Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.15292
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Enhancing Entity Extraction in E-Government Complaint Data using LDA-Assisted NER

Ahmad Khotibul Umam[1], Farrikh Al Zami[2], Ramadhan Rakhmat Sani[3], Asih Rohmani[4], Rama Aria Megantara[5]*, Dwi Puji Prabowo[6], Dewi Pergiwati[7], Iswahyudi[8]
[1,2,3,4)]Sistem Informasi, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
[5]Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
[67]Desain Komunikasi Visual, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
[8]Dinas Komunikasi dan Informatika Provinsi Jawa Tengah
ahmadumam246@gmail.com[1], alzami@dsn.dinus.ac.id[2], ramadhan_rs@dsn.dinus.ac.id[3], aseharsoyo@dsn.dinus.ac.id[4], aria@dsn.dinus.ac.id[5], dwi.puji.prabowo@dsn.dinus.ac.id[6], dewi.pergiwati@dsn.dinus.ac.id[7], iswahyudi@jatengprov.go.id[8]

Submitted :Sep 1, 2025 | **Accepted** : Sep 25, 2025 | **Published** : Oct 2, 2025

Abstract: With the rapid development of information technology, governments are increasingly challenged to provide digital channels that enhance public participation in governance. LaporGub, an official platform managed by the Central Java Provincial Government, accommodates citizens' aspirations and complaints, but faces challenges in processing large amounts of unstructured text. Manual analysis is time-consuming and error-prone, resulting in delayed responses and decreased service quality. Conventional Named Entity Recognition (NER) models struggle to handle informal Indonesian-language text, while transformer-based approaches require substantial computing resources that are not widely available in local government environments. Therefore, this study aims to develop a lightweight NER approach by integrating Latent Dirichlet Allocation (LDA) as a semantic pre-annotation tool to improve the accuracy of entity extraction in Indonesian e-government complaint data. To achieve this goal, a dataset of 53,858 complaint reports from the LaporGub platform (2022–2025) was processed using LDA topic modeling (k=10) to provide semantic context during annotation. Next, the enriched dataset was used to train a spaCy-based NER model targeting three entity types: LOCATION, ORGANIZATION, and PERSON, with a training-validation-test split ratio of 70:15:15 using stratified sampling. The evaluation showed that the proposed NER+LDA model achieved a precision of 90.03%, a recall of 81.86%, and an F1-score of 85.75%, representing improvements of +5.78, +2.55, and +4.04, respectively, compared to the baseline NER model (F1-score: 81.71%). Furthermore, the most significant improvements occurred in the detection of ORGANIZATION and PERSON entities. These findings confirm that the integration of LDA as a pre-annotation strategy effectively improves NER performance on informal complaint texts in Indonesia, thus offering a practical and resource-efficient alternative to transformer-based methods for e-government applications.

Keywords: Named Entity Recognition, Latent Dirichlet Allocation, Text Mining, Public Services, E-Government

INTRODUCTION

With the rapid development of information technology, governments are increasingly required to provide digital channels that facilitate public participation in governance. In Central Java, one such initiative is LaporGub, an official platform managed by the provincial government to collect aspirations and complaints from citizens (Sakir, 2024). However, the large and growing volume of daily submissions creates significant challenges for processing unstructured text. Manual analysis is both time-consuming and error-prone, often resulting in delayed responses and decreased service quality (Azzahra, 2024).

To address these challenges, automated approaches are required to extract structured insights from unstructured complaint reports. Named Entity Recognition (NER) is a natural language processing technique designed to identify and classify important entities such as names of people, organizations, and locations. NER has been applied successfully in domains such as cybersecurity (Srivastava et al., 2023), healthcare

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

(Kusumawardani & Kusumawati, 2024), and disaster-related news data (Shidik et al., 2024). its application in Indonesian public complaint systems remains limited. Complaint texts are often informal, full of abbreviations, and linguistically diverse, which significantly reduces the effectiveness of conventional NER models (Reddy et al., 2025).

Recent advances such as IndoBERT, a pre-trained transformer model for Indonesian, have demonstrated strong performance in NER tasks (Reddy et al., 2025), However, these models require large annotated corpora and high computational resources, limiting their practicality in local government contexts. Alternative enrichment strategies such as gazetteers or weak supervision approaches also face adaptability issues because they depend on curated lexicons or complex labeling pipelines. Thus, there remains a need for lightweight, flexible, and context-sensitive solutions that can effectively handle noisy and linguistically diverse complaint texts.

To address this challenge, the present study proposes integrating Latent Dirichlet Allocation (LDA) with NER for Indonesian public complaint data. LDA is a probabilistic topic modeling technique that identifies latent semantic structures within large corpora (Choirinnisa et al., 2025). Rather than functioning as a final classifier, LDA is employed as a semantic pre-annotation tool to enrich the dataset construction process. By grouping semantically related words into topics, LDA supports more consistent labeling of underrepresented entities such as ORGANIZATION and PERSON, as shown in previous studies on agriculture and social media, where combining LDA with NER improved annotation quality and entity extraction performance (Gangadharan & Gupta, 2020).

Although research on Indonesian NER has expanded across domains such as dataset reviews (Budi & Suryono, 2023) and complaint analysis using topic modeling (Khadija & Nurharjadmo, 2023), there is no research that systematically integrates LDA with NER for Indonesian public complaint data, particularly in e-government platforms such as LaporGub. Previous works either used NER in isolation or applied LDA solely for thematic analysis, without leveraging its semantic structures to support entity annotation. To fill this gap, this study introduces a novel LDA-assisted spaCy-based NER model. Here, LDA functions as a complementary semantic layer that enhances annotation consistency and improves recognition of low-frequency entities. The novelty lies in positioning LDA not as a classifier but as an auxiliary semantic component that strengthens lightweight NER pipelines, offering a resource-efficient and adaptive solution for e-government applications.

