

NLP Model For Narco Activity Detection

Ayushmaan Pandey¹, Ms. Chaitali Mhatre², Devansh Parmar³, Nishant Mishra⁴, Siddharth Nair⁵

^{1,3,4,5}Student, Department of Computer Engineering, Universal College of Engineering, Kaman, Maharashtra, India

²Assistant Professor, Department of Computer Engineering, Universal College of Engineering, Kaman, Maharashtra, India

Abstract - This paper presents a chat classification system using Natural Language Processing (NLP) and Machine Learning (ML) to detect suspicious conversations related to illegal drug transactions. With the rise of end-to-end encrypted messaging apps, cyber drug trafficking has become harder for law enforcement to monitor using traditional methods [3][5]. To address encryption and privacy challenges, the system is deployed in a simulated secure messaging environment. This approach responds to the increasing presence of illicit drug activity on platforms like Telegram and Instagram, which are difficult to analyze in real-time using conventional forensics [1][4][8].

The system replicates a standard chat platform to enable real-time analysis without bypassing encryption protocols. Conversations undergo preprocessing for cleaning and normalization, followed by conversion into numerical vectors using Term Frequency-Inverse Document Frequency (TF-IDF). These vectors are classified using a Logistic Regression model to determine whether a conversation is illicit or benign, similar to modern forensic frameworks [7][9].

If the probability exceeds a set threshold, the system flags the conversation and generates an alert on an admin dashboard. It also includes a feedback loop where admin input helps retrain and improve model accuracy over time. Additionally, an automated module generates forensic-style alerts for authorities [6][10]. Overall, the system offers a scalable and practical solution for early detection of digital drug trafficking.

Key Words: Natural Language Processing (NLP), Machine Learning (ML), Chat Classification, Drug Trafficking Detection, TF-IDF, Logistic Regression, Cybercrime Detection, Encrypted Messaging Analysis

1.INTRODUCTION

The digital landscape has significantly transformed modern communication, with instant messaging applications enabling fast and seamless global connectivity. However, the widespread use of these platforms has also created new opportunities for cybercrime and antisocial activities [1][3]. Encrypted messaging services such as Telegram and WhatsApp provide strong privacy and anonymity features, which are increasingly being misused for illegal activities, particularly the online trade of drugs and narcotics [4][8]. The largely unregulated nature of these platforms allows

traffickers to operate hidden marketplaces, often avoiding traditional law-enforcement monitoring while directly targeting vulnerable individuals [5].

Monitoring activity on such secure platforms presents a major challenge for law enforcement agencies (LEAs) and platform administrators. The massive volume, speed, and diversity of digital conversations make traditional manual investigation methods slow and inefficient, leading to massive backlogs in evidence analysis [3][10]. Moreover, individuals involved in illegal activities often rely on coded language, slang, and evolving conversational patterns to hide their intentions [5]. As a result, manual moderation and basic keyword-based filtering systems frequently fail to detect suspicious content effectively, often suffering from the significant delays inherent in traditional administrative reporting [9]. This delay in detection makes it difficult for authorities to respond quickly, allowing illegal drug-trafficking activities to persist and negatively affect community safety.

To address these challenges, Artificial Intelligence (AI), particularly Natural Language Processing (NLP) and Machine Learning (ML), has become an important tool in digital forensics and automated chat monitoring [3][7]. NLP techniques can analyze unstructured conversational text to identify contextual meaning, linguistic patterns, and semantic relationships, making them more effective than simple keyword searches [1][10]. By integrating NLP with supervised machine-learning models, systems can process large volumes of chat data, analyze linguistic features, and automatically flag potentially suspicious conversations for further investigation [2][5].

