JASMINE
Journal of Intelligent Systems and Machine Learning

Articles in Press

Deep Learning Model for Identification of Indonesian National Figure Entities on Social Media Using LSTM Architecture

Very Dwi Setiawan[1,*], Dwi Utari Iswavigra[2] and Mutia Ulfa[3]
[1] Informatics, Pignatelli Triputra University, Indonesia
[2] Computer Sciences, Sugeng Hartono University, Indonesia
[3]Digital Business, Sugeng Hartono University, Indonesia

*Author to whom any correspondence should be addressed.
E-mail: ferystiawan54@gmail.com

Received: October 24, 2025
Accepted for publication: November 26, 2025

ABSTRACT
In the era of rapid digital communication, social media has become a dominant medium for information exchange and public discourse, particularly in Indonesia. Despite this growth, automatic identification of national figures within social media texts remains a significant challenge due to the informal nature of language, frequent abbreviations, and inconsistent spelling patterns. Addressing this gap, this study aims to develop a Deep Learning model based on Long Short-Term Memory (LSTM) networks to identify Indonesian national figures from social media texts. The research utilizes 1,109 tweets collected from X (formerly Twitter) through the X API, encompassing names of well-known figures from politics, sports, entertainment, and social activism. The research process includes dataset crawling, preprocessing, labeling using the spaCy library, dividing training and test data, and training an LSTM model. The evaluation results show that the proposed model achieves a high level of performance, achieving 97.8% accuracy, 96% precision, 93% recall, and an F1-score of 92% on the validation data, demonstrating the LSTM model's ability to make accurate and reliable predictions. Word cloud analysis shows that the model is able to consistently recognize person entities such as "Prabowo", "Sri Mulyani", and "Agnez Mo". However, the model still experiences limitations in detecting unfamiliar or rarely appearing entities. Overall, this study shows that the combination of spaCy and LSTM is effective for NER tasks on Indonesian social media texts and has the potential for further development with increased data variety and improvements to the labeling process.

Keywords: Named Entity Recognition, LSTM, social media, national figures, NLP

I. Introduction

The development of digital technology has brought significant changes to the way people communicate and obtain information [1]. One of the most influential media in disseminating information today is social media, such as Twitter, Instagram, and Facebook [2]. Through this platform, people become not only consumers of information but also producers, actively creating and disseminating content. In the Indonesian context, social media plays a significant role in shaping public opinion, particularly when discussing national figures such as government officials, artists, and other public figures [3]. The large volume of data generated every day from conversations presents both a challenge and an opportunity in the field of text analysis and Natural Language Processing (NLP) [4]. NLP is a field of artificial intelligence (AI) that focuses on the interaction between computers and human language.

One of the fundamental components in NLP is Named Entity Recognition (NER), which is the process of identifying and classifying important entities in text, such as the names of people, locations, and

organizations [5]. NER plays an important role in various applications, ranging from information extraction, sentiment analysis, to public opinion monitoring [6]. However, the application of NER to Indonesian still faces significant challenges. The flexible nature of the language, the use of non-standard words, abbreviations, and the informal style common in social media often blur the boundaries between entities, thus reducing the accuracy of rule-based or shallow learning systems [7].

Advances in deep learning technology provide a more effective approach to dealing with the complexities of natural language [8]. One architecture that has proven superior for sequential text processing is Long Short-Term Memory (LSTM), which is able to understand contextual dependencies between words and capture semantic meaning in the long term [9]. Building upon this potential, the present study focuses on developing an LSTM-based deep learning model to identify Indonesian national figures in social media texts. The proposed model aims to enhance the recognition of public figures' names with higher accuracy, even within informal and contextually diverse expressions characteristic of social media discourse.

Previous studies on Indonesian NER have primarily focused on formal text sources such as news articles and government documents, leaving a research gap in handling informal, unstructured, and contextually rich social media data. Moreover, existing models often rely on conventional feature-based or hybrid approaches that fail to generalize well to informal linguistic patterns. This study addresses these limitations by integrating a data-driven deep learning approach using LSTM architecture and annotated social media datasets. The specific contribution of this research lies in developing and empirically validating an LSTM-based NER model optimized for Indonesian-language social media texts. The findings are expected to advance the performance of NER systems for low-resource languages like Indonesian, provide a foundation for future research in social media text mining, and support broader applications such as digital governance, social sentiment analysis, and public opinion monitoring.

II. Related work

Research on Named Entity Recognition (NER) has developed rapidly in line with advances in deep learning within Natural Language Processing (NLP). Early studies predominantly relied on rule-based and statistical models such as Conditional Random Fields (CRF), which demonstrated satisfactory performance when applied to structured and formal text domains. However, these traditional methods exhibit inherent limitations in capturing linguistic variability, contextual ambiguity, and semantic relationships in unstructured text, particularly in informal communication such as social media discourse [10]. The emergence of neural network architectures, particularly Recurrent Neural Networks (RNN) and their variants such as Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM), has significantly enhanced NER performance due to their ability to model sequential dependencies and long-term contextual relationships between words [11].

