



BiLSTM-Based Named Entity Recognition for Itihasa Digital Heritage: Towards Cultural Tourism Information Services

Ni Luh Wiwik Sri Rahayu Ginantra¹^{*}, Ni Made Novi Sintya Dewi¹,
and Wayan Gede Suka Parwita²

¹ Informatics Study Program, Institut Bisnis dan Teknologi Indonesia, Bali, Indonesia

² Business Administration Department, Politeknik Negeri Bali, Bali, Indonesia
wiwik@instiki.ac.id

Abstract. The Mahabharata and Ramayana are central Itihasa narratives that continue to shape literary, performative, and cultural traditions in Indonesia. Their digitization creates opportunities for cultural heritage documentation, but long narrative texts remain difficult to search and organize when entities are not represented explicitly. This study develops a Bidirectional Long Short-Term Memory (BiLSTM) model for Named Entity Recognition (NER) in digitized Itihasa texts. The model processes contextual information in both directions and assigns BIOES sequence labels to entity mentions. The corpus comprises 1,982 Ramayana sentences and 9,814 Mahabharata sentences, divided into 80% training and 20% testing sets. At 30 epochs, the reported token-level testing accuracy reached 99.79% for Ramayana and 99.95% for Mahabharata. These values show stable token classification, although entity-level precision, recall, and F1-score are required for a complete NER assessment. Beyond digital preservation, structured entities can support cultural-tourism search, heritage knowledge bases, interpretive applications, and smart-tourism services that connect characters, places, kingdoms, events, and other cultural concepts. The study therefore positions domain-specific NER as a technical layer for both digital heritage management and tourism-oriented access to classical narratives.

Keywords: BiLSTM, Cultural Tourism, Digital Heritage, Itihasa, Named Entity Recognition, Smart Tourism

1 Introduction

The Mahabharata and Ramayana form part of the Itihasa tradition and have influenced literature and cultural expression across Asia, including Indonesia (Purwandana, 2024). In the Nusantara context, both epics have been reinterpreted through *wayang*, dance, *kakawin*, and other performative and literary forms (Astiniasih, 2023; Paigasari, 2024). This continuing adaptation gives the narratives value not only as texts but also as components of living cultural heritage that require careful documentation and accessible digital representation

Itihasa narratives contain dense references to characters, places, kingdoms, weapons, and events that are important for literary analysis and cultural interpretation (Septiyani, 2019). When these references remain embedded in long prose, users must rely on

manual reading or keyword search, which limits structured exploration. Converting entity mentions into machine-readable records can therefore improve indexing, cross-referencing, and discovery across digital heritage collections.

This requirement is also relevant to cultural tourism. Digital tourism services increasingly depend on structured cultural information to support semantic search, knowledge graphs, recommendation, itinerary planning, and visitor interpretation. Cultural-tourism knowledge graphs, for example, link historical figures, events, intangible heritage, and tourism resources to answer cultural information needs (Fan & Chen, 2024). More broadly, machine-learning pipelines that use NER to enrich cultural and touristic data have been proposed for smart travel and tourism promotion (Deligiannis et al., 2025). In this setting, an *Itihasa* NER model can act as an upstream information-extraction component rather than as a tourism application by itself.

Named Entity Recognition (NER) is an NLP task that locates entity spans and assigns them to semantic categories (Ariyani, 2023). For Indonesian-language sequence labelling, recurrent and transformer-based models have been used across news, social media, and other domains (Irfan & Hidayatullah, 2021; Pradhana et al., 2025). BiLSTM remains a practical architecture when the objective is to model left and right context with a relatively compact network (Mas et al., 2025; Pradhana et al., 2025). However, performance reported on general Indonesian NER corpora does not automatically transfer to classical or culturally specialised narratives, where names, epithets, kingdoms, objects, and event expressions follow domain-specific patterns.

Accordingly, the contribution of this study lies primarily in the domain application and digital-heritage use case rather than in proposing a new neural architecture. The study develops a BiLSTM-based sequence labeller for Ramayana and Mahabharata narratives, documents the BIOES annotation workflow, and evaluates the model on separate test sets. The resulting structured entities are intended to support later heritage information systems, including cultural-tourism search, entity-linked storytelling, museum or destination interpretation, and smart-tourism knowledge services. This framing also avoids treating token accuracy alone as evidence of superiority over established Indonesian NER benchmarks.

