Technema: Journal of Intelligent Engineering and Computing

Vol 1 No 3 September 2026, Hal 44-59
ISSN: 3163-6020 (Print) ISSN: 3163-5997 (Electronic)
Open Access: https://sovereignresearch.org/technema

Implementation of the Indobert Model for Named Entity Recognition in Public Complaint Texts from the City of Semarang

Yossi Rifky Shabara[1*], Mustafa[2]
[1-2] Universitas Islam Sultan Agung, Semarang, Indonesia
email: yossishabara@gmail.com[1]

Article Info :
Received:
01-08-2026
Revised:
17-08-2026
Accepted:
24-08-2026

Abstract

Public complaint management in public services continues to face challenges because complaint data are generally presented as unstructured text containing informal language, abbreviations, and Indonesian-Javanese code-mixing. These characteristics complicate the identification of essential information and the determination of appropriate destination agencies when complaints are processed manually. This study aims to analyze the capability of the IndoBERT model for Named Entity Recognition (NER) in extracting relevant entities from public complaint texts in Semarang City to support complaint routing. A quantitative experimental approach was employed through complaint data collection, BIO-scheme annotation, text preprocessing, IndoBERT fine-tuning, and model evaluation using Precision, Recall, and F1-Score. The dataset consisted of 1,066 complaint sentences containing four target entities: service, location, problem category, and time. Strict-chunk evaluation using seqeval showed that the best IndoBERT checkpoint, obtained at Epoch 3, achieved a Macro F1-Score of 0.62, with the highest performance for LAYANAN (F1=0.68) and the lowest for LOKASI (F1=0.56). Error analysis further indicated that entity-boundary shifts, category ambiguity, and lexical variation in local and informal expressions remained important factors affecting recognition performance. The findings demonstrate that IndoBERT can transform heterogeneous complaint narratives into structured entity information and provide a foundation for improving the efficiency of public complaint management and routing within smart governance systems.

Keywords: *IndoBERT, Named Entity Recognition, Public Complaints, Semarang City, Smart Governance.*

©2026 Authors.. This work is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License.
(https://creativecommons.org/licenses/by-nc/4.0/)

INTRODUCTION

The rapid expansion of digital public services has transformed citizen-government interaction by shifting complaint submission from predominantly administrative procedures toward data-intensive, platform-mediated processes that continuously generate large volumes of textual feedback. Public complaint platforms therefore represent an increasingly important source of operational intelligence because the information submitted by citizens can reveal service failures, geographic patterns, administrative problems, and temporal characteristics that are difficult to capture through conventional structured databases. The analytical value of these data, however, is constrained by the fact that a substantial proportion of complaints remains in unstructured textual form, with variations in spelling, abbreviations, informal expressions, and heterogeneous linguistic practices that complicate automated information extraction. Indonesian digital communication presents an even more complex setting because users may combine Indonesian with regional languages and English within the same utterance, producing code-mixed expressions that introduce additional ambiguity into language identification and downstream NLP processing (Hidayatullah et al., 2023). Research on Indonesian public-sector text has increasingly demonstrated that transformer-based language models can provide stronger contextual representations than conventional approaches for understanding citizen-generated content, including complaint-related classification tasks (Agustina et al., 2024). The increasing availability of Indonesian NLP resources and benchmarks has also strengthened the foundation for developing context-sensitive computational models capable of processing linguistic phenomena specific to Indonesian rather than relying exclusively on models developed for high-resource languages (Cahyawijaya et al., 2021). Within this landscape, the challenge is no longer limited to determining whether a complaint belongs to a particular category, but increasingly concerns how specific pieces of information embedded within

each complaint can be identified, structured, and subsequently connected to public-service decision processes.

In public complaint management, this distinction is particularly consequential because a single complaint may simultaneously contain information concerning the type of service, location of an incident, problem category, responsible institution, and time of occurrence, making entity-level extraction a more informative analytical task than document-level classification alone. The operational importance of these elements is evident in the management of public complaints at the Pusat Pengelolaan Pengaduan Masyarakat (P3M) Kota Semarang, where complaint processing requires information to be interpreted and directed toward the appropriate administrative unit (Pitaloka Subagyo et al., 2023). Named Entity Recognition (NER) provides a suitable methodological framework because it transforms unstructured language into semantically labelled units that can subsequently support information retrieval, categorization, routing, and decision support. A systematic review of Indonesian NER research indicates that the field has developed across diverse datasets and entity definitions, yet considerable variation remains in annotation practices, datasets, model selection, and evaluation settings (Budi & Suryono, 2023). Such variation is not merely technical because differences in annotation schemes can materially influence how NER systems learn entity boundaries and labels, particularly when linguistic contexts are ambiguous or entity categories overlap (Alshammari & Alanazi, 2021). The growing application of transformer-based models has created an opportunity to address these challenges through contextual representations that capture relationships between words rather than treating tokens as independent units.

Recent studies provide evidence that IndoBERT has become an increasingly relevant architecture for Indonesian-language NLP because its contextual representation can be adapted to multiple downstream tasks involving complex semantic relationships. IndoBERT-based approaches have demonstrated competitive performance in Indonesian sentiment classification and related public-service text analysis, indicating that the model can capture contextual patterns in citizen-generated language (Hidayat & Gata, 2025). Similar evidence has emerged from public sentiment analysis of government-related discourse, where IndoBERT-based models have been applied to noisy user-generated content and shown their relevance for computational analysis of public opinion (Firdaus & Trisnawarman, 2025). IndoBERT has also been incorporated into more complex architectures for textual entailment, demonstrating that its contextual representations can be combined with additional machine-learning features to improve the recognition of semantic relations in Indonesian text (Tandi et al., 2025). These findings collectively suggest that the strength of IndoBERT lies not simply in its transformer architecture, but in its capacity to model contextual dependencies within Indonesian linguistic environments. For NER specifically, transformer-based approaches have achieved promising results across formal Indonesian domains, including legal documents, where entity recognition requires sensitivity to specialized terminology and contextual boundaries (Yulianti et al., 2024). More recent research has also applied IndoBERT to Indonesian fake-news texts by integrating part-of-speech information with BERT-based representations, suggesting that contextual transformer models can remain effective when entity identification is embedded within linguistically heterogeneous information environments (Cahyo et al., 2025). The expanding evidence base establishes a strong methodological rationale for investigating IndoBERT beyond conventional classification and formal-document settings, particularly in domains where the extracted entities have direct implications for information management.

Despite these developments, important empirical and conceptual gaps remain in the application of Indonesian transformer models to public complaint NER. Existing comparisons of IndoBERT and IndoLEM for health-information extraction have primarily relied on online news articles, whose grammatical structures, editorial conventions, and linguistic regularity differ substantially from citizen-generated complaints (Istiqomah & Novika, 2025). A recent study on folklore has demonstrated the implementation of IndoBERT for NER in Indonesian narrative text, but the characteristics of literary narratives cannot adequately represent the fragmented, abbreviated, context-dependent, and institutionally oriented language found in public complaints (Erna et al., 2026). Research addressing e-government complaint data has begun to examine entity extraction through an LDA-assisted NER approach, yet this direction does not resolve the specific question of how a pretrained Indonesian transformer model performs when directly applied to the entity structures embedded in municipal complaint narratives (Umam et al., 2025). Previous studies have also concentrated substantially on

sentiment or document-level classification, including public sentiment toward political and government-related issues, which provides useful evidence regarding text understanding but does not necessarily demonstrate the ability to identify and label the constituent entities required for automated complaint routing (Maharani & Latifa, 2025). The literature therefore presents an important inconsistency: IndoBERT has demonstrated broad effectiveness across Indonesian NLP tasks, while evidence regarding its entity-level performance on noisy, nonformal, geographically specific public complaint data remains comparatively limited. This gap becomes more significant when complaint texts contain code-mixing, inconsistent lexical forms, implicit references, and locally specific expressions that may challenge models trained or evaluated primarily on formal textual resources.