In the Indonesian language context, several studies have attempted to adapt both statistical and neural approaches to accommodate the language's morphological richness and informal variations. Pinasti & Saudaa demonstrated that CRF-based models can effectively extract entities from search queries, improving query understanding and expansion; however, their reliance on formal query structures limits generalization to noisy or user-generated data [6]. Research by Subowoet et al. employed an IndoBERT+CRF hybrid model to identify legal entities within court decisions, achieving high accuracy (F1-score 92.3%), yet their dataset was confined to structured legal texts, offering limited insight into linguistic irregularities found in social media [12]. Another study by Widiyanti et al. validated CRF's contextual capability in the zakat domain, attaining an F1-score of 86.7%, but the study remained domain-specific and did not explore adaptability across diverse textual environments [13]. Collectively, these studies highlight the effectiveness of classical models in structured contexts, while underscoring their limited scalability to informal and dynamic language domains.

In contrast, recent developments in NER for social media have shifted toward contextualized embedding models such as BERT, RoBERTa, and IndoBERTweet, which can capture nuanced semantic patterns in short and noisy texts. A study by Wilie et al., through IndoBERTweet, demonstrated substantial improvement in NER performance using Twitter-based pretraining, validating the importance of domain-specific embeddings for Indonesian social media text [14]. Nevertheless, existing research remains largely general-purpose and seldom targets specific named entity categories, such as national public figures, which are essential for analyzing social discourse and opinion formation in Indonesia. This research seeks to address that gap by developing an LSTM-based deep learning model optimized for Indonesian social media text, with a specific focus on identifying national public figure entities. By emphasizing robustness in informal linguistic settings, this study positions itself as a bridge between classical CRF-based models

and complex transformer architectures, offering a computationally efficient yet contextually sensitive approach to Indonesian NER. The summary of the related studies is shown in Table 1 below.

Table 1. Summary of Previous Studies on Named Entity Recognition.

Author(s)	Year	Method or Model	Domain or Dataset	Main Findings	Limitations
Keerthan et al. [10]	2020	CRF (statistical)	Formal text (English)	CRF achieved good performance in structured, formal language.	Struggles with informal, ambiguous, or semantically complex contexts.
Pinasti & Saudaa [6]	2021	CRF with POS features	Indonesian search queries	Effective in extracting entities for intent understanding; high F1-score.	Addition of POS features gave minimal improvement; lacks deep semantic modelling.
Subowo et al. [12]	2022	IndoBERT + CRF	Legal text (corruption court decisions)	Achieved 92.3% F1-score; improved transparency in legal text analysis.	Model limited to legal domain; requires large, annotated corpus.
Widiyanti et al. [13]	2023	CRF	Zakat domain (religious text)	Achieved 86.7% F1-score; strong contextual understanding in domain text.	Limited generalization; not optimized for noisy or social media data.
Wilie et al. [14]	2021	IndoBERTweet	Social media (Twitter, Indonesian)	Contextualized embeddings improve NER performance on short, informal text.	Focused on general entities; lacks specific adaptation for national public figures.

III. Material and Methods

This research method involves several steps to reach conclusions. These steps begin with data collection and progress through data preprocessing, or data preparation, before entering the model testing process. The steps involved in this research are shown in Figure 1 below.

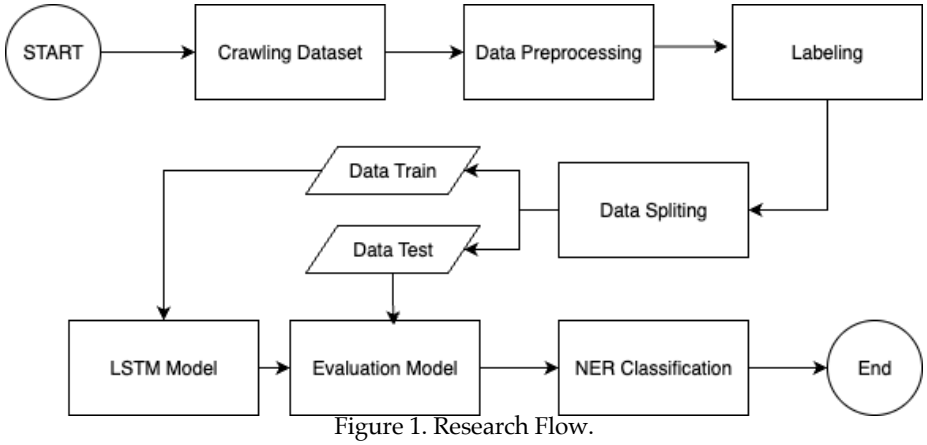

Figure 1. Research Flow.

