iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XI **Month of publication:** November 2025

DOI: https://doi.org/10.22214/ijraset.2025.75573

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

CyGuardNLP: Chat Log Analysis for Cybercrime Detection

Yashaswini N[1], Dr. S Vidhya[2]
[1]*PG Student, Dept. of Cyber Security, The Oxford College of Engineering, Bangalore, India*
[2]*Assistant Professor, Dept. of ISE, The Oxford College of Engineering, Bangalore, India*

Abstract: ***In the era of digital communication, chat platforms have increasingly become targets for cybercriminal activities such as phishing, fraud, and harassment. This paper presents CyGuardNLP, a deep learning-based system designed to analyze chat logs and detect various cybercrime categories. Leveraging a fine-tuned BERT model combined with Optical Character Recognition (OCR) for screenshot analysis, CyGuardNLP supports multiple input formats including raw text, CSV files, and images. Experimental results demonstrate the system's robustness and high accuracy, highlighting its potential for real-time cybercrime monitoring and safer online interactions.***
Keywords: Cybercrime detection, chat log analysis, BERT, OCR, deep learning, natural language processing.

I. INTRODUCTION

Instant messaging platforms such as WhatsApp, Telegram, and Slack have revolutionized how people interact, offering convenient and instant communication. However, these platforms have also given rise to new cyber threats. Cybercriminals exploit chats for phishing, scams, cyberbullying, and harassment, presenting a challenge for traditional detection systems that depend heavily on keyword matching or metadata, which is often sparse or absent in chats.

Natural Language Processing (NLP) methods, especially transformer-based models like BERT, have shown remarkable success in understanding context and subtle textual nuances. CyGuardNLP harnesses these advances by fine-tuning a BERT model to classify chat messages into multiple cybercrime categories. Since chat evidence is also often shared as screenshots, an OCR module based on Tesseract is integrated to extract text from images, enabling multimodal analysis. CyGuardNLP offers a comprehensive, user-friendly interface implemented using Gradio, allowing users to analyze individual messages, perform batch processing on CSV files, or analyze images in real-time with confidence scores and detailed outputs.

A. Problem Statement

Existing tools for cybersecurity emphasize network intrusion and email phishing but overlook chat-based threats, which are often subtle, fast-evolving, and context-dependent. Manually moderating large-scale chat environments is impractical and inconsistent. Therefore, an intelligent, automated system is required to detect and classify chat-based cybercrime across varied formats plain text, bulk chat logs, and screenshots using an interpretable deep learning approach.

B. Objectives

1) Develop a dataset combining multiple categories of cybercrime (phishing, harassment, spamming, cyberbullying) and benign communication.
2) Fine-tune a pre-trained BERT model to identify cybercrime with high contextual accuracy.
3) Integrate OCR technology to extract and analyze text from chat screenshots.

II. LITERATURE REVIEW

Early cybercrime detection focused mainly on keyword-based filtering and shallow text models, which are ineffective against informal, context-dependent chat messages. Recurrent Neural Networks (RNNs) attempted to capture sequential dependencies, but their ability to understand nuances remained limited.

Transformer models, particularly BERT, improved classification accuracies owing to their bidirectional context understanding. Multimodal approaches combining text and images have gained attention, but few systems integrate OCR for chat screenshots in cybercrime detection. CyGuardNLP addresses this gap by combining a fine-tuned BERT classifier with OCR to handle diverse input sources, classifying phishing, fraud, cyberbullying, and benign chats effectively.