The unresolved problem has both scientific and practical significance because inaccurate entity extraction can propagate errors into subsequent complaint classification, prioritization, routing, and administrative response processes. A model that correctly understands the general semantic topic of a complaint may still fail operationally if it cannot distinguish whether a phrase represents a location, service category, problem type, or temporal expression. Studies of Indonesian NLP have increasingly moved toward domain adaptation and task-specific modeling, including domain-specific sentiment and NER frameworks for user complaints, indicating that generic language models may require closer evaluation within the linguistic and semantic characteristics of particular service domains (Asri et al., 2025). At the same time, research comparing transformer architectures for Indonesian e-government service reviews demonstrates continuing interest in determining how pretrained language models perform in public-service contexts rather than assuming that general benchmark performance automatically transfers to government data (Utomo, 2025). The methodological issue is particularly important because model evaluation must distinguish between overall predictive performance and class-specific behavior, especially when complaint datasets contain uneven entity distributions and potentially ambiguous labels. Appropriate multi-class evaluation metrics are therefore necessary to establish whether a model performs consistently across entity categories rather than achieving apparently strong aggregate performance through dominant classes (Grandini et al., 2020). The practical challenge faced by municipal complaint-management systems is to convert heterogeneous citizen narratives into structured information without eliminating the contextual details that make those narratives actionable. A reliable NER model could bridge this gap by transforming free-form complaints into machine-readable entity structures that provide more precise inputs for downstream public-service workflows.

This study is positioned within that gap by treating public complaints from the City of Semarang as a domain-specific Indonesian NLP problem in which entity recognition is evaluated under the linguistic conditions of real citizen-generated text. Rather than extending the literature through another document-level classification experiment, the research focuses on the extraction of semantically meaningful entities that can support the transformation of complaint narratives into structured information. The study specifically examines the capability of the IndoBERT model to recognize entities representing service categories, locations, problem categories, and temporal information within public complaint texts. This positioning connects transformer-based Indonesian language modeling with the information requirements of municipal complaint management, while maintaining attention to the linguistic irregularities that distinguish public complaints from formal news, legal, or literary corpora. The research also responds to the broader development of Indonesian NER by testing whether a pretrained contextual model can provide useful entity-level representations in a domain characterized by nonstandard language and operationally meaningful local references. Its contribution is therefore not limited to reporting predictive performance, but also concerns the empirical assessment of whether IndoBERT can serve as a methodological bridge between unstructured citizen narratives and structured information required for public-service analytics. The study aims to implement and evaluate IndoBERT for Named Entity Recognition on public complaint texts from the City of Semarang, with particular attention to the identification of service, location, problem, and time entities. The expected contribution is a domain-specific empirical basis for Indonesian NER research and a methodological foundation for developing more structured, scalable, and context-aware public complaint processing systems that can support intelligent public-service management and smart-governance applications.

RESEARCH METHODS

This study employs an empirical experimental design to implement and evaluate the IndoBERT model for Named Entity Recognition (NER) on public complaint texts from the City of Semarang. The system architecture consists of data preparation, text preprocessing, tokenization, IndoBERT fine-tuning, BIO-based entity classification, entity extraction, and downstream complaint routing. The model uses IndoBERT-Base, a Transformer-based pretrained language model developed for Indonesian text, which generates contextual token representations through self-attention before passing them to a token-level classification layer for entity prediction (Cahyawijaya et al., 2021). The complaint data are annotated using the BIO scheme and contain four target entity categories, namely Problem, Location, Time, and Service, while tokens that do not belong to an entity are assigned the O label. Preprocessing is conducted to reduce irrelevant textual variations while preserving information required for entity identification, followed by IndoBERT subword tokenization and alignment between the tokenized sequence and its corresponding BIO labels. The dataset is separated into training, validation, and testing subsets, where the training data are used for model fine-tuning, validation data are used to monitor learning and select the appropriate model configuration, and testing data are retained for final evaluation. The overall processing mechanism begins with raw complaint input, followed by preprocessing and NER prediction, extraction of the four entity categories, mapping of extracted information to agency reference data, and generation of the recommended destination agency, as presented in Figure 1.

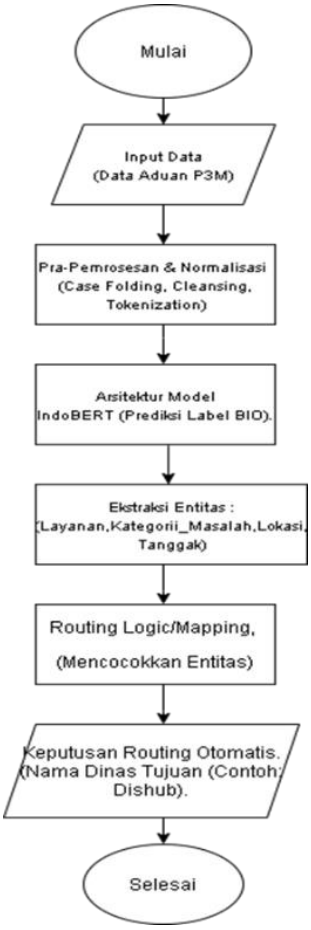

Figure 1. Data Processing and Automatic Complaint Routing Workflow

The performance of the trained IndoBERT model is evaluated using an independent test dataset that is not involved in model parameter optimization or model selection. Model performance is assessed using Precision, Recall, Accuracy, and F1-Score, allowing the evaluation to capture both the correctness and completeness of entity predictions (Grandini et al., 2020). Precision measures the proportion of predicted entities that are actually correct, while Recall measures the proportion of reference entities

that are successfully identified by the model. Accuracy represents the proportion of correctly classified instances across the evaluated predictions, whereas F1-Score represents the harmonic balance between Precision and Recall. The mathematical formulation of the evaluation metrics is presented below, with TP, TN, FP, and FN representing true positive, true negative, false positive, and false negative predictions.

Accuracy

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Precision

$$\frac{TP}{TP + FP}$$

Recall

$$\frac{TP}{TP + FN}$$

F1-Score

$$2\left(\frac{Precision\ x\ Recall}{Precision + Recall}\right)$$

Evaluation additionally applies strict matching, in which an entity prediction is considered correct only when both the predicted entity type and the complete entity boundary correspond exactly to the gold-standard BIO annotation, providing a stricter assessment of NER performance (Alshammari & Alanazi, 2021). A confusion matrix is generated to examine the distribution of correct and incorrect predictions among the Problem, Location, Time, Service, and O labels and to identify potential confusion between entity categories, as presented in Figure 2. The validation subset is used during training to monitor model performance and determine the best model configuration, while the final performance values are obtained exclusively from the independent test subset. The evaluation results are reported for both the overall NER task and individual entity categories to identify differences in recognition capability across Problem, Location, Time, and Service entities. This procedure provides a reproducible basis for determining the capability of IndoBERT to transform unstructured public complaint texts from Semarang into structured entity information that can support subsequent public-service complaint management.