2 Methodology

2.1 Dataset

The corpus consists of digitised narrative text from the Ramayana and Mahabharata sources from book title Ramayana and Mahabharata. The study reports 159,368 raw text “entries” during collection, although the report does not define the unit represented by an entry. For model development, the reproducible unit is the sentence sequence: 1,982 Ramayana sentences and 9,814 Mahabharata sentences, giving 11,796 sequences in total. These sequences were used for annotation, preprocessing, model training, and held-out testing. The two corpora are not balanced; the larger Mahabharata subset is therefore considered when interpreting differences in model behavior.

2.2 Research Stages

The workflow comprises literature review, corpus preparation, annotation, preprocessing, sequence padding, data splitting, BiLSTM training, prediction, and model evaluation. The stages are arranged to separate data preparation from model fitting and held-out assessment.

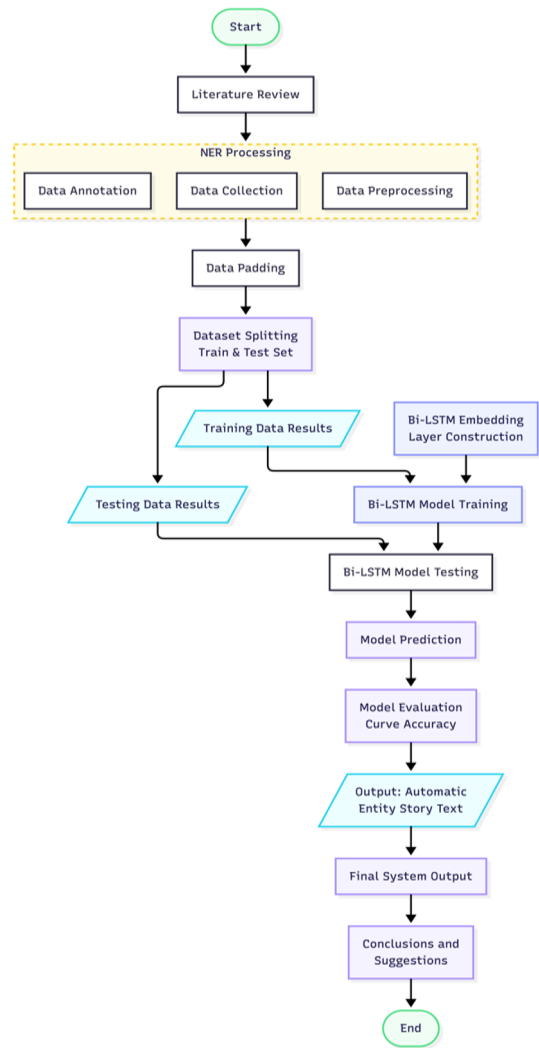


Figure 1. Research Methodology Workflow

Figure 1 summarizes the experimental workflow. The study begins with a review of NER, sequence labelling, and digital-heritage literature, followed by corpus collection and manual annotation. Preprocessing removes selected punctuation and converts the annotated text into padded token sequences. The resulting sequences are separated into training and testing subsets before the BiLSTM model is fitted. Predictions on the held-

out test data are then compared with the reference labels. Training and validation accuracy and loss are retained as optimisation diagnostics, whereas NER performance should be reported with entity-level precision, recall, and F1-score. The final output is structured entity information that can subsequently be integrated into digital heritage or tourism-oriented information services.

2.3 Model Evaluation

Model evaluation distinguishes token-level optimisation indicators from entity-level NER metrics. Accuracy and categorical cross-entropy loss are used to monitor model fitting. For final NER evaluation, predicted BIOES sequences should be assessed with the seqeval framework in strict IOBES/BIOES-compatible mode (Nakayama, 2018). An entity is counted as correct only when both its span boundary and entity type match the reference annotation. Precision is defined as $TP/(TP+FP)$, recall as $TP/(TP+FN)$, and F1-score as $2PR/(P+R)$. Results should be reported for each entity type together with micro-averaged precision, recall, and F1-score. This distinction is important because a high token accuracy can be dominated by the non-entity O class and therefore cannot, on its own, establish strong NER performance.

