Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 3, Juni 2025, Page. 1203-1216
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4593

Digital Forensic Chatbot Using DeepSeek LLM and NER for Automated Electronic Evidence Investigation

Nuurun Najmi Qonita[1], Maya Rini Handayani[2], Khothibul Umam*[3]

[1,2,3]Department of Information Technology, Science and Technology Faculty, Walisongo State Islamic University Semarang, Indonesia

Email: [3]khothibul_umam@walisongo.ac.id

Received : Apr 12, 2025; Revised : May 12, 2025; Accepted : May 13, 2025; Published : Jun 10, 2025

Abstract

The growing complexity of cybercrime necessitates efficient and accurate digital forensic tools for analyzing electronic evidence. This research presents an intelligent digital forensic chatbot powered by DeepSeek Large Language Model (LLM) and Named Entity Recognition (NER), designed to automate the analysis of various digital evidence, including system logs, emails, and image metadata. The chatbot is deployed on the Telegram platform, providing real-time interaction with investigators. The metric results show that the chatbot achieves a precision of 83.52%, a recall of 88.03%, and an F1-score of 85.71%. These results demonstrate the chatbot's effectiveness in accurately detecting forensic entities, significantly improving investigation efficiency. This study contributes to digital forensics by integrating LLM and NER for enhanced evidence analysis, offering a scalable and adaptive solution for automated cybercrime investigations. Future research may explore integrating anomaly detection and blockchain-based evidence integrity.

Keywords : *Chatbot, Cybercrime Investigation, DeepSeek LLM, Digital Forensics, Named Entity Recognition.*

1. INTRODUCTION

The rapid development of information and communication technology has brought significant changes in various fields, including in the realm of cyber and digital forensics. Digital forensics has become a critical field as cybercrime continues to evolve, presenting new challenges for investigators in identifying and analyzing electronic evidence. With the increasing amount of electronic data generated every day, digital forensic investigations face great challenges in extracting and analyzing evidence efficiently [1], [2]. Technological developments make it possible to gather information about crime situations including finding hidden links between organizations and individuals who commit crimes with artificial intelligence (Radulov in [3]). Traditional manual investigation methods are time-consuming and prone to errors, necessitating automated solutions. One approach that can be used to overcome this challenge is to apply artificial intelligence technology in the investigation process, specifically through the utilization of chatbots based on the Large Language Model (LLM) model [4]–[6].

Artificial intelligence-based chatbots have been widely applied in various sectors, such as customer service, text analysis, to document-based information retrieval systems [7], [8]. However, there is limited research integrating DeepSeek LLM with NER in a real-time chatbot for digital forensics, particularly on Telegram platforms. To date, there have been no original research publications focusing on the application of LLM in the digital forensics domain [9]. LLM models such as DeepSeek have the ability to understand the context of conversations as well as extract important information from digital documents, including system logs, emails, image metadata, and other digital communications

Vol. 6, No. 3, Juni 2025, Page. 1203-1216
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4593

[10]. In A Comparison of DeepSeek and Other LLMs by Gao et al. it was found that DeepSeek showed higher classification accuracy than Gemini, GPT, and Llama in most tests [11], [12]. One technique that can support the effectiveness of chatbots in the investigation process is Named Entity Recognition (NER), which aims to identify and categorize important entities in electronic evidence, such as IP addresses, suspicious domains, event timestamps, and user or device identities [13], [14].

This research aims to develop an intelligent digital forensic chatbot powered by DeepSeek Large Language Model (LLM) and Named Entity Recognition (NER), which can efficiently analyze diverse types of electronic evidence. Unlike previous studies that focus on rule-based or limited-domain chatbots, this study leverages advanced LLM capabilities for contextual understanding and accurate entity recognition, making it a versatile tool for forensic analysis. Chatbot was chosen because it provides a user-friendly interface, allowing users to interact intuitively through natural language [15], [16]. The implementation of this chatbot is expected to improve the efficiency of the investigation process by providing automated analysis of the electronic evidence provided [17]. With the integration of NER techniques and DeepSeek models, the chatbot will be able to accurately extract important information, provide investigation recommendations, and produce more structured forensic reports [18]. In addition, this research also evaluates the performance of the developed chatbot using appropriate evaluation metrics to measure accuracy, precision, and efficiency in analyzing electronic evidence. Thus, this research is expected to contribute to the development of a digital forensics system that is more intelligent and adaptive to the needs of modern investigations.

2. METHOD

This research aims to develop a digital forensic chatbot based on DeepSeek's Large Language Model (LLM) model equipped with Named Entity Recognition (NER) features to assist the investigation process of various types of electronic evidence. In the process, this research uses a methodological approach consisting of several main stages, namely: data collection from various digital sources, data preprocessing to ensure quality and consistency, development of a chatbot model that can understand the forensic context, implementation of the chatbot into a suitable platform, and testing and evaluation of system performance. The overall mechanism of this research process is depicted in figure 1.