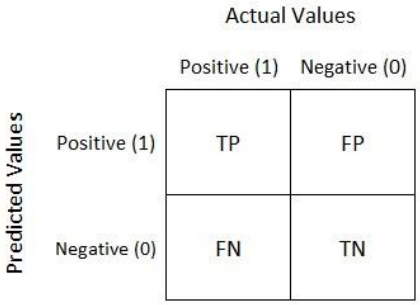

Figure 2. Confusion Matrix of IndoBERT NER Classification

RESULTS AND DISCUSSION

Dataset Collection and Characteristics

The dataset used in this study consists of public complaint texts collected from the Pusat Pengelolaan Pengaduan Masyarakat (P3M) of Semarang City through the LAPOR! Semarang channel, representing citizen-generated narratives submitted within a public-service complaint management environment. An initial corpus of 1,090 complaint sentences was obtained, after which data filtering was conducted to remove sentences that did not contain any target entity, resulting in a final dataset of

1,066 sentences. The use of complaint data from an operational public-service environment provides a more realistic testing context for NER because the textual characteristics differ from those commonly found in formally edited news, legal documents, or other standardized corpora. Previous research on the Semarang City complaint-management system indicates that complaint processing involves the identification and management of information submitted by citizens, making structured extraction from complaint narratives relevant to improving the efficiency of administrative handling (Pitaloka Subagyo et al., 2023). The dataset was also subjected to anonymization to remove sensitive information such as national identification numbers, telephone numbers, and full names before computational processing. This procedure ensures that the experimental corpus retains information relevant to entity recognition while reducing the exposure of personally identifiable information during model development and evaluation.

Table 1. Research Data Specifications

Data Collection Aspect	Description / Specification
Data Source	Pusat Pengelolaan Pengaduan Masyarakat (P3M) Kota Semarang through the LAPOR! Semarang channel
Number of Data	1,090 initial sentences; 1,066 sentences after filtering sentences without entities (pure-O)
Data Security	Anonymization was performed to remove sensitive reporter information (national identification numbers, telephone numbers, and full names) to maintain privacy ethics
Data Characteristics	Unstructured text, nonformal language, containing code-mixing (a mixture of Indonesian and Javanese/Semarang language), as well as many nonstandard abbreviations
Labeling Scheme (Annotation)	Manually performed at the token level using the BIO (Beginning, Inside, Outside) scheme
Target Entities (Classes)	Focused on four main entities: Service (related institution), Location (street/area names), Problem Category (type of complaint), and Date/Time of occurrence

As presented in Table 1, the corpus exhibits characteristics that are substantially different from standardized Indonesian-language datasets because the complaints were produced naturally by citizens rather than generated under controlled linguistic conditions. The presence of informal expressions, nonstandard abbreviations, and mixed Indonesian-Javanese forms introduces lexical and contextual variation that can affect the identification of entity boundaries and entity types. Code-mixed Indonesian-Javanese-English communication has previously been recognized as a distinct computational challenge because language switching may occur within short textual sequences and complicate automatic linguistic interpretation (Hidayatullah et al., 2023). The four entity categories were selected based on their functional relevance to complaint processing, namely Service, Location, Problem Category, and Date/Time, allowing the extracted information to represent both the substance and administrative context of a complaint. Manual token-level annotation using the BIO scheme provides explicit information about entity beginnings, internal tokens, and non-entity tokens, which is essential for training a token-classification model such as IndoBERT. Differences in annotation schemes can influence NER behavior because entity boundaries and label definitions determine the supervision received by the model during fine-tuning (Alshammari & Alanazi, 2021). The resulting annotation structure therefore establishes a direct relationship between the linguistic content of citizen complaints and the information requirements of an automated public-service processing system.

The final dataset of 1,066 sentences was subsequently divided into training, validation, and testing subsets using an 80:10:10 proportion to maintain a clear separation between model development and final performance assessment. The training subset contains 852 sentences and 10,631 tokens, while both the validation and testing subsets contain 107 sentences and 1,329 tokens each. This division provides the model with the largest proportion of available observations for parameter adaptation while reserving independent samples for checkpoint selection and unbiased evaluation. The separation between validation and testing data is particularly important in empirical NER research because

repeated exposure to the test data during model development could inflate the estimated generalization performance. Transformer-based Indonesian NER studies similarly emphasize the importance of evaluating models against datasets that represent the specific linguistic domain being investigated rather than assuming that performance from unrelated corpora will transfer directly (Yulianti et al., 2024). The distribution of the dataset used in the experimental stages is summarized in Table 2, which provides the sentence count, proportion, token count, and function of each subset. The resulting partition creates a reproducible experimental structure in which the final test set remains unavailable to the model during training and validation.

Table 2. Data Split

Subset	Number of Sentences	Percentage	Number of Tokens	Description
Training	852	80%	10,631	Data for model training
Validation	107	10%	1,329	Data for selecting the best checkpoint
Testing	107	10%	1,329	Final test data (never seen by the model)
Total	1,066	100%	13,289	Final dataset after filtering

The distribution in Table 2 indicates that the training subset contains sufficient observations to expose IndoBERT to recurring linguistic patterns while the validation and testing subsets retain identical sentence and token counts, facilitating consistent evaluation conditions. The use of 852 training sentences is particularly relevant because entity recognition requires the model to learn not only lexical associations but also contextual patterns that distinguish entity categories within variable sentence structures. The validation subset functions as an intermediate control during model development, allowing the best-performing checkpoint to be identified without directly incorporating information from the final test set. The independent testing subset then provides an objective basis for measuring whether the learned representations can generalize to previously unseen complaint texts. This experimental arrangement is consistent with the broader development of Indonesian NLP, where pretrained contextual models are adapted to downstream tasks through task-specific fine-tuning rather than being evaluated solely through their pretrained representations (Cahyawijaya et al., 2021). The dataset configuration also supports subsequent analysis of category-specific performance because the test corpus contains examples representing the four target entity classes. The resulting structure is particularly relevant for public complaint NER because the practical objective is not merely to classify sentences correctly, but to identify actionable information embedded within individual complaint narratives.

The composition of the corpus also reflects the domain-specific nature of public complaint data, in which a single sentence may contain multiple pieces of information related to a service, problem, location, and occurrence time. Such characteristics make the dataset suitable for evaluating whether IndoBERT can transform heterogeneous citizen narratives into structured entity representations that can subsequently support complaint routing. Previous research has shown that IndoBERT can effectively support Indonesian public-service text analysis, including complaint classification and sentiment-related applications, yet these tasks operate primarily at the document or sentence level rather than requiring precise entity boundary identification (Agustina et al., 2024). The present dataset extends this setting by requiring the model to distinguish multiple entity types within relatively short and linguistically irregular complaint sentences. The combination of manual BIO annotation, domain-specific entity definitions, and separated experimental subsets creates an empirical foundation for assessing entity-level performance under realistic public-service conditions. The dataset is also sufficiently structured to permit subsequent comparison of precision, recall, and F1-score across Service, Location, Problem Category, and Date/Time entities. This configuration provides the basis for examining whether differences in entity performance arise from the model itself or from the linguistic characteristics and distribution of the complaint corpus.