The objective of this study is to improve the performance of NER in analyzing Indonesian public complaint data by integrating LDA into the annotation process. The study develops a spaCy-based NER model specifically designed for complaint data, applies LDA as a semantic pre-annotation tool to enrich entity labeling, and evaluates the effectiveness of this integration using standard metrics such as precision, recall, and F1-score (Aditama & Wicaksono, 2025). In addition, the research analyzes the distribution of extracted entities to generate actionable insights that can support government decision-making and improve responsiveness in handling citizen complaints. This research contributes to the advancement of more adaptive and resource-efficient NLP methods for e-government in Indonesia. By integrating topic modeling with NER, it offers a practical alternative for improving the quality of entity extraction in unstructured and linguistically diverse complaint texts.

LITERATURE REVIEW

Research on Named Entity Recognition (NER) in Indonesian has developed considerably across diverse domains, although its application to public complaint texts remains limited. In the health domain, for example, the BiLSTM-CRF algorithm has been employed to extract medical entities, demonstrating that deep learning approaches can be effective when supported by expert annotations and sufficient domain knowledge (Kusumawardani & Kusumawati, 2024). In another case, deep learning models applied to disaster-related datasets showed advantages in recognizing unseen entities; however, such models are highly dependent on large training corpora to achieve optimal performance (Shidik et al., 2024).

Transformer-based architectures, particularly IndoBERT, have set strong baselines for Indonesian NER tasks and shown promising results in various contexts (Wafda, 2025; Reddy et al., 2025). Despite their success, these approaches require large-scale annotated corpora and substantial computational resources, which may not be feasible for resource-constrained government environments. By contrast, lightweight frameworks such as spaCy pipelines offer simpler deployment and lower computational cost, but tend to struggle with the noisy, informal, and domain-specific vocabulary typically found in public complaint reports.

Alongside NER developments, topic modeling techniques have also gained traction in Indonesian text analysis. Latent Dirichlet Allocation (LDA), for instance, has been successfully applied to study public opinion on social media (Choirinnisa et al., 2025) and to analyze customer complaint data (Khadija & Nurharjadmo, 2023). These studies demonstrate LDA's ability to identify latent semantic structures and group words into meaningful topics, thereby improving interpretability. However, most of these works remain limited to thematic analysis, and do not extend to entity extraction. A few studies outside the Indonesian context, such as in the agricultural domain, have explored integrating LDA with NER and found that topic-level information can improve annotation quality and enhance extraction results (Gangadharan & Gupta, 2020).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.15292
e-ISSN : 2541-2019
p-ISSN : 2541-044X

To consolidate these prior findings, Table 1 provides a summary of related works on Indonesian NER and topic modeling. The table highlights key domains of application, methods employed, and their respective limitations.

Table 1 Summary of Related Works

Authors & Year	Data Domain	Methods Applied	Main Contribution	Limitations
(Kusumawardani & Kusumawati, 2024)	Indonesian health consultation (medical text)	BiLSTM-CRF with manual annotation	Effectively recognized medical entities with expert annotation support	Required large annotated corpus and strong domain expertise
(Wafda, 2025)	Multi-domain NLP (Indonesia)	IndoBERT (transformer)	Demonstrated high performance of IndoBERT across various NLP tasks	Limited in recognizing local terms; requires substantial computational resources
(Reddy et al., 2025)	Multilingual NLP	IndoBERT with fine-tuning	Improved NER accuracy in informal text	Less adaptive to region-specific/local entities
(Shidik et al., 2024)	Indonesian disaster news data	NER from multi-source media using advanced deep learning	Provided comprehensive benchmark for disaster-related NER in Indonesian textsrecognizing unseen entities	Still limited by noisy data quality and domain-specific vocabulary
(Cahyo et al., 2025)	Indonesian fake news texts	BiLSTM-CRF, BiGRU, BERT for NER	Requires high computational resources and large annotated corpora	Requires high computational resources and large annotated corpora
(Budi & Suryono, 2023)	Various Indonesian datasets	Review of NER methods and datasets	Provided systematic overview of NER applications in Bahasa Indonesia, covering methods, domains, and entity labels	Did not focus on integration with topic modeling; highlights gaps in complaint-specific NER research
(Khadija & Nurharjadmo, 2023)	Indonesian customer complaint texts	LDA topic modeling with BERT embeddings	Demonstrated how topic modeling supports analysis of complaint data, improving thematic grouping	Did not integrate NER; semantic annotation of entities remains unexplored
(Choirinnisa et al., 2025)	Social media (Indonesian public opinion)	LDA topic modeling	Effectively identified dominant topics in unstructured text	Did not integrate with NER
(Nursyahrina et al., 2024)	Social media	LDA for semantic context	Improved vocabulary grouping and opinion analysis	Did not explore integration with entity extraction

Table 1 summarizes previous related works on Indonesian NER and topic modeling using both traditional and transformer-based approaches. However, no study has systematically integrated LDA with NER for Indonesian public data, particularly for citizen complaint texts in e-government platforms such as LaporGub. Previous works have either focused on NER in isolation or employed LDA merely for thematic analysis, without leveraging its semantic structures to enhance entity annotation. This gap motivates the present study to propose an LDA-assisted spaCy-based NER model, in which LDA functions as a semantic pre-annotation tool to enrich entity labeling. By providing contextual cues, this approach improves annotation consistency and enhances the recognition of underrepresented entities such as ORGANIZATION and PERSON. The novelty of this research lies in positioning LDA not as a final classifier but as a complementary semantic layer that strengthens lightweight NER pipelines, offering a resource-efficient and adaptive alternative for e-government applications (Budi & Suryono, 2023; Nursyahrina et al., 2024).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.15292
e-ISSN : 2541-2019
p-ISSN : 2541-044X

METHOD

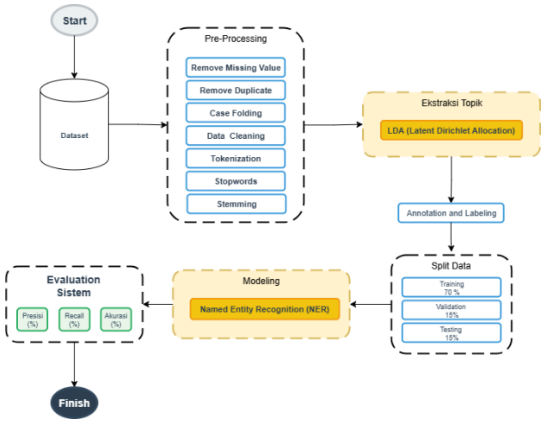

Fig. 1 Research method flow for NER–LDA integration.

