Equivalent: Jurnal Ilmiah Sosial Teknik
Volume 8, Issue 5, 1964-1977
e_ISSN: 2775-0833
https://jurnalequivalent.id/index.php/jequi/index
DOI: doi.org/10.59261/jequi.v8i5.463

Forensic Analysis of WhatsApp Conversations Using NER-Based Information Extraction Techniques

Febrianto*
Universitas Bina Nusantara, Indonesia

Tuga Mauritsius
Universitas Bina Nusantara, Indonesia

***Corresponding author:**
Febrianto, Universitas Bina Nusantara, Indonesia. ✉febrianto004@binus.ac.id

Article Info:

Article history:
Received: September 09, 2026
Revised: October 14, 2026
Accepted: October 16, 2026
Available Online: October 24, 2026

Keywords:
Digital Forensics;
WhatsApp;
Named Entity Recognition;
IndoBERT;
Narcotics;

This is an open access article under the CC BY-SA license.

Copyright © 2026 by Author. Published by Cerdika International Cyber Polytechnic

Abstract

Background: The widespread use of WhatsApp as a primary communication medium in narcotics-related crimes generates a massive volume of digital evidence, making manual forensic analysis inefficient, time-consuming, and prone to human error.
Objective: This study aims to design, develop, and evaluate a custom named entity recognition (NER) model based on IndoBERT Base P1 to automatically extract forensic information from WhatsApp conversations in narcotics-related cases.
Methods: This study adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, which comprises six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The dataset was derived from forensic extractions of WhatsApp conversations conducted by the Indonesian National Police Forensic Laboratory Center (Puslabfor Bareskrim Polri). It comprises 822,943 message rows from 859 pieces of digital evidence. Following annotation using the BIO (Beginning, Inside, Outside) tagging scheme, the dataset yielded 334,576 annotated instances across 18 custom forensic entity categories.
Results: The evaluation results show that the model achieved a precision of 83.44%, a recall of 95.81%, a weighted F1 score of 89.05%, and a macro F1 score of 89.15%. Training was completed in 5,090.05 seconds over three epochs. All aggregate evaluation metrics exceeded the researcher-defined operational criterion of 80% established for this study.
Conclusion: These findings indicate that the IndoBERT-based custom NER model can transform unstructured conversational data into structured investigative information and has the potential to serve as a digital forensic decision-support tool for narcotics crime investigations.

To cite this article: Febrianto, & Mauritsius, T. (2026). Forensic analysis of WhatsApp conversations using NER-based information extraction techniques. *Equivalent: Jurnal Ilmiah Sosial Teknik, 8*(5), 1964-1977. https://doi.org/10.59261/jequi.v8i5.463

INTRODUCTION

Digital transformation has significantly changed communication patterns among Indonesians, with WhatsApp becoming the country's most dominant instant messaging platform. By 2025, the number of WhatsApp users in Indonesia was projected to reach 112 million, representing 90.8% of the country's total internet user base. The widespread use of this platform makes its digital traces crucial for reconstructing crime chronologies because communication artifacts have significant evidentiary value and can serve as authentic documentation that cannot be unilaterally manipulated (Soni, 2024). The 2025 figure of 112 million users (90.8%) is treated in this study as a projection rather than an observed administrative count; accordingly, it is used only to indicate the scale of platform adoption and not as a definitive population statistic.

Narcotics crimes were selected as the focus of this research because they account for a substantial proportion of digital forensic examinations conducted nationally. According to the

recapitulation of forensic examinations conducted by the Forensic Laboratory Center (*Pusat Laboratorium Forensik*, Puslabfor) of the Criminal Investigation Agency (*Badan Reserse Kriminal*, Bareskrim) of the Indonesian National Police (*Kepolisian Negara Republik Indonesia*, Polri) during the 2024–2025 period, 585 of the 1,167 cases handled (50.13%) involved narcotics, with 859 items of evidence consisting of extracted electronic conversation texts that yielded 822,943 lines of communication data. The clandestine and organized nature of narcotics crimes makes participants highly dependent on digital communication to coordinate handover points, financial transactions, and the use of complex pseudonyms. Consequently, this historical data corpus has strategic value as training data for information extraction models. These figures are derived from the official internal recapitulation of digital forensic examinations conducted by Puslabfor Bareskrim Polri during 2024–2025 and are reported here as institutional administrative data rather than as estimates generated by the author.

The urgency of automating forensic extraction is further underscored by the scale of the national investigative workload. Pusiknas Bareskrim Polri reported 414,812 crime reports recorded nationwide from January 1 to December 24, 2025, demonstrating the volume of information that law enforcement institutions must process (Pusiknas Bareskrim Polri, 2025). In digital forensic investigations, manually reviewing tens of thousands of conversational records can therefore be time-consuming and cognitively demanding, increasing the risk that potentially relevant information will be overlooked.

Computational approaches to processing unstructured information in criminal investigations have become increasingly necessary because the volume, velocity, and variety of digital data far exceed investigators' capacity for manual analysis (Skipanes et al., 2025). Natural language processing (NLP), particularly named entity recognition (NER), offers considerable potential for automating the identification of information relevant to investigations in conversational text and has been applied in various WhatsApp forensic frameworks (Anwar & -, 2019; Mubarok Nahdli et al., 2024). Recent research has also integrated entity extraction results into knowledge graphs to reconstruct criminal communication networks on instant messaging platforms (Pozzi et al., 2025). Computational intelligence and machine learning approaches have demonstrated significant capabilities in extracting valuable information and supporting data-driven decision making from large-scale and heterogeneous datasets. Previous studies conducted by Mauritsius and colleagues successfully applied machine learning techniques in social media analytics, fraud detection, financial transaction analysis, user behavior prediction, and infrastructure demand forecasting, demonstrating the capability of data-driven approaches to solve complex analytical problems across various domains, including transportation and smart infrastructure planning (Binsar & Mauritsius, 2020; Mauritsius et al., 2020; Tenggino & Mauritsius, 2022; Mauritsius et al., 2025).