Dataset Exploration and Label Distribution

The exploratory analysis was conducted to characterize the statistical and linguistic structure of the complaint corpus before IndoBERT training and to identify factors that could influence token-

level entity recognition. The initial dataset contained 1,090 complaint sentences, of which 1,061 sentences, equivalent to 97.3%, contained at least one annotated entity, while 24 pure-O sentences were removed through sentence-level filtering, producing a final corpus of 1,066 sentences. The relatively high proportion of sentences containing entities indicates that the collected complaints generally carry information relevant to the defined NER task rather than consisting predominantly of statements without identifiable service-related information. The corpus contained 13,740 tokens in the initial dataset, with an average sentence length of 12.6 tokens and a vocabulary size of 2,847 unique words before WordPiece tokenization. These characteristics indicate that the dataset is relatively compact at the sentence level but contains considerable lexical variation, which is expected in citizen-generated complaints where different users may describe similar problems through different lexical forms. Indonesian NER research has identified variation in datasets, annotation structures, and linguistic domains as important factors affecting the development and evaluation of entity recognition systems (Budi & Suryono, 2023). The descriptive statistics therefore provide an initial indication that the model must learn entity patterns from a corpus combining relatively short sequences with heterogeneous lexical expressions.

Table 3. Quantitative Descriptive Statistics of the Initial Dataset

Statistical Metric	Value	Description
Initial number of sentences	1,090	Before filtering
Sentences containing entities	1,061	97.3% of total
Pure-O sentences removed	24	2.2% of total
Sentences after balancing	1,066	Final dataset
Total tokens (initial dataset)	13,740	All label classes
Average sentence length	12.6 tokens	Per sentence
Vocabulary size (unique)	2,847 words	Before WordPiece
Percentage of O-labeled tokens	61.4%	8,432 / 13,740 tokens
Percentage of entity tokens	38.6%	5,308 / 13,740 tokens
Normalization dictionary size	11 slang-formal pairs	Domain-specific

The statistics presented in Table 3 reveal an important imbalance between entity and non-entity tokens, with the O label accounting for 61.4% of the 13,740 tokens, while entity tokens account for 38.6%. This distribution is expected in sequence-labeling tasks because complaint sentences contain functional words and contextual expressions surrounding the actual entities, but the magnitude of the O class can still influence optimization and increase the tendency of a model to favor non-entity predictions. The removal of 24 pure-O sentences represents a relatively limited filtering operation, equivalent to 2.2% of the initial corpus, and was intended to ensure that the final dataset retained sentences containing information relevant to at least one target entity. The presence of 11 slang-formal normalization pairs further demonstrates that lexical normalization was domain-specific rather than based on a broad transformation of the corpus. This condition is relevant because citizen-generated language frequently contains spelling variation, abbreviations, and informal forms that may not correspond directly to standardized Indonesian vocabulary. Research involving Indonesian public and social-media text has similarly demonstrated the importance of processing user-generated linguistic variation when computational models are applied to citizen discourse (Aji et al., 2022). The statistical profile consequently suggests that the principal challenge is not merely dataset size, but the interaction between class distribution and heterogeneous lexical forms within a relatively small domain-specific corpus.

The dominance of the O class becomes particularly important when interpreting NER performance because high token-level accuracy can potentially conceal difficulties in identifying comparatively less frequent entity classes. In the present dataset, 8,432 of the 13,740 tokens were assigned to O, meaning that a model can obtain substantial correct classifications through accurate recognition of non-entity contexts even when some entity boundaries remain difficult to detect. This characteristic supports the use of strict entity-level evaluation in subsequent analysis, where an entity must be identified with both the correct class and complete boundary rather than receiving credit from isolated token predictions. The issue is consistent with the methodological concern that evaluation

choices and annotation structures can substantially affect the interpretation of NER performance (Alshammari & Alanazi, 2021). The 5,308 entity tokens provide sufficient representation for learning the four target categories, although their distribution across individual categories is not necessarily uniform. Variability in entity frequency can cause the model to learn highly recurrent lexical patterns more effectively than categories characterized by broader contextual variation. The problem is particularly relevant for location entities because local place names can appear in multiple lexical forms, abbreviations, and compositional structures that are not necessarily shared across different complaints. Such conditions create a basis for examining category-specific precision and recall rather than relying exclusively on aggregate performance.

The vocabulary characteristics further indicate that the corpus contains a combination of recurring public-service terminology and localized expressions that may influence contextual representation during IndoBERT fine-tuning. A vocabulary size of 2,847 unique words before WordPiece tokenization suggests that lexical diversity is substantial relative to the total number of sentences, although the average sentence length remains short at 12.6 tokens. WordPiece tokenization can address previously unseen or infrequent word forms by decomposing them into subword units, but this process may also fragment locally specific names and informal expressions into multiple tokens. Such fragmentation becomes particularly relevant when entity boundaries correspond to multiword place names or expressions whose constituent tokens individually have weak semantic indications of the target entity. Transformer-based Indonesian models have demonstrated the capacity to capture contextual information across such sequences, which provides a methodological basis for applying IndoBERT to domain-specific NER tasks (Cahyo et al., 2025). The dataset therefore presents a meaningful test environment in which contextual modeling must operate alongside lexical normalization and BIO-based annotation. The combination of short complaint sentences, diverse vocabulary, and nonstandard language can make local context more important than isolated lexical identity during entity prediction.

The linguistic heterogeneity of the corpus is also consistent with the broader characteristics of Indonesian digital communication, particularly where Indonesian interacts with regional languages and informal online writing conventions. Code-mixing between Indonesian and Javanese is particularly relevant for Semarang complaint data because local expressions may appear within otherwise Indonesian sentences and may carry information necessary for interpreting a location, service, or problem description. Previous corpus research on Indonesian-Javanese-English code-mixed communication demonstrates that mixed-language text requires explicit attention because conventional language-processing assumptions may not adequately represent its linguistic structure (Hidayatullah et al., 2023). In a public complaint context, the effect can extend beyond language identification because a mixed-language phrase may also influence the boundaries and semantic interpretation of an entity. The resulting corpus therefore differs from formal Indonesian datasets in which grammatical regularity and standardized vocabulary provide stronger lexical cues for entity recognition. The presence of nonstandard abbreviations and localized expressions may explain why contextual learning is essential for distinguishing entity classes that cannot be identified reliably through dictionary matching alone. This characteristic also reinforces the importance of domain-specific evaluation because performance obtained from formal news or legal datasets cannot automatically be interpreted as representative of municipal complaint texts. The dataset exploration consequently establishes linguistic and statistical conditions that provide a direct basis for interpreting the training behavior and category-level NER results reported in the following analysis.