III. METHODOLOGY

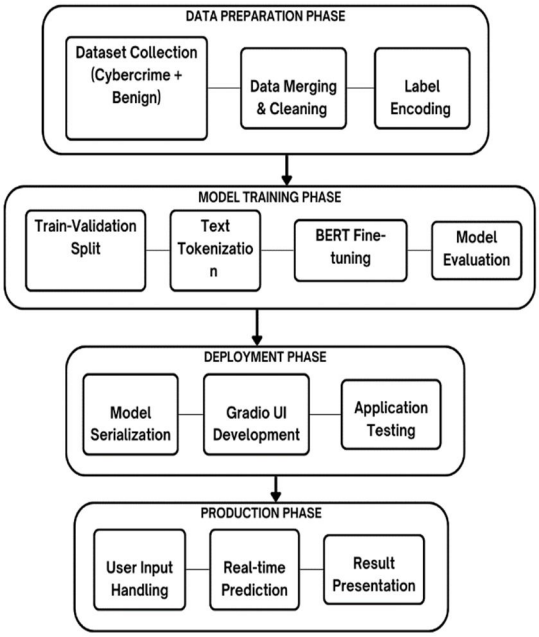


Fig.1 Training & Evaluation Process

A. Data Collection and Preparation

message	label
You smiled at me last week. Remember that. â€" Sam Martinez	Cyber Stalking
Come talk to me privately, I won't judge. I can help with anything.	Cyber Grooming
We need your bank details to process your onboarding bonus. Reply with account and routing. Please confirm.	Online Job Fraud
Earn $5,000 a week from home. Limited spots.	Spamming
No one likes you. Seriously, just shut up. (sent 7h ago)	Cyber Bullying
You qualify for manager role â€" pay registration fee and we guarantee placement. Please confirm.	Online Job Fraud
Someone is messaging your contacts pretending to be you asking for money.	Impersonation and Identity Theft
Thanks for the update â€" that change looks good to me.	Benign
I have screenshots from your camera. Pay $500 in crypto or I send them to everyone you know. Reply ASAP.	Online Sextortion
Security update: install patch from http://secure.example-downloads.com/patch (sent 2h ago)	Phishing

Fig.2 Messages with Label

A comprehensive dataset was assembled by combining labeled chat messages from multiple publicly available and proprietary sources, encompassing a range of cybercrime categories including phishing, spamming, harassment, and benign communications. The data included textual chat logs as well as chat screenshots captured from real conversations. To maintain the diversity and real-world relevance of the dataset, minimal pre-processing was applied. The inclusion of benign messages ensured that the model learned to distinguish harmful content from regular conversation, reducing false positives.

B. Preprocessing and Tokenization

To accommodate the informal and varied nature of chat messages, pre-processing steps were designed to retain critical contextual information. Instead of conventional heavy text normalization such as stemming or stop-word removal, messages were largely preserved in their original form. Tokenization employed BERT's WordPiece tokenizer, allowing effective handling of slang, abbreviations, and misspelled words by breaking them into subword units.

Input sequences were padded or truncated to a consistent length of 128 tokens for efficient batch processing. For image inputs, Tesseract OCR was used to extract text from screenshots, enabling transformation of unstructured visual data into processable textual form.

C. Model Fine-Tuning

A pretrained BERT-base model was fine-tuned with an added classification layer for multi-label cybercrime classification. Training employed AdamW optimizer and cross-entropy loss, with stratified train-test splits to maintain balanced class distribution.

D. System Architecture and Pipeline

The system architecture was designed for versatility and scalability, capable of processing three types of user inputs: plain text messages, bulk CSV files containing chat logs, and chat screenshots. Image inputs are first processed through the OCR pipeline to extract textual content. Then, all text inputs undergo tokenization and are passed to the fine-tuned BERT model, which outputs probability distributions over the predefined cybercrime categories. The highest probability class is selected as the prediction, with confidence scores provided for interpretability. The entire process is seamlessly integrated into a user-friendly Gradio interface, enabling real-time interaction for users with different technical backgrounds.