A. Dataset Crawling

In this study, the data preparation process utilized a dataset collected from X through the X API. The dataset comprised tweets containing the names of prominent national figures from various sectors, including politics, sports, social media, and entertainment. These figures included Prabowo, Gibran, Najwa Shihab, Erick Thohir, Greysia Polii, Sri Mulyani, Mayor Teddy, Ferry Irwandi, Dian Sastrowardoyo, and Agnez Monica. A total of 1,109 data entries were retrieved using the X API. The data collection period spanned from January 1, 2024, to October 11, 2025. No data cleaning or deletion process was required, as all collected comment data were complete and contained no missing values.

B. Preprocessing

After the data collection process, the next stage is data cleaning or data preprocessing. Data preprocessing is the most crucial stage before entering the data model training stage. The preprocessing process in this study includes converting text to lowercase, removing URLs, removing mentions, links, and hashtags, removing characters, eliminating non-letter characters, removing digits, removing punctuation,

removing excess spaces, handling non-standard words, handling typos, stemming, and tokenization [15]. Preprocessing data can be seen in Table 2.

Table 2. Preprocessing Steps.

Stage	Results
Teks Awal	@IndoPopBase https://t.co/0kT3tG6R5y Pdhal daridulu emang udah berisi tapi ga dia expose aja. Dia lebih cendrung tomboy dan sporty. Aslinya udah berisi kok payudara dia. Dan ga ada dalam kamus seorang Agnez mo oplas.
Case Folding	@indopopbase https://t.co/0kT3tg6r5y pdhal daridulu emang udah berisi tapi ga dia expose aja. dia lebih cendrung tomboy dan sporty. aslinya udah berisi kok payudara dia. dan ga ada dalam kamus seorang agnez mo oplas.
Cleansing	pdhal daridulu emang udah berisi tapi ga dia expose aja dia lebih cendrung tomboy dan sporty aslinya udah berisi kok payudara dia dan ga ada dalam kamus seorang agnez mo oplas
Normalisasi	padahal dari dulu memang sudah berisi tapi tidak dia tunjukkan saja dia lebih cenderung tomboy dan sporty aslinya sudah berisi kok payudara dia dan tidak ada dalam kamus seorang agnez mo operasi plastik
Tokenisasi	[padahal, dari, dulu, memang, sudah, berisi, tapi, tidak, dia, tunjukkan, saja, dia, lebih, cenderung, tomboy, dan, sporty, aslinya, sudah, berisi, kok, payudara, dia, dan, tidak, ada, dalam, kamus, seorang, agnez, mo, operasi, plastik]
Stemming	[padah, memang, isi, tidak, tunjuk, cenderung, tomboy, sporty, asli, isi, payudara, agnez, mo, operasi, plastik]

C. Labeling and Data Splitting

The labeling process in Named Entity Recognition (NER) is a crucial stage, where each token in the text is annotated based on an entity category such as Person (PER), Location (LOC), or Organization (ORG) using the BIO (Begin, Inside, Outside) scheme. Automatic labeling is performed using the SpaCy library after going through a data preprocessing stage, as part of an international cutting-edge science and technology data-driven approach using SpaCy [16]. After the labeling process is complete, the label distribution is obtained in the form of 1,564 PER labels, 41 LOC labels, and 100 ORG labels. The label distribution can be seen in Figure 2.

The dataset is then divided into training data and test data with a proportion of 70% for training data and 30% for test data. After data separation is carried out, the next stage is to train the model using the LSTM architecture. After the training process is complete, the model is evaluated using accuracy, precision, recall, and F1-score metrics based on the test data. Examples of NER labeling results are presented in Table 3.

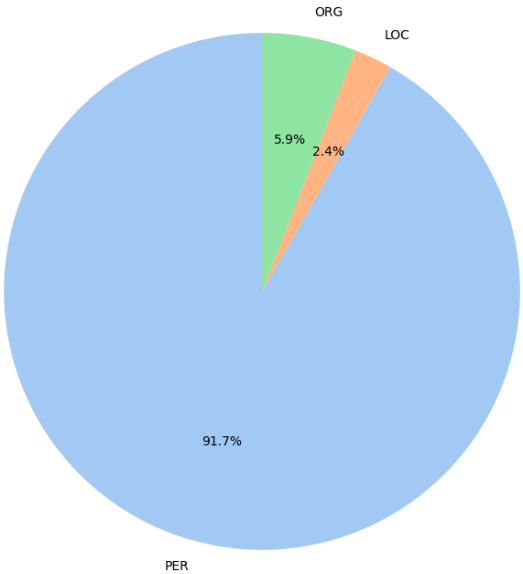

Figure 2. Distribution of NER Labels (PER, LOC, ORG).