Several previous studies have developed forensic and automation approaches for analyzing WhatsApp data. Fathiyana et al. (2023) and Irfandhia et al. (2022) compared WhatsApp and Telegram forensic artifacts on Android devices, while Sulisdyantoro & Marzuki (2023) identified WhatsApp digital evidence using the APK downgrade method. Gunes Eris & Akbal (2021) forensically analyzed WhatsApp Messenger message databases on Android devices, applied NER to WhatsApp data from Android devices for forensic purposes. Nor et al. (2023) developed an automated approach for correlating user entities using NER, while Pirzada et al. (2024) visualized WhatsApp conversation patterns using network graphs. Qonita et al. (2025) explored a forensic chatbot based on a large language model (LLM) to support digital investigations. However, most of these approaches use generic entity schemes, such as person, location, and date, or focus on a single platform without adapting to the specific vocabulary, pseudonyms, and evidentiary requirements of the Indonesian narcotics domain. Collectively, these studies demonstrate progress in forensic artifact recovery, generic entity extraction, network visualization, and LLM-assisted forensic analysis. Nevertheless, they do not resolve the domain-specific problem addressed in this study: extracting narcotics-related evidentiary entities from informal Indonesian-language WhatsApp conversations. Accordingly, this study shifts the focus from generic forensic recovery to domain-adapted information extraction using a custom entity taxonomy and an Indonesian pretrained language model.

This research gap motivates the development of a custom NER model based on an entity scheme specifically designed to meet the needs of narcotics investigations and trained using authentic data obtained from the National Police Criminal Investigation Center. As its base model, this study uses IndoBERT, a Transformer-based pretrained language model that has demonstrated strong performance in understanding Indonesian-language contexts compared with generic multilingual models (Wilie et al., 2020).

The novelty of this research lies in three aspects. First, it introduces 18 categories in a custom entity taxonomy designed specifically to support the evidentiary requirements of narcotics investigations. These categories encompass narcotics types and weights, packaging camouflage, the roles of actors in distribution networks, and digital assets such as cryptocurrency wallets and electronic money (*e-money*), which are not adequately represented by generic NER categories such as person, organization, and location. Second, the model uses training data derived directly from authentic evidence obtained through forensic acquisition by the National Police Criminal Investigation Center. This enables the model to learn linguistic patterns and pseudonymous terminology that reflect the communication practices of individuals involved in narcotics-related activities, rather than relying on synthetic corpora or simulated data. Third, the annotation process combines rule-based automation with manual validation by police personnel experienced in narcotics investigations and digital forensics, thereby strengthening the alignment of the gold-standard dataset with actual investigative requirements. Based on this rationale, this study aims to design, develop, and evaluate a custom NER model based on fine-tuning IndoBERT Base P1 to automatically extract structured forensic information from WhatsApp conversations associated with narcotics cases. The study follows the CRISP-DM framework and assesses the feasibility of implementing the model as a prototype analysis tool for investigators. Methodologically, the contribution should be understood as a domain-adaptation and empirical feasibility contribution rather than as evidence that the proposed model outperforms all alternative NER architectures. Because the experiment did not include controlled baseline comparisons with generic NER schemes, multilingual models, or ablation variants, no claim of performance superiority is made beyond the reported IndoBERT configuration.

METHOD

This study adopted the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework because it provided a systematic, structured, and iterative approach to developing data-driven analytical models and was relevant to contemporary data science projects involving unstructured data (Shimaoka et al., 2024). This framework consisted of six stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Together, these stages guided the research from identifying investigative needs to developing a prototype analytical tool. The applicability of the CRISP-DM framework has also been demonstrated in various machine learning and data mining studies conducted by Mauritsius and colleagues, including credit classification, banking analytics, and predictive modeling applications. These studies showed that CRISP-DM provides a structured methodology for transforming business problems into data-driven analytical solutions (Darmawan et al., 2020; Mauritsius & Braza, 2019).

Business Understanding

This stage identified the main challenges investigators faced when handling WhatsApp conversation evidence, namely: (1) the large volume of data; (2) unstructured text; (3) the use of informal language, abbreviations, and pseudonyms; (4) reliance on manual analysis, which was prone to human error; and (5) limited human resources relative to the volume of cases. Based on these challenges, an IndoBERT-based custom NER model was identified as a potential solution for generating structured investigative information.

Data Understanding and Entity Schema

The research dataset was derived from WhatsApp digital forensic acquisitions conducted by the National Police Criminal Investigation Center using standard forensic software (Cellebrite) licensed for academic research. Accordingly, this study used only secondary data and did not

independently acquire data from the suspect's device. The integrity and authenticity of digital evidence at the acquisition stage were important considerations because the chain of custody affected the admissibility of electronic evidence in court. The structure of the extracted data reflected the characteristics of message databases on Android and iOS devices commonly described in previous WhatsApp forensic studies (Khaled Zakarneh, 2024; Moreb, 2022; Salah et al., 2025).