The research method integrates Latent Dirichlet Allocation (LDA) and Named Entity Recognition (NER), as illustrated in Fig.1. The dataset consists of aspirations and complaints submitted by Central Java residents through the Laporgub platform between 2022 and March 2025, which was preprocessed to obtain clean and structured text. LDA was applied to extract frequently occurring topics, and the resulting thematic information was utilized to support dataset annotation and labeling. The annotated dataset was then divided into 70% for training, 15% for validation, and 15% for testing using a stratified strategy to preserve proportional distribution of entity types. NER modeling was performed on this dataset, and the system was evaluated using precision, recall, and F1-score metrics to measure performance.

Dataset

This study uses a dataset of Indonesian-language public complaints collected from the LaporGub platform between 2022 and March 2025. This dataset is stored in a structured CSV format to facilitate pre-processing and subsequent analysis. Most reports are written in an unstructured manner, often using informal Indonesian expressions, abbreviations, and colloquialisms, which require extensive pre-processing before modeling. To ensure research ethics, all data obtained from the platform is used exclusively for academic purposes without disclosing any personal or sensitive information of the reporters.

Pre-processing

The raw data often contained unstructured words, so it could not be directly used for modeling. Therefore, a pre-processing stage was required to clean and transform the data into a suitable format (Muhammad et al., 2022). The procedures included removing missing values to maintain data quality, eliminating duplicate entries to avoid redundancy, and applying case folding to ensure uniform representation of words such as “Semarang,” “semarang,” and “SEMARANG.” Data cleaning was conducted to remove characters such as numbers, punctuation, symbols, emoticons, and links. Tokenization was then applied, using NLTK for the LDA stage and spaCy for the NER stage. Stopword removal was performed to filter out frequent functional words (e.g., *yang, dan, atau*) that do not contribute to document representation (Jelita, 2024). Finally, stemming with the Sastrawi library was applied to reduce words to their root forms, for example *menjalankan* → *jalan* (Pardede & Darmawan, 2025).

Preprocessing was adjusted to the requirements of each modeling task. For LDA, a complete pipeline was used (case folding, cleaning, tokenization, stopword removal, and stemming with Sastrawi) to normalize linguistic variations, reduce noise, and improve topic coherence. For NER, a lighter approach was applied (removing missing values and duplicates, case folding, emoji removal, and punctuation removal) to preserve the original forms of entities such as person names, organizations, and locations. This dual strategy ensured semantic consistency for LDA and maintained entity authenticity for NER, allowing both tasks to operate under optimal conditions.

Topic Extraction with Latent Dirichlet Allocation (LDA)

In this study, topic modeling was performed using the Latent Dirichlet Allocation (LDA) method implemented through the Gensim library. The LDA approach, as a generative probabilistic model, assumes each document consists of multiple topics, where each word is associated with one of these topics (Nursyahrina et al., 2024). In this study, LDA was not applied as the final analysis result, but rather as a supporting tool during the annotation

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.15292
e-ISSN : 2541-2019
p-ISSN : 2541-044X

process. By mapping the topic distribution within the dataset, annotators gain a clearer semantic context, enabling more consistent and accurate labeling. An illustration of the LDA structure is presented in Fig. 2.

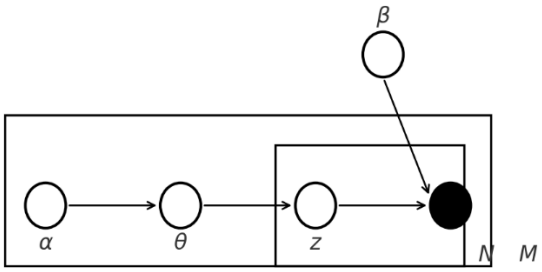

Fig. 2 LDA modeling process for topic distribution

Based on Fig. 2, the parameter α regulates how many topics typically appear in a document, θ reflects the topic distribution within each document, β indicates the distribution of words across topics, and z serves as the topic marker for each word (Choirinnisa et al., 2025). In this study, the number of topics (k) was optimized using the coherence score method, with k = 10 achieving the best balance (coherence = 0.48). To ensure topic quality, α was configured as symmetric for balanced allocation, while β (η) was set as asymmetric to allow greater flexibility in word assignments. The model was trained over 1,000 iterations with 20 passes through the corpus, and a fixed random seed was applied for reproducibility. Topic stability was validated by coherence consistency across multiple runs, ensuring that the generated topics were both semantically meaningful and computationally reliable to support the subsequent NER annotation process.

Data Annotation and Labeling

The annotation and labeling stage is an essential phase in forming a structured training and testing dataset from the LaporGub complaint corpus. This process, as the core of supervised learning, was carried out using a character-span-based annotation scheme to precisely mark the start and end positions of entities, in accordance with the requirements of modern libraries such as spaCy.

In the initial stage, a total of 877 entities were manually annotated from the original LaporGub complaint reports, covering three categories: PERSON, ORGANIZATION, and LOCATION. To enrich semantic context and address linguistic variation in complaint texts, Latent Dirichlet Allocation (LDA) was applied as a pre-annotation tool. This refinement increased the dataset to 899 entities, ensuring more consistent labeling of underrepresented categories such as PERSON and ORGANIZATION. The finalized dataset was then used as the gold standard for model training and evaluation.