3 Result and Discussion

3.1 Data Preprocessing and Labelling

Preprocessing removed selected punctuation characters before tokenisation. This step was applied consistently to the two corpora so that the model received comparable sequence inputs. Punctuation removal can also remove boundary cues that are useful for NER; therefore, the present preprocessing choice should be interpreted as part of the reported experimental configuration rather than as a general recommendation for literary-text NER (Fadli & Hidayatullah, 2021).

Entity annotation follows the BIOES convention, in which B marks the beginning of a multi-token entity, I marks an internal token, E marks the final token, S marks a single-token entity, and O marks a token outside any entity. The original study documents 17 semantic label categories for *Ramayana* and 19 for *Mahabharata*, yielding 36 dataset-specific category names across the two annotation dictionaries. The model summary separately reports `num_tags=36`. Because the archived report does not provide the final tag-to-index mapping after BIOES encoding, the two semantic label inventories and their documented lexicon sizes are presented without assuming a one-to-one correspondence between these categories and the network output indices.

The archived annotation examples show that non-entity tokens such as “Bagian” and “Pertama” are labelled O, while single-token entities use S-prefixed labels. Examples include RAMA as S-HERO, Ayodya as S-PLACE, Dewabrata as S-NICKNAME, Maharaja as S-POSITION, and dewi, Dewi, and Gangga as S-GOD. These examples mainly illustrate O and S tags; B, I, and E remain part of the BIOES scheme for multi-token entities but are not explicitly shown in the archived examples and are defined in Table 1; no unarchived multi-token sequence is invented solely to populate the example Table 1.

Table 1. BIOES Prefix Definitions

Prefix	Meaning	Boundary interpretation
B	Beginning	First token of a multi-token entity
I	Inside	Internal token of a multi-token entity
E	End	Final token of a multi-token entity
S	Single	A complete entity represented by one token
O	Outside	Token not belonging to a named entity

Ramayana Label Inventory. The Ramayana annotation dictionary contains 17 semantic categories with 166 category-level lexicon entries: STORY (1; Ramayana), STORY_PART (7; Bala Kanda and Ayodya Kanda), HERO (19; Rama, Dewi Shinta, and Hanuman), VILLAIN (13; Rahwana, Indrajid, and Surpanaka), SUPPORTING_CHARACTER (25; Wiswamitra, Kekayi, and Trijata), ROYAL_CHARACTER (9; Dasarata, Janaka, and Bharata), PLACE (17; Ayodya, Alengka, and Hutan Dandaka), GOD (13; Wisnu, Indra, and Sang Hyang), WEAPON (9; Gandhewa Waja and Panah Brahmastra), SACRED_OBJECT (6; Cupu Manik Asthagina and Lingkaran Sakti), ROLE (11; Prabu, Raja, and Begawan), CEREMONY (6; Sesaji Aswamedha and Tapa Brata), ANIMAL_FORM (12; Wanara, Garuda, and Naga Kuwara), AUTHOR (1; Empu Walmiki), BOOK (4; Kitab Ramayana and Kakawin Ramayana), CHILDREN (4; Kusa and Lawa), and EVENT (9; Penculikan Dewi Shinta and Perang Alengka). These labels operationalize narrative structure, character roles, locations, deities, weapons and sacred objects, ceremonies, creatures, authorship, textual works, offspring characters, and major events.

Mahabharata Label Inventory. The Mahabharata annotation dictionary contains 19 semantic categories with 412 category-level lexicon entries: STORY (1; Mahabharata), STORY_PART (17; Adi Parva, Vana Parva, and Karna Parva), HERO (12; Abimanyu, Drupadi, and Krishna), VILLAIN (32; Karna, Sengkuni, and Sisupala), SUPPORTING_CHARACTERS (107; Drupada, Gandari, and Pandu), PLAYING_CHARACTERS (6; Narada, Vyasa, and Sanjaya), NICKNAME (41; Achuta, Partha, and Werkudara), LOCATION (19; Kurukshetra, Mathura, and Upaplawya), GOD (22; Brahma, Indra, and Wisnu), WEAPON (15; Brahmastra, Gandewa, and Vajra), POSITION (15; Acharya, Raja, and Senapati), GOD_CEREMONY (6; Agnihotra, Aswamedha Yajna, and Rajasuya), CREATURES (7; Airawata, Garuda, and Naga Vasuki), SACRED_OBJECT (8; Akshyapatra, Amrit, and Chakra), EVENT (11; Kurukshetra War, Dice Game, and Exile of Pandavas), CLAN (6; Yadawa, Kurawa, and Pandawa), BATTLE_FORMATION (6; Garuda Wyuha and Chakra Wyuha), PANDAWA (5; Yudhistira, Bima, and Arjuna), and KORAWA (76; Duryudana, Dursasana, and Vikarna). The Mahabharata schema therefore extends the narrative categories with aliases, clans, battle formations, and group-specific character labels.