The dataset consisted of 859 CSV files containing a total of 822,943 message lines. To maintain consistency in the analysis, each file was reduced to three main attributes: timestamp, sender_phone (anonymized sender number), and message. Initial exploration revealed five distinctive characteristics of the conversations: informal language (e.g., "udh nyampe?"), abbreviations ("otw," "gk"), narcotics-related slang ("goods," "horses," "green"), variations in monetary notation ("500k," "37mt," "gocap"), and Indonesian-English code-switching ("share loc," "smoke transfer"). The 859 CSV files corresponded to 859 items of evidence recorded in the institutional extraction records. However, the available metadata did not establish that each file represented a fully independent device, conversation, or criminal case. This distinction was important because multiple files could have originated from related investigative sources.

The entity schema was developed specifically for this study (i.e., a custom entity schema) through a literature review, case document analysis, observation of communication patterns, and discussions with digital forensic examiners. This approach was adopted because generic entity categories used in conventional NER systems were considered inadequate for domain-specific investigative needs (Yadav & Bethard, 2018). The resulting schema comprised 18 entity categories, as presented in Table 1.

Table 1
Forensic Narcotics Custom Entity Scheme

Code	Entities	Description
PER	Person Name	Individual name/identity
DAT	Date	Date information
TIM	Time	Time information
LOC	Location	Location of the incident/transaction
MON	Money	Transaction amount
PHO	Phone Number	Contact number
BNK	Banks	Bank account name/number
EMN	E-Money	Name/e-money number
PKN	Vehicle Plate	Vehicle police number
KND	Vehicles	Vehicle characteristics
CRY	Cryptocurrency	Crypto wallets
EML	E-mail	Email address
SOC	Social Media	Social media accounts
JNR	Types of Narcotics	Types of narcotics
BNR	Narcotic Weight	Weight/amount of narcotics
BBK	Criminal Evidence	Supporting evidence
KMS	Camouflage	Packaging/camouflage features
ROL	Actor Roles	The role of actors in the network

Data Preparation

The data preparation stage consisted of five steps. First, data selection was performed on the initial 822,943 message lines to remove blank messages, messages containing only symbols or emojis (e.g., "👍"), WhatsApp system messages, and duplicates arising from device synchronization and forensic data export. These messages were generally considered unlikely to contain the target entities or contribute meaningfully to model training. Conversely, messages containing meaningful text and potentially relevant entities were prioritized for retention.

Second, data cleaning involved removing irrelevant characters, such as repetitive sequences of symbols (e.g., "########" and "********"), normalizing excessive spacing resulting from users' typing habits, removing hidden characters (\n, \r, and \t) introduced during forensic

data export, and standardizing the encoding of all files to UTF-8 to ensure compatibility with NLP libraries and IndoBERT tokenizers. Third, anonymization was performed on individuals' full names, phone numbers, bank account numbers, e-money and cryptocurrency account identifiers, street addresses, and email addresses to protect the confidentiality of the parties involved while preserving the linguistic context required for model training. This step was also intended to support compliance with the Personal Data Protection Law and was consistent with the legal and technical principles of electronic evidence protection.

Fourth, entity annotation was conducted in two stages. The first stage involved automatic pre-annotation using regular expressions to accelerate the initial labeling of the large-scale dataset while maintaining label consistency. The second stage involved manual validation and verification by four personnel affiliated with the National Police's Korbinmas Baharkam, the Bali Police Directorate of Narcotics, and the Corruption Eradication Commission. This process was intended to ensure the accuracy, consistency, and contextual appropriateness of each entity label. For example, the sentence "Febri transfer 37 million to BCA tomorrow at 10 o'clock" was annotated by labeling "Febri" as PER, "37 million" as MON, "BCA" as BNK, "tomorrow" as DAT, and "10 o'clock" as TIM. This process generated 334,576 annotated instances across 18 entity categories using the BIO (Beginning, Inside, Outside) labeling scheme. BIO is a common standard in NER research because it represents the boundaries of multi-token entities. For example, the phrase "Jl Sudirman No 125 Jakarta" was annotated as B-LOC, I-LOC, I-LOC, I-LOC, and I-LOC, enabling the model to learn entity boundaries more accurately than flat labeling schemes (Yadav & Bethard, 2018).

Fifth, tokenization was performed using the built-in IndoBERT Base P1 tokenizer, which employed a WordPiece-based subword approach. This approach allowed words to be divided into smaller units (e.g., "per," "##dagang," and "##an"), thereby helping to address out-of-vocabulary problems caused by abbreviations and unfamiliar terms in informal conversations. The dataset was then divided into training (80%; 267,660 instances), validation (10%; 33,458 instances), and test (10%; 33,458 instances) sets to assess the model's ability to generalize to previously unseen data. The reported 80:10:10 split counts referred to annotated instances after tokenization. Because the available source metadata did not document any case-level or evidence-level grouping constraints for the split, complete independence between the training, validation, and test partitions could not be guaranteed. Accordingly, the reported test performance should be interpreted in light of this potential data leakage risk. Future validation should therefore employ group-based partitioning at the evidence, conversation, or case level.

Modeling

This study used the indobenchmark/indobert-base-p1 model as its base model. It was a pretrained language model based on the Transformer architecture that had been trained on a large Indonesian-language corpus and had demonstrated strong performance on various Indonesian natural language understanding tasks compared with generic multilingual models (Wilie et al., 2020). Because IndoBERT had not been specifically adapted to the narcotics domain, the model was fine-tuned using the BIO-annotated dataset for token classification. The model architecture consisted of three components: the input layer, the IndoBERT encoder, and the token classification layer. The contextual embedding of each token produced by the Transformer encoder was passed to the classification layer to determine its entity label. The use of domain-specific model adaptation through fine-tuning is consistent with previous research conducted by Mauritsius and colleagues, which demonstrated that machine learning models achieve higher analytical performance when optimized using datasets and characteristics specific to the target domain (Binsar & Mauritsius, 2020; Tenggino & Mauritsius, 2022).