Split Data

After the annotation and labeling process, the dataset was divided into three subsets using a stratified fragmentation strategy: 70% for training, 15% for validation, and 15% for testing. This stratified split ensures that the distribution of each entity type locations, organizations, and people remains proportional across the subsets, thus reducing potential bias in the evaluation model. The training subset is used to optimize model parameters, while the validation subset plays a role in hyperparameter tuning and applying early stopping to prevent overfitting. Meanwhile, the testing subset is used to provide an objective assessment of the model's generalization ability. The 70:15:15 configuration was chosen because the presence of a validation subset allows for systematic monitoring of model performance throughout the training process, which is crucial in Named Entity Recognition (NER) tasks that frequently deal with class measurements. Thus, this split strikes the right balance between data utilization for the learning process and clarity of evaluation results.

Modeling Named Entity Recognition (NER)

Named Entity Recognition (NER) is a technique in Natural Language Processing (NLP) that aims to identify and classify important entities in text (Li et al., 2022). In this study, it was built using the spaCy library with a supervised learning approach . The initial model for Indonesian was initialized using spacy.blank("id"), after which an NER component was added along with the LOCATION, ORGANIZATION, and PERSON entity labels according to the annotated dataset. The annotated dataset was divided into 70% for training, 15% for validation, and 15% for testing using a stratified strategy to maintain a proportional distribution of entity types. The training set was used to optimize the model parameters, the validation set was applied to tune hyperparameters and apply early stopping, while the test set was reserved for final evaluation to ensure unbiased performance measurement.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.15292
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The training process was conducted over 30 epochs/iterations with a dropout rate of 0.4, utilizing a minibatch compounding strategy in which the batch size dynamically ranged from 4 to 32 in order to enhance training stability. To ensure reproducibility, a fixed random seed 42 was applied throughout the process. Furthermore, an early stopping mechanism with a patience of five iterations, determined by the F1-score on the validation set, was employed to minimize the risk of overfitting and to promote robust model convergence. The training progress was systematically monitored through loss values and validation performance at each iteration, and the best-performing model was preserved for subsequent evaluation.

Evaluation System

The evaluation was conducted to measure the accuracy of the Named Entity Recognition (NER) model in detecting location entities, organizations, and people's names according to labels on previously unseen test data (Yanti et al., 2021). Model performance is examined using three common metrics: precision, recall, and F1 score. Precision reflects the ratio of correctly identified entities to the total number of predicted entities, recall indicates the proportion of relevant entities successfully detected out of all entities that should have been identified, while F1 score provides a balanced measure by calculating the harmonic mean of precision and recall.

Precision

Shows how many entity predictions are correct out of all entities predicted by the model.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

Recall

Shows how many entities the model actually recognizes out of all the entities it should recognize.

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

F1-score

Shows how many entities the model actually recognizes out of all the entities it should recognize

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

RESULT

This section presents the research results in stages, starting from the data preprocessing stage, topic analysis using LDA, application of the NER and NER+LDA models, to evaluation of model performance and distribution of entities in the LaporGub report corpus.

Preprocessing Results

A total of 54,212 raw report entries were obtained from the LaporGub platform. The data contained various anomalies, including blank values, duplicate entries, and non-standard characters. After the cleaning process, two blank entries and 352 duplicate entries were removed, leaving 53,858 valid records, as summarized in Table 2.

Table 2. Dataset Cleaning Results

Steps	Before	After
Missing values	2	0
Duplicates	352	0
Total	54212	53858

The cleaned data then went through a series of text preprocessing steps, including case folding, non-alphabetic character cleaning, tokenization, stopword removal, and stemming using the Sastrawi library. Examples of preprocessing results are shown in Table 3.

Table 3 Preprocessing Results

Data Before Preprocessing	Data After Preprocessing
Pasar Jepara II dan Terminal Jepara kota menjijikkan	['pasar', 'jepara', 'ii', 'terminal', 'jepara', 'kota', 'jijik']

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.15292
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Pihak sekolah SMAN 1 Jekulo Kudus melarang siswanya untuk mengerjakan pakai hp iPhone mohon bapak gubernur bisa ditindak lanjuti secepatnya.	['sekolah', 'sman', 'jekulo', 'kudus', 'larang', 'siswa', 'pakai', 'hp', 'iphone', 'mohon', 'gubernur', 'tindak', 'cepat']

This transformation successfully simplifies lexical variation and retains important information such as location and institution names, resulting in a text corpus that is optimal for further analysis.

Topic Modeling Results with Latent Dirichlet Allocation

The latent Dirichlet alloaction topic modeling produced the highest coherence value. The results of the coherence value are shown in Fig 3.

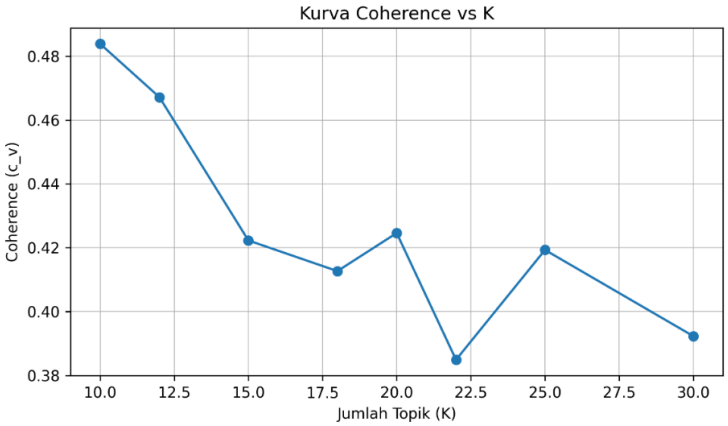

Fig.3 Coherence Score Results

In the coherence value graph, it can be seen that the highest graph value is on topic 10 with a value of 0.483875, so in this writing, 10 topics are used for coherence value. The highest coherence value can be used as a reference in this topic modeling. Topic modeling can find the value of words against topics and topics against documents. The results of the topic against words in.

