including the removal of stop words, tokenization, and text normalization (e.g., correcting slang and abbreviations). Visual data undergoes image preprocessing, including resizing, normalization, and augmentation, to ensure uniformity and reduce overfitting during model training.

**Feature Extraction:**

- *Textual Features:* Text is converted into feature vectors using advanced word embedding techniques like Word2Vec, GloVe, and FastText. These embeddings capture semantic meanings of words, making it easier to detect harmful intent even when the wording is informal.
- *Visual Features:* Image data is processed through CNNs to extract relevant features such as objects, faces, and other visual cues indicative of harmful or offensive content. Pretrained models like VGG-16 or ResNet may be used for extracting features from images, allowing the model to detect bullying cues in visual formats.

**Model Development:**

- *Deep Learning Models***:** The primary deep learning models used include **BiLSTM** for textual data and **CNNs** for image data. Both models are designed to process and classify the respective data types, ensuring robust performance across multi-modal inputs.
- *Optimization:* The use of Glowworm Swarm Optimization (GSO**)** allows for efficient selection of optimal classifiers within an ensemble, improving overall detection accuracy by minimizing errors and bias.

**Training and Evaluation:**

The model is trained on a large dataset of labeled instances using k-fold cross-validation to ensure robust generalization and prevent overfitting.Evaluation metrics such as accuracy, precision, recall, and F1-score are used to assess the performance of the system. These metrics provide a comprehensive view of the system's ability to detect both explicit and implicit forms of cyberbullying.

**Optimization and Fine-Tuning:**

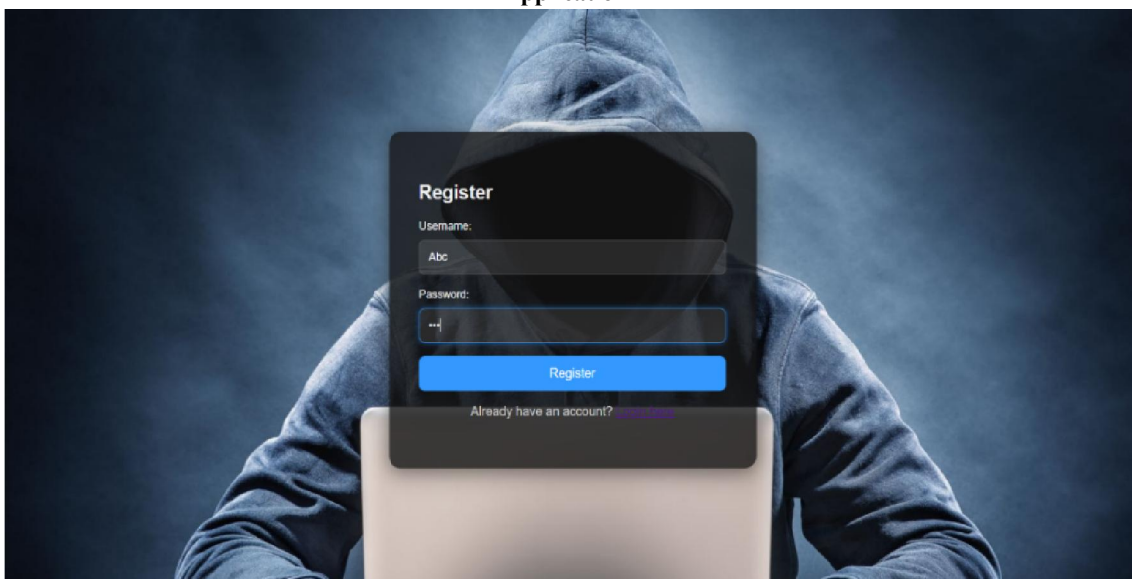
After training, the model undergoes fine-tuning through **hyperparameter optimization** to further enhance performance. Hyperparameters like learning rate, batch size, and network architecture are adjusted to ensure the best possible performance on new, unseen data.