The combined statistical and linguistic profile indicates that the complaint corpus provides a challenging but operationally relevant environment for evaluating IndoBERT-based NER. The 61.4% proportion of O tokens creates a class imbalance that requires careful interpretation of accuracy, while the 38.6% entity-token proportion provides substantial evidence for learning the semantic categories required by the application. The relatively short average sentence length can facilitate contextual modeling because relevant entities may occur within compact textual windows, although abbreviated and code-mixed expressions can reduce the reliability of surface-level lexical cues. The 2,847-word vocabulary and 11 domain-specific normalization pairs further indicate that the corpus contains sufficient lexical diversity to test whether contextual representations can generalize beyond repeated entity strings. Similar domain adaptation considerations have been observed in Indonesian public-service applications using IndoBERT, where citizen-generated text presents linguistic characteristics

that differ from standardized language resources (Hidayat & Gata, 2025). These findings establish a clear analytical basis for examining the subsequent training and validation curves, particularly whether the model can reduce learning error without merely exploiting the dominant O class. The dataset characteristics also provide an explanation framework for interpreting why certain entity categories may achieve stronger recognition performance than others in the final strict-chunk evaluation.

IndoBERT Training and Entity Extraction Performance

The IndoBERT model was fine-tuned for 40 epochs using the annotated complaint corpus, with the training process monitored through training loss and validation loss to assess learning behavior and generalization. The training loss was first recorded at Epoch 10 because the configured logging_steps exceeded the 54 optimization steps available within a single epoch, causing the initial logging point to occur only after nine epochs had passed. Validation loss, in contrast, was recorded at the end of each epoch because the evaluation strategy was configured at the epoch level. This distinction in logging frequency is important when interpreting the learning curve because the absence of early training-loss points does not indicate that optimization was inactive during the initial epochs. The model continued updating its parameters throughout the training process, while the recorded values only represent the points at which the logging mechanism captured the loss. Transformer-based fine-tuning is particularly dependent on monitoring the relationship between training and validation behavior because contextual representations must adapt to the target domain without excessively fitting the training corpus (Tandi et al., 2025). The resulting loss trajectories are presented in Figure 3, which provides the basis for assessing the model's optimization behavior across the 40 training epochs.

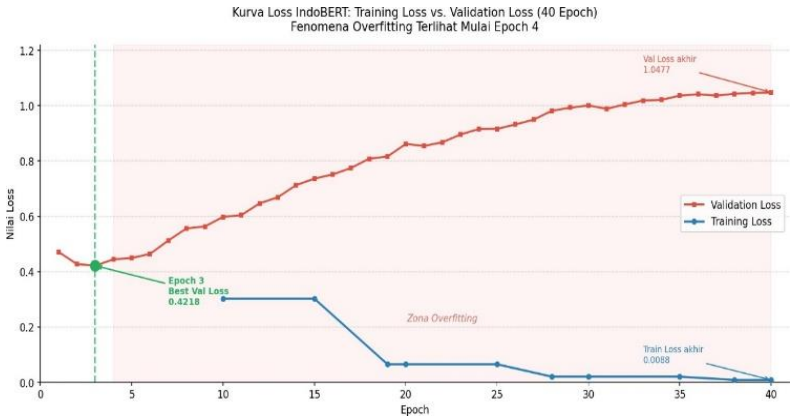

Figure 3. IndoBERT Loss Curve: Training Loss vs. Validation Loss (40 Epochs)

The training process demonstrates the adaptation of a pretrained Indonesian language representation to the specific semantic and lexical characteristics of municipal complaint texts. The validation loss provides an important complementary perspective because decreasing training loss alone cannot establish that the model has learned entity patterns that generalize to unseen complaints. The 40-epoch training configuration provides repeated exposure to the annotated corpus, allowing the token-classification layer and the underlying contextual representations to adjust to the four target entity categories. The interpretation of the loss curve must also account for the relatively limited size of the domain-specific corpus, since repeated optimization over a compact dataset may increase the possibility of learning corpus-specific patterns rather than broadly transferable linguistic representations. IndoBERT has previously demonstrated applicability across Indonesian NLP tasks involving public-sector and user-generated text, supporting its use as a pretrained foundation for domain-specific adaptation (Agustina et al., 2024). In the present task, the relevant learning objective is more granular because the model must distinguish entity boundaries within sentences rather than assigning a single label to an entire complaint. The relationship between the training and validation curves therefore provides an initial indication of whether contextual representations are being adapted in a controlled manner before final evaluation through strict entity-level metrics.

The final NER results demonstrate that IndoBERT successfully identified the four predefined entity categories, although performance differed substantially across categories. Table 5 shows that the LAYANAN category obtained the strongest F1-score of 0.68, supported by Precision of 0.64 and Recall

of 0.73, whereas LOKASI recorded the lowest F1-score at 0.56 because its Recall decreased to 0.51. KATEGORI_MASALAH and WAKTU both produced an F1-score of 0.61, with Recall values of 0.63 and 0.62, respectively, indicating moderate ability to recover annotated entities from the test data. The micro, macro, and weighted averages were all close to 0.61–0.62, suggesting that the overall model behavior was relatively consistent across aggregation strategies despite differences in individual entity categories. The results indicate that the model does not fail uniformly across entity types, but instead exhibits category-specific sensitivity to lexical and contextual characteristics. Comparable Indonesian NER research has shown that transformer-based models can provide effective contextual representations while their performance remains dependent on the domain and structure of the evaluated corpus (Istiqomah & Novika, 2025). The classification results therefore require interpretation at the entity level rather than being reduced to a single aggregate performance value.

Table 5. Classification Report Seqeval — Strict-Chunk Level Evaluation

Entity Category	Precision	Recall	F1-Score	Accuracy	Entity Support
KATEGORI_MASALAH	0.59	0.63	0.61	95.70%	404
LAYANAN	0.64	0.73	0.68	98.45%	173
LOKASI	0.64	0.51	0.56	97.14%	279
WAKTU	0.60	0.62	0.61	99.07%	90
micro avg	0.61	0.61	0.61	—	946
macro avg	0.61	0.62	0.62	—	946
weighted avg	0.61	0.61	0.61	—	946

The category-level results in Table 5 reveal that LAYANAN has the highest recognition capability, particularly in terms of Recall, which reaches 0.73, indicating that the model successfully retrieves a relatively large proportion of the service entities present in the reference annotations. This performance can be interpreted as an indication that service-related expressions possess comparatively stable lexical or contextual patterns, particularly when complaints explicitly mention government agencies, public facilities, or recognizable service terms. The lower Recall of LOKASI at 0.51 indicates that nearly half of the annotated location entities were not recovered under the strict matching criterion, suggesting greater difficulty in identifying complete location spans. Local place names may consist of multiple tokens, abbreviations, or combinations of generic geographic terms and specific names, making boundary detection more demanding than recognition of comparatively standardized service expressions. Transformer-based NER models have demonstrated strong contextual capabilities in Indonesian specialized domains, but domain-specific terminology and entity structures remain important determinants of recognition quality (Yulianti et al., 2024). The WAKTU category achieves the highest reported entity-level Accuracy at 99.07%, although its F1-score of 0.61 indicates that this value should not be interpreted as evidence of uniformly high entity extraction performance because the O class contributes substantially to token-level correctness. This distinction illustrates why strict Precision, Recall, and F1-score provide more informative evidence for entity extraction than accuracy alone.

