Improving SpaCy-based NER Accuracy using Conditional Random Fields (CRF) as Advanced Features for Javanese Legends

Kevin Dwi Mahendra[1*], Danang Arbian Sulistyo[2]

[1,2] *Institut Teknologi dan Bisnis Malang, Jl. Soekarno Hatta, Rembusari 1A, Malang, Jawa Timur, Indonesia*

Keywords

Named Entity Recognition (NER); Conditional Random Fields (CRF); SpaCy; Javanese Legends

****Correspondence Email:***
danangarbian@email.com

Abstract

Named Entity Recognition (NER) is a crucial task for information extraction, particularly for preserving the rich cultural data within Javanese legends. However, standard NER frameworks like SpaCy can face limitations when processing languages with unique linguistic characteristics. This research addresses this gap by exploring the effectiveness of integrating Conditional Random Fields (CRF) as an advanced feature extraction layer to enhance a SpaCy-based NER model. The proposed hybrid model leverages CRF's strength in sequence labeling to improve the contextual understanding of entities within Javanese narratives. Experimental results demonstrate a significant performance increase, with the model achieving a precision of 0.8923, recall of 0.8678, and an overall F1-score of 0.8803. This study confirms that augmenting SpaCy with a CRF layer provides a robust solution for improving NER accuracy on Javanese texts. Future work could involve incorporating more complex contextual embeddings or applying this model to other genres of traditional Indonesian literature to further validate its effectiveness and adaptability.

1. Introduction

In the era of digital information, the automated extraction of meaningful data from unstructured text has become a cornerstone of Natural Language Processing (NLP). A key subfield, Named Entity Recognition (NER), focuses on identifying and classifying named entities such as persons, locations, and organizations within text(Sulistyo, Wibawa, Prasetya, & Ahda, 2025b). This technology is not only pivotal for applications such as machine translation and information retrieval but also holds immense value for the digital preservation and analysis of cultural heritage(Sulistyo, Wibawa, Prasetya, & Ahda, 2025a).

Despite the power of state-of-the-art NER frameworks like SpaCy, their performance often diminishes when applied to languages with unique linguistic characteristics, such as Javanese. These texts can feature complex sentence structures, varying orthography, and context-dependent entity definitions that challenge standard models primarily trained on more widely-spoken languages. This performance gap highlights a critical need for specialized models tailored to the nuances of Javanese language and literature(Yanti et al., 2021). The motivation for this study is to bridge this gap by developing a more accurate and context-aware NER model for Javanese legends, thereby facilitating deeper computational analysis of this important cultural heritage.

To address these challenges, this paper proposes a hybrid approach that enhances the SpaCy framework by integrating Conditional Random Fields (CRF) as an advanced feature extraction layer. The purpose of this study is to investigate whether leveraging CRF's strength in sequential pattern recognition can significantly improve the accuracy of entity identification in Javanese narratives. Our methodology involves using SpaCy for foundational tokenization and feature engineering, then feeding these features into a CRF layer to model the dependencies between adjacent labels in a sequence. This allows the model to make more informed predictions by considering the entire sentence context(Holla et al., 2019).

The primary contribution of this research is the empirical demonstration that this hybrid SpaCy-CRF model substantially improves NER performance on Javanese legends. Our experiments show that the proposed model achieves a high F1-score of 0. 0.8803, with a precision of 0.8923 and a recall of 0.8678. These findings confirm that augmenting a robust framework like SpaCy with a sequence-labeling model like CRF is an effective strategy for handling the linguistic complexities of Javanese text. This paper details the model's architecture, the dataset preparation process, and a comprehensive analysis of the results, concluding with a discussion on the implications for future research in NLP for other local languages.

1.1 Literature Review

This chapter reviews the existing body of research relevant to Named Entity Recognition (NER), focusing on the evolution of its methodologies, the application of prominent frameworks, and the specific challenges encountered when applying these technologies to low-resource languages like Javanese. The goal is to critically evaluate prior work to identify the gap that motivates the present study.