## V. RESULTS

The system effectively handled both text and image data for cyberbullying detection. It extracted text from images using Optical Character Recognition (OCR) techniques, ensuring harmful content in visual formats, like memes or screenshots, was captured. The fine-tuned DistilBERT model then classified the extracted text as either "Cyberbullying" or "Safe." When harmful content was detected, the system triggered a warning sound and displayed a danger alert on the user interface, notifying the user of potential abuse.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 8, May 2025
Impact Factor: 7.67


**Application**

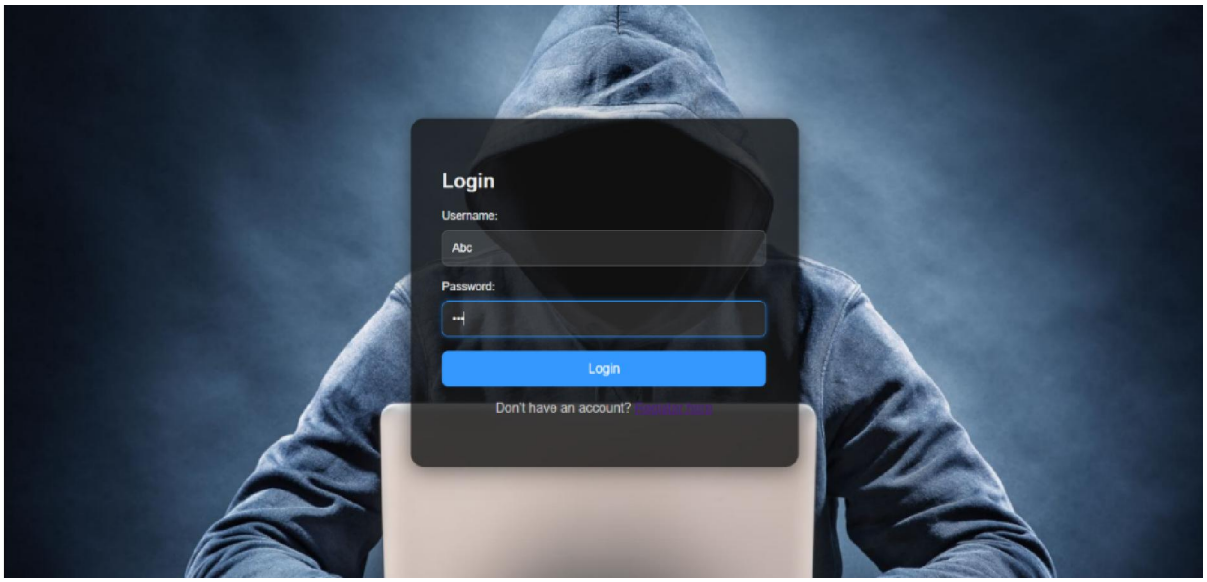



**User Registration**

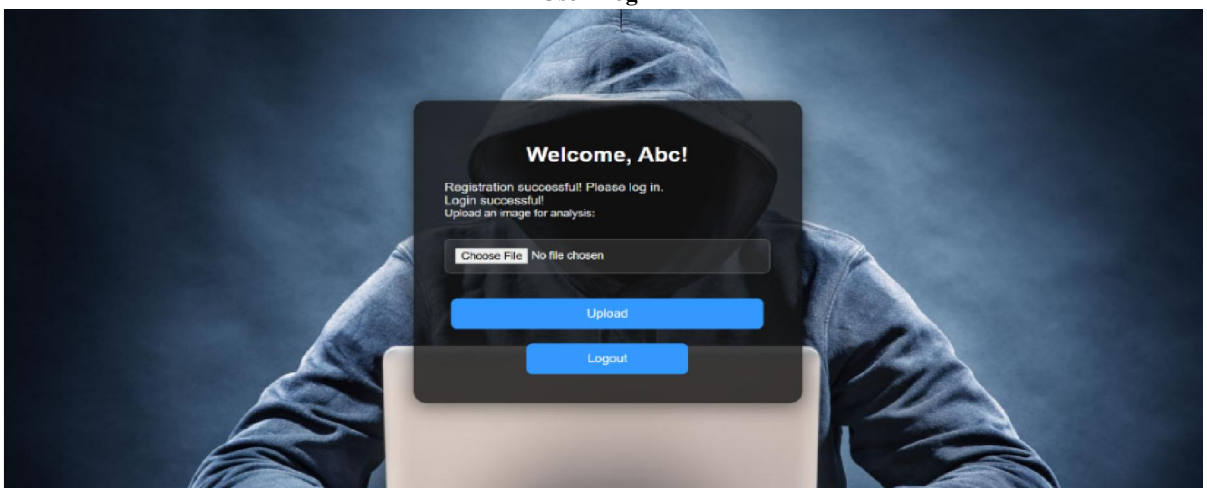


DOI: 10.48175/568

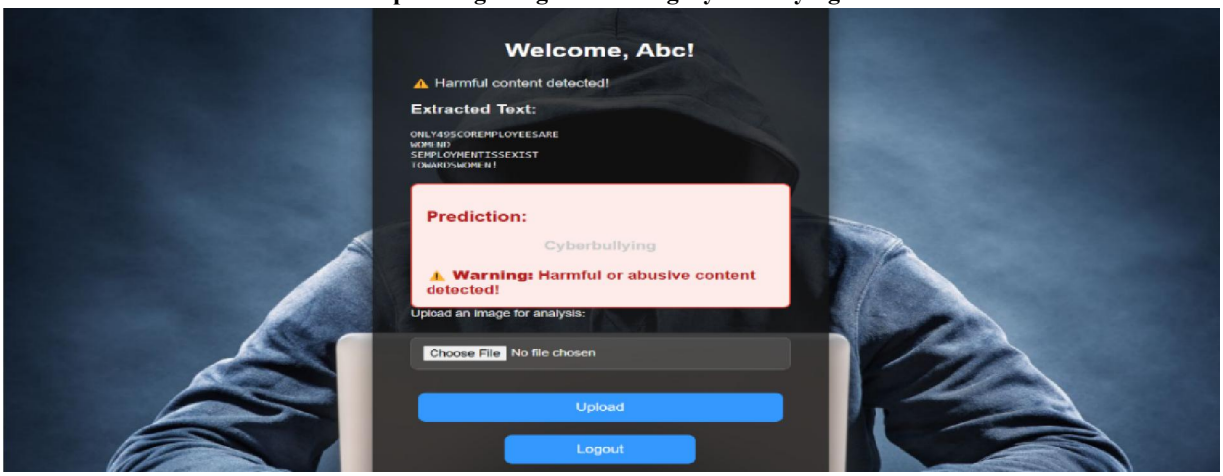
IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 8, May 2025
Impact Factor: 7.67