The hyperparameter configuration used in the experiment is presented in Table 2. A learning rate of 2×10^{-5} was selected because it is commonly used when fine-tuning Transformer models to maintain stable weight updates ($\theta_{t+1} = \theta_t - \alpha \nabla J(\theta_t)$, with θ denotes the model weights and α denotes the learning rate). A batch size of 16 was selected to balance GPU memory requirements and gradient stability, while three training epochs were used to reduce the risk of overfitting on the large-scale dataset. Cross-entropy loss was used as the loss function, and the model weights were optimized using AdamW with a weight decay of 0.01 to help reduce

overfitting through regularization.

Table 2
Fine-tuning hyperparameter configuration

Parameters	Value
Base model	IndoBERT Base P1
Learning rate	2×10^{-5}
Batch size	16
Epoch	3
Maximum sequence length	256 tokens
Weight decay	0.01
Optimizer	AdamW
GPU	NVIDIA Tesla T4 (15 GB)
System RAM	12.7 GB
Programming languages	Python 3

Evaluation

Evaluation was performed by comparing the ground-truth labels with the predicted labels on the test set, which was not used during model training. The evaluation metrics were precision, recall, F1 score, and accuracy, calculated using true positives (TP), false positives (FP), and false negatives (FN):

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1 score = 2 × (Precision × Recall) / (Precision + Recall)

F1 score was used as the primary indicator because it balanced precision and recall. Accuracy was not used as the primary metric because NER datasets are generally imbalanced, with tokens labeled Outside (O) often substantially outnumbering entity tokens. Error analysis was performed on false-positive and false-negative cases to identify the causes of misclassification. These errors were generally grouped into five categories. First, errors arose from nonstandard informal language (e.g., "gk," "udh," and "brp"). Second, errors resulted from narcotics-related slang whose meanings were highly context-dependent (e.g., "goods," "horses," and "patients"). Third, contextual ambiguity occurred when a term such as "goods" in the sentence "goods have arrived" could refer to ordinary objects or narcotics, depending on the conversational context. Fourth, errors occurred in complex, multi-token entities, such as full addresses, where a misclassification of a single token could cause the entire entity to be evaluated as incorrect. Fifth, limited training data for infrequent categories, such as EMN and BNK, made it more difficult for the model to learn consistent patterns for those categories. The researchers established a minimum threshold of 80% for precision, recall, and F1 score as an operational criterion for research success. This threshold was specific to the present study and was not intended to represent a universal benchmark for NER performance.

Deployment

The deployment phase focused on designing a workflow prototype rather than developing a fully operational system. The proposed workflow consisted of five components: input of forensically extracted conversation data, text preprocessing, the IndoBERT-NER model, entity extraction, and preparation of structured investigative reports. The prototype was intended to serve as a decision-support tool that helped investigators retrieve relevant information more efficiently rather than replace human verification entirely.

RESULTS AND DISCUSSION

Results

Dataset Annotation Distribution

The annotation process produced 334,576 entity instances across 18 categories. The distribution of entities is presented in Table 3. LOC (Location) had the largest number of instances (45,853), followed by PER (Person Name) with 37,254 and MON (Money) with 36,302. EMN (E-Money) had the smallest number of instances (4,921).

Table 3

Distribution of Entities on Annotated Datasets

Code	Entities	Quantity
LOC	Location	45.853
PER	Person Name	37.254
MON	Money	36.302
ROL	The Role of the Actor	32.192
BNR	Narcotics Weight	30.116
JNR	Types of Narcotics	19.120
TIM	Time	18.414
CRY	Cryptocurrency	16.438
PKN	Vehicle Plate	15.949
KMS	Narcotics Camouflage	15.824
BBK	Criminal Evidence	12.416
PHO	Phone Number	12.893
EML	E-mail	10.063
SOC	Social Media	9.463
DAT	Date	7.061
KND	Vehicles	5.225
BNK	Bank Account	5.072
EMN	E-Money	4.921
Total		334.576

Training Loss and Validation Loss

The training loss decreased from 0.089610 in epoch 1 to 0.088535 in epoch 2 and 0.085023 in epoch 3. Similarly, the validation loss decreased from 0.097193 to 0.085904 and 0.083966 across the same sequence of epochs. From the first to the third epoch, training loss decreased by 5.12%, while validation loss decreased by 13.61%. Beginning in the second epoch, the validation loss fell below the training loss. Both curves exhibited a consistent downward trend, with no apparent indication of overfitting through the final epoch.

Overall Model Performance

Table 4 presents the precision, recall, F1 score, and accuracy for each epoch, together with the macro-average and weighted-average metrics from the final evaluation. Precision decreased gradually from 87.12% in epoch 1 to 82.87% in epoch 3, whereas recall increased from 90.85% to 95.81% over the same period. The F1 score remained relatively stable, ranging from 88.94% to 88.87%, while accuracy remained above 94% throughout the three epochs.

Table 4

Summary of Model Performance per Epoch and Final Average

Metrics	Epoch 1	Epoch 2	Epoch 3	Macro Avg	Weighted Avg
Precision	87.12%	84.86%	82.87%	83.64%	83.44%
Recall	90.85%	93.49%	95.81%	95.89%	95.81%
F1 Score	88.94%	88.96%	88.87%	89.15%	89.05%
Accuracy	94.16%	94.18%	94.13%	–	–

Computing Efficiency

Fine-tuning IndoBERT Base P1 for three training epochs was completed in 5,090.05 seconds, or approximately 1 hour, 24 minutes, and 50 seconds, using an NVIDIA Tesla T4 GPU with 15 GB of memory. The average processing throughput was 157.76 samples per second and 9.86 training steps per second. The average duration was 28.28 minutes per epoch, and the estimated computational complexity was (9.38 \times 10^{15}) floating-point operations (FLOPs).