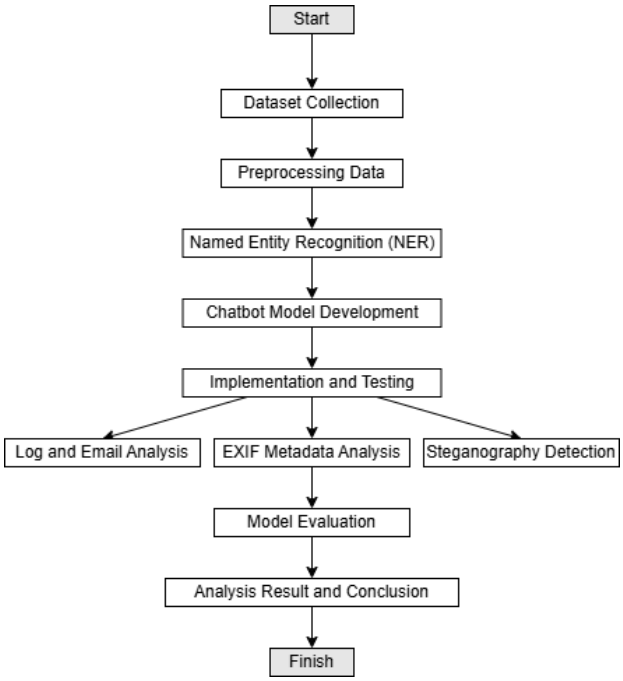

Figure 1. Research Flow of Chatbot Development

Vol. 6, No. 3, Juni 2025, Page. 1203-1216
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4593

2.1. Dataset Collection

The data used in this research includes various types of electronic evidence, such as system logs, emails, image metadata, as well as other digital communications. The datasets were obtained from open source and forensic databases available for research purposes. The collected data includes relevant information, such as IP address, suspicious domain, event timestamp, device ID, as well as image metadata containing GPS location information, camera type, and image capture timestamp.

2.2. Data Preprocessing

Data preprocessing aims to ensure that the data used in the model is clean, structured, and ready to be analyzed [19]. An input sentence entered by the user will be processed into data that is easier for the chatbot to understand and process. This stage is known as preprocessing, which is a crucial step in *Natural Language Processing* (NLP)-based systems because it determines the quality of the input that goes into the model. The *preprocessing* process in this research consists of three main parts, as follows:

1. Data Cleaning: This stage is carried out to ensure that the data used does not contain disturbances that can affect the performance of the model. This step includes removing duplicate data, handling empty data or null values, integrating data from various sources, and handling missing values. In addition, text normalization is also performed, such as removal of special characters, irrelevant punctuation, unimportant numbers, and conversion of letters to all lowercase so that data consistency is maintained.
2. Data Transformation: Aims to transform raw data into a format that can be understood by the model. At this stage, extraction of important features from various types of data such as system logs and image metadata (EXIF) is done. Next, the text is converted into tokens using tokenization techniques. This process breaks sentences into smaller parts (words or phrases) to facilitate analysis. In addition, encoding of categorical data is used to convert symbolic values into numerical forms that can be recognized by machine learning algorithms.
3. Named Entity Recognition (NER): The final stage in preprocessing is the application of Named Entity Recognition (NER). This technique is used to extract important entities from digital documents, such as usernames, IP addresses, locations, and timestamps of digital events. The NER model used has been trained with datasets relevant to the domain of cybersecurity and digital forensics, thus improving the accuracy in the identification of important entities that often appear in electronic evidence.

The development process utilized the Python programming language. For Named Entity Recognition (NER) tasks, the spaCy library was employed along with the pre-trained en_core_web_sm model. Text preprocessing and tokenization were handled using NLTK. The chatbot interfaced directly with DeepSeek LLM. Image metadata extraction and error level analysis (ELA) were conducted using Pillow, OpenCV, and piexif. Additional steganographic analysis was supported by the Stegano library.

2.3. Chatbot Model Development

The chatbot model in this research was developed using DeepSeek's Large Language Model (LLM) designed to intelligently understand and respond to electronic evidence-based text. The chatbot has the ability to analyze system logs and digital communications to identify suspicious patterns potentially related to cybercrime activities. In addition, the chatbot is designed to present relevant analysis results by referring to forensic databases as a reference in supporting the investigation process.

The DeepSeek LLM was selected due to its capability in understanding multilingual forensic texts, support for longer context windows, and strong performance on open-domain tasks with domain adaptation. Compared to other open-source LLMs such as GPT or Gemini, DeepSeek offers better balance between computational efficiency and forensic reasoning capabilities, especially in low-

Vol. 6, No. 3, Juni 2025, Page. 1203-1216
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4593

resource investigative scenarios. DeepSeek LLM was configured with a 16,384-token context window and 32-layer transformer architecture, optimized for low-latency inference. Instead of full fine-tuning, prompt engineering and entity injection from the NER module were used to adapt the model to forensic tasks, allowing more focused and relevant analysis.

To improve performance and accuracy in the analysis process, the chatbot model is integrated with a MySQL database used to store information related to digital evidence and the results of investigations that have been conducted. This integration allows the system to record, access and manage investigation data efficiently. In addition, the model was also developed to be able to interact in real-time through the Telegram platform, making it easier for users, especially digital investigators, to conduct analysis and make decisions directly through a familiar and accessible interface.

The integration of DeepSeek LLM and NER was performed using Python-based modules where the NER output is used to enrich the context passed to the LLM. Detected entities such as IPs, user IDs, timestamps, and suspicious domains are inserted as part of the prompt context, allowing DeepSeek to generate more precise and evidence-aware responses.

2.4. Implementation and Testing

In the implementation and testing phase, the developed chatbot was thoroughly tested. The trials were conducted to evaluate the accuracy, response speed and robustness of the system in the face of diverse types of electronic evidence and analysis requests from users. The developed chatbot was tested with various investigation case scenarios, including:

1. Log and Email Analysis: The chatbot is given a dataset of system logs and emails that contain indications of security threats, and evaluated for its ability to detect anomalies and suspicious entities.
2. Image Metadata Analysis: The chatbot was asked to extract EXIF metadata from images to identify location, timestamp, and device information.
3. Steganography Detection: Models were tested to detect the presence of hidden messages in images using histogram analysis and Error Level Analysis (ELA) techniques. The ELA method is used to visually detect image manipulation and has been developed in the field of digital forensic science [20].