1.1.1 Methodologies in Named Entity Recognition

Named Entity Recognition has been a central task in Natural Language Processing for decades, with methodologies evolving from manually crafted rules to sophisticated deep learning models. Early systems relied on rule-based and gazetteer-based approaches, which utilized handcrafted linguistic rules and extensive dictionaries to identify entities(Li et al., 2020).While effective in constrained domains, these systems suffer from poor scalability and brittleness, as they fail to generalize to new or unseen text and require significant manual effort to create and maintain.

The limitations of rule-based systems led to the dominance of statistical machine learning models. Among these, Conditional Random Fields (CRF) emerged as a particularly powerful method for sequence labeling tasks like NER. Unlike models that classify tokens independently, CRF considers the context of the entire sequence, modeling the dependencies between adjacent labels. This ability to capture contextual information makes CRF highly effective at resolving ambiguity and improving label consistency, which is why it has remained a foundational component in many NER systems.

1.1.2 The Role of NLP Frameworks: SpaCy

To democratize access to NLP technologies, open-source libraries like SpaCy have become indispensable. SpaCy is an industrial-strength framework known for its speed, efficiency, and production-ready models. Its statistical NER component provides a strong baseline that is easy to train and deploy. However, a critical evaluation reveals that SpaCy default architecture, while robust, is optimized for general-purpose tasks and well-resourced languages. When applied to niche domains or languages with unique grammatical structures, its performance can be suboptimal. The framework's reliance on localized context windows for feature generation may not fully capture the sequential dependencies that are critical for accuracy in complex narratives, a limitation that has been noted in various specialized applications(Keraghel et al., 2024)

1.1.3 NER for Low-Resource Languages and Javanese

The majority of NER research has concentrated on high-resource languages like English, Chinese, and German. Applying NER to low-resource languages, including many regional languages of Indonesia like Javanese, presents distinct challenges. These include the lack of large, publicly available annotated corpora, morphological complexity, and inconsistent orthography(Sulistyo, Prasetya, et al., 2025). Previous studies on NER for Bahasa Indonesia have often focused on formal domains like news articles, adapting existing models with some success(Karo et al., 2025).

However, research specifically targeting Javanese, and particularly the literary domain of legends, remains critically underdeveloped. Javanese legends possess a unique vocabulary, metaphorical expressions, and narrative structures that are not present in contemporary or formal text. Critically, no existing model has been specifically optimized to handle these nuances, creating a significant gap in the digital humanities and NLP landscape for Indonesian cultural heritage(Sulistyo, Wibawa, Prasetya, Ahda, et al., 2025).

1.1.4 Synthesis and Identified Research Gap

Synthesizing the literature reveals a clear tension. On one hand, modern NLP has powerful, accessible frameworks like SpaCy and highly accurate but resource-intensive deep learning models. On the other hand, there is a pressing need for effective NLP tools for low-resource, culturally significant languages like Javanese. A simple application of a tool like SpaCy is unlikely to yield optimal results due to its generalized architecture. Conversely, training a large transformer model from scratch is often infeasible due to data and resource constraints(Gurgurov et al., 2024).

This leads to the central research gap this study aims to fill: There is a need for a practical and effective method to improve NER accuracy for Javanese legends that leverages the strengths of existing tools while mitigating their weaknesses. This study posits that a hybrid model one that combines the efficient feature engineering of SpaCy with the superior sequential modeling capabilities of CRF offers a balanced and powerful solution. By integrating CRF as an advanced feature layer, we hypothesize that the model can better capture the contextual dependencies inherent in Javanese narratives, thereby overcoming the limitations of using SpaCy in isolation and providing a significant improvement in NER performance for this specific, under-researched domain.

2. Research Methods

This section outlines the systematic procedures followed in this study, from data acquisition and preparation to model training and evaluation. The methodology is designed as a quantitative experimental study to rigorously assess the performance of the proposed Named Entity Recognition (NER) model. This ensures that the research is replicable and the findings are empirically verifiable.