Confusion Matrix and Evaluation Per Entity

Analysis of the confusion matrix revealed a high concentration of values along the main diagonal for most categories, indicating relatively high classification performance. The entities with the highest recognition performance were SOC (Social Media), EML (E-Mail Address), CRY (Cryptocurrency), PHO (Phone Number), and DAT (Date), all of which achieved F1 scores above 94%. These results may be attributable to their relatively consistent and structured writing patterns. Classification errors occurred most frequently between BNK (Bank Account) and EMN (E-Money), which are both associated with financial transactions, and between KND (Vehicle) and LOC (Location), particularly when vehicles or other means of transportation were used as location indicators in conversations.

Table 5 summarizes the highest F1 score achieved by each entity category and the corresponding epoch. At their respective best-performing epochs, all 18 categories reportedly exceeded the researcher-defined 80% performance threshold. However, these best-epoch results should not be conflated with final-epoch performance. In the final epoch (epoch 3), which was used for the production model, PKN achieved an F1 score of 71.81%, falling below the 80% threshold.

Table 5

Summary of the Best F1-Score per Entity

Entities	Best F1	Epoch
SOC	99.66%	2
EML	99.23%	1/2
PHO	96.53%	2
CRY	95.56%	2
DAT	94.90%	2
JNR	92.94%	3
LOC	92.92%	1
BNR	92.24%	2
ROL	90.76%	3
TIM	91.14%	2
KND	89.64%	1
MON	86.24%	2
BNK	85.84%	3
PER	83.96%	1
KMS	83.60%	3
EMN	82.90%	3
BBK	84.63%	3
PKN	81.03%	2

To provide a more detailed assessment of model performance at the end of training, Table 6 presents the precision, recall, and F1-score for each entity category in the third epoch, which was selected as the final epoch for the production model. The categories with the lowest precision were PKN (65.79%), EMN (71.43%), and KMS (72.87%). The lowest recall values were observed for PKN (79.05%), PER (90.77%), TIM (90.90%), and KND (90.91%). Due to its comparatively low precision and recall, PKN obtained the lowest F1-score among all entity categories (71.81%). In contrast, most categories achieved recall values above 95%. Notably, BNK and EMN attained high recall values of

96.30% and 98.76%, respectively, despite their relatively low precision, indicating that the model successfully identified most relevant entities while producing a higher number of false positive predictions.

Table 6
Precision, Recall, and F1-Score per Entity in Final Epoch (Epoch 3)

Entities	Precision	Recall	F1 Score
SOC	99.11%	99.79%	99.45%
EML	97.38%	98.65%	98.01%
PHO	94.35%	98.31%	96.29%
CRY	91.36%	99.87%	95.42%
DAT	90.52%	99.30%	94.71%
JNR	88.58%	97.75%	92.94%
LOC	89.42%	96.58%	92.86%
BNR	85.94%	98.95%	91.99%
ROL	85.23%	97.06%	90.76%
TIM	85.75%	90.90%	88.25%
KND	81.95%	90.91%	86.20%
MON	75.67%	97.61%	85.25%
BNK	77.43%	96.30%	85.84%
PER	77.91%	90.77%	83.85%
KMS	72.87%	98.03%	83.60%
EMN	71.43%	98.76%	82.90%
BBK	74.81%	97.41%	84.63%
PKN	65.79%	79.05%	71.81%

Implementation of Entity Extraction and Confidence Score

The implementation results showed that each token or phrase identified as an entity was assigned a category label and a confidence score. The categories with the highest average confidence scores were BNR (1.0000), LOC (0.9185), MON (0.9137), and ROL (0.9132), whereas BNK (0.3982) and BBK (0.5932) had the lowest average confidence scores. Based on the distribution of predicted confidence levels, 134 entity predictions (60.91%) were classified as high-confidence predictions (≥90%), 20 (9.09%) as moderate-confidence predictions (70%–89%), and 66 (30.00%) as low-confidence predictions (<70%).

Discussion

The results indicate that fine-tuning IndoBERT Base P1 is an effective approach for developing a custom named entity recognition (NER) system capable of extracting forensic information from unstructured, informal, and deliberately obfuscated WhatsApp conversations in Indonesian. The weighted F1 score of 89.05% and macro F1 score of 89.15% support the potential of Indonesian-specific pretrained language models to capture local linguistic context, consistent with previous research on Indonesian language modeling (Wilie et al., 2020). These findings are also consistent with developments in NER research emphasizing the importance of Transformer-based contextual representations for addressing morphological and lexical variation in natural language (Yadav & Bethard, 2018). The findings of this study are in line with previous studies by Mauritsius and colleagues, which demonstrated that machine learning approaches are effective in transforming large and complex datasets into meaningful information that supports analysis and decision-making processes. The consistently strong performance achieved across various application domains indicates the robustness of data-driven analytical approaches in addressing real-world challenges (Binsar & Mauritsius, 2020; Tenggino & Mauritsius, 2022).