2.5. Model Performance Evaluation

Evaluation of chatbot performance is done using some of the following metrics:

1. Named Entity Recognition (NER) Accuracy: Measured using Precision, Recall, and F1-Score to assess how well a chatbot recognizes entities in electronic evidence. The implementation of NER enables faster and more efficient analysis and retrieval of information, significantly reducing the time required to process and understand large document content [21]–[23]. Precision is calculated as TP / (TP + FP), where TP is the number of correct entity detections and FP is the number of incorrect detections. Recall is calculated as TP / (TP + FN), and F1- Score as the harmonic mean of Precision and Recall.
2. LLM DeepSeek Model Performance: Tested based on the quality of chatbot responses in providing analysis and investigation recommendations using BLEU Score and ROUGE Score. BLEU Score is a simple way to measure the extent to which our answers match existing reference answers [24]. According to Moradi et al. [25], ROUGE works by measuring the overlap of n-grams between the generated text and the original text written by humans, where a larger score indicates a higher similarity between the two answers.
3. Evaluation of Image Metadata Analysis in Digital Forensics: Image forensics aims to extract information from digital images to determine their authenticity. The main focus is to detect,

Vol. 6, No. 3, Juni 2025, Page. 1203-1216
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4593

identify, or trace manipulated images, and protect the integrity and authenticity of images from potential misuse [26]. In this context, the success rate of the chatbot in extracting and interpreting image metadata is evaluated and compared with standard forensic tools commonly used in the field of digital forensics.

3. RESULT

This section describes in detail the results of the implementation of DeepSeek's LLM-based digital forensics chatbot with Named Entity Recognition (NER) in analyzing electronic evidence, both in text and image form. The evaluation was conducted based on the accuracy of electronic evidence identification, the effectiveness of the chatbot in supporting investigations, as well as system performance in terms of responsiveness and integration with forensic databases. In addition, evaluation metrics were calculated to assess the extent to which the system can accurately recognize digital evidence.

3.1. Implementation and Testing of Chatbot on Text Analysis.

3.1.1. Testing Entity Identification in Electronic Evidence

This test aims to evaluate the chatbot's ability to identify important entities from various types of text-based electronic evidence. The datasets used include system logs, emails (ham and spam), and image data that stores EXIF information. These datasets were chosen because they are often found in digital and cyber forensics cases.

The Named Entity Recognition (NER) model integrated in the chatbot is tasked with recognizing crucial entities such as IP addresses, email addresses, suspicious domains, device IDs, usernames, and credit card numbers from the digital evidence content.

The test results in table 1 show that the NER model implemented in the chatbot can recognize important entities with an average F1-Score of 82.48%, based on the evaluation of four types of digital evidence. Precision and Recall metrics also show high performance, with an average of 78.41% and 87.85%, respectively.

Table 1. Evaluation of NER Performance by Electronic Evidence Type

No	Types of Digital Evidence	Precision (%)	Recall (%)	F1-Score (%)
1	System Logs	89.2	85.4	87.2
2	Ham Email	77.78	98.72	87.01
3	Email Spam	65.17	84.06	73.42
4	EXIF Image Metadata	83.50	81.20	82.30
	Average	78.41	87.85	82.48

The variation in performance metrics across evidence types reflects the characteristics of each dataset. System logs show high precision and F1-Score due to their structured format, which aids accurate entity recognition. Ham emails achieve exceptionally high recall (98.72%) as their formal and consistent language allows the model to generalize well, although the lower precision (77.78%) suggests occasional over-detection. In contrast, spam emails result in lower precision (65.17%) due to noisy and obfuscated language, though recall remains relatively high (84.06%), indicating that the model can still capture many potential entities, albeit with more false positives. EXIF metadata demonstrates balanced performance as its structured key-value format simplifies entity extraction.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 3, Juni 2025, Page. 1203-1216
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4593

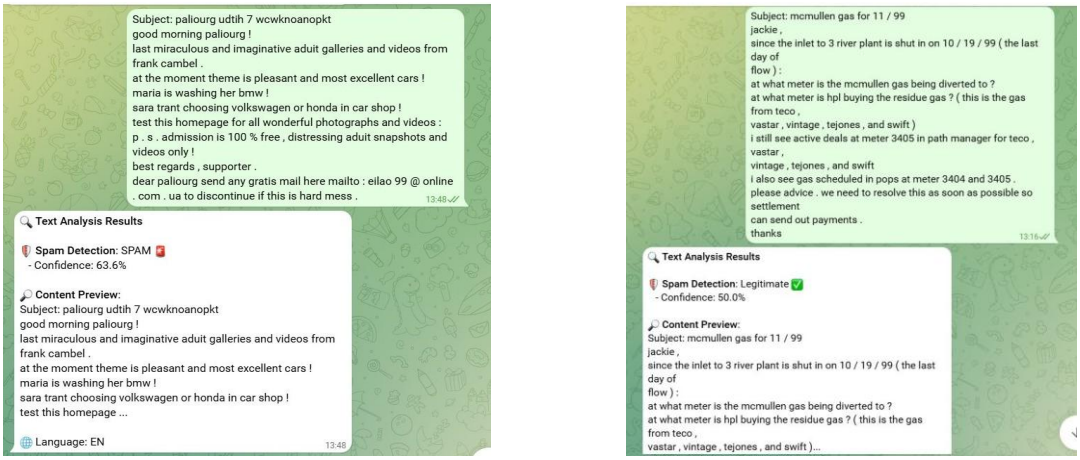

Figure 2. Spam Email Analysis (left) and Ham Email Analysis (right) Results

Figure 2 on the left shows the results of the analysis of an email that is indicated as spam. This email contains irrelevant sentences and promotion of adult content, as well as the use of unnatural language. The chatbot automatically detected the email as spam with a confidence level of 63.6%. In addition, the system also displays a content preview of the email content to strengthen the proof of classification. This feature shows that the chatbot is able to identify common characteristics of spam, such as the use of promotional keywords, unstructured sentences, and suspicious hyperlinks.