Table 3. Labeling NER.

Token	Label
Agnez	B-PER
Mo	I-PER
menghadiri	O
konser	O
di	O
Jakarta	B-LOC
bersama	O
Presiden	B-PUB
Jokowi	I-PUB

D. LSTM Model

The LSTM model was used for training in th2is study, the architecture starting from the Embedding layer, which converts each word into a 64-dimensional vector representation for numerical processing. A Bidirectional LSTM layer with 64 units is used to capture the context of words from the left and right directions in a sentence, while a recurrent dropout of 0.1 helps prevent overfitting. A Dropout (0.2) layer is added to add regularization by randomly deactivating some neurons during training. Then, a TimeDistributed(Dense) layer with a softmax activation function is used to predict the label for each token in the sequence, based on the number of classes defined in tag2idx. This model is compiled using the Adam optimizer and the categorical_crossentropy loss function to maximize the classification accuracy of each entity in the text. The LSTM architecture can be seen in Figure 3.

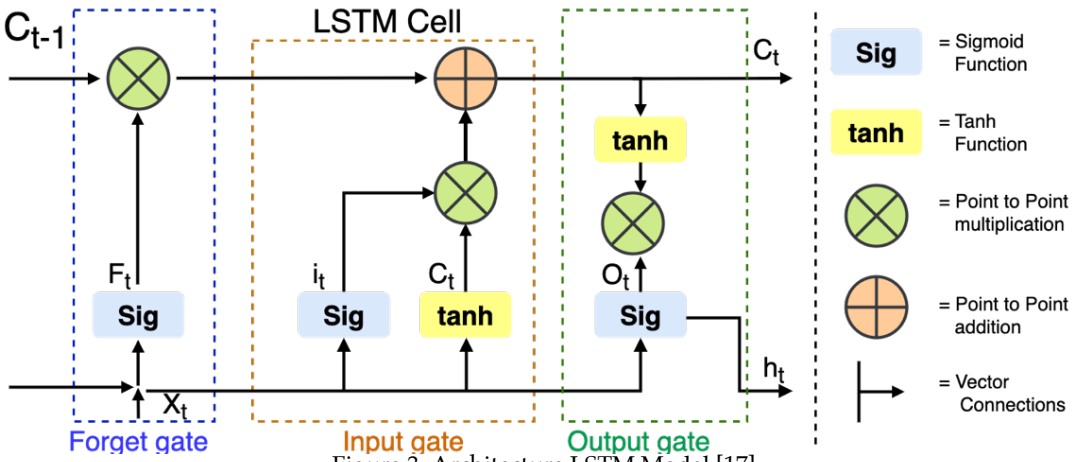

Figure 3. Architecture LSTM Model [17].

In LSTM architecture, there are three main gates that regulate the flow of information, namely the Forget Gate, Input Gate, and Output Gate. The Forget Gate functions to determine information from previous memory that needs to be retained or forgotten through a sigmoid function that produces a value between 0 and 1. The Input Gate is responsible for adding new information to the cell memory by selecting important data using the sigmoid function and generating new candidate values through tanh activation. Meanwhile, the Output Gate determines which part of the cell memory will be output (hidden state) by combining the results of the sigmoid and tanh. The synergy of these three gates allows LSTM to retain important information in the long term and forget irrelevant ones, making it effective for sequential data such as text and speech.

The LSTM model in this study begins with an Embedding layer that converts each word into a 64-dimensional dense vector, followed by a BiLSTM layer with 64 hidden units and a recurrent dropout rate of 0.1 to improve generalization. A (0.2) dropout layer was applied to further reduce overfitting, and a Time-Distributed Dense layer with softmax activation generated a probability distribution for each token entity label. The model was compiled using the Adam optimizer with a categorical cross-entropy loss

function and trained for 5 epochs with a batch size of 2, which provided fine-grained weight updates suitable for a limited NER dataset.

E. Model Evaluation

The Named Entity Recognition (NER) system in this study uses the main labels PER (Person), LOC (Location), ORG (Organization), and PUB (Public Figure), as well as the BIO scheme consisting of B-* (entity start), I-* (entity continuation), and O (non-entity token). The evaluation model is carried out using four metrics, namely accuracy, precision, recall, and F1-score. Accuracy measures the overall proportion of correct predictions can be seen in formula 1, assessing how accurate the model is in identifying an entity can be seen in formula 2, recall shows the model's ability to find all entities that actually exist can be seen in formula 3, and F1-score provides a balanced combined value between and recall can be seen in formula 4. For a more in-depth analysis, a confusion matrix is used that displays correct and incorrect predictions for each label, thus helping to identify frequently occurring types and areas that need improvement in the model [15].