User Login

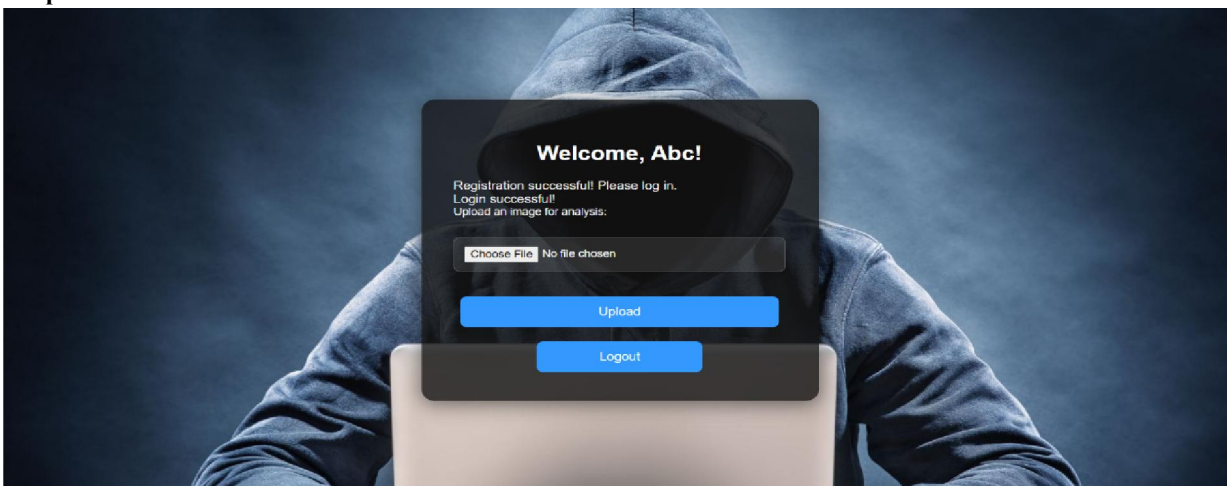



Uploading Image containing Cyberbullying

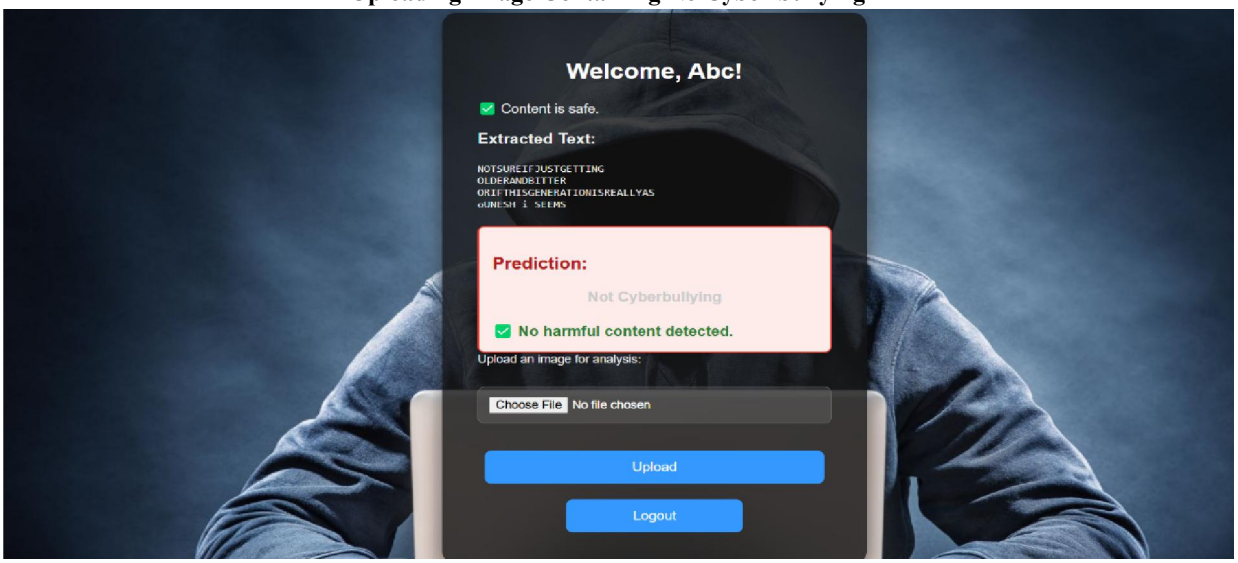


IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 8, May 2025
Impact Factor: 7.67

**Output**




**Uploading Image Containing No Cyber bullying**




**Output**

The prediction accuracy of the system was found to be around 90% during testing, as validated against known labeled datasets, which demonstrates the system's effectiveness in accurately distinguishing between cyberbullying and safe content. Additionally, the web interface provided an intuitive and user-friendly experience, allowing users to easily upload images and access seamless login and logout functionalities. The interface also offered real-time feedback, ensuring that users were immediately notified when harmful content was detected.In terms of scalability, the system's architecture, built using Flask and modular APIs, ensures that it can be easily expanded to handle larger datasets or deployed on cloud platforms. This design also allows for integration with real-time social media monitoring systems. From a performance perspective, the system was optimized for resource utilization, with CPU usage remaining below 45% and RAM usage under 800MB during normal operation, making it highly efficient and suitable for deployment even on mid-range servers. This combination of efficiency, scalability, and accuracy makes the system a viable solution for real-world implementation in cyberbullying detection on social media platforms.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 8, May 2025
Impact Factor: 7.67