The confusion matrix provides a more granular representation of the token-level prediction behavior and clarifies the sources of error that are less visible in the aggregate classification report. As shown in Figure 4, the O label dominates the matrix because non-entity tokens occur substantially more frequently than individual entity labels, producing a large number of correct O predictions. The most prominent systematic error involves B-KATEGORI_MASALAH being predicted as O, representing false-negative behavior in which the model fails to recognize the beginning of a problem-category entity. The reverse pattern, in which O is predicted as B-KATEGORI_MASALAH, represents false-positive behavior and indicates that contextual expressions surrounding problem descriptions may sometimes resemble the lexical patterns learned for problem entities. These errors are consistent with the challenge of distinguishing entity boundaries in noisy user-generated text, particularly when a problem description is expressed through informal or abbreviated language. The use of strict entity evaluation makes such boundary errors consequential because an incomplete or incorrectly delimited entity cannot receive full credit even when some of its constituent tokens are correctly classified (Alshammari & Alanazi, 2021). The confusion matrix therefore indicates that future optimization

should focus not only on overall classification accuracy but also on improving boundary detection and reducing confusion between problem-related expressions and contextual non-entity tokens.

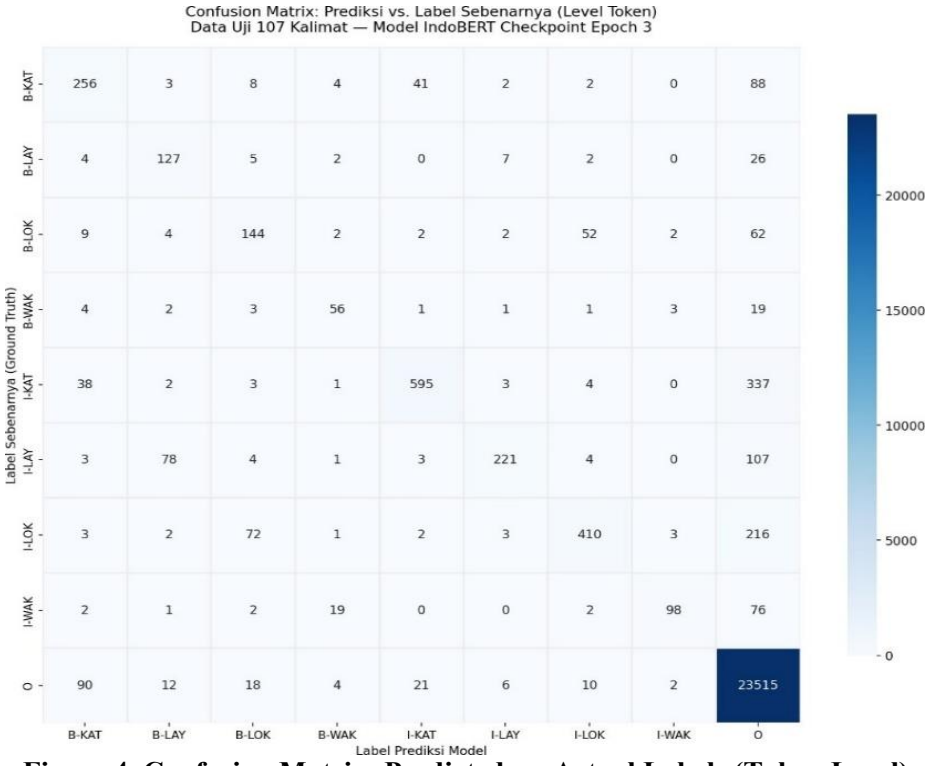

Figure 4. Confusion Matrix: Predicted vs. Actual Labels (Token Level)

The presence of 23,515 correct O predictions in the token-level confusion matrix further demonstrates the influence of class distribution on aggregate classification behavior. A high number of correctly predicted O tokens is useful because it indicates that the model can distinguish ordinary contextual language from entity-bearing expressions, but it can simultaneously obscure weaknesses in less frequent entity categories if accuracy is interpreted without complementary metrics. The strict-chunk results provide a more conservative assessment because the model must correctly identify both the semantic category and the complete span of an entity. The overall micro-average F1-score of 0.61 therefore indicates moderate entity extraction capability rather than near-complete recognition, despite the considerably higher token-level accuracy values reported for individual categories. This pattern is consistent with the broader observation that NER evaluation is sensitive to the definition of correctness and the relationship between token-level and entity-level assessment (Grandini et al., 2020). The results indicate that IndoBERT has learned meaningful entity representations from the Semarang complaint corpus, but the model remains sensitive to variations in entity expression and boundary structure. These findings establish the basis for evaluating how the trained model is subsequently integrated into an application environment for practical complaint-information extraction.

IndoBERT NER Application and Research Implications

The trained IndoBERT model was subsequently integrated into a Streamlit-based application to demonstrate the practical use of the NER model for processing public complaint texts. The application adopts a modular architecture consisting of a backend component, a frontend component, and the trained IndoBERT model as the computational core. The backend is implemented through ner_service.py, which manages model loading, preprocessing, inference, and aggregation of BIO predictions into complete entity spans. The frontend is implemented through app.py and provides an interactive interface through which users can submit complaint text and inspect the resulting entity annotations. The trained model is stored through the required model artifacts, including config.json, model.safetensors, tokenizer.json, and tokenizer_config.json, allowing the trained configuration and tokenizer to be reproduced independently from the application interface. The application architecture

separates computational inference from presentation logic, which facilitates maintenance and allows the NER component to be reused in other complaint-processing environments. This design is consistent with the increasing use of domain-specific NLP frameworks for extracting structured information from user complaints, where NER functions as an information-extraction layer rather than merely a standalone classification task (Asri et al., 2025).

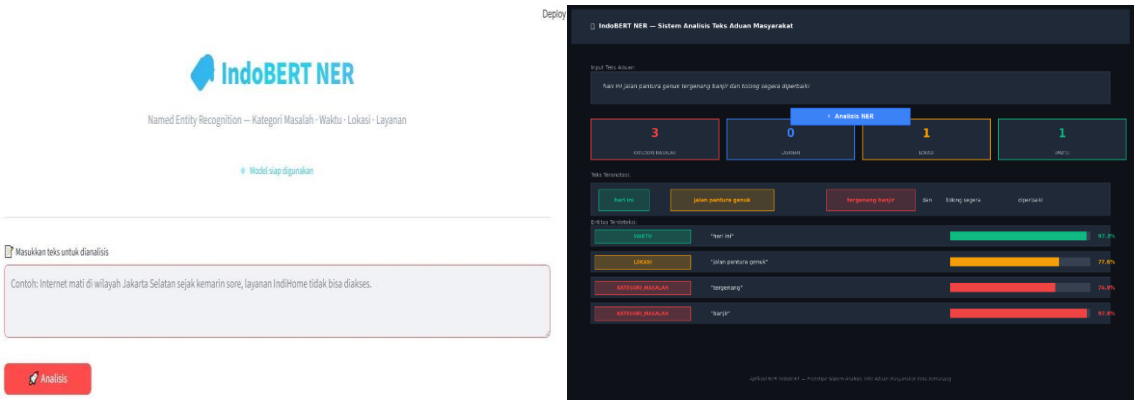

Figure 5. IndoBERT NER Application Interface

The application produces token-level annotations that are grouped according to the four target entities, namely KATEGORI_MASALAH, LAYANAN, LOKASI, and WAKTU, enabling users to inspect how the model interprets each complaint. The visualization assigns distinguishable representations to entity categories and provides a summary of the number of entities detected for each class. This output converts the model's numerical predictions into information that can be interpreted directly by complaint-management personnel without requiring users to examine BIO labels or model probabilities. The practical value of this representation lies in its ability to expose the semantic components of an unstructured complaint while retaining the original textual context. Such an approach extends the role of IndoBERT from predictive modeling toward decision-support functionality because extracted entities can subsequently become inputs for categorization, information retrieval, or complaint-routing processes. Previous work on e-government complaint data has demonstrated that entity extraction can be incorporated into broader analytical workflows to facilitate the transformation of complaint narratives into structured information (Umam et al., 2025). The application therefore represents an operational layer connecting the empirical NER model with the information requirements of municipal public-service management.