In response to this growing concern, this research proposes **Chat Classifier**, an automated monitoring system designed to detect potential drug-trafficking activities within digital communications. Because direct monitoring of real messaging platforms is restricted by privacy regulations and end-to-end encryption mechanisms, the system is implemented and evaluated within a simulated secure messaging environment [5][8]. The proposed framework applies NLP techniques to preprocess conversation logs and convert textual data into numerical representations using the Term Frequency-Inverse Document Frequency (TF-IDF) method. These features are then processed using a Logistic Regression classifier, selected for its efficiency in binary classification and its ability to produce probabilistic risk scores [7]. The

model classifies conversations as illicit or normal. By automatically identifying suspicious messages and notifying administrators through a centralized web-based dashboard, the Chat Classifier aims to significantly reduce detection time and support proactive efforts, including the generation of automated forensic alerts to combat online drug-trafficking activities [6][7].

1.1 Project Idea

The rapid growth of digital communication platforms has created new opportunities for illegal activities such as online drug trafficking. Many traffickers use encrypted messaging services and coded language to conduct transactions, making it difficult for authorities to detect suspicious communications through traditional monitoring methods. As the volume of online conversations increases, manual analysis becomes inefficient and time-consuming.

The objective of this project is to develop a Chat Classifier system that uses Natural Language Processing (NLP) and Machine Learning (ML) techniques to automatically analyze chat conversations and identify patterns related to illegal drug transactions. By applying a Logistic Regression model along with TF-IDF-based text processing, the system can classify conversations as illicit or benign and generate alerts for administrators to review potentially suspicious activities.

2. EXISTING SYSTEMS

Existing systems for monitoring illegal activities in digital communications largely depend on manual surveillance and basic keyword filtering techniques. In many cases, human analysts are required to read through large volumes of chat logs or online messages to identify suspicious patterns. This approach is time-consuming, inefficient, and highly dependent on the experience and judgment of the analyst. As the volume of online communication continues to grow rapidly, manual monitoring becomes increasingly difficult to manage effectively.

Some systems also use simple rule-based filtering methods that flag conversations containing predefined keywords related to drugs or illegal activities. However, these systems often fail to understand the context of conversations and may produce a large number of false positives or miss conversations that use coded language or slang. Due to these limitations, existing systems are not always capable of accurately detecting illicit activities in real-time, highlighting the need for more intelligent and automated solutions using machine learning and natural language processing techniques.

2.1 Literature Survey

The rapid proliferation of instant messaging applications has transformed digital communication, but has also facilitated a rise in cybercrimes. This has prompted researchers to explore Artificial Intelligence (AI) and Natural Language Processing (NLP) as mechanisms for automated content moderation and digital forensics.

2.1.1 Cybercrime on Encrypted Platforms

Recent studies extensively document the exploitation of encrypted messaging applications for illicit activities. Nali et al. [4] identified a thriving, unregulated marketplace on Telegram for the illegal sale of cannabis and nicotine products, noting that the platform's anonymity makes manual monitoring nearly impossible. Building upon this, Sonawane et al. [8] developed an AI pipeline to detect drug-related content, finding that over 90% of flagged messages on Telegram bots were active seller announcements. Sreeram and Bansal [5] emphasized that while end-to-end encryption safeguards user privacy, it simultaneously creates a formidable barrier for law enforcement agencies attempting to intercept illicit communications.

2.1.2 NLP and Machine Learning for Content Classification

To combat malicious digital communication, researchers rely heavily on NLP and Machine Learning (ML). Naz and Illahi [1] reviewed NLP techniques for detecting harmful User-Generated Content (UGC), concluding that current ML models are highly effective but predominantly reactive. Armoogum et al. [2] and Tuarob et al. [9] demonstrated the efficacy of using the Term Frequency-Inverse Document Frequency (TF-IDF) technique alongside traditional machine learning algorithms to classify abusive and crime-related text, proving TF-IDF remains a robust feature extraction method. In the realm of digital forensics, Mogaji et al. [7] designed an intelligent chat monitoring system to detect suspicious activities, emphasizing the critical need for real-time analysis. Similarly, Bokolo and Liu [3], and Xi et al. [10] highlighted how NLP can be utilized to identify evidentiary topics in noisy, colloquial forensic data.