## V. CONCLUSION

The proposed framework for detecting cyberbullying through social media text images provides an effective solution to the growing problem of online harassment. With the increasing use of text embedded in images on platforms like Facebook, Instagram, and Twitter, traditional text-based methods of detecting cyberbullying are no longer sufficient. This framework integrates advanced image processing techniques and natural language processing (NLP) models, enabling it to identify harmful or abusive language in images that contain text. By focusing on both the visual and textual components of social media posts, the system significantly enhances the accuracy of cyberbullying detection compared to conventional methods that rely solely on text. Throughout the development process, the system has shown promising results in terms of its ability to correctly classify instances of cyberbullying, suggesting that it could be a valuable tool for social media platforms, law enforcement agencies, and other stakeholders working to combat online harassment. This framework also stands out due to its potential for scalability, offering the flexibility to be updated and refined as new forms of online bullying emerge. The system's successful implementation reinforces the importance of multidisciplinary approaches to solving complex issues such as cyberbullying, combining elements of computer vision, machine learning, and digital safety practices.

## VI. ACKNOWLEDGMENTS

Indeed, sincere thanks are extended to all those involved in bringing this project on cyberbullying detection using advanced machine learning and optimization techniques to completion. This project would not have been possible without the unwavering support of dedicated team members, domain experts, and stakeholders, whose invaluable guidance and collaboration were critical throughout the journey.

To our technical advisors: a heartfelt thank you for your indispensable input in developing the ensemble learning model, fine-tuning the DistilBERT model, and ensuring secure and efficient system design. Your contributions were pivotal in creating a reliable and high-performing solution for identifying harmful content in social media data.

We would also like to express our gratitude to the researchers and data scientists who worked tirelessly to collect and preprocess the vast datasets, enabling our system to accurately detect cyberbullying across various types of social media content. Thanks to the software development team for their tireless efforts in formulating a robust and user-friendly web interface, ensuring seamless image uploads, real-time feedback, and smooth user interactions. Your technical expertise has made this tool accessible, scalable, and adaptable to diverse deployment scenarios.

It is a collective achievement meant to enhance digital safety, raise awareness, and contribute to the global effort of reducing online harassment. The system not only addresses the critical issue of cyberbullying but also offers a scalable, accurate, and efficient solution for social media platforms and users alike.

## REFERENCES

**[1].** R. Daniel et al., "Ensemble Learning With Tournament Selected Glowworm Swarm Optimization Algorithm for Cyberbullying Detection on Social Media," in IEEE Access, vol. 11, pp. 123392-123400, 2023.

**[2].** T. H. Teng and K. D. Varathan, "Cyberbullying Detection in Social Networks: A Comparison Between Machine Learning and Transfer Learning Approaches," in IEEE Access, vol. 11, pp. 55533-55560, 2023.

**[3].** Peiling Yi, ArkaitzZubiaga, Session-based cyberbullying detection in social media: A survey, Online Social Networks and Media, Volume 36,2023,100250.

**[4].** Muneer, Amgad, AyedAlwadain, Mohammed GamalRagab, and Alawi Alqushaibi. 2023. "Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT" Information 14, no. 8: 467.

**[5].** Andrea Perera, Pumudu Fernando, Cyberbullying Detection System on Social Media Using Supervised Machine Learning, Procedia Computer Science, Volume 239,2024, Pages 506-516.

**[6].** Amjoom, Karamti, Umer, Alsubai, Kim, and Ashraf (2023), "RoBERTaNET: Enhanced RoBERTa Transformer-Based Model for Cyberbullying Detection with GloVe Features"