Lexicon Interpretation. The totals above represent documented category-level lexicon entries after exact duplicate strings within each category are collapsed; they are not

corpus mention frequencies. Because the study does not report class-wise mention counts, a true mention-level entity distribution cannot be reconstructed without the original annotated files. Accordingly, lexicon size is used only to document the annotation inventory and should not be interpreted as entity prevalence in the corpus.

3.2 Data Split

The annotated sequences were divided using an 80:20 random split: Ramayana yielded 1,587 training and 397 testing sentences from 1,982 sequences, whereas Mahabharata yielded 7,851 training and 1,963 testing sentences from 9,814 sequences. The split reserves unseen sequences for held-out evaluation, although reproducibility would be strengthened by reporting the random seed and checking duplicate or near-duplicate sequences before splitting (Muningsih, 2022).

3.3 Building the BiLSTM Model

The model uses an embedding layer followed by a Bidirectional LSTM and a TimeDistributed dense output layer. Bidirectional processing allows the representation of each token to incorporate both preceding and following context, which is useful when an entity can only be interpreted from neighbouring words. The embedding dimension is 64, and padded sequences have a maximum length of 75 tokens.

The Bidirectional LSTM contains 100 recurrent units and uses `return_sequences=True` so that the network produces a representation for every token position. A recurrent dropout value of 0.1 is applied during training to reduce overfitting (Prabowo et al., 2026). The TimeDistributed dense layer applies a softmax classifier to each token representation. Its output dimension is 36, corresponding to the tag classes in the training label dictionary. The model is optimized with Adam and categorical cross-entropy, while token accuracy is monitored during training. Because the output space depends on the project-specific label inventory, the full mapping between the 36 indices and BIOES tags must be reported alongside the annotation definitions. In summary, the model uses a 64-dimensional embedding layer for sequences padded to 75 tokens (400,768 parameters), followed by a Bidirectional LSTM with 100 recurrent units and `return_sequences=True` (132,000 parameters). A TimeDistributed dense layer classifies each token into 36 tag classes using softmax (7,236 parameters). The complete model contains 540,004 parameters and is trained with Adam and categorical cross-entropy.

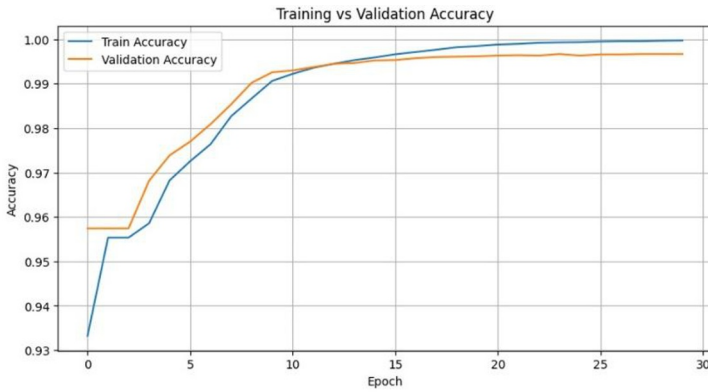
3.4 Discussion

The BiLSTM models were trained for 15 and 30 epochs on the respective annotated corpora. Training and validation trajectories are reported to show optimisation behaviour across the three schedules; held-out test results are discussed separately in Section 3.5.

Ramayana. The Ramayana model was trained for 15 and 30 epochs. Figure 2 presents the corresponding training and validation accuracy curves.



(a)



(b)

Figure 2. (a) Training and Validation Accuracy for the Ramayana NER Model for 15 epoch; (b) Training and Validation Accuracy for the Ramayana NER Model for 30 epoch.

At 15 epochs, the values increased to 99.60% and 99.57%, respectively. At 30 epochs, training accuracy reached 99.97% and validation accuracy reached 99.66%, validation loss 1.66%. The narrowing gap between the two curves suggests stable token-level optimisation, although these values should not be interpreted as entity-level NER scores.