2.1 Research Framework

The workflow of this study is depicted in the flowchart in Figure 1. The process begins with the collection of a corpus of Javanese legends from the sastra.org digital archive. These texts then undergo manual annotation to create the ground truth dataset. This annotated corpus is subsequently split into 80% training , 10% validation, and 10% testing sets(Ghiffari et al., 2024). The proposed hybrid model, which uses SpaCy for linguistic feature generation and a CRF for advanced extraction, is trained on the training data. Finally, the model's performance is evaluated against the test set using standard evaluation metrics.

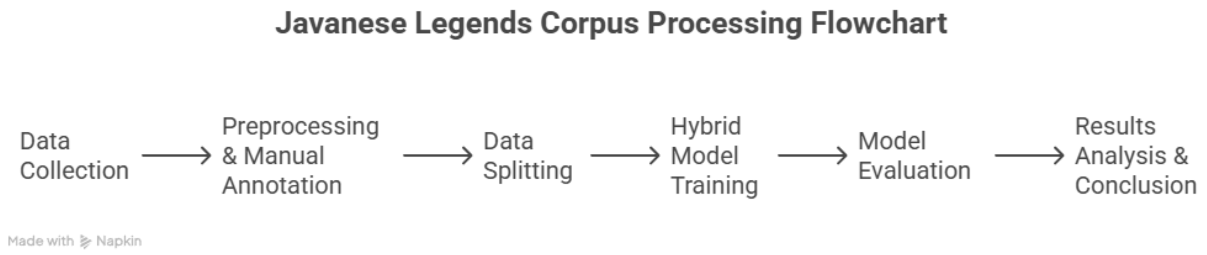

Javanese Legends Corpus Processing Flowchart
Data Collection
Preprocessing & Manual Annotation
Data Splitting
Hybrid Model Training
Model Evaluation
Results Analysis & Conclusion
Made with Napkin

Fig 1. Flowchart

2.2 Data Sourcing and Preparation

The research corpus, consisting of Javanese legends, was sourced from the sastra.org digital archive. A total of 100 stories were selected from this archive for the study. To create the ground-truth dataset for model training, a meticulous process of manual annotation was conducted using a specialized NER annotation application. During this phase, annotators identified and labeled relevant spans of text, such as character names or locations, according to the predefined entity categories. The direct result of this process was a corpus with whole phrases marked with a single entity label(Tentua et al., 2024).

2.3 Measurement and Annotation Scheme

The core measurement task of this research is the identification and classification of named entities as shown in table 1. To maintain consistency, a clear annotation scheme was defined with six entity types relevant to the context of Javanese legends:

Table 1. labelling example

Entity Label	Definition	Examples in Javanese Legend Context
PERSON (PER)	Names of characters, deities, or mythological beings.	Sang Arjuna, Gatotkaca, Semar, Petruk
LOCATION (LOC)	Names of geographical places, or places with a specific name.	Ngardi Wilis, Wukir Rêtawu, Girirêtna, Kawah Căndradimuka
ORGANIZATION (ORG)	Names of groups, dynasties, or kingdoms.	Pandhawa, Kurawa, Ngastina
TIME (TIME)	References to specific temporal markers (e.g., eras, named days).	Dalu, Pitung dalu, Siyang
EVENT (EVENT)	Names of significant historical or cultural events and rituals. This includes micro-events that are part of, or trigger, larger macro-events.	Prang Bharatayudha, Sekaten, Ruwatan
LEGENDARY_OBJECT (LGO)	Names of specific heirlooms, artifacts, or mythological objects.	Pasopati, Keris Mpu Gandring, Jamus Kalimasada

2.4 Proposed Model Architecture: A SpaCy-CRF Hybrid

This study proposes a hybrid model architecture that integrates the strengths of both SpaCy and Conditional Random Fields. The foundational layer of the model utilizes the SpaCy framework to perform efficient tokenization and extract a rich set of linguistic features from the input text(Keraghel et al., 2024). These features subsequently serve as input for a Conditional Random Fields layer, which is responsible for advanced sequence labeling.

Before the CRF layer could be trained, a critical data transformation step was necessary. The span-level labels from the manual annotation phase were programmatically converted into the BIO (Beginning, Inside, Outside) tagging scheme. This format is essential for the CRF model, as it operates on a token level and relies on this scheme to learn the transitional probabilities between adjacent tags(Pan et al., 2024).This process enables the model to ultimately produce more contextually coherent and accurate entity predictions. The architecture of this research shown in figure 2 below.