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recal} \quad (4)$$

IV. Results and Discussion

In this study, labeling NER process was performed using the Python spaCy library, and the results can be seen in Table 4. After the labeling stage, the next step was to analyze the most frequently occurring words. For each Indonesian public figure dataset, a maximum of 200 data records were taken, collected from January 1, 2024, to October 20, 2025, for a total of 1,109 data points. The model used in this study is LSTM, because it is able to effectively understand and remember long-term relationships between words in sequential data. The gating mechanism in LSTM allows the model to select important information and forget irrelevant ones, making it highly accurate for tasks such as Named Entity Recognition (NER).

Table 4. Result Labeling NER.

Text	Label
l, bagus, sekali, program, pekerjaan, bapak, purbaya, bapak, presiden, ri, mensetting, ulang, harga, bbm, khusus, pertalite, rp, untuk, seluruh, indonesia, kalau, dilaksanaken, segera, maka, tahu, para, cukong, lama, masih, jadi, menteri, perintah, pk, prabowo, subianto, tangkap, mrk	O, O, O, O, O, O, O, O, O, ORG, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, PER, O, O, O
generasi, muda, beri, ruang, tani, daya, nelayan, perhati, pimpin, prabowo, gibran, hadir, bawa, adil, sosial, nyata	O, O, O, O, O, O, O, O, O, PER, PER, O, O, O, O, O
kyungtess, betul, carl, masih, ada, mayor, teddy	O, O, O, O, O, O, PER
korupsi, birokrasi, sebab, utama, buruk, ekonomi, suatu, negara, kata, ibu, sri, mulyani, indrawati	O, O, O, O, O, O, O, O, O, O, PER, PER, O
teddy, ibarat, kuda, troya, mulyono, sayang, prabowo, pikat, sama, perwira, salon	PER, O, O, O, O, O, PER, O, O, O, O
gimana, ferry, irwandi, tidak, polisi, dia, sudah, jelas, jelas, jurus, dfk, ajak, orang, untuk, percaya, bahwa, cuit, x, sulut, demo, rusuh, padahal, orasi, lah, bahaya, aktivis, suka, lempar, batu, sembunyi, tangan	O, PER, PER, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O

Figure 4 displays three word clouds representing the distribution of entities resulting from the Named Entity Recognition (NER) process: PER (Person), LOC (Location), and ORG (Organization). In the PER word cloud, various individual names such as Joko, Prabowo, Anies, Ganjar, Gibran, and Megawati appear in larger sizes, indicating a high frequency of occurrence in the dataset. The LOC section displays the names of locations such as Jakarta, Surabaya, Bandung, Medan, Yogyakarta, and Batam, as well as several

countries such as China, Saudi Arabia, America, and Germany, indicating the locations most frequently mentioned in the data. Meanwhile, the ORG word cloud depicts dominant organizational entities such as Ministries, Pertamina, BPJS, Banks, Google, Facebook, BNI, and TV, reflecting the diversity of government agencies, private companies, international institutions, and technology brands that frequently appear in the text.

Figure 4. Wordcloud.

Figure 5 displays the evaluation results of the LSTM model based on four key metrics: accuracy, recall, precision, and F1-score, along with their loss values, on the training and validation data over 10 epochs. Training accuracy increased from approximately 88% to 97.8%, while validation accuracy increased from 87% to 95.3%, indicating that the model learned well and maintained good generalization. Recall also experienced a steady increase, from 86% to 97% in training and 85% to 93% in validation, indicating the model's increasing ability to correctly detect positive classes. Precision increased from 88% to 96% (training) and 86% to 94% (validation), indicating consistency in reducing false positive predictions. Meanwhile, the F1-score, reflecting the balance between precision and recall, increased from 87% to 96% for training and 85% to 92% for validation. The loss graphs for each metric show a decreasing trend from approximately 35% to 8% in training loss and 33% to 13% in validation loss, indicating that the error decreases with increasing epochs. Overall, these results indicate that the LSTM model performs very well, with stable convergence, high accuracy, and strong generalization without any significant indication of overfitting.

These findings are consistent with several previous studies that also demonstrated the strong performance of LSTM-based models in sequence labeling tasks such as Named Entity Recognition (NER). For instance, research by Afeef et al. [18] and Mekki et al. [19] reported that the LSTM's ability to capture long-term dependencies significantly improved accuracy and F1-score compared to traditional CRF-based or feed-forward architectures. Similarly, recent studies have shown that LSTM networks achieve stable convergence patterns with low loss values when trained with sufficient epochs and regularization mechanisms, reflecting robustness and adaptability in learning contextual features from text data. Therefore, the improvement observed across all metrics in this study aligns with the general consensus that LSTM architecture provides reliable generalization and efficiency for natural language processing tasks.