Topic modeling using Latent Dirichlet Allocation (LDA) yielded optimal coherence at k = 10 topics with a coherence score of 0,483875. These topics were used as a thematic dictionary in the entity annotation stage, helping annotators understand the semantic context. The main topic extraction results are presented in Table 4.

Table 4 LDA Top 10 Topic Results

Topic	Top 10 Words
Topic 1	jam, layan, sakit, puskesmas, obat, makan, dri, rs, bpjs, rumah_sakit
Topic 2	kerja, usaha, tugas, gaji, istri, semarang, bayar, karyawan, samsat, hak
Topic 3	ganjar, jawa, daftar, jateng, salah, masuk, semarang, brebes, laksana, atur
Topic 4	anak, sekolah, biaya, siswa, bayar, rb, sma, kelas, sd, rp
Topic 5	bantu, rumah, tindak, keluarga, bangun, butuh, tanah, tinggal, keluh, dana
Topic 6	uang, proses, urus, bayar, surat, ambil, ktp, kantor, selesai, sulit
Topic 7	dpt, cair, pkh, magelang, pati, thn, tpi, suwun, karanganyar, mas
Topic 8	jual, beli, pasar, harga, dagang, vaksin, parkir, tertib, minyak_goreng, pupuk
Topic 9	jalan, rusak, arah, kondisi, jalan_rusak, tindak, dusun, jalan_raya, rusak_parah, jembatan
Topic 10	air, banjir, sungai, sawah, blora, sragen, nggak, longsor, pdam, tambang

Modeling Results

The trained NER model was then applied to the entire clean (unannotated) corpus to evaluate the overall entity distribution. The entity label distributions showed significant differences between the baseline model and the LDA-enhanced NER model, as shown in Fig. 4.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.15292
e-ISSN : 2541-2019
p-ISSN : 2541-044X

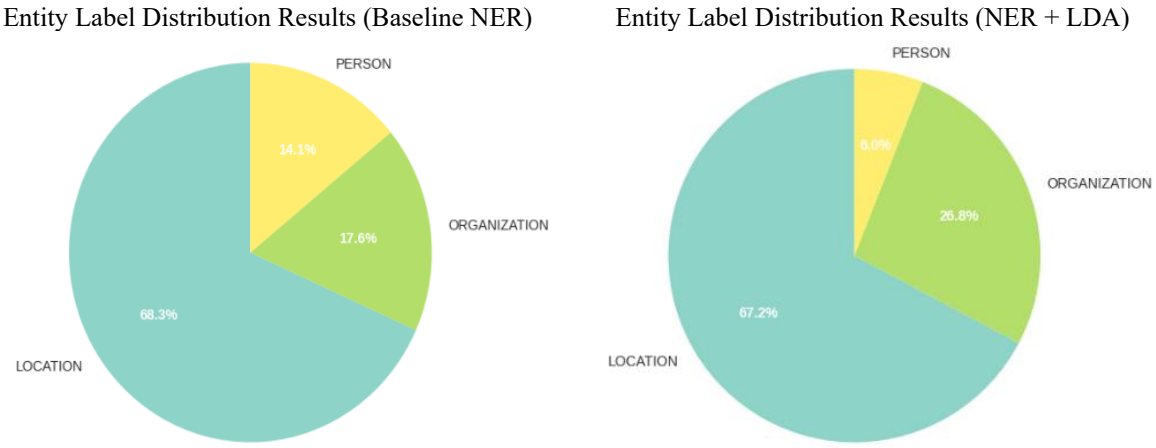

Fig. 4 Entity Distribution Results (Baseline NER) & (NER + LDA)

Fig. 4 compares the distribution of entities identified using the baseline NER model and the proposed NER+LDA model. In the baseline NER results, LOCATION dominates with 113,909 mentions (68.3%), followed by ORGANIZATION with 29,420 (17.6%) and PERSON with 23,478 (14.1%). This indicates that public complaint data is strongly spatial in nature, frequently referencing cities, districts, or sites. Organizations are mostly mentioned in relation to government institutions and services, while person entities generally refer to officials.

In contrast, the NER+LDA model shows a shift in proportions: LOCATION remains dominant with 95,725 mentions (67.2%), but ORGANIZATION increases significantly to 38,135 mentions (26.8%), while PERSON decreases to 8,500 mentions (6.0%). This suggests that integrating LDA enriches the detection of organizational entities—such as government agencies and service providers—while maintaining emphasis on spatial aspects. The smaller proportion of person entities reflects that individuals, often officials or public figures, are less frequently referenced compared to places and institutions.

Frequently Appearing Entities

Frequency analysis reveals the entities most frequently mentioned in public reports. The top ten entities are dominated by the names of districts/cities and government institutions, as shown in Fig 5 and Fig 6.

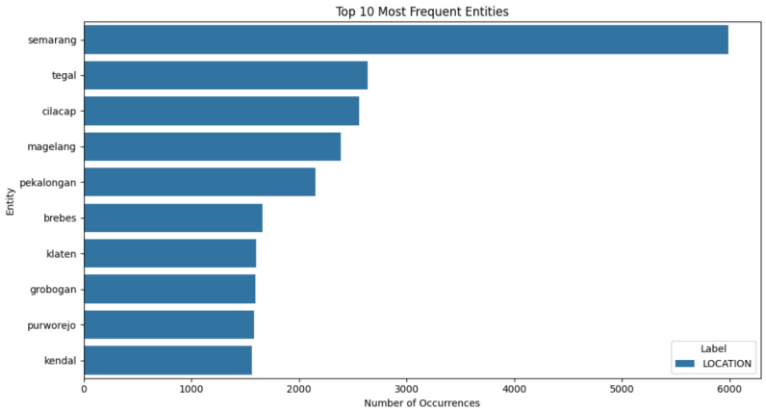

Fig. 5 Top 10 Most Frequent Entities (Baseline NER)