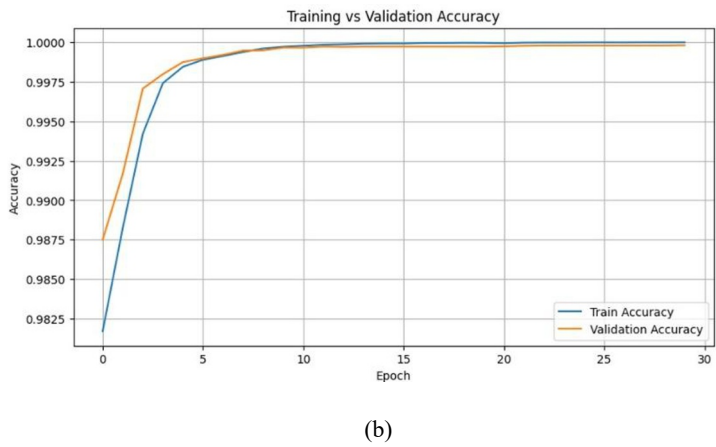
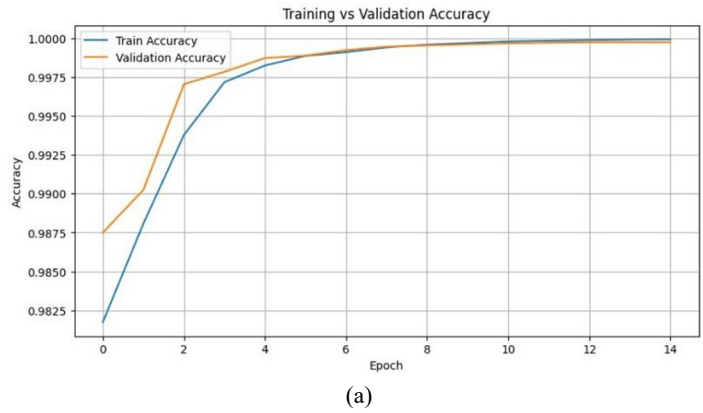
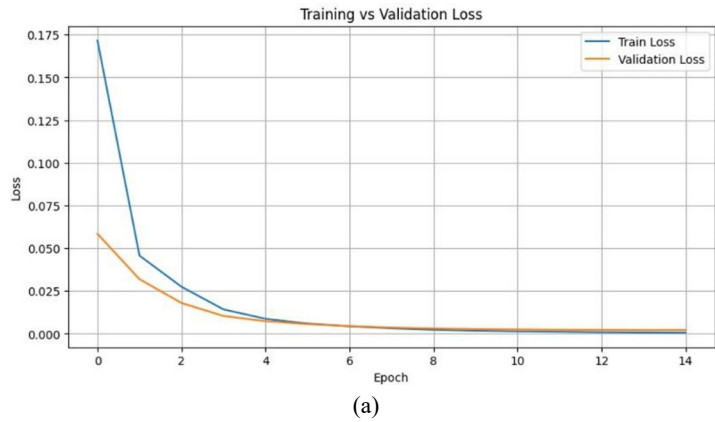


Figure 3. (a) Training and Validation Accuracy for the Mahabharata NER Model at 15 epoch; (b) Training and Validation Accuracy for the Mahabharata NER Model at 30 epoch.



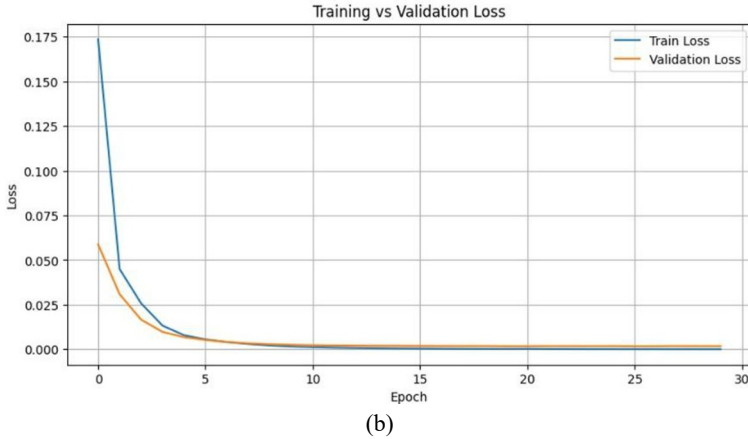


Figure 4. (a). Training and Validation Loss for the Mahabharata NER Model at 15 epoch; (b) Training and Validation Loss for the Mahabharata NER Model at 30 epoch.