The performance differences across entity categories provide important evidence regarding how the linguistic structure of Semarang complaint texts influences IndoBERT's recognition behavior. LAYANAN achieved the highest F1-score of 0.68 and Recall of 0.73, indicating that service-related expressions were comparatively easier for the model to recover under the strict entity-matching criterion. This pattern can be associated with the relatively stable terminology used to refer to government agencies, public facilities, and service types, which provides contextual cues that can be learned during fine-tuning. KATEGORI_MASALAH and WAKTU both achieved F1-scores of 0.61, suggesting moderate recognition performance despite differences in their linguistic realization. The comparatively lower LOKASI F1-score of 0.56, together with its Recall of only 0.51, indicates that location extraction remains the principal weakness of the current model. Indonesian NER research has shown that transformer-based models can perform effectively in specialized domains, but entity characteristics and corpus-specific linguistic patterns continue to affect recognition outcomes (Istiqomah & Novika, 2025). The observed variation therefore indicates that model capability cannot be separated from the lexical diversity, entity structure, and contextual ambiguity associated with each category.

The lower recognition of LOKASI can be interpreted in relation to the linguistic complexity of geographically specific expressions in citizen-generated complaints. Location entities may contain combinations of street names, neighborhood names, landmarks, administrative areas, abbreviations, or colloquial references, creating entity boundaries that are less predictable than those of standardized service terms. WordPiece tokenization can further fragment rare or locally specific expressions into

subword units, potentially reducing the consistency of token-level patterns available to the classification layer. This issue becomes more pronounced when local expressions are combined with Indonesian or Javanese forms, since code-mixing may introduce lexical structures that differ from the dominant patterns represented in the training corpus. Research on Indonesian-Javanese-English code-mixed text confirms that mixed-language expressions introduce additional complexity for computational language processing because linguistic boundaries may not correspond neatly to conventional language categories (Hidayatullah et al., 2023). The result observed for LOKASI therefore suggests that additional domain-specific normalization, richer location dictionaries, or more extensive annotated examples could potentially improve entity-boundary recognition. The finding is also important from an application perspective because accurate location extraction can determine whether a complaint can be associated with a particular service area or administrative responsibility.

The distinction between token-level and entity-level performance also has implications for interpreting the practical reliability of the application. Although the reported category accuracies are relatively high, ranging from 95.70% for KATEGORI_MASALAH to 99.07% for WAKTU, the corresponding strict-chunk F1-scores remain between 0.56 and 0.68. This difference demonstrates that correct classification of individual tokens does not necessarily guarantee successful extraction of complete entities, particularly when an entity consists of multiple tokens or when its boundary is ambiguous. The use of strict matching is therefore appropriate for this application because a partially extracted location or service name may be insufficient for downstream routing even if several component tokens are correctly classified. The distinction between classification accuracy and other evaluation measures is particularly important in multi-class machine-learning applications because class distribution can substantially influence aggregate accuracy (Grandini et al., 2020). The results indicate that the application should use extracted entities as structured decision-support information rather than treating every predicted entity as unquestionably correct. Such interpretation is consistent with the broader development of IndoBERT-based Indonesian NLP systems, where model outputs are increasingly incorporated into practical information-processing workflows while remaining dependent on task-specific evaluation (Utomo, 2025).

From a public-service perspective, the implemented system demonstrates the feasibility of using IndoBERT to transform heterogeneous citizen complaints into structured information that can support more systematic complaint management. The extracted entities can provide an intermediate representation for identifying the nature of a problem, the affected service, the location of an incident, and the time associated with its occurrence. These representations can subsequently be linked to agency reference information to support more consistent routing decisions, reducing dependence on manual interpretation when complaint volumes increase. Research on Semarang's P3M system emphasizes the importance of effective complaint-management processes, while computational extraction offers a complementary mechanism for organizing information contained within citizen submissions (Pitaloka Subagyo et al., 2023). The current findings also demonstrate that IndoBERT is capable of operating on nonformal complaint language containing abbreviations, localized expressions, and code-mixing, although the moderate F1-scores indicate that the system remains subject to recognition errors. The principal methodological implication is that domain-specific NER should be evaluated through entity-level measures and error analysis rather than relying exclusively on overall accuracy. The application consequently provides a practical proof of concept for integrating Indonesian transformer-based NER into intelligent complaint-management workflows while identifying location recognition and entity-boundary accuracy as important directions for subsequent model improvement.

CONCLUSION

The implementation and evaluation demonstrate that IndoBERT can be applied to Named Entity Recognition (NER) in 1,066 public complaint texts from the City of Semarang containing nonformal language, abbreviations, and code-mixing, with four target entities comprising LAYANAN, LOKASI, KATEGORI_MASALAH, and WAKTU. The best model achieved a Macro F1-Score of 0.62 under strict-chunk evaluation, indicating that IndoBERT was able to extract meaningful structured information from heterogeneous complaint narratives, although its recognition capability remained moderate. Performance varied across entity categories, with LAYANAN obtaining the highest F1-score of 0.68, while LOKASI recorded the lowest F1-score of 0.56, reflecting the greater linguistic variability and boundary complexity of local geographical expressions. Error analysis indicates that entity

boundary shifts, ambiguity between entity categories, and vocabulary variation associated with informal and local expressions constitute important sources of prediction errors. The training process also indicated a tendency toward overfitting related to the relatively limited size of the domain-specific dataset compared with the capacity of the IndoBERT architecture, while the best-checkpoint selection mechanism helped retain the most appropriate model configuration for final evaluation. The integration of the trained model into a Streamlit prototype further demonstrates the feasibility of presenting extracted entities in an operational interface and using them as structured information for potential automated complaint routing. These findings position IndoBERT-based NER as a promising foundation for intelligent public complaint processing in Semarang, while its current performance indicates that deployment in a production environment requires additional optimization and validation.