Meanwhile, the right side shows the results of the analysis of a legitimate email. This email contains discussions related to gas distribution management and payments, which are legitimate business message content. The chatbot classifies this email as legitimate with a confidence score of 50.0%, which although not very high, still shows the system's ability to distinguish genuine messages from spam. The preview provided also shows that this email has a clear sentence structure and a valid communication context.

These two results reflect how the digital forensics chatbot was able to perform an initial classification of the type of email being analyzed using a confidence score-based approach as well as understanding the context of the message content. This capability is critical in the digital investigation process, particularly in sifting through relevant electronic evidence automatically and efficiently.

3.1.2. Digital System Log Auto-Analysis Testing

This test aims to evaluate the chatbot's ability to automatically analyze digital system log files. The analysis process involved identifying important entities such as the total number of log lines, error messages, warnings, as well as relevant information such as the IP address and system module associated with a particular event. A total of 100 log files were tested in this scenario, with the results shown in table 2.

Table 2. Summary statistics of log analysis results

No.	Sample Quantity	Log Row Average	Average Error	Average Warning	Average Response Time (seconds)
1	100	169.66	3.47	0.27	0.0538

The average number of errors detected per log file (3.47) indicates that the system can filter out critical anomalies efficiently. Meanwhile, the low number of warnings (0.27) indicates that the system focuses on digging out high-severity issues, which is useful in prioritizing investigation tasks. The fast response time (0.0538 seconds per file) reflects the system's efficiency in real-time parsing. The system performs well with the ability to accurately detect and classify critical entities. All analysis results are

Vol. 6, No. 3, Juni 2025, Page. 1203-1216
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4593

displayed in a single chatbot response, including a summary of analyzed log lines, the number and examples of error and warning messages, and the analysis time. This capability not only speeds up the digital incident investigation process but also provides an efficient and informative early view of potential system disruptions. The test results for system log analysis can be seen in figure 3.

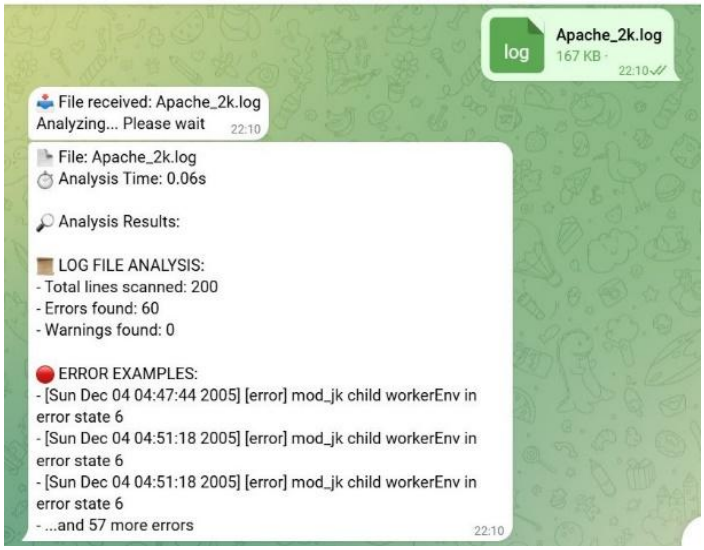

Figure 3. System Log Analysis Results

3.2. Implementation and Testing of Chatbot on Image Analysis

3.2.1. EXIF Metadata Identification Testing

Image metadata analysis is one of the important stages in the digital forensic investigation process that aims to reveal the hidden information behind an image file. Metadata, especially EXIF (Exchangeable Image File Format) metadata, stores important information related to digital images, such as the date and time the image was taken, the GPS location (if the location feature was enabled during capture), the type of camera or device used, and other technical parameters such as ISO, shutter speed, and aperture.

In this research, the chatbot developed is equipped with the ability to automatically extract and interpret EXIF metadata. This process begins when the user uploads an image as input. After the image is received by the system, the initial stage is to read the image file structure and extract the EXIF metadata embedded in it. At this stage, the chatbot uses image processing libraries that are capable of accessing low-level information from digital files, without requiring visual manipulation of the image itself.

Once the raw data from the EXIF has been retrieved, the next process is parsing, which is sorting and categorizing the information that is deemed relevant to the digital forensic investigation. The information is then converted into an easy-to-understand text format, including important elements such as GPS coordinates (if available), the precise time the image was taken (hour, date, even time zone), and the type of hardware and software used. The end result of this process is then displayed by the chatbot to the user in the form of a structured reply. This information is very useful in investigations as it can help uncover the context or whereabouts of the perpetrator when the event occurred.

From testing 100 image samples, the chatbot successfully extracted metadata with an accuracy rate of 94%. This achievement shows that the system is able to perform metadata analysis well, even under conditions of device variation and diverse image formats. The analysis process is visualized in figure 4, which illustrates the flow from image input, metadata extraction, information parsing, to the presentation of results by the chatbot.

Vol. 6, No. 3, Juni 2025, Page. 1203-1216
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4593

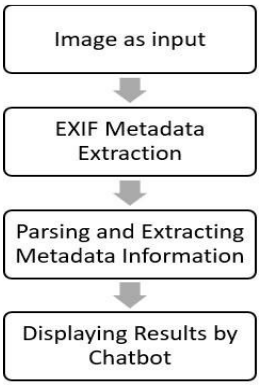

Figure 4. Flowchart of metadata analysis process by chatbot

Steganogra Table 3 presents the evaluation results of the chatbot's ability to identify image metadata based on EXIF (Exchangeable Image File Format) information. The evaluation was conducted on four main metadata parameters common to digital images, namely the date and time of capture, exposure time, aperture, and ISO.