Mahabharata. The Mahabharata model was trained for 15, and 30 epochs. Figure 3 shows that training accuracy 99.99% at 15 epochs and 100.00% at 30 epochs, while validation accuracy 99.97% and 99.98% across the same schedules.

Figure 4 shows the corresponding loss curves. Training and validation loss were 0.04% and 0.21% at 15 epochs, and 0.004% and 0.18% at 30 epochs. The validation loss remains low across all three schedules.

3.5 Model Evaluation

Ramayana. On the held-out test set, token-level accuracy at 15 epochs and 99.79% at 30 epochs. Test loss 1.71% and 1.03%, respectively. The 30-epoch model is therefore selected as the best Ramayana configuration among the three reported schedules.

Mahabharata. On the held-out test set, token-level accuracy was 99.94% at 15 epochs, and 99.95% at 30 epochs. The corresponding losses were 0.37%, and 0.38%. The 30-epoch model provides the highest reported token accuracy, although the 15-epoch model has a slightly lower loss.

Entity-level evaluation. Token accuracy is retained for comparison with the original experiment, but it is not sufficient for NER. The study does not report seqeval outputs and does not include saved y_{true}/y_{pred} sequences from the held-out sets. Consequently, entity-level precision, recall, F1-score, and support cannot be reconstructed from the reported accuracy and loss values. Table 2 therefore records these values as not reported (NR) rather than estimating them. A complete seqeval assessment in strict BIOES/IOBES-compatible mode (Nakayama, 2018) requires recovering the saved predictions or re-running inference from the archived checkpoints.

Table 2. Entity-Level Precision, Recall, and F1-score using Seqeval

Dataset / entity type	Precision	Recall	F1-score	Support
Ramayana – micro average	NR	NR	NR	NR
Ramayana – per entity type	NR	NR	NR	NR
Mahabharata – micro average	NR	NR	NR	NR
Mahabharata – per entity type	NR	NR	NR	NR

As shown in Table 2, NR denotes values not reported in the archived research and therefore not derivable from token accuracy or loss alone.

3.6 Qualitative Error Analysis

The research provides six free-text testing examples that complement the aggregate accuracy results. Several predictions are consistent with the documented schema, including Prabu as ROLE, Dasarata as ROYAL_CHARACTER, Ayodya as PLACE, Sengkuni as VILLAIN, and Yudhistira/Bima as PANDAWA. The same screenshots also reveal errors that token accuracy does not expose clearly. In the second Ramayana test sentence, the multi-word name “Raden Rama Wijaya” is only partially highlighted as “Rama” in the prediction display, indicating a span-boundary problem. In the third Ramayana test sentence, “Rahwana” is displayed as SUPPORTING_CHARACTER even though the Ramayana label dictionary lists Rahwana under VILLAIN. These cases show why entity-span metrics and an annotation-consistency audit are necessary before interpreting the very high token accuracy as uniformly high NER quality.

Benchmark context A controlled baseline on the same *Itihasa* split would provide the strongest comparison. Published Indonesian NER studies provide only descriptive external context because their datasets, entity inventories, and evaluation protocols differ from the present *Itihasa* experiment. Reported results include an F1-score of 93.6% for CRF on Indonesian flood-news NER; entity F1-scores of $90.85\% \pm 0.08$ for BiLSTM-CRF with FastText, $94.90\% \pm 0.31$ for BiLSTM-CRF with IndoBERT features, and $93.13\% \pm 0.31$ for fine-tuned IndoBERT on standardized Indonesian news NER; and an entity F1-score of 84.11% for BiLSTM-CRF with FastText on Indonesian tweets. In comparison, the current BiLSTM experiment reports token accuracy of 99.79% for Ramayana and 99.95% for Mahabharata; these values are not directly comparable with entity-level F1-scores.

Reproducibility limitation. The archived study is sufficient to recover the dataset-specific entity dictionaries, BIOES examples, sentence split, model architecture, optimization curves, and qualitative test cases. It is not sufficient to recover entity-level seqeval scores or to create a controlled baseline on the same split, because the necessary prediction sequences, annotated files, and alternative-model outputs are not included. These quantities should not be estimated from token accuracy.