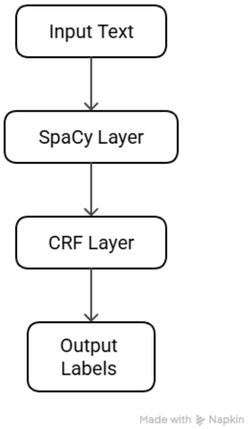


Fig 2. Architecture of the Hybrid SpaCy-CRF Model

2.5 Evaluation Metrics

The model's performance was quantitatively evaluated using three standard metrics for NER tasks:

1. Precision: This metric measures the accuracy of the positive predictions,

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

2. Recall: This metric measures the model's ability to find all relevant instances,

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

3. F1-Score: This is the harmonic mean of Precision and Recall, providing a single, balanced measure of the model's overall performance. It is the primary metric used to report the final result.

$$F1 - Score = 2\frac{Precision\ \times Recall}{Precision + Recall}$$

The generalization ability of the model was rigorously assessed through a final evaluation on a dedicated test set, which consisted of data withheld from all training stages(Warto et al., 2023).

3. Result and Discussion

This chapter presents the empirical results of the experiments and provides a comprehensive discussion of their implications. The section is divided into two parts: first, a summary of the performance metrics for both the baseline SpaCy model and the proposed SpaCy-CRF hybrid model, and second, an in-depth analysis of these findings.

3.1 Experimental Results

To evaluate the effectiveness of integrating a Conditional Random Fields (CRF) layer, we compared the performance of a standard SpaCy NER model against our proposed hybrid SpaCy-CRF model. Both models were trained and evaluated on the same annotated dataset of Javanese legends. The experiments were repeated five times with different random seeds to ensure the stability and reliability of the results. The performance was measured using Precision, Recall, and F1-Score.

The results for the baseline SpaCy model are summarized in Table 2. The model's performance shows some variance across different training runs, achieving an average F1-Score of approximately 0.268.

Table 2. Performance of the Baseline SpaCy Model

Run	Precision	Recall	F1-Score
1	0.7014	0.1727	0.2690
2	0.7138	0.1706	0.2670
3	0.7016	0.1757	0.2705
4	0.6926	0.1764	0.2709
5	0.7170	0.1697	0.2661

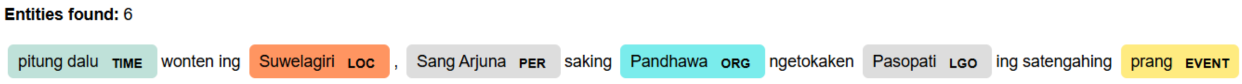
Table 3 presents the results for the proposed SpaCy-CRF hybrid model. The integration of the CRF layer yielded a dramatic and consistent improvement in performance across all metrics.

Table 3. Performance of the Hybrid SpaCy-CRF Model

Run	Precision	Recall	F1-Score
1	0.8923	0.8678	0.8803
2	0.8923	0.8678	0.8803
3	0.8923	0.8678	0.8803
4	0.8923	0.8678	0.8803
5	0.8923	0.8678	0.8803

As shown, the hybrid model achieved a stable F1-Score of 0.8803, a significant increase of nearly 0.60 points compared to the baseline model's average. This demonstrates the substantial contribution of the CRF layer.

Beyond the quantitative metrics, a qualitative analysis was conducted to observe the model's practical performance on specific sentences. Figure 3 provides an example of the hybrid model's output on a complex sentence from the test set.