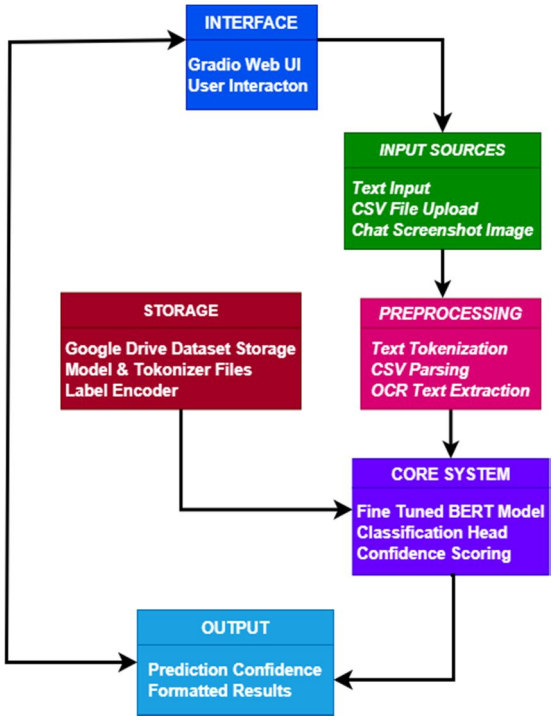


Fig.3 System Architecture

IV. IMPLEMENTATION

The implementation phase involved constructing a modular and scalable system architecture using Python programming language, with extensive reliance on libraries such as PyTorch for deep learning, HuggingFace Transformers for NLP modelling, pytesseract for OCR, and Gradio for the user interface. The training leveraged GPU acceleration to handle the computational demands of fine-tuning the BERT model efficiently, significantly reducing required training times.

The system's modular design allows each component—the data processing pipeline, OCR extraction, BERT-based classification, and user interface—to function independently. This separation of concerns not only facilitates ease of debugging and updates but also supports future scalability. Special attention was given to pre-processing parameters used in OCR, optimizing text extraction accuracy from diverse chat screenshot qualities. The error handling mechanisms were carefully designed to ensure robustness when dealing with noisy, informal, or corrupted inputs typical in real-world chat data. Overall, the implementation prioritizes practical usability alongside high model performance.

V. RESULTS AND EVALUATION

The performance of the CyGuardNLP system was assessed comprehensively across three different input modalities: text input, CSV batch input, and screenshot input. Evaluation metrics such as accuracy, precision, recall, and F1-score were used to measure classification effectiveness in detecting cybercrime categories. The results demonstrate the model's robustness and practical applicability in diverse usage scenarios.

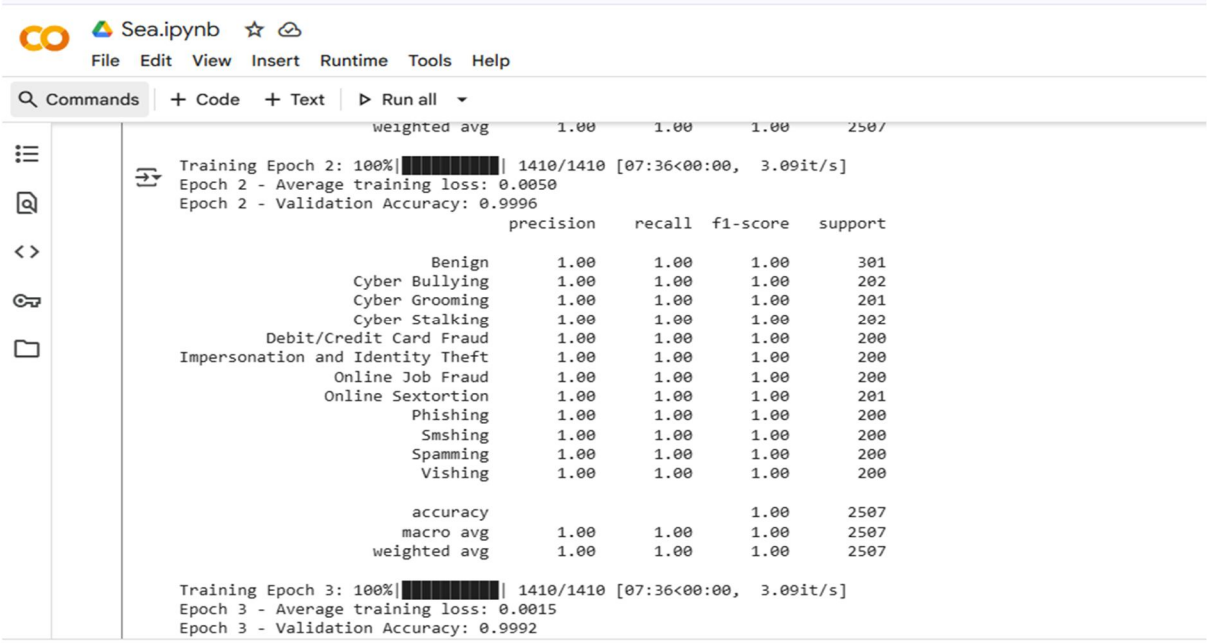


Fig.4 Model Training

A. Text Input Results

For single text messages entered directly, the system delivered rapid and accurate classification of cybercrime-related content. The fine-tuned BERT model effectively captured subtle linguistic cues, enabling it to differentiate between benign and malicious intents even in short, informal messages containing slang or abbreviations. Precision and recall scores were high across all cybercrime categories, indicating reliable detection without excessive false alarms. This makes the system suitable for real-time monitoring where immediate decisions on message safety are required.