4 Conclusion

This study developed a BiLSTM-based NER model for digitized Ramayana and Mahabharata texts. The 30-epoch models achieved token-level test accuracies of 99.79% for Ramayana and 99.95% for Mahabharata, indicating stable token

classification performance. However, the results should be interpreted carefully because entity-level precision, recall, and F1-score could not be evaluated from the available archived data.

The main contribution of this study is the application of NER to Itihasa digital heritage, enabling important entities such as characters, places, events, and cultural concepts to be extracted and structured. These entities can support digital heritage management and future cultural-tourism applications, including semantic search, knowledge graphs, and interactive heritage information services.

Future work should evaluate the model using entity-level metrics, improve annotation consistency, and compare the BiLSTM model with other NER approaches such as BiLSTM-CRF or transformer-based models on the same dataset.

References

- Ariyani, R. (2023). Literature review information extraction pada teks biomedical dengan pendekatan machine learning dan named entity recognition. *JUIT: Jurnal Utilitas Inovasi dan Teknologi*, 2(1), 83–88.
- Astiniasih, N. M. D. (2023). Hindu Nusantara: Perkembangan teks dan konteks. *ŚRUTI: Jurnal Agama Hindu*, 3(2), 145–154.
- Deligiannis, K., Tryfonopoulos, C., Raftopoulou, P., Vassilakis, C., Kaffes, V., & Skiadopoulos, S. (2025). A machine learning framework for harvesting and harmonizing cultural and touristic data. *Information*, 16(12), 1038. <https://doi.org/10.3390/info16121038>
- Fadli, H. F., & Hidayatullah, A. F. (2021). Identifikasi cyberbullying pada media sosial Twitter menggunakan metode LSTM dan BiLSTM. *Automata*, 2(1).
- Fan, Z., & Chen, C. (2024). CuPe-KG: Cultural perspective-based knowledge graph construction of tourism resources via pretrained language models. *Information Processing & Management*, 61(3), 103646. <https://doi.org/10.1016/j.ipm.2024.103646>
- Irfan, M., & Hidayatullah, A. F. (2021). Tinjauan literatur named entity recognition dengan machine learning dan deep learning pada ulasan wisata. *Automata*, 2(2).
- Mas, A., Krismono, B., & Priyanto, D. (2025). Prediksi gender berdasarkan nama menggunakan kombinasi model IndoBERT, convolutional neural network (CNN), dan bidirectional long short-term memory (BiLSTM). *JTIM: Jurnal Teknologi Informasi dan Multimedia*, 7(3), 448–460.
- Muningsih, E. (2022). Kombinasi metode K-means dan decision tree dengan perbandingan kriteria dan split data. *Jurnal TeknoInfo*, 16(1), 113–118.
- Nakayama, H. (2018). *segeval: A Python framework for sequence labeling evaluation*.
- Paigasari, L. G. H. (2024). Peran perempuan dalam Itihasa: Refleksi tokoh Sita dalam cerita Ramayana. *PRAMANA: Jurnal Hasil Penelitian*, 4(1), 10–19.

- Prabowo, T., Sarif, M. I., Sebayang, A., Ferdillah, T. D., & Azuan, M. (2026). Analisis sentimen masyarakat terhadap isu korupsi dana bencana di Indonesia menggunakan metode bidirectional long short-term memory (Bi-LSTM). *JUKTISI: Jurnal Komputer, Teknologi Informasi dan Sistem Komputer*, 4(3), 1787–1795.
- Pradhana, A. H., Daniati, E., & Najibulloh, M. (2025). Penerapan Bi-LSTM untuk named entity recognition pada teks bahasa Indonesia. *Indonesian Journal of Computer Science Research*, 4(2), 96–106.
- Purwandana, I. P. R. A. (2024). Pengaruh epik Mahabharata dalam karya sastra modern (perspektif susastra Hindu). *PRAMANA: Jurnal Hasil Penelitian*, 4(1), 20–28.
- Septiyani, V. I. (2019). *Hubungan intertekstual novel Ramayana karya C. Rajagopalachari dan Rahwana Tattwa karya Agus Sunyoto serta ideologinya*.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