The trade-off between declining precision (87.12% to 82.87%) and increasing recall (90.85% to 95.81%) across epochs suggests that the model became more inclined to identify candidate entities as training progressed. In forensic investigations, this tendency may be advantageous because false negatives (that is, potentially evidentiary information that remains undetected) can be more detrimental to an investigation than false positives, which produce

additional entity candidates that can subsequently be filtered through manual verification. These findings are consistent with the argument that information-processing systems used in criminal investigations should prioritize minimizing missed information rather than simply maximizing accuracy (Skipanes et al., 2025).

The strong performance of the SOC, EML, PHO, CRY, and DAT categories indicates that the IndoBERT-NER model performs particularly well on entities with structured and consistent writing patterns. A similar characteristic has been reported in research on NER-based automated entity correlation, which found that digital identities with standardized formats are easier to extract than free-form textual descriptions (Dauletov et al., 2026; Nor et al., 2023). By contrast, the lower performance observed for PKN, PER, MON, BBK, KMS, and EMN suggests that substantial variation in entity representations, semantic overlap between categories, and dependence on implicit sentence context remain challenging for machine learning-based NER systems (Yadav & Bethard, 2018). Such challenges are particularly relevant to criminal communication, where coded language and pseudonyms are frequently used (Skipanes et al., 2025).

The per-entity performance results for the final epoch (Table 6) reinforce this pattern and reveal several operationally relevant findings. PKN (Vehicle License Plate) was the lowest-performing category in terms of both precision (65.79%) and recall (79.05%), resulting in the lowest F1 score among all categories (71.81%). This performance may be attributable to the considerable variation in Indonesian vehicle license plate formats and their resemblance to numerical patterns in other entity categories, such as PHO and MON. Consequently, the model may have difficulty learning consistent entity boundaries from the 15,949 annotated instances reported for this category. In contrast, the financial categories BNK and EMN exhibited high recall values (96.30% and 98.76%, respectively) but comparatively low precision values (77.43% and 71.43%, respectively). MON also exhibited a similar pattern, with recall of 97.61% and precision of 75.67%. These results suggest that the model may overpredict these financially related categories, leading to confusion between semantically adjacent entity classes. This fine-grained classification challenge has also been widely reported in NER research involving overlapping entity taxonomies (Yadav & Bethard, 2018). From an investigative perspective, high recall combined with lower precision may still be beneficial because investigators can review a more comprehensive set of candidate financial transactions, although this approach requires additional manual screening to eliminate false positives.

Compared with previous research on WhatsApp forensics that generally applied generic entity schemes, such as person, location, and date Subbulakshmi et al. (2024), this study contributes a custom taxonomy comprising 18 entity categories tailored to the evidentiary requirements of narcotics investigations. These categories include narcotics types, vehicle license plates, narcotics-related camouflage, and the roles of actors within distribution networks. This approach supports the argument that generic natural language processing (NLP) pipelines may be insufficient for criminal communication analysis without adequate domain adaptation. The resulting structured entity output, particularly the combination of PER, ROL, LOC, and PHO, also provides a foundation for further development of relationship extraction and knowledge graph construction to map suspected perpetrator networks, consistent with research integrating NLP and knowledge graphs for criminal communication analysis on instant messaging platforms.

In terms of computational efficiency, the training time of 5,090.05 seconds on a single NVIDIA Tesla T4 GPU demonstrates that discriminative fine-tuning of IndoBERT can be computationally feasible compared with generative approaches based on large language models (LLMs) that are being explored for similar applications, including LLM-based forensic chatbots (Qonita et al., 2025). These characteristics may facilitate adoption in digital forensic laboratories with limited computing resources while maintaining strong entity extraction performance. This result further supports previous findings reported by Mauritsius and colleagues, where machine learning models could be effectively implemented to support practical decision-making and analytical processes while maintaining computational efficiency and scalability in real-world environments (Darmawan et al., 2020; Mauritsius et al., 2020).

Nevertheless, the distribution of confidence scores indicates that approximately 30% of entity predictions fall into the low-confidence category (<70%). Therefore, model outputs should

not be treated as conclusive evidence without human verification. This finding reinforces the model's role as a decision-support tool that accelerates the identification and triage of potentially relevant investigative information rather than replacing manual review. This position is consistent with the principle that electronic evidence requires legally and technically accountable verification and documentation (Arshad et al., 2018; Haryati, 2026). The traceability provided by documented confidence scores also supports the principle of chain of custody in digital evidence handling, which requires each stage of evidence processing to be auditable and accountable.

This study has several limitations. First, the dataset covers only one platform (WhatsApp) and one crime type (narcotics), so generalization to other instant messaging platforms, such as Telegram and Signal, or other communication platforms, such as Instagram, as well as to crimes other than narcotics offenses, requires further evaluation. Second, the limited number of training epochs (three) may have affected the performance of categories with relatively few training instances, such as EMN and BNK. The observed fluctuations in precision and recall across epochs suggest that performance might be improved through additional training, hyperparameter optimization, or class-balancing techniques, although these possibilities require empirical testing. Third, the study does not include entity linking, coreference resolution, or relationship extraction. Consequently, the model outputs independent entity predictions without explicitly representing relationships between entities, such as the relationship between a particular PER entity and a ROL, LOC, or MON entity within the same transaction event. Fourth, although four police personnel with relevant investigative backgrounds validated the annotations, differences in interpretation among annotators may remain. Because inter-annotator agreement was not quantitatively measured, label consistency in ambiguous cases cannot be established independently of the validators' judgments. Fifth, the evaluation does not include controlled baseline comparisons or ablation studies; therefore, the contributions of the custom entity taxonomy, the IndoBERT backbone, and other modeling choices cannot yet be isolated. Sixth, because the partitioning metadata do not establish source-level independence across evidence items, conversations, devices, or cases, the possibility of information leakage between dataset splits cannot be completely excluded.