Entities found: 6

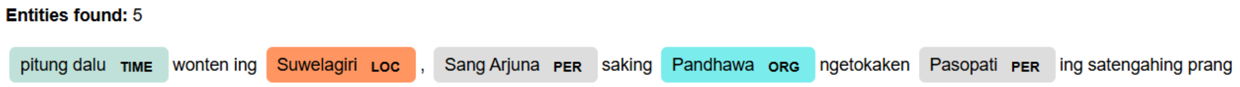
pitung dalu TIME wonten ing Suwelagiri LOC , Sang Arjuna PER saking Pandhawa ORG ngetokaken Pasopati LGO ing satengahing prang EVENT

Fig 3. Sample output from the SpaCy-CRF model

In this example, the model successfully identified all six named entities within the sentence: pitung dalu (TIME), Suwelagiri (LOCATION), Sang Arjuna (PERSON), Pandhawa (ORGANIZATION), Pasopati

(LEGENDARY_OBJECT), and prang (EVENT). The ability to correctly classify multiple, distinct entity types in close proximity demonstrates the model's nuanced contextual understanding and practical effectiveness.

To provide a direct comparison and highlight the baseline model's weaknesses, Figure 4 shows the output of the standard SpaCy model (without CRF) on the exact same sentence.



Entities found: 5

pitung dalu TIME wonten ing Suwelagiri LOC , Sang Arjuna PER saking Pandhawa ORG ngetokaken Pasopati PER ing satengahing prang

Fig 4. Sample output from the baseline SpaCy model (without CRF)

In stark contrast to the hybrid model's correct output, the baseline model exhibits significant failures. It incorrectly classifies Pasopati (LEGENDARY_OBJECT) as a PER (PERSON). Furthermore, it completely fails to identify prang (war) as an EVENT. This qualitative example provides a clear visual demonstration of the low recall and contextual deficiencies discussed in the quantitative results, which are the exact problems solved by the integration of the CRF layer.

3.2 Discussion

The experimental results clearly validate the central hypothesis of this study: integrating a CRF layer as an advanced feature extraction mechanism significantly improves the accuracy of a SpaCy-based NER model for Javanese legends.

The baseline SpaCy model, while achieving high precision, suffered from extremely low recall. This indicates that the model was very conservative in its predictions; it correctly identified entities when it made a prediction but failed to identify the vast majority of entities present in the text. This is a common issue in models that rely on local context and struggle with the linguistic ambiguity found in literary texts. The resulting F1-Score of around 0.27 is insufficient for any practical application.

In stark contrast, the SpaCy-CRF model demonstrates a dramatic improvement in performance. The F1-Score saw a substantial increase from an average of 0.27 on the baseline model to a stable 0.8803 on the hybrid model. This massive improvement is the core finding of this study and is directly attributable to the CRF layer's ability to address SpaCy's primary weakness: its inability to understand sequential context. Whereas the baseline model predicts each token in isolation leading to many missed entities (low recall) the CRF evaluates the probability of the entire sequence of labels in a sentence. The CRF layer learns contextual rules, such as the fact that an B-PERSON tag is highly likely to follow a I-PERSON tag. By doing so, the CRF effectively "corrects" illogical predictions and drastically boosts the model's ability to find previously missed entities, which explains the significant jump in the recall score and, consequently, the overall F1-Score.

A notable finding is that the results for the SpaCy-CRF model are identical across all five experimental runs. This is an expected outcome due to the deterministic nature of the CRF algorithm. This perfect consistency strongly highlights that the performance improvement achieved is not only quantitatively significant but also highly stable and reliable. This stability confirms the model's reliability, proving that this is not just a onetime improvement, but a robust and trustworthy solution for practical, real-world applications.

In conclusion, the findings strongly support the use of a hybrid architecture to improve NER performance. The SpaCy framework serves as a good foundation, while the CRF layer provides the crucial mechanism for sequential learning that results in a significant and robust improvement in overall accuracy.

4. Conclusions

This research has successfully demonstrated a definitive method for improving the accuracy of a SpaCy-based Named Entity Recognition model for the specific domain of Javanese legends. The core conclusion is that the integration of a Conditional Random Fields (CRF) layer as an advanced feature extraction mechanism effectively resolves the primary weaknesses of the standard SpaCy model, particularly its low recall.

The evidence for this improvement is unequivocal: the F1-Score saw a substantial increase from an average of approximately 0.27 with the baseline model to a stable and robust 0.8803 with the proposed hybrid model. This confirms that augmenting SpaCy with a sequential learning layer like CRF is a highly effective strategy for this task.