2.1.3 Research Gaps and Proposed Contribution

Despite significant advancements in AI-driven forensics, existing solutions face two major limitations:

1. **Encryption and Privacy Barriers:** Implementing centralized real-time monitoring on commercial apps like WhatsApp or Telegram is severely restricted by end-to-end encryption and strict data privacy laws, making external data scraping largely impractical.
2. **Computational Overhead and Reactivity:** Many existing frameworks either rely on highly

complex, resource-heavy deep learning models that are difficult to scale, or they remain purely reactive, analyzing crimes only after they occur.

How the Proposed System Overcomes These Problems: The proposed "Chat Classifier" system is specifically designed to overcome these existing barriers through two strategic implementations:

1. **Circumventing Encryption via a Simulated Environment:** Rather than attempting to penetrate external encrypted networks, the system is purposefully deployed within a simulated "dummy" social media chat environment. This allows the system to ingest raw conversation logs directly at the source in real time, entirely bypassing the legal and technical restrictions of commercial end-to-end encryption protocols.
2. **Proactive, Lightweight Detection:** To avoid the computational drag of heavy deep learning algorithms, the system utilizes TF-IDF for text vectorization paired with a highly efficient Logistic Regression classification model. Instead of reacting post-incident, this lightweight approach proactively calculates a probability score for ongoing conversations. The moment drug-related keywords, slang, or transactional patterns exceed a predefined threshold, the system immediately flags the threat and alerts administrators on a web-based dashboard, enabling rapid, early-stage intervention.

3. PROPOSED SYSTEMS

The proposed system is a Chat Classifier designed to automatically detect suspicious conversations related to illegal drug trafficking using Natural Language Processing (NLP) and Machine Learning (ML) techniques. The system analyzes chat messages exchanged within a simulated messaging environment and identifies patterns or keywords that may indicate illicit activities. By automating the monitoring process, the system reduces the dependence on manual analysis and improves the speed and efficiency of detecting suspicious communications.

In this system, conversation logs are first collected and processed using text preprocessing techniques such as cleaning, tokenization, and removal of unnecessary words. The processed text is then converted into numerical form using the TF-IDF vectorization method. A Logistic Regression model is trained on labeled datasets containing both illicit and normal conversations. When a new message is received, the trained model analyzes the conversation and generates a probability score indicating whether the message is illicit or benign.

If the probability score crosses a predefined threshold, the system flags the conversation and generates an alert for the administrator. The administrator can review the flagged message through a web-based dashboard and decide whether it represents a genuine threat or a false positive. Feedback from the administrator can later be used to retrain and improve the model's accuracy. This automated approach allows the system to monitor large volumes of communication efficiently while assisting investigators in identifying suspicious activities at an early stage.

3.1 Algorithm

The proposed system uses Logistic Regression as the primary machine learning algorithm for classifying conversations as illicit or benign. Logistic Regression is a supervised learning algorithm commonly used for binary classification problems where the output belongs to one of two categories. In this project, the algorithm predicts the probability that a given conversation is related to illegal drug transactions.

Before applying the algorithm, the conversation text is preprocessed using Natural Language Processing (NLP) techniques. This includes cleaning the text, removing unnecessary characters, and converting the messages into a structured format. After preprocessing, the text data is converted into numerical vectors using the TF-IDF (Term Frequency–Inverse Document Frequency) method. This technique measures the importance of words within the conversation and allows the machine learning model to analyze textual data effectively.

The Logistic Regression model is then trained using a labeled dataset containing examples of both illicit and normal conversations. During training, the algorithm learns the relationship between textual features and the classification labels. When a new conversation is received, it is transformed into a TF-IDF vector and passed to the trained model, which generates a probability score between 0 and 1. If the probability exceeds a predefined threshold, the system classifies the conversation as illicit and generates an alert for further review. This framework allows the system to automatically analyze conversations and identify suspicious patterns efficiently.