The visualization in Fig. 5 shows the Top 10 most frequent location entities found in the public complaint dataset. Semarang appears as the most dominant entity with 5,990 occurrences (3.59%), which is significantly higher compared to other regions. This reflects the high volume of public reports centered around the provincial capital, likely due to its large population and complex urban infrastructure. The next frequent entities include Tegal (2,637; 1.58%), Cilacap (2,561; 1.54%), Magelang (2,389; 1.43%), and Pekalongan (2,150; 1.29%), all of which represent major districts/cities in Central Java that serve as economic or regional hubs. Meanwhile, Brebes, Klaten, Grobogan, Purworejo, and Kendal each contribute less than 1% individually but together highlight that complaints are spread widely across multiple regions. Overall, the data suggests that although Semarang dominates in

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.15292
e-ISSN : 2541-2019
p-ISSN : 2541-044X

frequency, attention should also be directed toward other regencies with considerable report volumes, as these indicate region-specific public service or infrastructure challenges requiring government intervention.

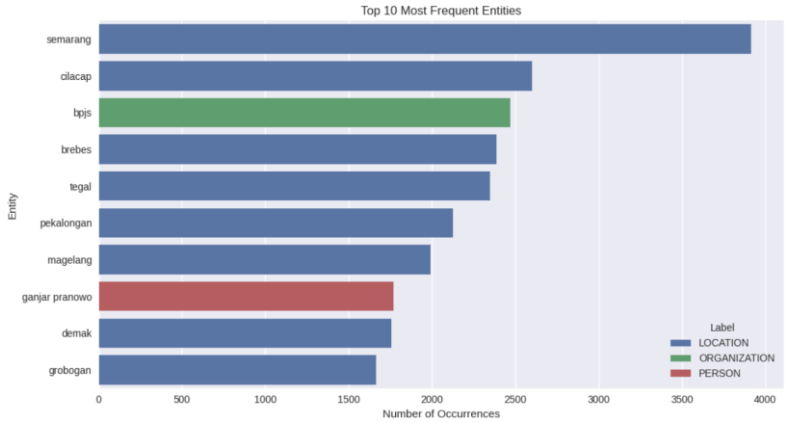

Fig. 6 Top 10 Most Frequent Entities (NER + LDA)

Fig. 6 shows the entity extraction results after applying NER enhanced with LDA, showing the 10 most frequently mentioned entities. Semarang City remains the dominant entity with 3,915 occurrences (2.75%), followed by Cilacap (2,600; 1.83%). BPJS emerges as the primary organizational entity with 2,469 mentions (1.73%), reflecting the high volume of public complaints related to health insurance services. Other frequently mentioned locations include Brebes (2,389; 1.68%), Tegal (2,349; 1.65%), Pekalongan (2,125; 1.49%), Magelang (1,993; 1.40%), Demak (1,755; 1.23%), and Grobogan (1,665; 1.17%), which represent hotspots for public problems at the regional level. Importantly, the emergence of the private entity Ganjar Pranowo (1,814; 1.27%) indicates that public figures also feature prominently in citizen reports. Overall, these findings confirm that LDA integration enriches entity detection, encompassing not only locations but also organizations and individuals, thus providing the government with more comprehensive insights to effectively address public issues..

Comparative Evaluation of Models

Quantitative evaluation of both models was conducted using precision, recall, and F1-score metrics on a 10% subset of the test data. The comparative results are presented in Table 5.

Table 5 Model Performance Comparison

Metrix	**NER Model (Without LDA)**	**NER + LDA Model (Proposed)**	**Performance Improvement**
Precision	84.25 %	90.03 %	+5.78
Recall	79.31 %	81.86 %	+2.55
F1-Score	81.71 %	85.75 %	+4.05

The comparative evaluation demonstrates that the proposed NER+LDA model achieved consistent improvements across all metrics compared to the baseline NER model. Precision increased from 84.25% to 90.03% (+5.78), reflecting the model's enhanced ability to reduce false positives by leveraging the semantic context provided by topic modeling. Recall also improved, rising from 79.31% to 81.86% (+2.55). Although the gain in recall was smaller than the increase in precision, this indicates that LDA contributed to capturing entities that were previously overlooked, particularly those with lower frequency or context-dependent expressions. Consequently, the F1-score improved from 81.71% to 85.75% (+4.04), confirming a balanced enhancement in both precision and recall. These findings suggest that the integration of LDA provides not only greater accuracy in entity classification but also broader coverage in entity detection, enabling a more comprehensive understanding of public complaints. In practical terms, this improvement enhances the government's ability to identify diverse issues across regions, institutions, and individuals, thus supporting more informed and responsive policy decisions.

DISCUSSIONS

The results of this study indicate that the integration of the Latent Dirichlet Allocation (LDA) method with Named Entity Recognition (NER) significantly improves the accuracy of entity extraction in Indonesian public complaint data. The proposed NER+LDA model achieved a Precision of 90.03%, Recall of 81.86%, and F1-score of 85.75%, higher than the baseline model without LDA (Precision 84.25%, Recall 79.31%, F1 81.71%). This improvement proves that the topic distribution from LDA is able to provide additional semantic context that

*name of corresponding author

strengthens the entity identification process. This finding is in line with Khadija & Nurharjadmo (2023) who showed that LDA can support customer complaint analysis through better semantic clustering.

Further entity distribution analysis shows that the LOCATION category remains dominant (67.2%), followed by ORGANIZATION (26.8%), and PERSON (6.0%). This pattern is consistent with Shidik et al. (2024), who found that spatial references were the most frequently appearing entity in NER tasks related to disasters in Indonesia. This can be explained because public reports generally focus on specific locations such as cities, districts, or specific infrastructure. The dominance of the ORGANIZATION entity indicates that citizens tend to associate complaints with institutions or service providers (e.g., BPJS or government agencies). This finding aligns with Kusumawardani & Kusumawati (2024), who also found a dominance of institutional entities in health consultation services. Meanwhile, the relatively small proportion of PERSON entities is understandable because individual names are usually only mentioned when referring to specific officials or public figures. This is consistent with Choirinnisa et al. (2025), who found that in public opinion on social media, individual entities appear primarily when referring to relevant public figures. This distribution reflects the natural characteristics of public complaint data, but it also poses challenges for supervised learning, as models tend to be biased toward the majority class.