For future work aimed at further improving model performance, several directions are suggested. First, future work could focus on improving the model's generalization capabilities by training it on a more extensive and varied corpus of Javanese literary texts. Additionally, a valuable next step would be to benchmark the current hybrid model against other advanced architectures, such as Transformer-based models, to identify new avenues for accuracy improvement. Finally, the effectiveness of this hybrid methodology as a general strategy could be validated by applying it to other low-resource local languages.

In summary, this study provides a validated, practical solution that significantly improves NER accuracy, paving the way for more reliable downstream applications in the digital analysis of cultural heritage texts.

5. References

Ghiffari, F. A. Al, Alfina, I., & Azizah, K. (2024). *Cross-lingual Transfer Learning for Javanese Dependency Parsing*.

Gurgurov, D., Hartmann, M., & Ostermann, S. (2024). *Adapting Multilingual LLMs to Low-Resource Languages with Knowledge Graphs via Adapters*. https://github.com/d-gurgurov/

Holla, A., Gaind, B., Katta, V. R., Kundu, A., & Kamalesh, S. (2019). *Hybrid NER System for Multi-Source Offer Feeds*.

Karo, I. M. K., Dewi, S., & Nasution, A. S. (2025). Named Entity Recognition on Indonesian Online News Based on Bidirectional LSTM-CRF. *2025 4th International Conference on Electronics Representation and Algorithm (ICERA)*, 251–256. https://doi.org/10.1109/ICERA66156.2025.11087367

Keraghel, I., Morbieu, S., & Nadif, M. (2024). *Recent Advances in Named Entity Recognition: A Comprehensive Survey and Comparative Study*.

Li, J., Sun, A., Han, J., & Li, C. (2020). *A Survey on Deep Learning for Named Entity Recognition*. https://doi.org/10.1109/TKDE.2020.2981314

Pan, H., Zhang, Q., Caragea, C., Dragut, E., & Latecki, L. J. (2024). *SciDMT: A Large-Scale Corpus for Detecting Scientific Mentions*.

Sulistyo, D. A., Prasetya, D. D., Ahda, F. A., & Wibawa, A. P. (2025). Pivoted Low Resource Multilingual Translation with NER Optimization. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *24*(5), 1–16. https://doi.org/10.1145/3727876

Sulistyo, D. A., Wibawa, A. P., Prasetya, D. D., & Ahda, F. A. (2025a). An enhanced pivot-based neural machine translation for low-resource languages. *International Journal of Advances in Intelligent Informatics*, *11*(2), 258. https://doi.org/10.26555/ijain.v11i2.2115

Sulistyo, D. A., Wibawa, A. P., Prasetya, D. D., & Ahda, F. A. (2025b). Indonesian cross-linguistic named entity recognition. *Research Methods in Applied Linguistics*, *4*(3), 100236. https://doi.org/10.1016/j.rmal.2025.100236

Sulistyo, D. A., Wibawa, A. P., Prasetya, D. D., Ahda, F. A., Arya Astawa, I. N. G., & Andika Dwiyanto, F. (2025). Multilingual Parallel Corpus for Indonesian Low-Resource Languages. *JOIV : International Journal on Informatics Visualization*, *9*(5), 2176. https://doi.org/10.62527/joiv.9.5.3412

Tentua, M. N., Suprapto, & Afiahayati. (2024). NERSkill.Id: Annotated dataset of Indonesian's skill entity recognition. *Data in Brief*, *53*, 110192. https://doi.org/10.1016/j.dib.2024.110192

Warto, Muljono, Purwanto, & Noersasongko, E. (2023). Improving Named Entity Recognition in Bahasa Indonesia with Transformer-Word2Vec-CNN-Attention Model. *International Journal of Intelligent Engineering and Systems*, *16*(4), 655–668. https://doi.org/10.22266/ijies2023.0831.53

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). In *Indonesian Journal of Information Systems (IJIS)* (Vol. 4, Issue 1). https://doi.org/10.24002/ijis.v4i1.4677