3.2 Design Details

The design details of the proposed system describe the overall structure and organization of the Chat Classifier application. It explains how different components of the system interact with each other to analyze chat conversations and detect suspicious activities. The system is designed in a modular manner to ensure that each component performs a specific function within the overall workflow.

The architecture of the system consists of multiple interconnected modules that handle tasks such as data collection, text preprocessing, conversation analysis, and alert generation. Initially, conversation logs from the simulated chat environment are received by the backend system. The messages are then processed using Natural Language Processing techniques to clean and prepare the text for analysis. After preprocessing, the data is passed to the machine learning model, which evaluates the conversation and generates a probability score indicating whether the message is illicit or benign.

Based on the generated score, the system determines whether the conversation should be flagged. If the score exceeds a predefined threshold, the system generates an alert and stores the information in the database. The administrator can then access the system through a web-based dashboard to review flagged conversations and monitor system performance. This modular design ensures efficient processing, easy maintenance, and scalability of the system as the volume of data increases.

3.2.1 System Architecture

The system architecture of the proposed Chat Classifier illustrates how different components interact to analyze and detect suspicious conversations. Initially, conversation messages from the simulated chat platform are sent to the backend system for processing. The backend applies text preprocessing and Natural Language Processing techniques to clean and prepare the data. The processed text is then converted into numerical vectors using the TF-IDF method and analyzed by the trained Logistic Regression model. Based on the generated probability score, the system determines whether the conversation is illicit or benign. If a conversation is identified as suspicious, an alert is generated and displayed to the administrator through the web-based dashboard for further review.

Natural Language Processing (NLP) and Machine Learning (ML). The system follows several steps, including data collection, preprocessing, feature extraction, model training, and prediction. These steps help in analyzing chat messages and identifying patterns related to illegal drug transactions.

First, a labeled dataset containing both illicit and normal conversations is collected. The data is then preprocessed using NLP techniques such as text cleaning, tokenization, and removal of stop words. After preprocessing, the text data is converted into numerical vectors using the TF-IDF technique so that it can be processed by the machine learning model.

The Logistic Regression algorithm is then trained on the processed dataset to learn patterns that distinguish suspicious conversations from normal ones. When a new conversation is received, it goes through the same preprocessing and TF-IDF conversion process before being analyzed by the trained model. The model generates a probability score, and if it exceeds a predefined threshold, the conversation is flagged, and an alert is generated for the administrator.

4. RESULTS

The results of the proposed Chat Classifier system demonstrate that the machine learning model is able to analyze and classify chat conversations effectively. After training the Logistic Regression model on the labeled dataset, the system was tested using new conversation samples to evaluate its performance. The model successfully identifies patterns associated with suspicious or illicit drug-related communication and classifies messages as either illicit or benign.

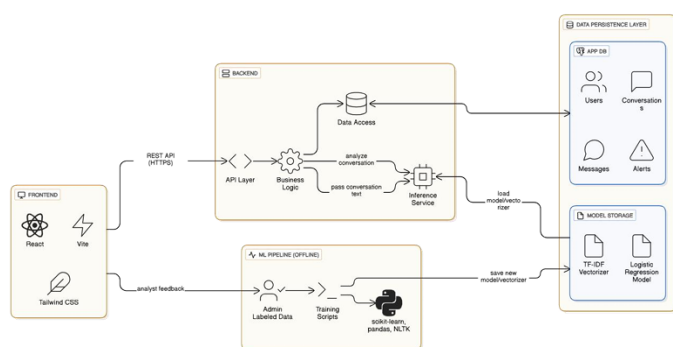


Fig -1: System Architecture

3.3 Methodology

The methodology of the proposed system explains the process used to detect suspicious conversations using