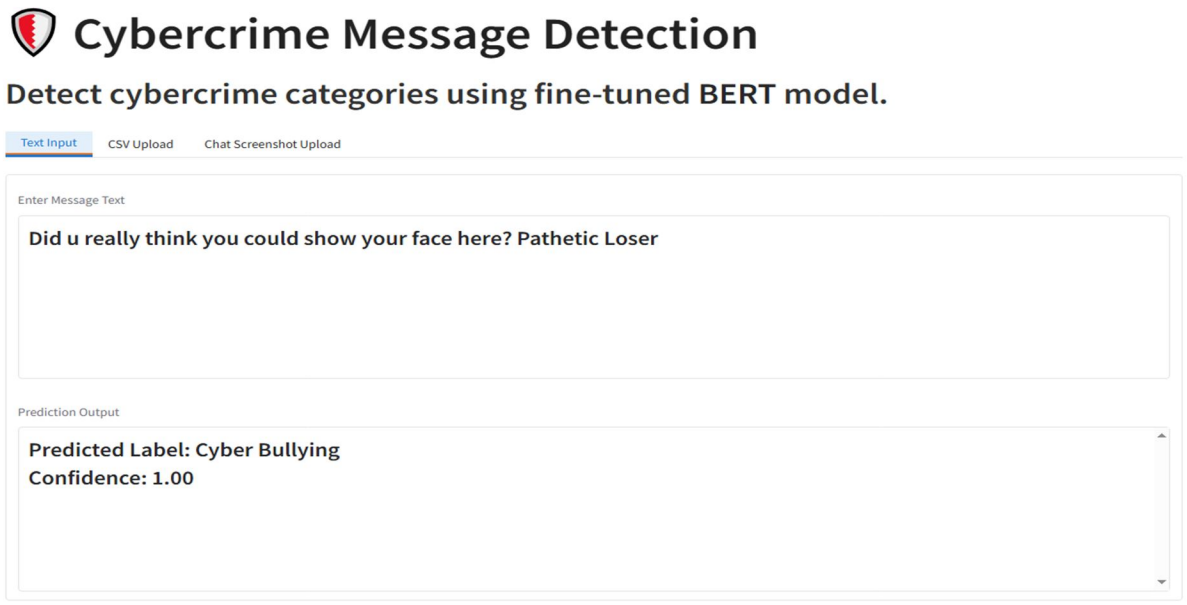


Fig.4 Text Input Result

B. CSV Input Results

Batch processing of large chat logs through CSV files validated the system's scalability. The model maintained consistent classification accuracy while handling thousands of messages simultaneously. This capability supports forensic investigations and organizational monitoring where voluminous chat data must be reviewed. Confidence scoring provided additional insight into prediction certainty, allowing prioritization of suspicious content for further analysis. Notably, minor drops in recall for rare classes were observed, which can be addressed with larger balanced datasets in future work.