Table 3. EXIF Metadata Testing Results

No.	Metadata Parameters	Extraction Accuracy (%)
1	Date & Time	77%
2	Exposure Time	99%
3	Aperture	100%
3	ISO	100%
	Average	94%

From the test results, it can be seen that the chatbot successfully extracted the aperture and ISO information with a perfect accuracy rate of 100%. The exposure time parameter also showed a very high accuracy of 99%. However, for the date and time of capture parameter, the accuracy rate was relatively lower than the other parameters, reaching only 77%. This difference in accuracy may be due to variations in the metadata storage format on various camera devices or the condition of image files that have undergone compression or modification.

Overall, the chatbot was able to identify image metadata with an average accuracy rate of 94%, which indicates that the system is reliable enough in extracting important information from image files to support the digital forensic investigation process. The results of the image metadata analysis can be seen in figure 5.

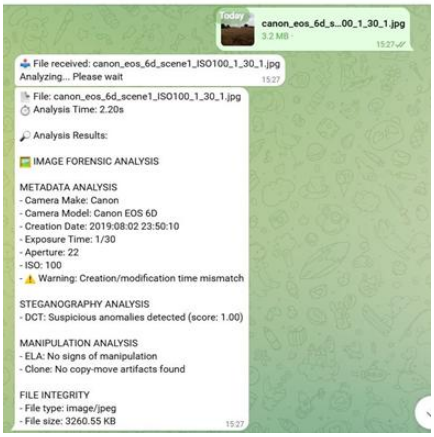

Figure 5. EXIF Metadata Analysis Result

Vol. 6, No. 3, Juni 2025, Page. 1203-1216
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4593

3.2.2. Steganography Detection and Image Manipulation Testing

This test aims to evaluate the chatbot's ability to detect images that have undergone digital manipulation or contain hidden messages (steganography). Two main approaches were used in this process, namely:

- Error Level Analysis (ELA): This method is used to detect potential manipulation in digital images by analyzing the difference in compression levels. Areas that have been edited will show different error levels compared to other areas that have not been manipulated.
- Least Significant Bit (LSB) Analysis: This method is used to detect the presence of a hidden message in an image by examining the least significant bit of the pixel data. This technique is commonly used in steganography as it does not cause significant visual changes to the image.

A total of 100 images containing steganographic and/or manipulation elements were tested using both methods. The analysis results show that the model performs well in detecting the presence of anomalies in table 4. The detection accuracy of the hidden messages is high, supported by satisfactory precision, recall, and F1-score values.

Table 4. Steganography Detection Evaluation Results

No.	TP	FP	FN	Precision (%)	Recall (%)	F1-Score (%)
1	70	19	11	78.65%	86.42%	82.35%

The Precision value indicates how precise the model is in performing the detection, while Recall reflects the model's ability to capture all problematic images. The combination of both is represented by F1-Score, which in this test reached 82.35%, indicating that the system is quite reliable in detecting potential hidden visual threats. Figure 6 presents the results of the steganography analysis as well as the image manipulation.

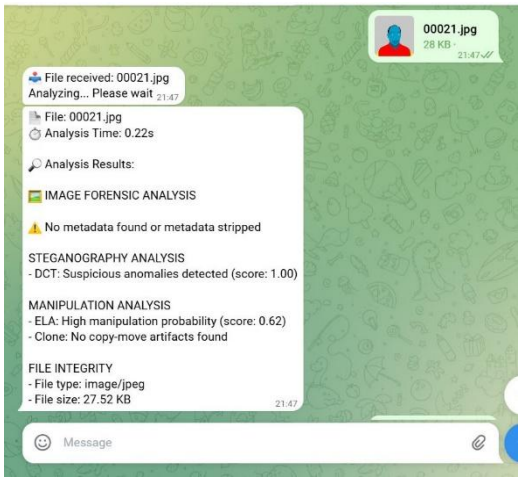

Figure 6. Steganography Detection and Manipulation Analysis Results

3.3. Chatbot Performance Evaluation

To measure the effectiveness of chatbots in digital forensic investigations, tests were conducted on chatbot response time, analysis accuracy, and user satisfaction.

3.3.1. Chatbot Response Time Evaluation

This test evaluates the chatbot's response speed in handling three types of forensic digital investigation commands: text entity identification, image metadata analysis, and steganography detection. The test results in table 5 show that the system responds fastest to text-based commands, with

Vol. 6, No. 3, Juni 2025, Page. 1203-1216
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4593

an average time of only 0.0538 seconds, reflecting high efficiency in text processing. When asked to analyze image metadata, the chatbot took longer, at 4.7224 seconds, along with the need to read and extract information from image files. The slowest response occurred when detecting steganography, at 12.6498 seconds, as this analysis required a more complex process. Overall, the chatbot recorded an average response time of 5.8087 seconds. These results show that although the response varies depending on the complexity of the command, the system is still able to provide analysis results quickly and responsively in the context of real-time digital investigations.