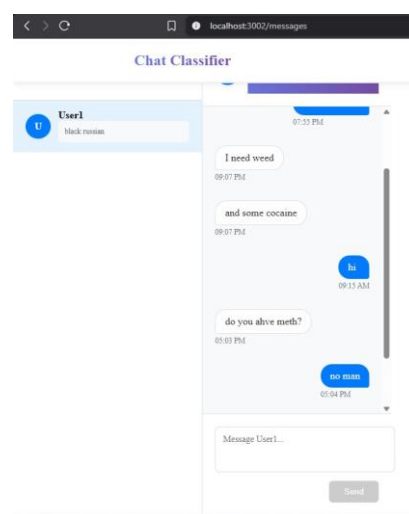


Fig - 2: User Chat Interface and Real-time Monitoring

This interface represents the front-end chat environment where users communicate. The system actively monitors

the text stream, identifying high-risk narcotic-related keywords like "heroin," "weed," "cocaine," and "meth" as they are typed. This serves as the primary data ingestion point for the classification engine to detect illicit activity during live conversations.

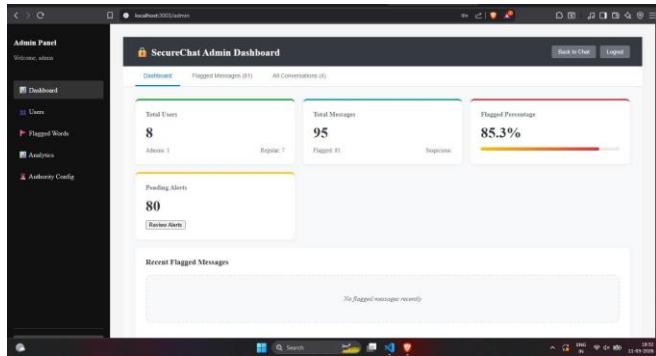


Fig - 3: SecureChat Admin Dashboard Overview

The main Admin Dashboard provides a high-level summary of system-wide activity, displaying critical metrics such as the total number of users and messages. It features a prominent "Flagged Percentage" gauge, currently showing a high rate of 85.3%, and a "Pending Alerts" counter to notify administrators of urgent security breaches. This centralized view allows authorities to quickly assess the overall threat level of the platform.

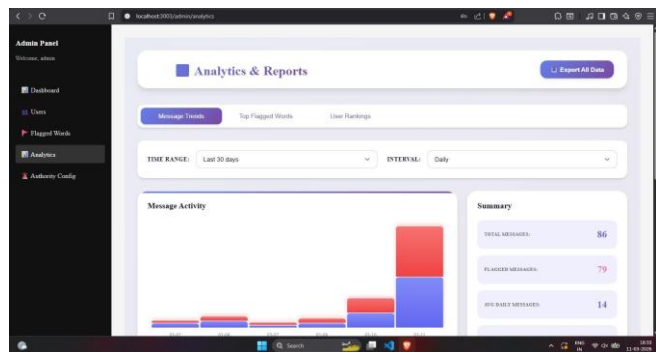


Fig - 4: Analytics and Message Activity Reports

This section provides deep insights into communication trends through visual bar charts, comparing total message volume against flagged illicit messages. It includes a summary sidebar with statistical data, such as average daily messages and total flagged counts, helping administrators identify periods of peak suspicious activity. The "Export All Data" feature ensures that these logs can be extracted for formal legal documentation.