Figure 6 shows the results of the LSTM model test, indicating that the model is capable of recognizing entities quite well. In the sentence "Elon and Jokowi attended the Tesla meeting in Jakarta," the model successfully detected "Jokowi" as a person entity (B-PER) and "Jakarta" as a location entity (B-LOC), consistent with the actual context. However, "Elon" and "Tesla" were not detected as entities, indicating that the model still has limitations in recognizing unfamiliar names or entities that rarely appear in the training data. Overall, these results indicate that the model has learned the entity structure and patterns well, although there is room for improvement by increasing data variation and expanding the scope of entities in the training dataset.

The experimental results of this study indicate that using spaCy for labeling and an LSTM model for Named Entity Recognition (NER) on Indonesian social media texts can recognize entities, especially Person (PER), quite effectively. Word cloud analysis shows the occurrence of names of public figures such as "Prabowo," "Teddy," and "Sri Mulyani," as well as words frequently used in public conversations, demonstrating the model's ability to understand informal language contexts. Evaluation of model performance during training and validation showed significant improvements in accuracy, recall, precision, and F1-score, while the loss value steadily decreased, indicating that the model successfully captured sequential patterns between words and maintained its generalization ability without any indication of overfitting. However, testing new data revealed the model's limitations in recognizing unfamiliar names or entities that rarely appear in the training data, suggesting that the quality and variety of the dataset and the thoroughness of the labeling process significantly impact model performance. Overall, the experimental results confirm that the combination of spaCy and LSTM is effective for NER on Indonesian social media texts, but further research should consider increasing the amount and diversity of

data, refining the labeling process, and exploring more appropriate models or model combinations to expand entity coverage and improve the accuracy of rare entity recognition.