Future development should prioritize expanding the annotated corpus to at least 5,000 complaint sentences to provide broader linguistic coverage and reduce the imbalance between model capacity and available training data. A larger corpus should include greater representation of local place names, informal expressions, abbreviations, spelling variations, and code-mixed Indonesian-Javanese complaints so that the model can learn more robust contextual patterns. The NER architecture can also be enhanced by incorporating a Conditional Random Field (CRF) layer above the IndoBERT token-classification output to improve sequential label consistency and reduce entity-boundary errors. Domain-specific normalization should be expanded through a corpus-driven vocabulary mechanism that continuously identifies and maps Semarang-specific dialects, abbreviations, and informal forms into representations that are more consistent with the annotated corpus. Future experiments should compare the proposed configuration with alternative transformer architectures and systematically evaluate whether additional preprocessing, data augmentation, class-balancing strategies, or CRF-based decoding can improve entity-level F1-score. The Streamlit prototype should also be developed into a more comprehensive complaint-routing framework by connecting extracted LAYANAN, LOKASI, KATEGORI_MASALAH, and WAKTU entities with validated agency-reference data and evaluating routing accuracy using real operational scenarios. Further validation using larger and temporally diverse complaint datasets is required before deployment so that the system's robustness, generalization capability, privacy protection, and reliability can be established under actual public-service conditions.

REFERENCES

Agustina, N., Naseer, M., Gusdevi, H., & Rismayadi, D. A. (2024, November). Development of a public complaint classification model to support E-Government using IndoBERT. In *2024 6th International Conference on Cybernetics and Intelligent System (ICORIS)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICORIS63540.2024.10903819

Aji, S., Maryani, I., & Muningsih, E. (2022). Analisis sentiment masyarakat menggunakan penggabungan algoritma Naive Bayes dan Particle Swarm Optimization. *IJCIT (Indonesian Journal on Computer and Information Technology), 7*(2), 119–126. https://doi.org/10.31294/ijcit.v7i2.14086

Alshammari, N., & Alanazi, S. (2021). The impact of using different annotation schemes on named entity recognition. *Egyptian Informatics Journal, 22*(3), 295–302. https://doi.org/10.1016/j.eij.2020.10.004

Asri, Y., Kuswardani, D., Suliyanti, W. N., & Fadhilah, N. A. (2025, October). A Domain-Specific Framework for Sentiment Analysis and Named Entity Recognition of PLN Mobile User Complaints. In *2025 International Conference on Advanced Technologies in Energy and Informatic (ICATEI)* (pp. 597-602). IEEE. https://doi.org/10.1109/ICATEI67676.2025.11405393

Budi, I., & Suryono, R. R. (2023). Application of named entity recognition method for Indonesian datasets: A review. *Bulletin of Electrical Engineering and Informatics, 12*(2), 969–978. https://doi.org/10.11591/eei.v12i2.4529

Cahyawijaya, S., Winata, G. I., Wilie, B., Vincentio, K., Li, X., Kuncoro, A., Ruder, S., Lim, Z. Y., Bahar, S., Khodra, M., Purwarianti, A., & Fung, P. (2021). IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 8875–8898). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.699

.

Cahyo, P. W., Aesyi, U. S., Setianto, W. A., & Sulaiman, T. (2025). A novel named entity recognition approach of Indonesian fake news using part of speech and BERT model on presidential election. *International Journal of Information Management Data Insights, 5*(2), 100354. https://doi.org/10.1016/j.jjimei.2025.100354

Erna, E. D., Baskoro, D., & Firliana, R. (2026). Implementation of the IndoBERT Model on Named Entity Recognition for Entity Identification in the Folklore of the Tale of Wayang Arjuna & Purusara. *The Indonesian Journal of Computer Science Research*, *5*(2), 131-137. https://doi.org/10.59095/ijcsr.v5i2.283

Firdaus, M. P., & Trisnawarman, D. (2025). Public sentiment analysis of the public housing savings program using the IndoBERT Lite model on YouTube comments. *MALCOM: Indonesian Journal of Machine Learning and Computer Science, 5*(1), 359–368. https://doi.org/10.57152/malcom.v5i1.1744

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for multi-class classification: An overview*. arXiv. https://arxiv.org/abs/2008.05756

Hidayat, R., & Gata, W. (2025). Pemanfaatan IndoBERT untuk analisis sentimen ulasan aplikasi Depok Single Window. *Information System for Educators and Professionals: Journal of Information System, 10*(1), 13–24. https://doi.org/10.51211/isbi.v10i1.3334

Hidayatullah, A. F., Apong, R. A., Lai, D. T. C., & Qazi, A. (2023). Corpus creation and language identification for code-mixed Indonesian-Javanese-English tweets. *PeerJ Computer Science, 9*, e1312. https://doi.org/10.7717/peerj-cs.1312

Istiqomah, N., & Novika, F. (2025). Perbandingan kinerja model NER IndoBERT dan IndoLEM dalam ekstraksi informasi kesehatan pascabencana dari berita daring di Indonesia: Comparative performance of IndoBERT and IndoLEM baseline models for post-disaster health information extraction from Indonesian online news. *Journal of Computer Science and Informatics Engineering, 4*(3), 158–174. https://doi.org/10.55537/cosie.v4i3.1173

Maharani, W., & Latifa, A. Z. (2025). Analyzing public sentiment on the relocation of Indonesia's capital to Kalimantan as the Ibu Kota Nusantara using logistic regression. *Jurnal Teknik Informatika (JUTIF), 6*(2), 575–592. https://doi.org/10.52436/1.jutif.2025.6.2.4230

Pitaloka Subagyo, A. D., Nurcahyanto, H., & Marom, A. (2023). Analisis sistem pengelolaan pengaduan masyarakat pada Pusat Pengelolaan Pengaduan Masyarakat (P3M) Kota Semarang. *Journal of Management and Public Policy, 12*(2), 621–634. https://doi.org/10.14710/jppmr.v12i2.38488

Putra Pratama, I. W. B. S., Putra, M. A. P., & Paramitha, A. A. I. I. (2026). Deteksi berita hoaks bahasa Indonesia menggunakan model hybrid IndoBERT-BiLSTM dengan teknik hyperparameter tuning. *JATI (Jurnal Mahasiswa Teknik Informatika), 10*(2), 2830–2838. https://doi.org/10.36040/jati.v10i2.17872

Tandi, T. Y., Abidin, T. F., & Riza, H. (2025). Incorporation of IndoBERT and machine learning features to improve the performance of Indonesian textual entailment recognition. *Journal of Information Systems Engineering and Business Intelligence, 11*(2), 173–186. https://doi.org/10.20473/jisebi.11.2.173-186

Umam, A. K. ., Alzami, F., Sani, R. R. ., Rohmani, A. ., Prabowo, D. P. ., Pergiwati, D. ., Megantara, R. A. ., & Iswahyudi, I. (2025). Enhancing Entity Extraction in E-Government Complaint Data using LDA-Assisted NER. *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, *9*(4), 1878-1888. https://doi.org/10.33395/sinkron.v9i4.15292

Utomo, V. G. (2025). Benchmarking IndoBERT and Transformer Models for Sentiment Classification on Indonesian E-Government Service Reviews. *Jurnal Transformatika*, *23*(1), 86-95. https://doi.org/10.26623/transformatika.v23i1.12095

Yulianti, E., Bhary, N., Abdurrohman, J., Dwitilas, F. W., Nuranti, E. Q., & Husin, H. S. (2024). Named entity recognition on Indonesian legal documents: A dataset and study using transformer-based models. *International Journal of Electrical and Computer Engineering, 14*(5), 5489–5501. https://doi.org/10.11591/ijece.v14i5.pp5489-5501