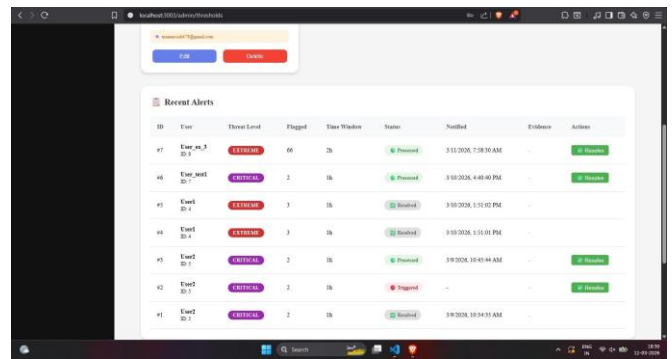


Fig - 5: Real-time Alert Monitoring and Evidence Log

The Recent Alerts log provides a chronological record of every security trigger generated by the system, categorized by threat levels like "EXTREME" or "CRITICAL." Each entry includes the user ID, the number of flagged messages, and a status indicator showing if the alert has been "Processed," "Triggered," or "Resolved." This screen acts as a digital evidence locker, allowing

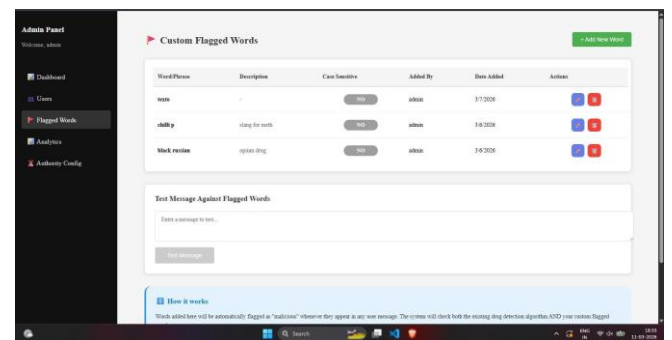


Fig - 6: Custom Flagged Words and Slang Configuration

This administrative tool allows for the dynamic updating of the detection engine by adding custom keywords and localized drug slang. It includes a management table showing words like "chilli p" (meth slang) and "black russian" (opium drug), along with their descriptions and case-sensitivity settings. A "Test Message" sandbox is also provided to verify if the detection algorithm correctly flags specific phrases before they are deployed live.

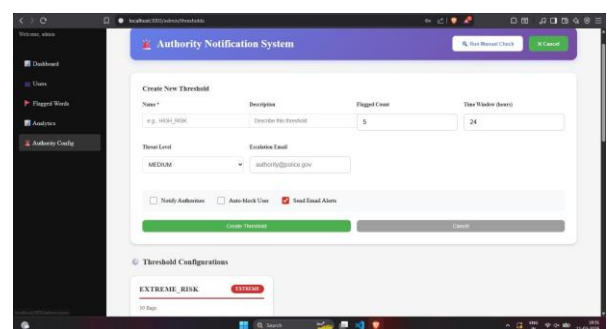


Fig - 7: Authority Notification System Configuration

This feature enables the setup of automated escalation protocols by creating risk thresholds, such as "EXTREME_RISK." Administrators can define specific conditions, like a user hitting 50 flags within a 2-hour window, to trigger immediate actions. These actions include notifying external authorities, auto-blocking the user, and sending automated email alerts to official law enforcement addresses.

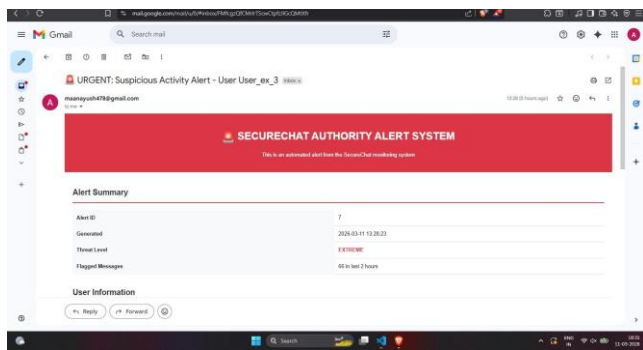


Fig -8: Automated Authority Email Alert System

When a high-risk threshold is breached, the system automatically generates an "URGENT" email notification to the configured law enforcement agency. The email contains a summary of the threat, including an Alert ID, the specific threat level (e.g., EXTREME), and the total count of flagged messages within a timeframe. This ensures that response teams are notified instantly of suspicious activity without requiring manual intervention from a human moderator.