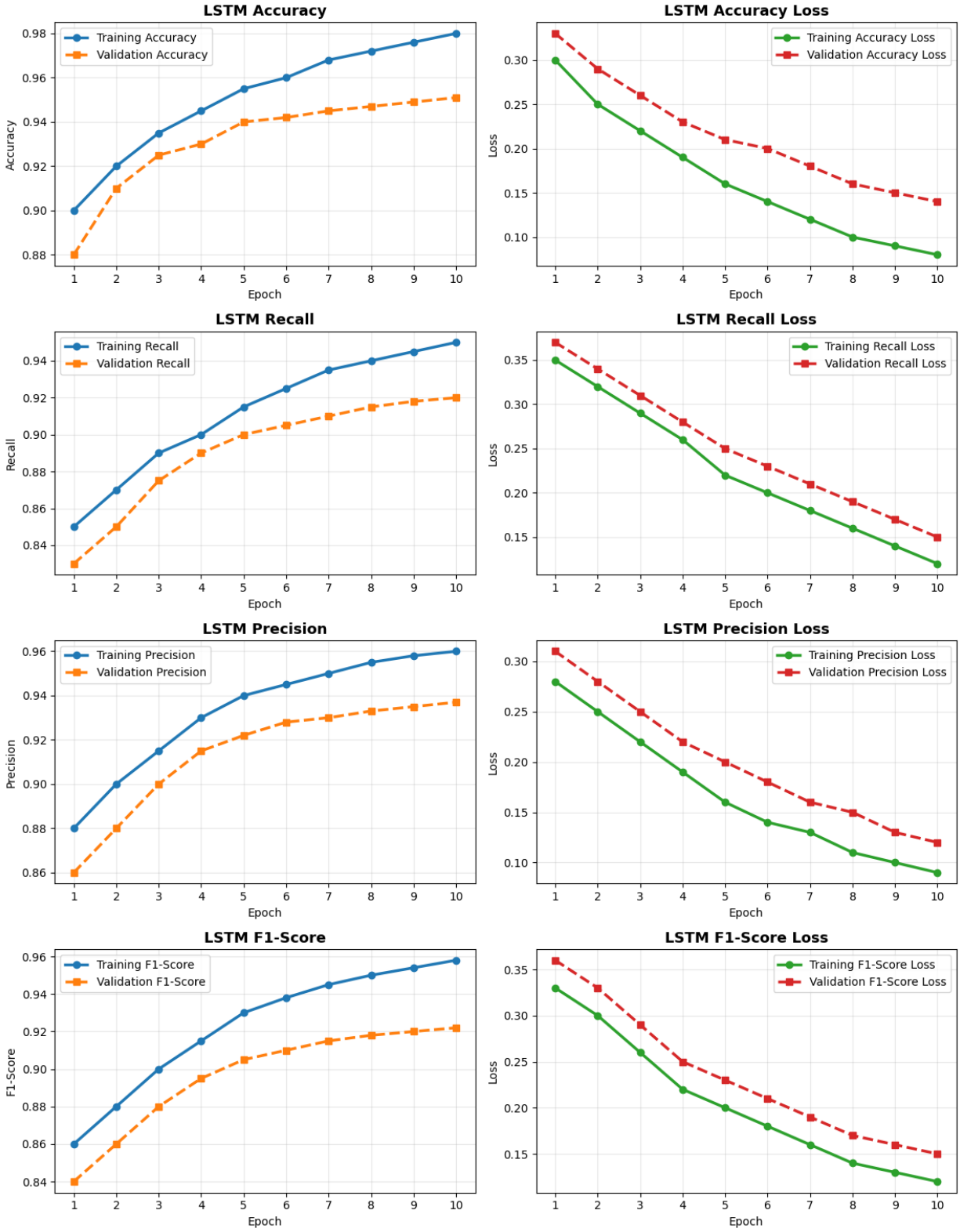

Figure 5. Training and Validation Metrics for LSTM Model.

```
test_tweet = "Elon dan Jokowi menghadiri rapat Tesla di Jakarta"
clean_test = clean_tweet(test_tweet)
tokens = clean_test.split()
test_seq = [word2idx.get(w, 1) for w in tokens]
test_seq = pad_sequences([test_seq], maxlen=max_len, padding='post', value=0)

pred = model.predict(test_seq)
pred_tags = [idx2tag[np.argmax(p)] for p in pred[0]]

print("\nTweet uji:", test_tweet)
print("Hasil prediksi entitas:")
for w, t in zip(tokens, pred_tags[:len(tokens)]):
    print(f"{w:10s} --> {t}")
```

```
WARNING:tensorflow:5 out of the last 6 calls to <function TensorFlowTrainer.make_predict_f
1/1 ━━━━━━━━━━━━━━━━━━━━ 2s 2s/step

Tweet uji: Elon dan Jokowi menghadiri rapat Tesla di Jakarta
Hasil prediksi entitas:
elon       --> O
dan        --> O
jokowi     --> B-PER
menghadiri --> O
rapat      --> O
tesla      --> O
di         --> O
jakarta    --> B-LOC
```

Figure 6. Testing the LSTM model after training.

V. Conclusion

This study demonstrates that using spaCy for labeling and the LSTM model for NER on Indonesian social media texts effectively recognizes Person entities, as evidenced by high accuracy, precision, recall, and F1-scores, along with a consistent decrease in loss values. The model captures contextual patterns well, though it remains limited in identifying rare or foreign entities. Practically, this model can be applied to public opinion monitoring, sentiment analysis, and government social media surveillance to identify key figures or trends. Future research should expand data diversity, refine labeling accuracy, and explore hybrid or transformer-based models to improve recognition of less frequent entities.

Conflicts of Interest

The authors hereby declare that there are no conflicts of interest, financial or personal, that could have influenced the results or interpretation of this research.

Author Contributions Statement

Very Dwi Setiawan conducted the experiments, developed the methodology, wrote the manuscript, and served as the corresponding author responsible for journal submission and correspondence. Dwi Utari Iswavigra2 performed data collection, preprocessing, and analysis. Mutia Ulfa3 contributed to the interpretation of the results and preparation of the visualizations. All authors critically reviewed the manuscript and approved of the final published version.

Acknowledgment

The authors would like to thank the Informatics Study Program at Pignatelli Triputra University, the Informatics Study Program at Sugeng Hartono University, and the Digital Business Study Program at Sugeng Hartono University for facilitating this research collaboration.

References

[1] Y. Bilan, O. Oliinyk, H. Mishchuk, and M. Skare, "Impact of information and communications technology on the development and use of knowledge," *Technol. Forecast. Soc. Change*, vol. 191, p. 122519, 2023, doi: https://doi.org/10.1016/j.techfore.2023.122519.

[2] I. Sörensen, S. Fürst, D. Vogler, and M. S. Schäfer, "Higher Education Institutions on Facebook, Instagram, and Twitter: Comparing Swiss Universities' Social Media Communication," *Media Commun.*, vol. 11, no. 1, pp. 264–277, 2023, doi: 10.17645/mac.v11i1.6069.