The increased precision values achieved indicate that LDA integration helps minimize misclassification of ambiguous entities. The distribution of words within topics provides a richer semantic context, allowing the model to label more selectively. Meanwhile, although the recall increase is relatively smaller, these results indicate that entities with low frequency or unclear context that were previously difficult to recognize can be more consistently detected. These findings support the review by Li et al. (2022), which emphasized that enriching semantic context is a crucial factor in reducing both false positive and false negative errors in NER tasks.

Furthermore, LDA integration also facilitated the detection of new entities previously missed by the baseline model. For example, the BPJS entity as ORGANIZATION and Ganjar Pranowo as PERSON began to appear more consistently after the addition of topic context. This aligns with Reddy et al. (2025) who emphasized the importance of entity enrichment strategies to improve the detection of rare or domain-specific terms in local languages. From a practical perspective, the ability to detect these new entities is crucial because it not only generates a location-based complaint map but also provides more comprehensive insights into the institutional and individual actors frequently mentioned in public reports. This allows the government to formulate more responsive policies, prioritize location-based interventions, strengthen coordination with relevant organizations, and ensure the accountability of public officials.

Data Quality and Ethical Framework

The dataset demonstrates comprehensive temporal coverage spanning from 2022 to March 2025, providing a robust foundation for analyzing citizen complaint patterns across multiple years. All regencies and cities within Central Java Province are represented in the LaporGub platform, ensuring complete geographic coverage without regional exclusions. The complaints are submitted in standard Indonesian language, maintaining linguistic consistency throughout the corpus and reducing preprocessing complexity for NER tasks. After data cleaning procedures that removed 2 missing values and 352 duplicate entries, the final dataset of 53,858 records provides sufficient volume and diversity for reliable entity extraction analysis.

In matter of Data Anonymization and Privacy, The Central Java Provincial Government has implemented privacy protection measures by removing all sender identification data from the complaint records prior to research access. This preprocessing ensures that individual complainants cannot be identified while preserving the substantive content necessary for entity extraction analysis.

In matter of Informed Consent Framework, Given the public nature of the LaporGub platform where citizens voluntarily submit complaints with the expectation of government review and action, we recommend adopting an implied consent model for research purposes. Citizens submitting complaints through official government channels have reasonable expectation that their submissions may be analyzed to improve public services. However, we recommend that government platforms include explicit language in their terms of service stating that anonymized complaint data may be used for service improvement research and policy analysis.

In matter of Data Retention and Management, All research data is retained in its current form without deletion to ensure reproducibility of results and enable future comparative studies. The dataset remains under secure institutional custody with access restricted to authorized research personnel only.

Regarding Algorithmic Ethics and Bias Mitigation, The imbalanced distribution of entity types (LOCATION 67.2%, ORGANIZATION 26.8%, PERSON 6.0%) reflects the natural characteristics of public complaint data but raises important fairness considerations. To address potential algorithmic bias, we implemented stratified sampling in our train-validation-test split to ensure proportional representation of all entity categories. The dominance of location entities may lead to better recognition accuracy for spatial references compared to organizational or personal entities, which could systematically underrepresent complaints involving specific institutions or

*name of corresponding author

individuals. Future implementations should incorporate bias-aware evaluation metrics and consider ensemble approaches to ensure equitable performance across all entity types.

Research Limitations and Future Work.

Although the findings are promising, this study has several limitations. First, the preprocessing stage did not fully normalize informal language such as slang, non-standard abbreviations, and regional dialects that frequently appear in citizen complaint texts, which may affect the consistency of entity recognition. Second, the error analysis revealed the presence of false positives and false negatives, for example when LOCATION entities were occasionally misclassified as ORGANIZATION or PERSON, indicating that the model still faces challenges in handling contextual ambiguity. Third, the imbalanced distribution of entities tends to bias the model toward predicting LOCATION, which may reduce the fairness of extraction results and limit the detection of minority categories.

For future research, several directions can be pursued. Expanding the dataset to include complaint reports from other provinces or national platforms such as SP4N LAPOR would allow validation across more diverse linguistic and regional contexts. In addition, the integration of gazetteer-based resources or hybrid approaches could be explored as complementary techniques to LDA, particularly for strengthening the recognition of ambiguous LOCATION and ORGANIZATION entities. Another possible extension is the application of statistical significance testing and fairness-aware learning strategies to ensure more reliable evaluation and balanced performance across entity categories. By addressing these limitations, future studies can establish a more robust, generalizable, and context-aware NER framework to support e-government applications in Indonesia.

CONCLUSION

This study demonstrates that integrating Latent Dirichlet Allocation (LDA) with Named Entity Recognition (NER) can significantly enhance entity extraction in unstructured Indonesian public complaint data. By employing LDA as a semantic pre-annotation method, the proposed model achieved consistent improvements with a Precision of 90.03%, Recall of 81.86%, and F1-score of 85.75%, outperforming the baseline NER model (Precision 84.25%, Recall 79.31%, and F1-score 81.71%).

The dominance of LOCATION entities confirms the inherently spatial nature of citizen complaints, while the improved recognition of ORGANIZATION and PERSON categories after LDA integration highlights the importance of topic-based semantic context in strengthening annotation consistency. These findings validate that combining LDA and NER not only improves technical performance but also provides actionable insights for government agencies, enabling more accurate identification of complaint patterns across regions, institutions, and individuals.