However, the current implementation of the system has certain limitations. The model analyzes only the current message bubble and does not consider the previous context of the conversation. Because of this, the system may fail to detect suspicious intent that becomes clear only when multiple messages are analyzed together. Additionally, individuals involved in illicit communication may use predefined code words or disguised language agreed upon between communicators. Such coded conversations may bypass the detection mechanism since the system primarily relies on known patterns present in the training data. Therefore, while the model demonstrates effective message-level classification, future improvements should focus on incorporating conversation context analysis and more advanced language models to better detect hidden or coded communication patterns.

5. CONCLUSIONS

The Chat Classifier proposed system illustrates the implementation of Natural Language Processing and Machine Learning for identifying suspicious talks about illegal drugs. Using TF-IDF and Logistic Regression, the system can identify chat messages as illicit or benign to minimize manual checking and improve detection

efficiency. However, at present, the current system just evaluates the given message and does not take into account the prior conversation context. This method works because it makes it so that communicators can avoid being caught by using coded language or code words that everybody agrees on. Future improvements can entail context-based analysis and even more advanced models to enhance the accuracy and reliability of the system.

REFERENCES

- [1] I. Naz, R. Illahi, "Harmful Content on Social Media Detection Using NLP," *Advances*, 2023, pp. 49-59.
- [2] S. Armoogum, D. A. Dewi, V. Armoogum, N. Melanie, T. B. Kurniawan, "Unveiling Criminal Activity: a Social Media Mining Approach to Crime Prediction," *Journal of Applied Data Sciences*, 2024, pp. 1482-1494.
- [3] B. G. Bokolo, Q. Liu, "Artificial Intelligence in Social Media Forensics: A Comprehensive Survey and Analysis," *Electronics*, 2024, pp. 1-24.
- [4] M. C. Nali, V. Purushothaman, Z. Li, M. Z. Larsen, R. E. Cuomo, J. Yang, T. K. Mackey, "Identification and Characterization of Illegal Sales of Cannabis and Nicotine Delivery Products on Telegram Messaging Platform," *Nicotine and Tobacco Research*, 2024, pp. 771-779.
- [5] K. Y. Sreeram, K. Bansal, "Algorithmic Chat Monitoring for Mitigating Crime in Telegram: A Multi-Pronged Approach to Prevention and Forensics," *International Journal of Innovative Research in Technology (IJIRT)*, 2024.
- [6] H. C. Aydogan, B. Yikar, H. Balandız, S. Özsoy, "Assessing ChatGPT-4's ability to generate forensic reports: a study of artificial intelligence in forensics," *Egyptian Journal of Forensic Sciences*, 2025, pp. 1-12.
- [7] S. A. Mogaji, T. T. Odufuwa, C. A. Afolalu, O. O. Faboya, M. V. Ige, O. D. Idowu, "Development of an Intelligent Chat Monitoring and Suspicious Activity Detection System," *FUOYE Journal of Pure and Applied Sciences*, 2025, pp. 212-224.
- [8] M. S. Sonawane, D. P. Patil, A. Kumar, S. Antoniv, "AI Detection of Drug Activity in Telegram Bots and Groups for Security Applications," *18th International Scientific and Technical Conference on Instrumentation Engineering*, 2025, pp. 452-455.
- [9] S. Tuarob, P. Tatiyamaneekul, S. Pongpaichet, T. Tawichsri, T. Noraset, "Beyond administrative reports: a deep learning framework for classifying and monitoring crime and accidents leveraging large-scale

online news," Neural Computing and Applications, 2025, pp. 7183-7205.

- [10] J. Xi, M. Siegel, D. Labudde, M. Spranger, "Towards a joint semantic analysis in mobile forensics environments," Forensic Science International: Digital Investigation, 2025, 301846.