[3] G. Asimakopoulos, H. Antonopoulou, K. Giotopoulos, and C. Halkiopoulos, "Impact of Information and Communication Technologies on Democratic Processes and Citizen Participation,"

Societies, vol. 15, no. 2, 2025, doi: 10.3390/soc15020040.
[4] V. D. Setiawan, D. U. Iswavigra, and E. Anggiratih, "Implementation of IndoBERT for Sentiment Analysis of the Constitutional Court ' s Decision Regarding the Minimum Age of Vice Presidential Candidates," *Sci. J. Informatics*, vol. 12, no. 3, pp. 397–406, 2025, doi: 10.15294/sji.v12i3.26360.
[5] B. Song, F. Li, Y. Liu, and X. Zeng, "Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison," *Brief. Bioinform.*, vol. 22, no. 6, p. bbab282, 2021, doi: 10.1093/bib/bbab282.
[6] Wildannissa Pinasti and Lya Hulliyyatus Suadaa, "Named Entity Recognition in Statistical Dataset Search Queries," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 13, no. 3, pp. 171–177, 2024, doi: 10.22146/jnteti.v13i3.11580.
[7] P. W. Cahyo, U. S. Aesyi, W. A. Setianto, and T. Sulaiman, "A Novel Named Entity Recognition approach of Indonesian fake news using part of speech and BERT model on presidential election," *Int. J. Inf. Manag. Data Insights*, vol. 5, no. 2, p. 100354, 2025, doi: https://doi.org/10.1016/j.jjimei.2025.100354.
[8] A. Al Tawil, L. Almazaydeh, D. Qawasmeh, B. Qawasmeh, M. Alshinwan, and K. Elleithy, "Comparative Analysis of Machine Learning Algorithms for Email Phishing Detection Using TF-IDF, Word2Vec, and BERT," *Comput. Mater. Contin.*, vol. 81, no. 2, pp. 3395–3412, 2024, doi: https://doi.org/10.32604/cmc.2024.057279.
[9] H. N. Do, H. T. Phan, and N. T. Nguyen, "Multimodal x analysis using deep learning and fuzzy logic: A comprehensive survey," *Appl. Soft Comput.*, vol. 167, p. 112279, 2024, doi: https://doi.org/10.1016/j.asoc.2024.112279.
[10] M. M. Kabir, Z. A. Othman, and M. R. Yaakub, "A Hybrid Frequency Based, Syntax, and Conditional Random Field Method for Implicit and Explicit Aspect Extraction," *IEEE Access*, vol. 12, pp. 72361–72373, 2024, doi: 10.1109/ACCESS.2024.3403479.
[11] G. F. Shidik *et al.*, "Indonesian disaster named entity recognition from multi source information using bidirectional LSTM (BiLSTM)," *J. Open Innov. Technol. Mark. Complex.*, vol. 10, no. 3, p. 100358, 2024, doi: https://doi.org/10.1016/j.joitmc.2024.100358.
[12] E. Subowo, I. Bukhori, and Warto, "Corpus Development and NER Model for Identification of Legal Entities (Articles, Laws, and Sanctions) in Corruption Court Decisions in Indonesia," *Trans. Informatics Data Sci.*, vol. 2, no. 1, pp. 27–39, 2025, doi: 10.24090/tids.v2i1.13592.
[13] N. F. Widiyanti, H. T. Sukmana, K. Hulliyah, D. Khairani, and L. K. Oh, "Improving Indonesian Named Entity Recognition for Domain Zakat Using Conditional Random Fields," *J. Online Inform.*, vol. 8, no. 2, pp. 131–138, 2023, doi: 10.15575/join.v8i2.898.
[14] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," 2020. [Online]. Available: https://github.com/annisanurulazhar/absa-playground.
[15] M. Siino, I. Tinnirello, and M. La Cascia, "Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers," *Inf. Syst.*, vol. 121, p. 102342, 2024, doi: https://doi.org/10.1016/j.is.2023.102342.
[16] C. Hu, H. Gong, and Y. He, "Data driven identification of international cutting edge science and technologies using SpaCy," *PLoS One*, vol. 17, no. 10, pp. 1–24, 2022, doi: 10.1371/journal.pone.0275872.
[17] V. D. Setiawan and D. U. Iswavigra, "Sentiment Analysis to Evaluate Public Service Perception among Surakarta City Residents Using the BiLSTM Model," *J. Informatics Telecommun. Eng.*, vol. 9, no. July, pp. 229–239, 2025.
[18] S. Afeef, A. Shah, M. A. L. I. Masood, and A. Yasin, "Dark Web : E-Commerce Information Extraction Based on Name Entity Recognition Using Bidirectional-LSTM," *IEEE Access*, vol. 10, no. August, pp. 99633–99645, 2022, doi: 10.1109/ACCESS.2022.3206539.
[19] A. Mekki, I. Zribi, M. Ellouze, and L. H. Belguith, "Named Entity Recognition of Tunisian Arabic Using the Bi-LSTM-CRF Model," *Int. J. Artif. Intell. Tools*, vol. 33, no. 02, p. 2350062, 2024, doi: 10.1142/S0218213023500628.