Although the results are promising, several limitations remain. The dataset is limited to complaints from a single provincial platform, and linguistic challenges such as slang and non-standard abbreviations persist. Future research should expand to broader datasets, refine preprocessing strategies for informal language, and explore complementary resources such as gazetteers or fairness-aware approaches to ensure greater robustness and generalizability.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Dinas Komunikasi dan Informasi Provinsi Jawa Tengah (Communication and Information Agency of Central Java Province) for providing access to the *Lapor Gub* complaint dataset, which made this research possible. Appreciation is also extended to colleagues, mentors, and academic peers who provided valuable insights and constructive feedback throughout the development of this study. Finally, the authors acknowledge the continuous support from their respective institutions, whose encouragement and resources contributed significantly to the successful completion of this work.

REFERENCES

Aditama, A. R., & Wicaksono, A. F. (2025). Classification of customer complaints on social media for e-commerce in Indonesia. *International Journal of Electrical and Computer Engineering (IJECE)*, *15*(3), 2977–2985. https://doi.org/10.11591/ijece.v15i3.pp2977-2985

Azzahra, M. D. (2024). *Analisis Implementasi Chatbot Sebagai Sarana Komunikasi dan Efisiensi Layanan Pelanggan Terhadap Peningkatan Kinerja PT Pelindo Terminal Petikemas Semarang* [Thesis, Universitas Islam Indonesia]. https://dspace.uii.ac.id/handle/123456789/51150

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

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*

*name of corresponding author

Journal of Information Management Data Insights, *5*(2), 100354. https://doi.org/10.1016/j.jjimei.2025.100354

Choirinnisa, D., Alzami, F., Indrayani, H., Rohmani, A., Nugraini, S. H., Zulfiningrumi, R., & Susanti, F. (2025). LDA Topic Modeling: Twitter-Based Public Opinion on Indonesian Ministry of Finance. *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, *9*(2), 849–863. https://doi.org/10.33395/sinkron.v9i2.14719

Gangadharan, V., & Gupta, D. (2020). Recognizing Named Entities in Agriculture Documents using LDA based Topic Modelling Techniques. *Procedia Computer Science*, *171*, 1337–1345. https://doi.org/10.1016/j.procs.2020.04.143

Jelita, M. (2024). Text Mining dengan Topic Modelling LDA dari Pertanyaan Gelar Wicara Literasi Perpustakaan Nasional RI. *Media Pustakawan*, *31*(3), 253–265. https://doi.org/10.37014/medpus.v31i3.5237

Khadija, M. A., & Nurharjadmo, W. (2023). Enhancing Indonesian customer complaint analysis: LDA topic modelling with BERT embeddings. *SINERGI*, *28*(1), 153–162. https://doi.org/10.22441/sinergi.2024.1.015

Kusumawardani, R. P., & Kusumawati, K. N. (2024). Named entity recognition in the medical domain for Indonesian language health consultation services using bidirectional-lstm-crf algorithm. *Procedia Computer Science*, *245*, 1146–1156. https://doi.org/10.1016/j.procs.2024.10.344

Li, J., Sun, A., Han, J., & Li, C. (2022). A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, *34*(1), 50–70. https://doi.org/10.1109/TKDE.2020.2981314

Muhammad, F., Maghfur, N. M., & Voutama, A. (2022). Sentiment Analysis Dataset On COVID-19 Variant News: Kumpulan Data Analisis Sentimen pada Berita Varian COVID-19. *Systematics*, *4*(1), 382–391. https://doi.org/10.35706/sys.v4i1.6347

Nursyahrina, Defit, S., & Sovia, R. (2024). Metode BERTopic dan LDA untuk Analisis Tren Penelitian Bidang Ilmu Komputer. *Jurnal KomtekInfo*, 332–341. https://doi.org/10.35134/komtekinfo.v11i4.580

Pardede, J., & Darmawan, D. (2025). Perbandingan Algoritma Stemming Porter, Sastrawi, Idris, Dan Arifin & Setiono Pada Dokumen Teks Bahasa Indonesia. *Jurnal Teknologi Informasi dan Ilmu Komputer*, *12*(1), 69–76. https://doi.org/10.25126/jtiik.2025128860

Reddy, S. K., Sheshadri, S. K., Avatapalli, K. L., & Gupta, D. (2025). Empirical Study on Efficiency of Different Language Modeling Techniques using Masking of Named Entities for Indic Languages. *Procedia Computer Science*, *258*, 146–159. https://doi.org/10.1016/j.procs.2025.04.228

Sakir, A. R. (2024). Tinjauan Literatur: Pemanfaatan Teknologi Informasi untuk Meningkatkan Mutu Pelayanan Publik. *Jurnal Administrasi Publik dan Bisnis*, *6*(2), 165–171. https://doi.org/10.36917/japabis.v6i2.170

Shidik, G. F., Saputra, F. O., Saraswati, G. W., Winarsih, N. A. S., Rohman, M. S., Pramunendar, R. A., Kusuma, E. J., Ratmana, D. O., Venus, V., Andono, P. N., & Hasibuan, Z. A. (2024). Indonesian disaster named entity recognition from multi source information using bidirectional LSTM (BiLSTM). *Journal of Open Innovation: Technology, Market, and Complexity*, *10*(3), 100358. https://doi.org/10.1016/j.joitmc.2024.100358

Srivastava, S., Paul, B., & Gupta, D. (2023). Study of Word Embeddings for Enhanced Cyber Security Named Entity Recognition. *Procedia Computer Science*, *218*, 449–460. https://doi.org/10.1016/j.procs.2023.01.027

Wafda, A. (2025). *Aspect-Based Sentiment Analysis terhadap Cuitan Platform X tentang Kurikulum Merdeka Menggunakan IndoBERT* [Thesis, Universitas Islam Indonesia]. https://dspace.uii.ac.id/handle/123456789/55157

Yanti, R. M., Santoso, I., & Suadaa, L. H. (2021). Application of Named Entity Recognition via Twitter on SpaCy in Indonesian (Case Study: Power Failure in the Special Region of Yogyakarta). *Indonesian Journal of Information Systems*, *4*(1), 76–86. https://doi.org/10.24002/ijis.v4i1.4677

*name of corresponding author