CONCLUSION

This study designed, developed, and evaluated a custom named entity recognition (NER) model based on fine-tuning IndoBERT Base P1 to extract forensic information from WhatsApp conversations associated with narcotics cases, following the CRISP-DM framework. The pipeline transformed 822,943 raw message rows into 334,576 annotated entity instances across 18 custom forensic entity categories. At the aggregate level, the model achieved a precision of 83.44%, a recall of 95.81%, a weighted F1 score of 89.05%, and a macro F1 score of 89.15%. These aggregate results exceeded the researcher-defined operational threshold of 80%; however, performance varied across categories. Notably, PKN achieved a final-epoch F1 score of 71.81%. The findings therefore support the feasibility of the proposed IndoBERT-based custom NER pipeline while highlighting the need for category-specific improvements and stricter source-independent validation.

These findings indicate that an IndoBERT-based custom NER approach can transform unstructured conversational text into structured investigative information with strong aggregate performance. Nevertheless, its outputs should not be treated as conclusive evidence without human verification. Because approximately 30% of predictions fall into the low-confidence category and some entity classes perform substantially below the aggregate level, the model is best positioned as a digital forensic decision-support tool that accelerates information triage, supports chronology reconstruction, and assists investigators while preserving expert review and evidentiary accountability.

Future research should expand the diversity and volume of the dataset, develop relationship extraction capabilities, and integrate knowledge graphs to map networks of actors more comprehensively. It should also extend the analysis to digital communication platforms beyond WhatsApp and compare IndoBERT with other language models, such as multilingual BERT (mBERT), XLM-RoBERTa, and large language models, to identify the most effective approach for extracting forensic information from Indonesian-language communications. Future

experiments should additionally use group-based training, validation, and test partitions at the evidence or case level and include controlled baselines and ablation studies to quantify the contributions of the custom entity taxonomy, the IndoBERT backbone, the preprocessing strategy, and class-balancing procedures.

ACKNOWLEDGEMENT

The author expresses his deepest appreciation and gratitude to the Chief of the National Police of the Republic of Indonesia, the Assistant Chief of Police for Human Resources (AS HR Kapolri), and the Human Resources Staff of the National Police of the Republic of Indonesia (SSDM Polri) for their trust, opportunity, and support through the educational scholarship program that has allowed the author to continue and complete his studies in the Master Program in Information Systems Management at Universitas Bina Nusantara.

The author also expresses his sincere respect and gratitude to Prof. Drs. Tuga Mauritsius, M.Si., Ph.D. as a supervisor who has provided direction, guidance, constructive input, and academic support during the research process and preparation of this thesis.

The author also expressed his gratitude to the fellow annotation validators for the contribution, time, and support provided in the process of verifying and validating the dataset used in this study.

Finally, the author expresses his sincere gratitude and appreciation to his parents, wife, family, and all parties who have given prayers, support, motivation, and enthusiasm during their education and completed this research. Without their support and sacrifice, the completion of this thesis would not have been able to be realized properly.

AUTHOR CONTRIBUTION STATEMENT

Febrianto contributes to all stages of research, including conceptualization, methodology design, data collection and preparation, model development, result evaluation, analysis, visualization, as well as thesis writing and editing. The validation process of dataset annotations is carried out with validators from Korbinmas Baharkam Polri personnel, the Bali Police Directorate of Narcotics, and the Corruption Eradication Commission. All scientific decisions and interpretation of research results are the responsibility of the author.

REFERENCES

Tenggino, D., & Mauritsius, T. (2022). Evaluation of Factors Affecting Intention to Use QRIS Payment Transaction. ICIC Express Letters, 16(4), 343-349

Binsar, F., & Mauritsius, T. (2020). Mining of Social Media on Covid-19 Big Data Infodemic in Indonesia. Journal of Computer Science, 16(11), 1598-1609. https://doi.org/10.3844/jcssp.2020.1598.1609

Tuga, Mauritsius & Alatas, Sofia & Binsar, Faisal & Jayadi, Riyanto & Legowo, Nilo. (2020). Promo Abuse Modeling in E-Commerce Using Machine Learning Approach. 1-6. 10.1109/ICOT51877.2020.9468744

Darmawan, T., Birawa, A. S., Eryanto, E., & Mauritsius, T. (2020). Credit Classification Using CRISP-DM Method On Bank ABC Customers. International Journal of Emerging Trends in Engineering Research. 8. 2375-2380. 10.30534/ijeter/2020/28862020

Mauritsius, T., et al. (2025). An analytics framework for predicting electric vehicle charging station demand using spatial and interpretive methods. Decision Analytics Journal, 17, 100648. https://doi.org/10.1016/j.dajour.2025.100648

Anwar, N., & -, S. (2019). Forensic Authentication of WhatsApp Messenger Using the Information Retrieval Approach. *International Journal of Cyber-Security and Digital Forensics*, *8*(3). https://doi.org/10.17781/p002609

Arshad, H., Jantan, A. Bin, & Abiodun, O. I. (2018). Digital forensics: Review of issues in scientific validation of digital evidence. *Journal of Information Processing Systems*, *14*(2). https://doi.org/10.3745/JIPS.03.0095

Dauletov, A., Muminov, B., Matyakubova, N., Matyakubova, T., Akhmedova, K., Kholmatova, Z., & Buriev, B. (2026). A Hybrid Model for Standardized, Flexible, and Intelligent Metadata-Based Description of Electronic Documents in Digital Library and Archival Information

Systems. *Information (Switzerland)*, *17*(6). https://doi.org/10.3390/info17060590