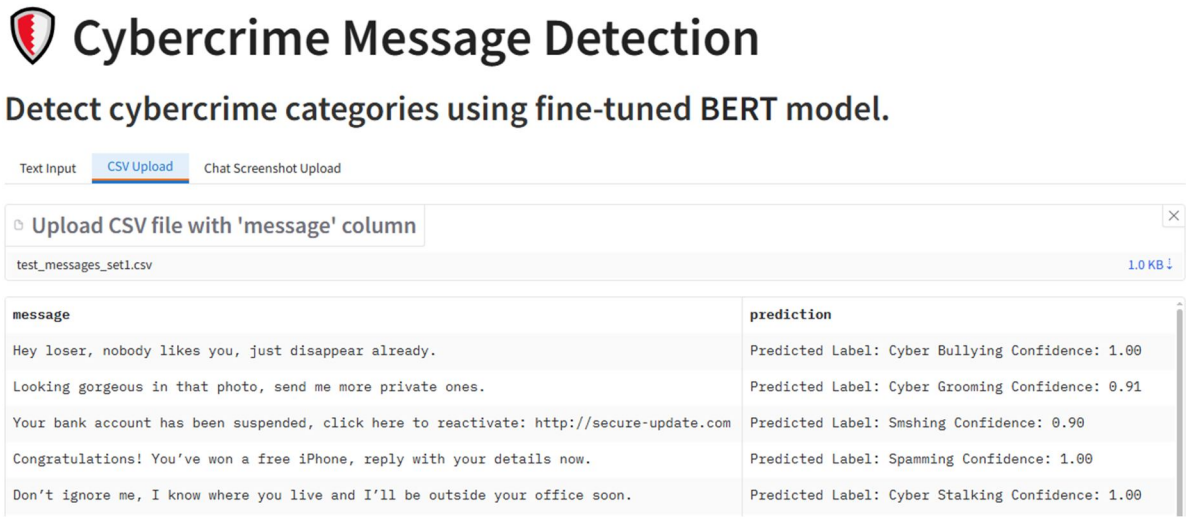


Fig.6 CSV Messages Prediction Result

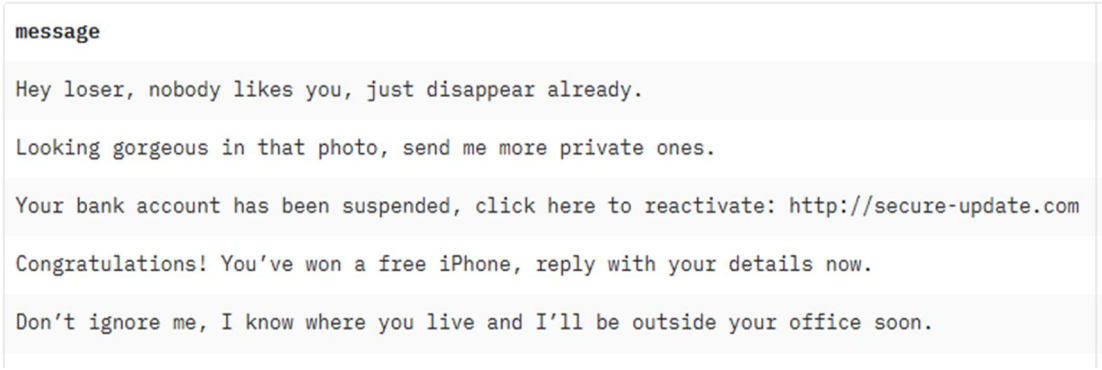


Fig CSV file Uploaded Messages

Fig CSV Messages Prediction

C. Screenshot Input Results

The integration of OCR enabled the analysis of chat screenshots often shared as evidence in cybercrime investigations. Text extraction through Tesseract proved effective for images of varying quality, with pre-processing steps improving readability and minimizing noise. The classification accuracy on OCR-extracted text closely matched that of direct text input, confirming the viability of a multimodal approach. Challenges were primarily related to low-resolution or highly distorted screenshots, suggesting avenues for future enhancement with advanced OCR techniques and image pre-processing.