Table 5. Chatbot Response Time Evaluation Results

No.	Command Type	Average Response Time (seconds)
1	Identify entities in the text	0.0538
2	Image metadata analysis	4.7224
3	Steganography detection	12.6498
	Average	5.8087

3.3.2. Accuracy Evaluation with Confusion Matrix Metric

The accuracy of the chatbot model was evaluated using confusion matrix and calculated using Precision (1), Recall (2), and F1-Score (3) metrics.

$$Precision = \frac{TP}{TP+FP} = \frac{375}{375+74} = \frac{375}{449} = 0.8352 \times 100\% = 83.52\% \quad (1)$$

$$Recall = \frac{TP}{TP+FN} = \frac{375}{375+51} = \frac{375}{426} = 0.8803 \times 100\% = 88.03\% \quad (2)$$

$$F1 - Score = \frac{2 \times 0.8352 \times 0.8803}{0.8352+0.8803} = \frac{1.4716}{1.7155} = 0.8581 \times 100\% = 85.81\% \quad (3)$$

Evaluation of the chatbot's performance in detecting digital entities was conducted using a confusion matrix involving three main metrics, namely Precision, Recall, and F1-Score. Based on the results of combining data from all types of digital evidence, 375 data were classified as True Positive (TP), 74 as False Positive (FP), and 51 as False Negative (FN).

The Precision value of 83.52% indicates that most of the entities recognized by the chatbot are indeed the correct entities. Furthermore, the Recall value reaches 88.03%, indicating that the chatbot is able to recognize most of the entities that actually exist in the data. To get a balanced picture of overall performance between Precision and Recall, the F1-Score metric is used. The F1-Score value obtained is 85.81%, which indicates that the system has a fairly stable and reliable detection performance in analyzing various types of digital evidence.

In general, these three metrics show that the chatbot has a good ability to perform digital entity classification with a relatively low error rate. This evaluation provides a strong basis for measuring the effectiveness of chatbots in the context of digital forensic investigations.

Figure 7 displays the overall performance of the Named Entity Recognition (NER) model integrated in the digital forensic chatbot after testing five types of digital evidence, namely system logs, ham emails, spam emails, image metadata (EXIF), and image steganography. The confusion matrix is compiled based on the aggregation of True Positive (TP), False Positive (FP), and False Negative (FN) values from each type of evidence tested, thus providing a comprehensive representation of the model's ability to detect digital entities accurately and consistently.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 3, Juni 2025, Page. 1203-1216
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4593

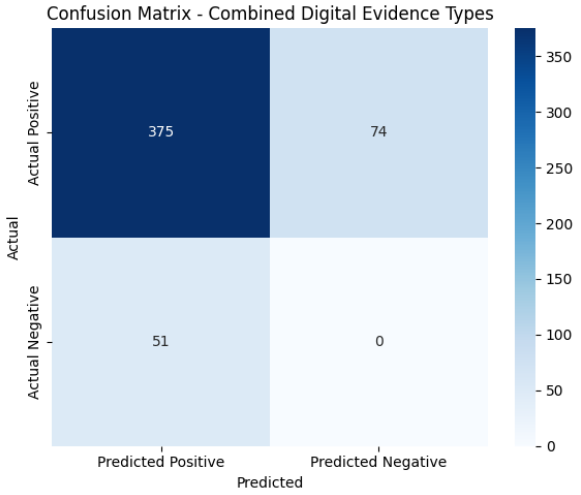

Figure 7. Confusion Matrix - All Digital Evidence Types Combined

True Positive (TP) in this context represents the number of entities or evidences that are correctly identified by the system, according to the predefined labels or ground truth. Conversely, False Positive (FP) indicates the number of entities that should not have been detected but were mistakenly recognized by the model as important entities. The False Negative (FN) indicates the number of important entities that the model failed to recognize despite actually existing in the data. These three parameters become the main components in forming evaluation metrics such as Precision, Recall, and F1-Score.

Table 6. Combined Evaluation of All Types of Digital Evidence Based on Confusion Matrix

No.	Command Type	Average
1	True Positive	375
2	False Positive	74
3	False Negative	51
4	Precision (%)	83.52
5	Recall (%)	88.03
6	F1-Score (%)	85.71

Based on the combined results of all tests in table 6, a total of 375 True Positive, 74 False Positive, and 51 False Negative were obtained. From these values, the overall Precision of the model is 83.52%, indicating that most of the entities recognized by the chatbot are indeed true. Recall reached 88.03%, indicating that the model was able to capture most of the entities that were indeed present in the data. The F1-Score value, which represents the harmonization between Precision and Recall, was recorded at 85.71%. These values indicate that the system has a high level of accuracy and is quite balanced between the ability to detect entities and minimize classification errors.

Thus, the confusion matrix visualization not only serves as a tool to understand the distribution of the model's classification results, but also serves as strong evidence that the developed digital forensics chatbot has a performance worthy of being used in real investigation scenarios. The model is able to reliably identify digital evidence from a variety of data types, making it a potential tool in digital forensics processes that require high speed and accuracy.

NER performance varies depending on the structure and consistency of the evidence type. Structured formats like system logs and EXIF metadata allow more accurate extraction, leading to higher F1-scores. Ham emails also perform well due to formal language, while spam emails show lower precision due to obfuscated or irregular text. These differences highlight how evidence format affects entity recognition accuracy in forensic analysis.

Vol. 6, No. 3, Juni 2025, Page. 1203-1216
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4593

4. DISCUSSIONS

This research shows that the developed digital forensics chatbot is able to identify important entities in various types of digital evidence with a fairly high performance, shown through an average Precision of 83.52%, Recall 88.03%, and F1-Score 85.71%. The high Recall value indicates the chatbot's ability to detect most of the relevant entities, which is an important aspect in the context of digital forensic investigations where missing critical information can impact the investigation process.

When compared to the research of K. Y. Sreeram and Kajal Bansal (2024) in the paper "Algorithmic Chat Monitoring for Mitigating Crime in Telegram: A Multi-Pronged Approach to Prevention and Forensics," their approach has similarities in terms of objectives, namely to assist in crime investigation and prevention through automatic digital data analysis. However, their focus is more on monitoring public conversations on Telegram and developing special algorithms to monitor and analyze conversation content in criminal groups such as detecting crime patterns based on historical data from Telegram groups and proposing prevention algorithms. While this study focuses on the development of an interactive chatbot-based system that can analyze and respond to evidence data in real-time, the approach used is broader because it is not limited to the Telegram platform alone, but includes various types of electronic evidence such as system logs, emails, image metadata, and images containing steganography. Thus, the contribution of this study lies in the integration of NLP, Named Entity Recognition (NER), and digital forensic analysis capabilities into a single chatbot interface that is practical for use by investigators or cybersecurity analysts. The use of DeepSeek LLM and Named Entity Recognition (NER) in this work enables a broader application, particularly in assisting investigators during evidence triage and contextual entity extraction.

In practical terms, this chatbot system has the potential to be deployed within law enforcement agencies or cybersecurity units as a decision-support tool. For instance, investigators can upload suspect emails or logs and immediately receive entity highlights, classification suggestions, and metadata summaries, streamlining the early phase of digital investigations.

Nonetheless, several limitations remain. The NER model showed lower precision in spam email analysis, suggesting a higher rate of false positives, likely caused by informal or obfuscated language. Additionally, the model may exhibit biases depending on the diversity and balance of training data, especially for image steganography or multilingual evidence. Current system responses are also limited to predefined entity types, which may not cover novel or evolving cybercrime indicators.

For future development, the chatbot could be enhanced through integration with threat intelligence platforms to enrich its analysis with contextual threat data. Another direction is improving model adaptability through continual learning and fine-tuning on updated forensic datasets to handle more complex and adversarial evidence formats.

5. CONCLUSION

This study successfully developed a digital forensic chatbot based on DeepSeek Large Language Model (LLM) and Named Entity Recognition (NER), capable of analyzing various types of electronic evidence, including system logs, emails, and image-based data such as metadata and steganography. By integrating Named Entity Recognition (NER), log analysis, and image forensics techniques such as Error Level Analysis (ELA) and Least Significant Bit (LSB), the chatbot performs well in detecting important entities and hidden information relevant to cybercrime investigations.

Test results show that the chatbot achieved an overall F1-Score value of 85.71%, with a Precision of 83.52% and a Recall of 88.03%. These values indicate that the system is quite reliable in recognizing valid digital artifacts, while being able to minimize detection errors. The chatbot's ability to operate via the Telegram platform also makes it an easily accessible tool for investigators in real-time situations, thus providing significant support in the early stages of forensic analysis.

Vol. 6, No. 3, Juni 2025, Page. 1203-1216
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4593

These results also confirm the potential of integration between LLM-based natural language processing and multimodal digital evidence analysis as a new approach to investigative process automation. However, limitations of this study include the reliance on domain-specific datasets and the exclusive use of the Telegram interface, which may constrain generalizability and integration with other forensic platforms. Future research may focus on enhancing chatbots with anomaly detection to expand their capabilities to handle real-time conversation monitoring, and integration with threat intelligence platforms to enrich their analysis with contextual threat data.

CONFLICT OF INTEREST

The author declares that there is no conflict of interest between the author and the object of research in the preparation and implementation of this research.

ACKNOWLEDGEMENT

The authors would like to thank Walisongo State Islamic University Semarang for providing support in the development of DeepSeek-based digital forensics chatbot. This research was conducted independently without receiving grants or funding from any party.

REFERENCES

[1] A. Fahrudin, G. Z. Muflih, and T. Informatika, "Analisis Forensik Digital Pada Pesan Whatsapp Yang Terenkripsi Dengan Pretty Good Privacy (PGP) Menggunakan Framework DFRWS," vol. 9, no. 1, pp. 780–787, 2025.

[2] H. A. F. Muhyidin and L. Venica, "Pengembangan Chatbot untuk Meningkatkan Pengetahuan dan Kesadaran Keamanan Siber Menggunakan Long Short-Term Memory," *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 5, no. 2, p. 152, 2023, doi: 10.36499/jinrpl.v5i2.8818.

[3] M. arif Budiman, "Penggunaan Agen Berbasis Intelijen Untuk Menangani Kejahatan Siber," *J. Innov. Res. Knowl.*, vol. 1, no. 8, pp. 455–462, 2022.

[4] A. Wickramasekara, F. Breitinger, and M. Scanlon, "Exploring the Potential of Large Language Models for Improving Digital Forensic Investigation Efficiency," 2024, doi: 10.1016/j.fsidi.2024.301859.

[5] H. C. Aydogan, B. Yıkar, H. Balandız, and S. Özsoy, "Assessing ChatGPT-4's ability to generate forensic reports: a study of artificial intelligence in forensics," *Egypt. J. Forensic Sci.*, vol. 15, no. 1, pp. 1–12, 2025, doi: 10.1186/s41935-025-00445-1.

[6] M. A. A. A. Dimas Rhoyhan Budi Satrio, Umar Mukhtar, "Penerapan Kecerdasan Buatan Dalam E-Commerce : Efisiensi," vol. 9, no. 1, pp. 788–800, 2025.

[7] Andika Isma, R. Rosidah, Sigit Sahalik Rahman, N. Nasrullah, Arif Setiawan Syam, and Novita Sari, "Analisis Penggunaan Chatbot Berbasis AI pada Model Hybrid di Jurusan Teknik Informatika dan Komputer," *J. Vocat. Informatics Comput. Educ.*, vol. 1, no. 2, pp. 79–92, 2023, doi: 10.61220/voice.v1i2.20239.

[8] J. T. Danang Kurniawan, "Penerapan Teknologi Langchain dan LLM pada Sistem Question Answering Berbasis Chatbot Telegram : Literature Review," pp. 95–104, 2025.

[9] K. Y. Sreeram and K. Bansal, "Algorithmic Chat Monitoring for Mitigating Crime in Telegram : A Multi - Pronged Approach to Prevention and Forensics," vol. 11, no. 1, pp. 14–21, 2024.

[10] DeepSeek-AI *et al.*, "DeepSeek LLM: Scaling Open-Source Language Models with Longtermism," no. April, 2024, [Online]. Available: http://arxiv.org/abs/2401.02954

[11] N. Rachmat and D. P. Kesuma, "Implementasi Large Language Models Gemini Pada Pengembangan Aplikasi Chatbot Berbasis Android," *J. Ilmu Komput.*, vol. 4, no. 1, p. 2024, 2024, doi: 10.31314/juik.v4i1.2831.

[12] T. Gao, J. Jin, Z. T. Ke, and G. Moryoussef, "A Comparison of DeepSeek and Other LLMs," no. February, 2025, doi: 10.48550/arXiv.2502.03688.

[13] A. Yudhana, I. Riadi, and R. Y. Prasongko, "Forensik WhatsApp Menggunakan Metode Digital Forensic Research Workshop (DFRWS)," *J. Inform. J. Pengemb. IT*, vol. 7, no. 1, pp. 43–48,

Vol. 6, No. 3, Juni 2025, Page. 1203-1216
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4593

2022, doi: 10.30591/jpit.v7i1.3639.

[14] A. Simorangkir, P. Intani Sihite, C. Lusia Kiareni, R. Priskila, and V. Handrianus Pranatawijaya, "Pemodelan Chatbot Rekomendasi Hotel Dengan Menggunakan Natural Language Processing," *Simtek J. Sist. Inf. dan Tek. Komput.*, vol. 9, no. 1, pp. 46–50, 2024, doi: 10.51876/simtek.v9i1.371.

[15] A. Lidén and K. Nilros, "Perceived benefits and limitations of chatbots in higher education," *Linnaeus Univ.*, 2020.

[16] R. A. Sanjaya and E. Winarno, "Pengembangan Chatbot Informasi Pariwisata di Kabupaten Pati Menggunakan Metode Natural Language Processing Berbasis Dialogflow," *Jutisi J. Ilm. Tek. Inform. dan Sist. Inf.*, vol. 13, no. 1, p. 368, 2024, doi: 10.35889/jutisi.v13i1.1828.

[17] G. P. Nugraha, L. H. Suadaa, N. Wilantika, "Pengembangan Aplikasi Chatbot dengan Large Language Model untuk Text-to-SQL Generation," *in National Seminar on Official Statistics*, 2024, pp. 831-840.

[18] E. K. W. Ibrahim Ahmad Assegaf, Muhammad Taufik Syastra, Rifky Kurniawan, Tiawan, Muhamad Soleh Fajari, Mawarseh, Retno Novarini, Ahmad Karim Harahap, Elfina Maulid, Yulia Irfayanti, Sutrisno, "Kata Kunci : Chatbot, OpenAi , Model GTP-3.5 Turbo, IDS Digital College .," pp. 785–793, 2024.

[19] M. Scanlon, F. Breitinger, C. Hargreaves, J. N. Hilgert, and J. Sheppard, "ChatGPT for digital forensic investigation: The good, the bad, and the unknown," *Forensic Sci. Int. Digit. Investig.*, vol. 46, no. S, p. 301609, 2023, doi: 10.1016/j.fsidi.2023.301609.

[20] F. M. Rahman, R. R. J. Putra, and Y. Wihardi, "Analisis Statistik dan Implementasi Image Masking Berdasarkan Hasil Error Level Analysis pada Gambar Digital," *JATIKOM J. Apl. dan Teor. Ilmu Komput.*, vol. 3, no. 1, pp. 22–28, 2020.

[21] A. Arrizal, "Named Entity Recognition (NER) Pada Teks Berbahasa Indonesia Dengan Fine-Tuning Indobert," Sriwijaya University, 2024. [Online]. Available: https://repository.unsri.ac.id.

[22] A. N. Fajari and A. Baizal, "Chatbot-based Culinary Tourism Recommender System Using Named Entity Recognition," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 7, no. 4, pp. 1131–1138, 2022, doi: 10.29100/jipi.v7i4.3210.

[23] Warto *et al.*, *Systematic Literature Review on Named Entity Recognition: Approach, Method, and Application*, vol. 12, no. 4. 2024. doi: 10.19139/soic-2310-5070-1631.

[24] Y. Heryanto, F. Farahdinna, and S. Wijanarko, "Evaluasi Responsivitas dan Akurasi: Perbandingan Kinerja ChatGPT dan Google BARD dalam Menjawab Pertanyaan seputar Python," *J. Ris. Sist. Inf. Dan Tek. Inform. (JURASIK*, vol. 9, no. 1, pp. 248–256, 2024, [Online]. Available: https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

[25] A. T. U. B. Lubis, N. S. Harahap, S. Agustian, M. Irsyad, and I. Afrianty, "Question Answering System pada Chatbot Telegram Menggunakan Large Language Models (LLM) dan Langchain (Studi Kasus UU Kesehatan)," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 3, pp. 955–964, 2024, doi: 10.57152/malcom.v4i3.1378.

[26] K. Eka Purnama, C. Rozikin, and A. Ali Ridha, "Analisis Forensic Citra Digital Menggunakan Teknik Error Level Analysis Dan Metadata Berdasarkan Metode Nist," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 2, pp. 1100–1107, 2023, doi: 10.36040/jati.v7i2.6660.