Fathiyana, R. Z., -, Y., Cahyadi, N., & Hidayat, D. J. (2023). Comparative Study and Analysis of Forensic Artifacts of WhatsApp and Telegram on Android Devices. *Journal of Informatics and Communication Technology (JICT)*, *4*(2). https://doi.org/10.52661/j_ict.v4i2.143

Gunes Eris, F., & Akbal, E. (2021). Forensic Analysis of Popular Social Media Applications on Android Smartphones. *Balkan Journal of Electrical and Computer Engineering*, *9*(4). https://doi.org/10.17694/bajece.761271

Haryati, E. D. W. and E. S. (2026). The Legal Position of Electronic Systems in Digital Forensic Practices under Law Number 1 of 2023 concerning the Criminal Code of Indonesia . *Jurnal Ilmiah Multidisiplin Indonesia (JIM-ID)*, *5*(5), 970–980.

Irfandhia, M. D., Satrya, G. B., & Nuha, H. H. (2022). Forensic Investigation Analysis of WhatsApp Messenger and Telegram Messenger on Android Based Device. *2022 1st International Conference on Software Engineering and Information Technology, ICoSEIT 2022*. https://doi.org/10.1109/ICoSEIT55604.2022.10030029

Khaled Zakarneh, S. (2024). Mobile Forensic Investigation on iOS & Android Smartphones: Case Study Investigation on WhatsApp. *International Journal of Applied Sciences and Smart Technologies*, *6*(1). https://doi.org/10.24071/ijasst.v6i1.6770

Moreb, M. (2022). Mobile Forensic Investigation for WhatsApp. In *Practical Forensic Analysis of Artifacts on iOS and Android Devices*. https://doi.org/10.1007/978-1-4842-8026-3_9

Mubarok Nahdli, M. F., Biddinika, M. K., & Riadi, I. (2024). Forensic Analysis of the WhatsApp Application Using the National Institute of Justice Framework. *International Journal of Advances in Data and Information Systems*, *5*(2). https://doi.org/10.59395/ijadis.v5i2.1328

Nor, G., Abuhmida, Dr. M., & Llewellyn, Dr. E. (2023). Improving Digital Forensic Investigations through Automated User Entity Correlation. *International Journal for Research in Applied Science and Engineering Technology*, *11*(5). https://doi.org/10.22214/ijraset.2023.51446

Pirzada, S., Ab Rahman, N. H., Cahyani, N. D. W., & Othman, M. F. (2024). A Framework of Forensic Analysis and Visualization: Using WhatsApp Chat Data as a Case Study. *International Journal on Informatics Visualization*, *8*(3–2). https://doi.org/10.62527/joiv.8.3-2.2868

Pozzi, R., Barbera, V., Alva Principe, R., Giardini, D., Rubini, R., & Palmonari, M. (2025). Combining Knowledge Graphs and NLP to Analyze Instant Messaging Data in Criminal Investigations. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *15437 LNCS*. https://doi.org/10.1007/978-981-96-0567-5_30

Pusiknas Bareskrim Polri. (2025, December 30). Refleksi akhir tahun: Data kepolisian dan tantangan penegakan hukum. https://pusiknas.polri.go.id/detail_artikel/refleksi_akhir_tahun%3A_data_kepolisian_dan_tantangan_penegakan_hukum

Qonita, N. N., Handayani, M. R., & Umam, K. (2025). Digital Forensic Chatbot Using DeepSeek LLM and NER for Automated Electronic Evidence Investigation. *Jurnal Teknik Informatika (Jutif)*, *6*(3). https://doi.org/10.52436/1.jutif.2025.6.3.4593

Salah, S., Amro, B., & Moreb, M. (2025). WhatsApp Messenger Forensics: Extracting Insights for Enhanced Learning Analytics in Smart Education. *2025 International Conference on Smart Learning Courses, SCME 2025*. https://doi.org/10.1109/SCME62582.2025.11104884

Shimaoka, A. M., Ferreira, R. C., & Goldman, A. (2024). The evolution of CRISP-DM for data science: Methods, processes and frameworks. *SBC Computing Reviews*, *4*(1), 28–43.

Skipanes, M., Demartini, G., Franke, K., & Nissen, A. B. (2025). Information analysis in criminal investigations: methods, challenges, and computational opportunities processing unstructured text. *Policing (Oxford)*, *19*. https://doi.org/10.1093/police/paaf005

Soni, N. (2024). Forensic Analysis of WhatsApp: A Review of Techniques, Challenges, and Future Directions. *Journal of Forensic Science and Research*, *8*(1). https://doi.org/10.29328/journal.jfsr.1001059

Subbulakshmi, N., Ariffa Begum, S., Kumar, A., Kumar, S., Satyarthi, S., & Singh, A. (2024). Forensic Investigation and WhatsApp Chat Analysis Using Named Entity Recognition with Web Based Visualization. *5th International Conference on Sustainable Communication Networks*

and Application, ICSCNA 2024 - Proceedings. https://doi.org/10.1109/ICSCNA63714.2024.10864198

Sulisdyantoro, D., & Marzuki, M. I. (2023). Identification of Whatsapp Digital Evidence on Android Smartphones using The Android Backup APK (Application Package Kit) Downgrade Method. *Journal of Integrated and Advanced Engineering (JIAE)*, *3*(1). https://doi.org/10.51662/jiae.v3i1.70

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL-IJCNLP 2020*. https://doi.org/10.18653/v1/2020.aacl-main.85

Yadav, V., & Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. *COLING 2018 - 27th International Conference on Computational Linguistics, Proceedings*.