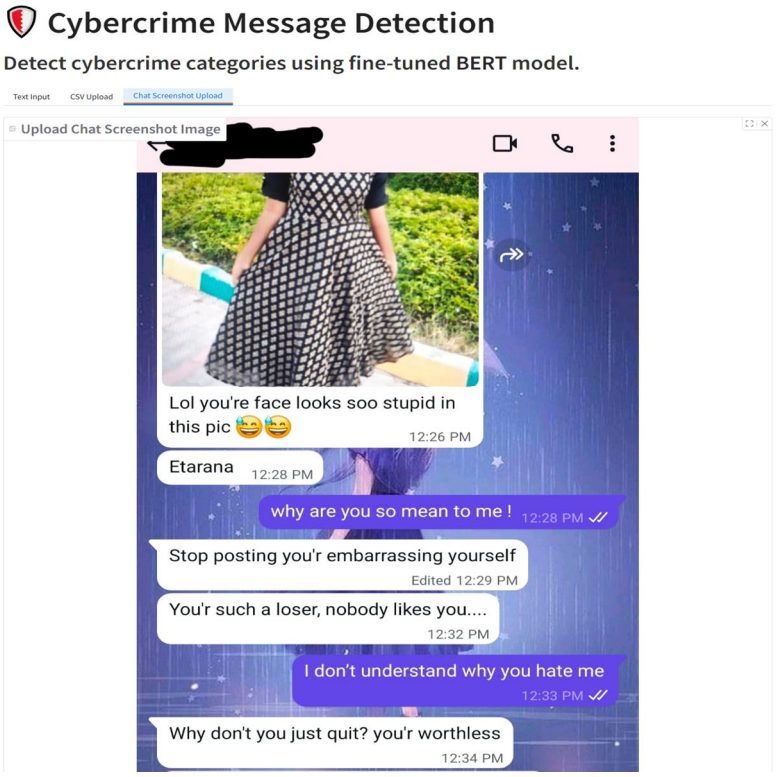


Fig .7 Message Screenshot Input

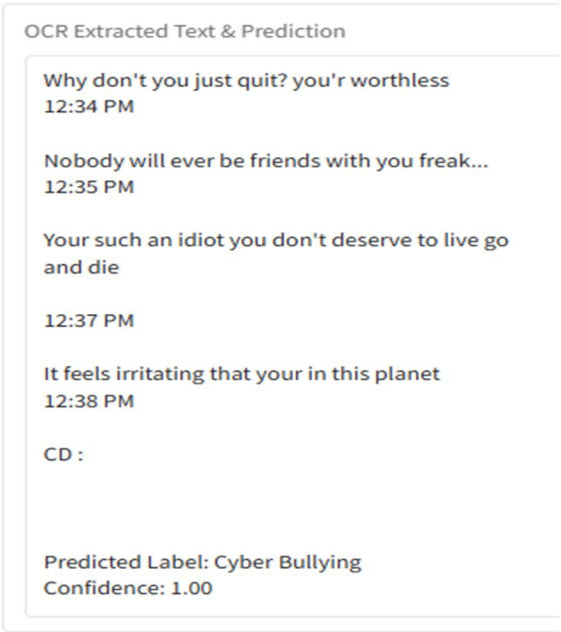


Fig.8 Message Screenshot Result

VI. CONCLUSION

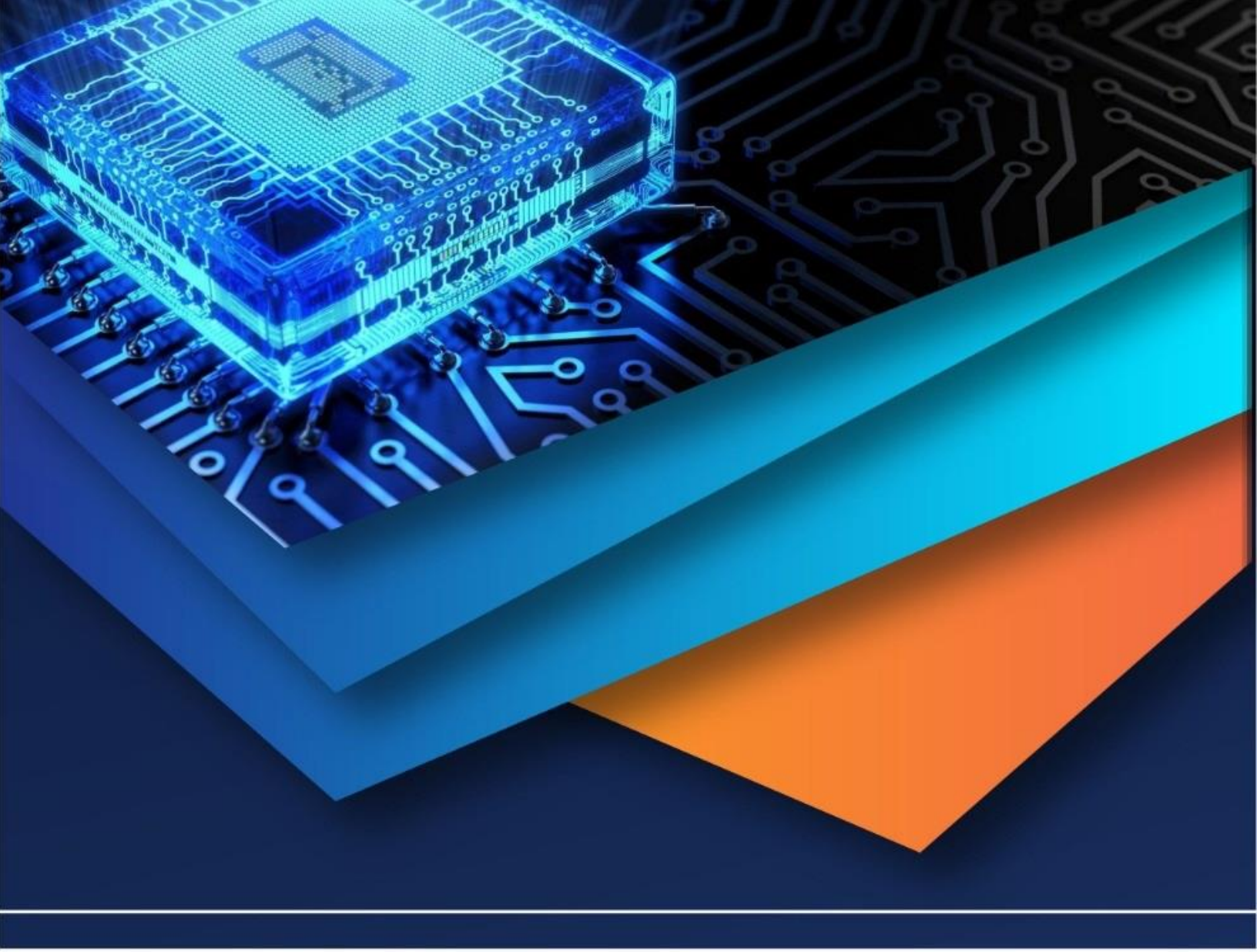
The CyGuardNLP framework demonstrates a comprehensive and effective approach for detecting cybercrime activities within chat-based communications. By integrating a fine-tuned BERT transformer model with OCR technology, the system is capable of processing diverse input formats including raw text, batch CSV data, and chat screenshots. The fusion of deep contextual understanding from NLP and visual text extraction enhances detection accuracy and flexibility in real-world applications. These results highlight the feasibility of employing advanced machine learning techniques for proactive cybercrime monitoring, contributing significantly to online safety and digital forensics. Furthermore, the modular design ensures that components can be updated or expanded independently, promoting adaptability and long-term sustainability of the system.

VII. FUTURE WORK

Further advancements could explore expanding the system's linguistic capabilities by incorporating multilingual models to accommodate increasingly diverse online user populations. Enhancing the contextual analysis by considering sequences of messages or conversations may improve detection of subtle or evolving cyber threats. Additionally, research into robust adversarial training techniques could strengthen resilience against sophisticated evasion strategies employed by cybercriminals. Optimizing the OCR component to handle non-standard languages, fonts, or low-quality images would broaden the scope of the system's applicability. Finally, integration with enterprise-grade cybersecurity infrastructure and real-time large-scale deployment pipelines would facilitate practical adoption and continuous protection in dynamic online environments.

REFERENCES

[1] Silva Sifath, Tania Islam, md Erfan, Samart Kumar Dey, md. Minhaj Ul Islam, md Samsuddoha, Tazizur Rahman (2024). "Recruitment Neural Network Based Multiclass Cyberbullying Classification", Natural Language ProcessingJournal,Volume9,2024,https://doi.org/10.1016/j.nlp.2024.100111.
[2] Ogunleye B, Dharmaraj B, "The Use of a Large Model for Cyberbullying Detection," Analytics,2023,2,694-707. https://doi.org/10.3390/analytics2030038.
[3] Kumar Y, Huang K, Perez A, Li J J, Morreale P et al. "Bias and Cyberbullying Detection and Data Generation Using Transformer AI Models and Top Large Language Models", Electronics 2024, 13, 3431.https://doi.org/10.3390/electronics13173431.
[4] Maity, K., Bhattacharya, S., & Saha, S. (2023). A Deep Learning Framework for the Detection of Malay Hate Speech. *IEEE Access*. DOI:10.1109/ACCESS.2023.3298808.
[5] Khan, S., Kamal, A., Fazil, M., & Alshara, M. (2022). HCovBi-Caps: Hate Speech Detection Using Convolutional and Bi-Directional GRU with Capsule Network. IEEEAccess.DOI:10.1109/ACCESS.2022.3143799.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY