Volume XI Issue 3 Year 2026 | Page 767-781 | ISSN: 2527-9866
Received: 22-05-2026 | Revised: 21-06-2026 | Accepted: 25-07-2026

Comparative Evaluation of Hand-Engineered Features for CRF-Based Named Entity Recognition in Indonesian Clinical Research Articles

Nurmaya[1*], Rofi Tulus[2], Nova Eka Diana[3], Herika Hayurani[4], Sri Puji Utami Atmoko[5], Yurika Sandra[6]
[1,2,3,4,5]Informatics Engineering Department, Faculty of Information Technology, YARSI University, Indonesia, 10510
[6] Medical Department, Faculty of Medicine, YARSI University, Indonesia, 10510
e-mail:nurmaya@yarsi.ac.id[1], rofituluss2@gmail.com[2], nova.diana@yarsi.ac.id[3],herika.hayurani@yarsi.ac.id[4], puji.atmoko@yarsi.ac.id[5], yurika.sandra@yarsi.ac.id[6]
*Correspondence: nurmaya@yarsi.ac.id

Abstract: *Extracting clinical entities such as signs, symptoms, diagnoses, and treatments from Indonesian clinical research makes clinical information more accessible to readers with varying levels of medical knowledge. Conditional Random Fields (CRFs) combined with hand-engineered features have proven effective for health domain Named Entity Recognition (NER). However, studies comparing hand-engineered feature combinations for Clinical NER in Indonesian clinical research articles remain limited. This study compares five hand-engineered feature combinations using a CRF model to identify the most effective combination for recognizing seven clinical entity types, using a newly annotated corpus of 29 Indonesian internal medicine research articles. The dataset was split into 80% training and 20% testing data. Model performance was evaluated using weighted-average and macro-average Precision, Recall, and F1-score, reported as Mean ± SD across five random states (5, 10, 15, 20, 25) and training epochs (50,100,150). The combination of contextual, prefix and suffix, and orthographic features achieved the best weighted F1-score of 78.90% at Epoch 50, though macro-F1 was considerably lower at 64.32% reflecting weaker performance on minority entity types. This gap indicates a need for further corpus expansion and data balancing.*

Keywords: *Clinical Named Entity Recognition, Conditional Random Fields, Hand-Engineered Features. Indonesian Language. Clinical Research Articles*

1. Introduction

Clinical research articles disseminate the latest clinical information derived from the evidence-based studies, case reports, and the literature reviews, thereby raising awareness of the newly identified diseases or updated treatment approach. These publications offer valuable information to the society, particularly medical professional and students. However, understanding scientific articles within a short time remain challenging, especially for health information seekers with limited health literacy or educational backgrounds [1]. Understanding of scientific texts requires advanced reading skills, and individual perspective may further influence interpretation [2], often leading to varied understanding of the same content. These challenges highlight the importance of automatically extracting key clinical information from clinical research articles by using Clinical Named Entity Recognition (NER). Clinical entities extraction from medical news could be carried out automatically by NER, resulting in disease, symptom, and drug entities [3]. By facilitating the extraction of essential information, Clinical NER, a natural language processing task for identifying clinical entities, supports more efficient knowledge acquisition and enables readers to quickly grasp valuable insights [4] that can be applied in clinical research and decision-making processes.

Previous studies have widely used deep learning and conventional machine learning for developing clinical NER models [5], [6]. For example, [7] and [8] demonstrated the promising performance in NER tasks using Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). However, these deep learning models require high computational resource [4]. In comparison, Conditional Random Fields (CRFs), a machine learning method with lower computational requirements, have been implemented in many systems with diverse feature extraction techniques [9], [10]. Hand-engineered feature extraction techniques, including part-of-speech (POS) tags, orthographic, and contextual features, as well as automatic embedding approaches based on pre-trained models such as BERT, transform the text corpus into symbolic features and numerical vectors, respectively. Both approaches have demonstrated effectiveness when integrated with CRF for NER tasks [11], [12]. For example, [13] employed a combination of six hand-engineered feature extraction techniques (contextual, normalized, n-gram, affixes, part-of-speech (POS), and orthographic features) as inputs to a CRF model for disease entity extraction on the NCBI disease corpus. This approach achieved F-scores of 94% and 88% on the training and testing sets, respectively, although performance dropped considerably on minority classes such as Composite Mention due to limited training samples. Considering linguistic differences from English, [14] developed a CRF-based NER model for the Urdu language by incorporating POS tags and contextual word windows, demonstrating an improvement of up to 3% in F1-score compared to a baseline approach employing bootstrap learning and resource-sharing techniques.

Despite the growth of the Clinical NER research, the availability of Indonesian clinical entity resources remains limited compared to those in English, particularly in the domain of clinical research articles which report the latest findings in the medical field. The need for clinical NER in the Indonesian language has increased alongside the growing number of clinical research articles published in Indonesian and the rise in health information seeking since COVID-19 pandemic. Reference [3] developed an Indonesian medical NER system to extract disease, symptom, and drug entities from unstructured content of health news articles written in Indonesian due to the increasing need for online medical consultation services and the limited availability of physicians. Their approach combined Word2Vec and IndoBERT embeddings with BiLSTM and CRF methods, and while the IndoBERT-BiLSTM-CRF configuration outperformed static embeddings, the overall macro-F1 remained modest, i.e., 0.3297, reflecting the difficulty of extracting medical entities from informal news text. Using online news headlines as the data source, [15] extracted the medical entities such as person, organization, location and disease. Their approach employed feature extraction technique including letter case, suffix patterns, title case formatting, numeric token identification, and POS-Tagging, combined with three models, including BiLSTM, BiLSTM-CRF, and IndoBERT. IndoBERT outperformed the other models with an F1-score of 79.64%. Nevertheless, all three models had difficulty to recognize less frequent sequential entities such as I-ORG due to class imbalance. On the other hand, [16] implemented a medical NER system for Indonesian health consultation platforms. Using a dataset of questions and answers (Q&A) on pregnancy and children's topics, the BiLSTM-CRF model achieved an accuracy of 99.68% in predicting medical entities such as Action, Anatomy, Physiology, Disease, Pathogen, Protein, Measurement, Species, and Drugs. However, this high accuracy is because of the dominance of non-entity tokens in the dataset. Informal language used by non-expert users also tends to rely on general terms rather than medical vocabulary, limiting the recognition of specific clinical entities [17].

These findings indicate that existing Indonesian medical NER studies still limited by informal or short-context text sources, such as news headlines with imbalance entity distribution [3], [15] and consultation Q&A filled with common, non-medical words [16], which limit both entity coverage and model precision. This pattern is consistent with prior findings that clinical notes

and informal text tend to be more inconsistent and more ambiguous than structured scientific literature, making entity boundaries harder to detect [18]. Using formal and structured text such as research articles may help address these challenges, as the language is written by medical professionals and follows consistent, standardized terminology. Furthermore, differing from previous studies that focused on broad medical entities, this study targets clinical entities directly related to clinical diagnosis and treatment. Clinical research articles follow structured content guidelines set by publishers, which facilitates the extraction of the seven entities proposed in this study, including organ, sign&symptom, disease, physical examination, supporting examination, pharmacological treatment, and non-pharmacological treatment. Therefore, developing a clinical NER model for the Indonesian language based on clinical research articles is a necessary and promising direction.

When applying machine learning method to build such models, appropriate feature extraction techniques should be carefully selected, as an annotated corpus for training Indonesian Clinical NER model is limited, and the writing style of scientific articles is more complex than that of health news and Q&A texts, which are generally simpler and more informal. Although transformer-based models such as IndoBERT have demonstrated promising result in Indonesian NER tasks, as shown by [15], such models typically require substantially more parameters and computational resources for training and fine-tuning compared to CFR-based approach [19]. Given the limited availability of annotated resources for Indonesian clinical text, this study adopts CRF combined with hand-engineered feature extraction techniques, which have proven effective when integrated with the CRF model in prior biomedical and disease NER work such as [13]. A comparative evaluation of five feature combinations, including contextual, POS tag, prefix and suffix, stopword, and orthographic, was conducted to identify the best performing combination for Indonesian Clinical NER in clinical research articles.

The comparison was also conducted across different epoch settings and random state values to examine model stability and to ensure that the model was neither underfitting nor overfitting during training, as reflected in Precision, Recall, and F1-score values. This study contributes to (1) the construction of a new annotated corpus derived from Indonesian clinical research articles, (2) the identification of feature extraction techniques that are effective in this specific domain, and (3) the development of a Clinical NER model for Indonesian language.

2.Methods

This study investigates the performance of various hand-engineered feature extraction on clinical entity extraction from Indonesian clinical research articles. Figure 1 illustrates the proposed procedure for Indonesian Clinical NER in clinical research articles and the evaluation of the hand-engineered feature extraction techniques, which consist of six main stages:

1. Data Collection: Clinical research articles written in Indonesian were collected from online academic journal platforms. The collected articles were converted into a text file.
2. Data Pre-processing: The text file was then pre-processed through tokenization to segment the text into individual tokens.
3. Data Annotation (Tagging and Labelling): Each token was manually annotated through a tagging process and clinical entity labels resulting corpus as a dataset for this study.
4. Feature Extraction: Hand-engineered feature extraction techniques were applied to each token to generate feature representations.
5. Training and Testing: The extracted features were used as inputs to a CRF-based NER model. The performance of the feature extraction combinations, along with the hyperparameter tuning and stability performance were evaluated based on the clinical entity extraction outputs in training and testing phase.

6. Evaluation: The performance of different feature combinations, along with hyperparameter tuning and its consistency, were measured from the Precision, Recall, and F1-score values.

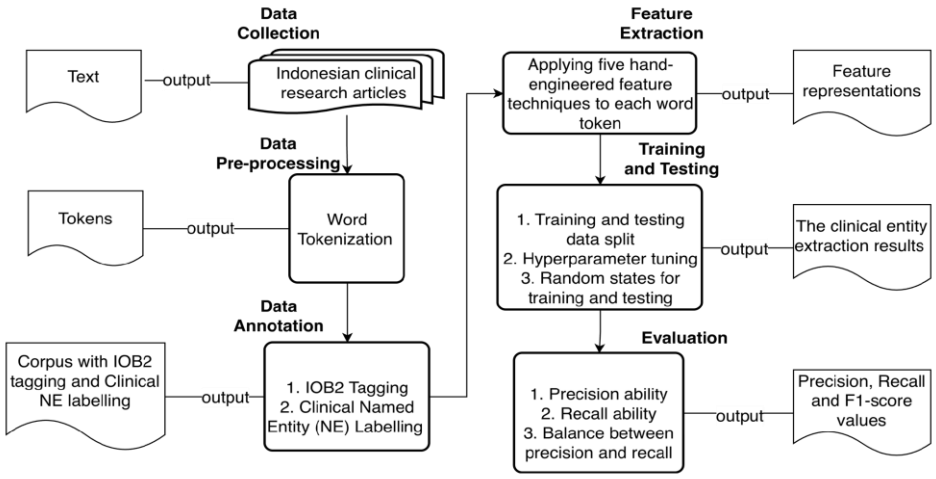

Figure 1. Proposed procedure for Indonesian Clinical NER in clinical research articles and the evaluation of the hand-engineered feature extraction techniques.

Data Collection

Forty internal medicine research articles were collected from an Indonesian online internal medicine journal (available at https://scholarhub.ui.ac.id/jpdi/) and converted into a text file. The journal was selected based on the interview with a general practitioner who also serves as a lecturer at a private medical faculty. The journal presents comprehensive discussions on clinical issue, especially disease, therapeutic approaches, and diagnostic examinations. According to the interview, clinical reasoning in medical practice commonly involves physical and supporting examination of organs, signs and symptoms reported by patients, diagnosis of the underlying disease, the application of pharmacological and non-pharmacological treatment. Thus, this study defined seven types of clinical entities in Indonesian for extraction such as *organ* (organ), *sign and symptom*, *penyakit* (disease), *pemeriksaan fisik* (physical examination), *pemeriksaan penunjang* (supporting examination), *terapi farmakologi* (pharmacological treatment) and *terapi non-farmakologi* (non-pharmacological treatment). Disease refers to a diagnosed condition, whereas Sign and Symptom refers to a complaint or finding prior to diagnosis. Physical Examination refers to direct clinician findings, while Supporting Examination refers to laboratory or imaging findings. Pharmacological Treatment refers to treatment involving medication, while Non-Pharmacological Treatment refers to treatment not involving medication. Entity classification was determined based on the contextual function of a term rather than the term in isolation. For instance, a term such as “anemia” can be classified as either Disease or Sign and Symptom depending on context.

Data Pre-Processing and Data Annotation

The sentences in the text file were split into individual words using a blank Indonesian tokenizer from spaCy [20]. General punctuation marks, e.g., commas, periods, parentheses, were tokenized as separate tokens. Numbers were kept in their original format, including Indonesian decimal and thousand separators. Abbreviations followed by a period, e.g., “dr.,” were kept as single tokens to prevent incorrect sentence splitting detection, and alphanumeric clinical codes, e.g., “CD3,” were kept as single tokens. Hyphenated words, including reduplicated Indonesian terms, e.g., “*bersama-sama*”, were tokenized as separate tokens, while Latin-derived medical terms were annotated directly as part of their corresponding entity type. Slash-based laboratory

units, e.g., "sel/µl," were generally kept as single tokens. Each word was annotated using the Inside-Outside-Begin (IOB2) tagging scheme [21] combined with the clinical entity types. This allowed multi-word entities to span consecutive tokens, e.g., "*kelenjar getah bening* (lymph node)" labelled as B-ORG, I-ORG, I-ORG. B represents the leading token of an entity followed by other tokens, I stands for tokens inside an entity while O corresponds to tokens outside any entity. The annotation was conducted manually by eleven medical students from a private university, supervised by one general practitioner and a responsible research team member. The articles were distributed among the eleven students by the general practitioner for annotation. Since each article was annotated by only one annotator, the inter-annotator agreement was not computed. To maintain annotation consistency, the students were provided with annotation guidelines prior to the annotation process. During annotation, the assigned annotators consulted the general practitioner and the research team member responsible for the labelling process whenever there was uncertainty in determining the appropriate entity type. The completed annotations were subsequently reviewed and verified by both the general practitioner and the responsible research team member to ensure correctness and consistency. From the forty collected articles, twenty-nine were fully annotated and used to construct the corpus, while the remaining were excluded due to incomplete annotation. The resulting corpus consists of 2049 sentences and 49444 tokens in total. Table 1 describes the annotation results of the clinical named entities (NE) and their corresponding with IOB2 labels and distribution. The total number for the clinical named entities with B and I labels is 7045 and 42399 was annotated with O label.

Table 1. Clinical Named Entity Types, IOB2 Lables, and Their Frequency Distribution in Corpus

Named Entity (NE)	Abbreviation	NE with IOB2	Number of NE per IOB2 Label
Organ (Organ)	ORG	B-ORG	186
		I-ORG	100
Sign and Symptom	SS	B-SS	381
		I-SS	678
Penyakit (Disease)	P	B-P	1505
		I-P	905
Pemeriksaan Fisik (Physical Examination)	PF	B-PF	27
		I-PF	80
Pemeriksaan Penunjang (Supporting Examination)	PP	B-PP	622
		I-PP	708
Terapi Farmakologi (Pharmacological Treatment)	TF	B-TF	1067
		I-TF	555
Terapi Non-Farmakologi (Non-Pharmacological Treatment)	TNF	B-TNF	129
		I-TNF	102

The label "O" represents tokens that are not part of named entity and no semantic meaning, such as "yang" (with) or "ke" (to).

Feature Extraction

The labelled words were transformed into feature representation using five techniques:

1. Contextual: Contextual feature captured the words surrounding a NE within a window of ±1 token, including the lowercase form and capitalization (uppercase and title case) of the preceding and following words [13]. For example, in "*Diabetes Melitus Tipe 2*" (Type 2 Diabetes Melitus), "Diabetes" is the preceding word, "Melitus" is the NE, and

"*Tipe*" (Type) is the following word. Beginning-of-sentence (BOS) and end-of-sentence (EOS) markers were used when no preceding or following word was available.

2. POS-Tagging: Indonesian POS-tagging was automatically assigned to each word using the pre-trained model developed by [22] using the corpus produced by [23]. The tags consist of proposition, adjective, verb, noun and so on. For example, the word "Diabetes" was tagged as NN (noun). POS-Tagging feature also assists in identifying the boundaries of phrases within sentences.
3. Prefix and Suffix: A prefix consists of a few letters at the beginning of a word, while a suffix consists of a few letters at the end. Prefixes were extracted as the first 1 to 5 characters of each word, and suffixes were extracted as the last 2 to 5 characters, resulting in nine character-length variants for prefix and suffix extraction. Each prefix and suffix produce a new meaning. In the medical field, prefixes and suffixes are commonly used as medical terms. For example, in the word "abnormal", the prefix "ab-" means "away from" or "not" indicating "far from normal." In gastralgia, the suffix "-algia" means "pain," referring "pain in the epigastrium."
4. Stopwords: A binary feature indicating whether a word matched an entry in a predefined Indonesian stopword list [24] (e.g., *adalah*, *atau*, *ke*). This feature was assigned independently of the word's entity label. Words matching the stopword list retained their original annotated label, e.g., B-SS or I-SS rather than being automatically labelled as "O", as stopwords may still appear within multi-word clinical entities, e.g., conjunctions within a Sign and Symptom span.
5. Orthographic: The orthographic feature set consisted of four indicators capturing the written form of each word. These included whether the word was entirely uppercase, such as "DM," the abbreviation for "Diabetes Melitus." Additional indicators captured whether only the first letter was capitalized, whether the word consisted entirely of digits, and whether the word ended with a digit while containing at least of alphabetic characters, e.g., "DM2". This feature has been shown to contribute effectively to the extraction of biological named entities from unstructured text [25].

Testing and Training

CRF is a supervised learning model that is useful for segmentation and labelling. It is based on an undirected graphical model, where nodes correspond to class label sequences, i.e., named entity types. Each feature function $f_k(y_t, y_{t-1}, x_t)$, as shown in Eq. (1) and Eq. (2), assesses whether a specific property of a word matches a specific label. The weighted sum of all active feature functions for each possible label is used to predict the most possible named entity label.

$$P(x) = \frac{1}{Z(x)} \prod_{t=1}^{T} exp\left(\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t)\right) \quad (1)$$

$$\text{where } Z(X) = \sum_y \prod_{t=1}^{T} exp\left(\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t)\right) \quad (2)$$

Five feature extraction techniques were combined to evaluate their effect on CRF performance, as shown in Table 2. The contextual feature was used as the benchmark for comparison. Epoch values and random states were adjusted to identify the optimal performance of the NER model and to ensure consistent result, respectively. In this study, epoch values of 50,100, and 150 were tested, while random states of 5, 10, 15, 20, and 25 were used. Following previous studies that performed data splitting at the sentence level for training and testing [15], [26], this study applied the same approach by splitting the dataset into 80% for training and 20% for testing at the sentence level.

Table 2. The combination of feature extraction techniques

Feature ID	Feature Extraction
Feature 1	Contextual
Feature 2	Contextual + Prefix and Suffix
Feature 3	Contextual + Prefix and Suffix + Orthographic
Feature 4	Contextual + Prefix and Suffix + Orthographic + Stopwords
Feature 5	Contextual + Prefix and Suffix + Orthographic + Stopwords + POS-Tagging

Evaluation and Data Analysis

The contribution of feature extraction techniques to the performance of CRF was evaluated using Precision, Recall, and F1-score as the evaluation metrics. The percentage of true positive entities among all positive predictions represents the Precision of the extraction while Recall shows the percentage of true positive entities within the total actual positive entities. F1-score calculates the combined ratio of Precision and Recall.

This study used weighted average to compute the overall Precision, Recall, and F1-score for all entity types by considering the support of each entity as the weight, where support refers to the number of tokens annotated with IOB2 labels for each named entity type in the test set. For example, given F1-score values of 95.30% for B-P (support = 362) and 71.88% for B-ORG (support = 36) at random state 5, the weighted average F1-score is computed as: Weighted Average Precision = (95.30 × 362 + 71.88 × 36) / (362 + 36). This ensures that entities with higher support, i.e., those appearing more frequently in the test set, have a greater influence on the overall score. This approach is appropriate for uneven datasets, as the weighted F1-score has been shown to be more representative under such conditions [27].

In addition to this token-level evaluation, entity-level performance was further analyzed by combining the predicted B- and I- tags into complete entity spans. These combinations were then evaluated using strict span matching. The predicted entity is considered correct only if both its boundary and entity type exactly match the ground truth annotation. For example, given a ground truth entity "*nyeri dada sebelah kiri (left-sided chest pain)*" labelled as B-SS, I-SS, I-SS, and I-SS, corresponding to entity type SS. The prediction is considered correct only if it identifies the exact same boundary, i.e., all four tokens, and assigns the same entity type, i.e., SS. Strict span matching was adopted because partial extraction, such as capturing only "*nyeri dada (chest pain)*" as SS without "*sebelah kiri (left-sided)*", could miss clinically relevant details that are essential for an accurate diagnosis.

3. Results and Discussion

Table 3 summarizes the descriptive statistics (Mean ± Standard Deviation (SD)) of the weighted average and macro-average Precision, Recall, and F1-score for each feature combination at different training epochs, calculated from five random states (5, 10, 15, 20, and 25). Using epoch 50 as the baseline, the CRF-based NER model with five feature combinations showed that Features 2 and 3 achieved the highest performance compared to the other features, with comparable Mean ± SD of F1-score of 77.10% ± 2.06% and 77.70% ± 2.63% respectively. Feature 3 achieved slightly higher performance than Feature 2 across all metrics. Compared to Feature 2 and Feature 3, Feature 4 and Feature 5 achieved lower performance, with Mean ± SD of F1-score 76.64% ± 2.68% and 76.51% ± 1.80%, respectively. Meanwhile, Feature 1 yielded the lowest performance at epoch 50 with a Mean ± SD of F1-score of 73.45% ± 1.28%. To further assess whether this weighted-average performance was evenly distributed across entity types, macro-average Precision, Recall, and F1-score were also calculated, treating all entity types with equal weight regardless of their frequency in the dataset. As shown in Table 3,

macro-average F1-scores were lower than their weighted-average counterparts across all feature combinations and training epochs. Feature 3 remained the best-performing configuration under macro-average evaluation as well, achieving a macro-average F1-score of 65.49% ± 5.74% at epoch 50 resulting a gap of 12.21 percentage points below its weighted-average F1-score of 77.70% ± 2.63%. Similarly, Feature 2 showed a difference between weighted F-1 score of 77.10% and the macro F1-score of 64.37%, resulting in a 12.73-point gap. This difference indicates that the relatively high weighted-average scores were largely driven by majority entity classes with greater support, while performance on minority entity types was considerably weaker. This finding is further examined through per-entity analysis in Tables 4 to 6.

Table 3. Mean ± Standard Deviation (SD) of Token-Level Weighted-Average and Macro-Average Precision, Recall, and F1-score for Each Feature Combination at Different Training Epochs

Epoch	Features	Weighted-Average			Macro-Average		
		Precision	Recall	F1-score	Precision	Recall	F1-score
50	Feature 1	86.84±2.50	65.19±1.71	73.45±1.28	83±6.96	52.16±3.51	61.61±3.97
	Feature 2	84.94±3.45	72.22±2.47	77.10±2.06	79.56±7.59	58.08±4.39	64.37±4.60
	Feature 3	85.42±2.84	72.80±3.07	77.70±2.63	80.73±8.46	58.97±5.23	65.49±5.74
	Feature 4	83.84±3.81	72.58±3.20	76.64±2.68	80.17±9.36	57.84±6.01	63.67±5.74
	Feature 5	84.21±2.57	71.44±1.90	76.51±1.80	78.36±7.21	57.24±4.45	63.54±4.40
100	Feature 1	85.94±2.73	65.57±1.59	73.60±1.60	78.77±7.14	52.29±4.47	61.17±4.67
	Feature 2	84.27±2.51	72.60±2.31	77.25±1.87	77.46±6.26	59.11±4.22	64.64±2.99
	Feature 3	83.74±3.63	72.95±3.30	77.15±3.03	78.97±8.76	60.37±5.60	66.21±5.96
	Feature 4	83.04±3.35	73.13±3.60	76.78±2.83	79.26±6.60	59.32±4.91	65.01±5.10
	Feature 5	83.61±2.16	72.12±3.01	76.65±2.34	75.73±6.29	59.36±5.76	63.44±3.54
150	Feature 1	86.23±2.64	65.73±2.04	73.69±1.86	81.70±6.63	52.78±4.50	61.88±4.55
	Feature 2	84.20±2.93	72.74±3.02	77.33±2.54	77.2±5.91	59.83±4.50	64.89±3.60
	Feature 3	83.49±2.60	73.05±3.10	77.12±2.51	79.22±7.04	59.55±4.55	65.7±4.64
	Feature 4	82.37±3.89	72.26±3.06	76.19±2.86	73.86±7.45	58.86±5.21	62.41±4.42
	Feature 5	82.81±2.53	71.74±2.75	76.04±2.06	73.56±3.32	57.55±3.70	61.39±2.40

As shown in Tables 4, 5, and 6, several minority entity types exhibited markedly lower token-level F1-scores compared to majority entities, particularly B-PF (31.87% ± 23.67%), I-PF (34.20% ± 40.79%), and I-TNF (41.15% ± 16.50%) for feature 1 at epoch 50, compared to majority entities such as B-P (88.83% ± 2.59%). This difference explains the gap between weighted-average and macro-average scores observed in Tabel 3, as these low-performing minority entities carry equal weight in the macro-average calculation despite their limited representation in the dataset. Beyond this baseline disparity, Precision, Recall, and F1-score generally improved from Feature 1 to Feature 2 and Feature 3 across most entity types at epoch 50. However, B-PF remained unchanged across all three features, while I-PF showed a slight decrease from Feature 1 to Feature 2 before returning to its original value in Feature 3. This is likely due to their limited corpus support (B-PF = 27 tokens, I-PF = 80 tokens), as shown in Table 1, resulting in very few testing instances per run which may affect the reliability of the evaluation metrics. A comparable issue was observed for I-ORG, which showed a consistent decrease in F1-score from Feature 1 (53.76% ± 16.91%) to Feature 2 (50.01% ± 14.30%) and further to Feature 3 (49.54% ± 12.01%). This is possibly because I-ORG only appears 100 times in the corpus, as shown in Table 1, suggesting that the decrease is more likely a result of how the data was randomly partitioned across the five runs than a genuine weakness of Feature 2 or Feature 3. While most entities showed improvement from Feature 1 to Feature 3, B-TF and I-P showed slight decreases in F1-score of 0.27% and 1.21% respectively, from Feature 2 to Feature 3. Nevertheless, their scores in Feature 3 remained higher than those in Feature 1, indicating that the overall improvement trend is maintained. These findings suggest that Feature 2

substantially improves the model's ability to recognize clinical entities compared to the baseline Feature 1, primarily through the addition of prefix and suffix information.

Table 4. Mean ± Standard Deviation (SD) of Per-Tag (Token-level) Precision, Recall, and F1-score for Feature 1 at Epoch 50

Entity	Mean (%) ± SD (%)			Mean ± SD
	Precision	Recall	F1-score	Support
B-ORG	85.47% ± 7.60%	57.86% ± 8.64%	68.58% ± 6.89%	37.6 ± 7.3
I-ORG	85.09% ± 23.79%	40.83% ± 14.42%	53.76% ± 16.91%	16.8 ± 7.2
B-SS	79.84% ± 11.12%	44.80% ± 5.50%	57.09% ± 5.56%	70.2 ± 8.2
I-SS	65.10% ± 5.96%	43.17% ± 16.73%	50.16% ± 9.07%	121.2 ± 16.1
B-P	94.52% ± 2.29%	83.86% ± 3.80%	88.83% ± 2.59%	310.6 ± 34.4
I-P	89.24% ± 1.79%	72.67% ± 5.27%	80.03% ± 3.40%	181.8 ± 27.9
B-PF	80.00% ± 44.72%	20.97% ± 18.09%	31.87% ± 23.67%	6.0 ± 3.2
I-PF	64.00% ± 40.99%	30.33% ± 41.63%	34.20% ± 40.79%	13.8 ± 13.7
B-PP	88.21% ± 6.24%	66.13% ± 3.99%	75.46% ± 3.46%	125.0 ± 15.0
I-PP	79.90% ± 5.25%	57.24% ± 7.23%	66.29% ± 4.49%	145.0 ± 22.9
B-TF	92.69% ± 1.59%	70.17% ± 2.82%	79.86% ± 2.33%	221.2 ± 10.5
I-TF	86.82% ± 4.28%	70.09% ± 5.99%	77.39% ± 3.82%	115.4 ± 20.4
B-TNF	90.42% ± 7.53%	43.13% ± 9.92%	57.84% ± 8.72%	28.2 ± 5.5
I-TNF	80.73% ± 15.57%	28.98% ± 14.50%	41.15% ± 16.50%	24.0 ± 10.1

Support refers to the mean number of instances per entity in the test set across 5 runs with different random states (5, 10, 15, 20, and 25)

Table 5. Mean ± Standard Deviation (SD) of Per-Tag (Token-level) Precision, Recall, and F1-score for Feature 2 at Epoch 50

Entity	Mean (%) ± SD (%)			Mean ± SD
	Precision	Recall	F1-score	Support
B-ORG	82.85% ± 8.80%	59.64% ± 6.65%	68.90% ± 4.28%	37.6 ± 7.3
I-ORG	66.98% ± 27.06%	44.05% ± 16.12%	50.01% ± 14.30%	16.8 ± 7.2
B-SS	72.82% ± 7.10%	47.35% ± 5.89%	57.30% ± 6.10%	70.2 ± 8.2
I-SS	61.04% ± 10.12%	49.04% ± 15.38%	52.77% ± 7.81%	121.2 ± 16.1
B-P	92.99% ± 2.43%	88.62% ± 2.80%	90.71% ± 1.55%	310.6 ± 34.4
I-P	86.29% ± 2.45%	79.65% ± 3.80%	82.75% ± 1.13%	181.8 ± 27.9
B-PF	80.00% ± 44.72%	20.97% ± 18.09%	31.87% ± 23.67%	6.0 ± 3.2
I-PF	49.95% ± 39.06%	32.40% ± 40.19%	33.92% ± 38.46%	13.8 ± 13.7
B-PP	88.78% ± 6.44%	73.96% ± 4.96%	80.46% ± 2.73%	125.0 ± 15.0
I-PP	76.28% ± 8.14%	68.60% ± 10.21%	71.32% ± 3.05%	145.0 ± 22.9
B-TF	92.07% ± 2.00%	80.91% ± 2.65%	86.12% ± 2.20%	221.2 ± 10.5
I-TF	89.37% ± 6.52%	76.90% ± 2.27%	82.53% ± 2.64%	115.4 ± 20.4
B-TNF	90.19% ± 7.00%	49.54% ± 10.20%	63.44% ± 8.70%	28.2 ± 5.5
I-TNF	84.21% ± 18.29%	41.52% ± 29.16%	49.04% ± 18.78%	24.0 ± 10.1

Support refers to the mean number of instances per entity in the test set across 5 runs with different random states (5, 10, 15, 20, and 25)

Entity-level performance was additionally evaluated using strict span matching to complement this token-level analysis. The entity-level is considered correct if both the entity boundary and type exactly match the ground truth. Table 7 presents the Mean ± SD of entity-level weighted-average and macro-average Precision, Recall, and F1-score for each feature combination across training epochs. Consistent with the token-level results, Feature 3 achieved the highest entity-level macro-F1 across most epochs, reaching 65.11% ± 5.92% at epoch 100, the highest among

all configurations, reaffirming its selection as the recommended feature combination. Most macro-F1 scores at the entity level were only slightly lower than the corresponding token-level macro-F1scores in Table 3. For this recommended configuration, the entity-level macro-F1 was just 1.10 percentage points below its token-level of 66.21% ± 5.96%. This small difference suggests that, in this study, entity boundaries were relatively precise once an entity was correctly detected, indicating that the model's limitations lie more in detection sensitiviy than in boundaries precision. The gap between the weighted-average and macro-average F1-scores at the entity level was larger than at the token level, suggesting that the stricter boundary-matching criterion is more sensitive to performance disparities across entity types, particularly for minority classes. To further examine the source of this disparity, an entity-level confusion matrix was constructed for the best-performing configuration, i.e., Feature 3 at epoch 100, aggregated across the five random state as shown in Table 8, allowing a more detailed breakdown of prediction errors for each entity type.

Table 6. Mean ± Standard Deviation (SD) of Per-Tag (Token-level) Precision, Recall, and F1-score for Feature 3 at Epoch 50

Entity	Mean (%) ± SD (%)			Mean ± SD
	Precision	Recall	F1-score	Support
B-ORG	84.44% ± 6.86%	60.61% ± 7.40%	70.11% ± 4.13%	37.6 ± 7.3
I-ORG	63.17% ± 24.93%	44.76% ± 15.11%	49.54% ± 12.01%	16.8 ± 7.2
B-SS	73.17% ± 5.25%	48.42% ± 8.28%	58.11% ± 7.43%	70.2 ± 8.2
I-SS	62.59% ± 8.79%	49.67% ± 14.43%	54.02% ± 7.62%	121.2 ± 16.1
B-P	92.98% ± 1.93%	88.85% ± 2.73%	90.83% ± 1.63%	310.6 ± 34.4
I-P	85.31% ± 3.09%	78.25% ± 3.89%	81.54% ± 1.95%	181.8 ± 27.9
B-PF	80.00% ± 44.72%	20.97% ± 18.09%	31.87% ± 23.67%	6.0 ± 3.2
I-PF	64.00% ± 40.99%	30.33% ± 41.63%	34.20% ± 40.79%	13.8 ± 13.7
B-PP	89.63% ± 6.01%	76.11% ± 5.13%	82.18% ± 4.10%	125.0 ± 15.0
I-PP	79.61% ± 6.64%	70.50% ± 7.36%	74.28% ± 2.05%	145.0 ± 22.9
B-TF	91.32% ± 1.95%	81.01% ± 2.39%	85.85% ± 1.91%	221.2 ± 10.5
I-TF	88.95% ± 6.50%	77.50% ± 4.75%	82.73% ± 4.61%	115.4 ± 20.4
B-TNF	91.76% ± 6.82%	53.71% ± 13.05%	67.11% ± 10.87%	28.2 ± 5.5
I-TNF	83.35% ± 18.02%	44.87% ± 22.00%	54.47% ± 14.80%	24.0 ± 10.1

Support refers to the mean number of instances per entity in the test set across 5 runs with different random states (5, 10, 15, 20, and 25)

The matrix reveals that errors for minority entities were overwhelmingly attributable to false negatives from missed detections rather than false positives from misclassification into other entity types. Specifically, of the 31 ground-truth PF instances, only 4 (12.9%) were correctly identified by the model (true positives), while 27 (87.1%) were false negatives, with the model instead predicting these tokens as "O". No PF instances contributed to false positives through misclassification into another entity type. Similarly, for TNF, 71 of 141 instances (50.4%) were correctly identified, while 69 (48.9%) false negatives, and a single instance (0.7%) was classified as P, contributing one false negative for TNF and one false positive for P. Across the matrix overall, misclassification between distinct entity types was comparatively rare relative to missed detections. For example, while 187 SS instances were missed entirely (false negatives), only 17 were misclassified as P, contributing 17 false negatives for SS and 17 false positives for P. In this study, this finding indicates the most errors came from false negatives due to failure to detect entities rather than from false positives caused by confusion between entity types. These results are consistent with the relatively high precision but low recall for these entities in Tables 4 to 6, e.g., B-PF precision of 80.00% but recall of only 20.97%. This result likely due to the limited number of training instances for entities such as PF and TNF as

shown in Table 1. This limitation constrains the CRF model's ability to learn the features that distinguish these classes apart from each other. This result aligns with prior findings that class imbalance in NER tends to make models predicting the dominant "O" label for minority entities, instead of confusing them with other entity types [28].

Table 7. Mean ± Standard Deviation (SD) of Entity-Level Weighted Average and Macro-Average Precision, Recall, and F1-score for Each Feature Combination at Different Training Epochs

Epoch	Features	Weighted-Average			Macro-Average		
		Precision	Recall	F1-score	Precision	Recall	F1-score
50	Feature 1	86.1±2.91	67.05±2.13	74.84±2.28	78.14±7.31	51.36±3.42	60.54±4.26
	Feature 2	84.97±2.90	73.45±1.81	78.34±2.06	74.54±6.79	55.85±3.77	62.82±4.60
	Feature 3	85.17±2.55	74.25±2.30	78.90±2.35	77.44±8.21	57.27±4.82	64.32±5.55
	Feature 4	84.67±2.44	73.97±1.93	78.46±1.93	78.9±8.62	56.94±4.10	64.32±4.77
	Feature 5	84.97±1.90	73.12±1.80	78.15±1.62	76.22±5.59	56.12±3.04	63.50±3.68
100	Feature 1	85.62±2.59	67.31±1.51	74.85±1.67	75.00±5.86	51.31±3.15	59.95±3.81
	Feature 2	84.57±2.26	73.8±1.67	78.44±1.80	73.40±4.62	56.83±3.71	63.18±3.66
	Feature 3	84.41±2.19	74.34±1.95	78.66±1.93	75.35±6.09	58.94±6.17	65.11±5.92
	Feature 4	84.31±2.32	74.37±1.82	78.57±1.73	76.11±6.54	57.09±3.61	63.79±4.26
	Feature 5	84.59±2.14	73.92±2.14	78.48±1.99	74.12±3.73	58.51±5.94	64.23±4.53
150	Feature 1	85.76±2.48	67.36±1.90	74.92±1.95	77.74±7.04	51.60±3.63	60.63±4.22
	Feature 2	84.81±2.23	74.15±1.71	78.75±1.75	73.67±4.28	57.18±3.71	63.52±3.57
	Feature 3	84.15±1.94	74.17±2.11	78.44±1.88	75.31±5.12	57.14±3.66	63.86±3.98
	Feature 4	84.23±2.54	74.11±1.95	78.42±1.99	72.97±4.21	56.57±4.10	62.75±4.05
	Feature 5	84.69±1.94	74.15±1.96	78.63±1.82	73.59±3.38	56.92±3.21	63.06±2.84

Table 8. Entity-Level Confusion Matrix for Feature 3 at Epoch 100 (Aggregated Across Five Random States)

Actual	Predicted							
	ORG	P	PF	PP	SS	TF	TNF	O
ORG	116	3	0	0	0	0	0	69
P	0	1323	0	6	2	1	0	227
PF	0	0	4	0	0	0	0	27
PP	0	3	0	424	0	3	2	193
SS	0	17	0	0	147	0	0	187
TF	0	2	0	1	0	892	0	211
TNF	0	1	0	0	0	0	71	69
O	29	139	3	101	83	102	9	-

As a supplementary robustness check, consistency across random states was examined for Features 2 and 3, the two best-performing configurations, at epochs 50, 100, and 150. The weighted-average Precision, Recall, and F1-score values, calculated from token-level predictions for each random state, are presented in Tables 9 and 10, while the corresponding interquartile ranges are presented in Tables 11 to 13. These results showed that Feature 2 was more consistent, with all values falling within the interquartile range across epochs. Meanwhile, Feature 3 showed more sensitivity to data partitioning, with outliers existed in Precision and F1-score at random states 15 and 25 across all three epochs. These results indicate that Feature 2 offers greater stability, whereas Feature 3 provides marginally higher and more balanced performance across entity types, though this difference was not statistically tested and may reflect variance across random states rather than a robust improvement. Feature 3 was therefore

chosen as the recommended configuration. However, its sensitivity to data partitioning is noted as a limitation for future work. These findings also suggest that contextual features alone are insufficient for the CRF-based model to effectively recognize clinical entities, and that combining contextual, prefix, and suffix features improves recognition by capturing both semantic meaning and word structure [26], as such clinical terms are often morphologically composed of informative prefixes, roots, and suffixes [29]. The further addition of orthographic features in Feature 3, capturing capitalization, digit patterns, and character-level formatting, provided a marginal additional improvement over Feature 2. This is likely because these features help the model recognize clinical term that look visually different from ordinary text, such as abbreviations, dosage notations, and standardized medical codes. This finding is consistent with previous studies confirming the effectiveness of orthographic features for clinical entity recognition, as biomedical texts frequently contain medical terms written in abbreviated form [25]. In contrast, the addition of POS-tagging and the stopword feature in Features 4 and 5 did not improve performance. For POS-tagging, this may be explained by the POS tagger having been trained on a general-domain Indonesian corpus [23] rather than clinical text, though its accuracy on clinical text was not directly evaluated in this study. For the stopwords feature, this may be due to the limited discriminative value of the binary stopword indicator when stopwords frequently appear within multi-word clinical entities, as describe in the Method (Feature Extraction subsection). The best-performing configuration as shown in Table 7, i.e., Feature 3, achieved a weighted F1-score of 78.90%, though macro-F1 was considerably lower at 64.32%, reflecting weaker performance on low-support entity types. Nevertheless, this level of performance on formal clinical text is consistent with the premise outlined in the Introduction, that standardized, professionally written text provides a more favourable context for entity recognition than informal or short-context sources [18].

Table 9. Token-Level Weighted Average Precision, Recall, and F1-score per Random State for Feature 2 at Different Training Epochs (50, 100, and 150)

Epoch	Metric	RS 5	RS 10	RS 15	RS 20	RS 25
50	Precision (%)	83.68	86.22	89.96	84.23	80.61
	Recall (%)	70.85	73.81	71.35	75.64	69.47
	F1-score (%)	75.96	78.69	78.46	78.38	73.99
100	Precision (%)	83.14	85.55	87.79	83.68	81.17
	Recall (%)	71.68	74.57	71.84	75.29	69.62
	F1-score (%)	76.41	79.23	78.24	77.93	74.43
150	Precision (%)	83.23	85.84	87.98	83.77	80.17
	Recall (%)	71.75	74.79	72.77	76.12	68.28
	F1-score (%)	76.54	79.50	78.87	78.48	73.24

Table 10. Token-Level Weighted Average Precision, Recall, and F1-score per Random State for Feature 3 at Different Training Epochs (50, 100, and 150)

Epoch	Metric	RS 5	RS 10	RS 15	RS 20	RS 25
50	Precision (%)	84.57	86.09	89.26	85.80	81.40
	Recall (%)	71.62	74.79	73.76	75.78	68.06
	F1-score (%)	76.96	79.26	80.00	78.84	73.44
100	Precision (%)	84.05	83.85	88.52	84.00	78.29
	Recall (%)	72.45	73.89	74.82	76.06	67.54
	F1-score (%)	77.17	77.92	80.12	78.44	72.09
150	Precision (%)	83.18	83.20	87.33	83.72	80.03
	Recall (%)	71.88	74.57	74.18	76.33	68.28
	F1-score (%)	76.41	77.94	79.39	78.76	73.09

Despite the positive results achieved by the feature combinations, further improvement in Recall is needed, as Recall scores were consistently lower than Precision across all features,

which limits the overall F1-score. This may be attributed to the limited number of instances for certain entity types in the corpus, which likely affects the model's ability to correctly identify all occurrences of those entities. Expanding the corpus with additional annotated data could potentially improve overall performance, particularly for entities with low support. Another limitation is that the corpus was derived from a single internal medicine journal, limiting generalizability to other clinical domains, journals, and writing styles. Future work should include data from multiple journals and domains to improve generalizability. A major limitation of this study is the single-annotator process. Thus, inter-annotator agreement could not be calculated, and annotator bias may not be fully eliminated despite supervision by a general practitioner. Future work should involve multiple annotators to enable inter-annotator agreement assessment. Additionally, sentence-level splitting may allow sentences from the same article to appear in both training and testing sets, potentially causing data leakage and exaggerating performance. Therefore, future studies should explore article-level splitting for a more robust evaluation. Moreover, investigating automated feature learning approaches is necessary to improve the efficiency of NER model development, as the present study relies on hand-engineered features. Such future work could use the current findings as a baseline for evaluating feature effectiveness.

Table 11. The Quartile, IQR, and Outlier Threshold of Token-Level Precision across all the random states for Feature 2 and 3 at different training epochs

Epoch	Features	Precision (%)				
		Q1	Q3	IQR	Q1 – (1.5 x IQR)	Q3 + (1.5 x IQR)
50	Feature 2	83.68	86.22	2.54	79.87	90.03
	Feature 3	84.57	86.09	1.52	82.29	88.37
100	Feature 2	83.14	85.55	2.41	79.53	89.16
	Feature 3	83.85	84.05	0.20	83.55	84.35
150	Feature 2	83.23	85.84	2.61	79.32	89.76
	Feature 3	83.18	83.72	0.54	82.37	84.53

Q1 = First Quartile; Q3 = Third Quartile; IQR = Interquartile Range; Q1 – (1.5 x IQR) = Lower Outlier Threshold; Q3 – (1.5 x IQR) = Upper Outlier Threshold

Table 12. The Quartile, IQR, and Outlier Threshold of Token-Level Recall across all the random states for Feature 2 and 3 at different training epochs

Epoch	Features	Recall (%)				
		Q1	Q3	IQR	Q1 – (1.5 x IQR)	Q3 + (1.5 x IQR)
50	Feature 2	70.85	73.81	2.96	66.41	78.25
	Feature 3	71.62	74.79	3.17	66.87	79.55
100	Feature 2	71.68	74.57	2.89	67.35	78.90
	Feature 3	72.45	74.82	2.37	68.90	78.37
150	Feature 2	71.75	74.79	3.04	67.19	79.35
	Feature 3	71.88	74.57	2.69	67.84	78.60

Q1 = First Quartile; Q3 = Third Quartile; IQR = Interquartile Range; Q1 – (1.5 x IQR) = Lower Outlier Threshold; Q3 – (1.5 x IQR) = Upper Outlier Threshold

Table 13. The Quartile, IQR, and Outlier Threshold of Token-Level F1-score across all the random states for Feature 2 and 3 at different training epochs

Epoch	Features	F1-score (%)				
		Q1	Q3	IQR	Q1 –(1.5 x IQR)	Q3 + (1.5 x IQR)
50	Feature 2	75.96	78.46	2.50	72.21	82.21
	Feature 3	76.96	79.26	2.30	73.51	82.71
100	Feature 2	76.41	78.24	1.83	73.66	80.98
	Feature 3	77.17	78.44	1.27	75.27	80.34
150	Feature 2	76.54	78.87	2.33	73.05	82.36

Epoch	Features	F1-score (%)				
		Q1	Q3	IQR	Q1 –(1.5 x IQR)	Q3 + (1.5 x IQR)
	Feature 3	76.41	78.76	2.35	72.88	82.29

Q1 = First Quartile; Q3 = Third Quartile; IQR = Interquartile Range; Q1 – (1.5 x IQR) = Lower Outlier Threshold; Q3 – (1.5 x IQR) = Upper Outlier Threshold

4. Conclusion

This study compared five hand-engineered feature combinations integrated with a CRF model for clinical NER in Indonesian clinical research articles, targeting seven clinical entity types. The combination of contextual, prefix and suffix, and orthographic features (Feature 3) achieved the best weighted F1-score of 77.70% and 78.90% at the token-level and entity-level, respectively, while contextual and prefix and suffix features (Feature 2) showed greater stability across random states. While combining contextual and morphological features is a promising approach for Indonesian clinical entity extraction, this study results highlight a few critical constraints. The low macro-average F1-score reveals that minority entities still suffer from class imbalance. Furthermore, because each article was annotated by a single annotator under general practitioner supervision, this study could not measure inter-annotator agreement. The dataset itself is taken from a single internal medicine journal using sentence-level splits, which creates potential data leakage and limits border availability. Addressing this limitation, future works should cover multi-annotator validation, article-level data splitting, corpus expansion across multiple journals and domains, and statistical testing to strengthen the robustness of feature comparisons.

References

[1] X. Jia, Y. Pang, and L. S. Liu, "Online Health Information Seeking Behavior: A Systematic Review," *Healthcare*, vol. 9, no. 12, p. 1740, Dec. 2021, doi: 10.3390/healthcare9121740.

[2] G. Balsarkar, "Art of Reading an Article in the Journal," *The Journal of Obstetrics and Gynecology of India*, vol. 72, no. 1, pp. 1–5, Feb. 2022, doi: 10.1007/s13224-021-01613-8.

[3] D. Ignasius, I. Novita Dewi, M. Bernadette Chayeenee Norman, and R. Rakhmat Sani, "Medical Named Entity Recognition from Indonesian Health-News using BiLSTM-CRF with Static and Contextual Embeddings," *Journal of Applied Informatics and Computing*, vol. 9, no. 6, pp. 2974–2985, 2025.

[4] V. Kocaman and D. Talby, "Accurate Clinical and Biomedical Named Entity Recognition at Scale," *Software Impacts*, vol. 13, p. 100373, Aug. 2022, doi: 10.1016/j.simpa.2022.100373.

[5] A. Dash, S. Darshana, D. K. Yadav, and V. Gupta, "A clinical named entity recognition model using pretrained word embedding and deep neural networks," *Decision Analytics Journal*, vol. 10, p. 100426, Mar. 2024, doi: 10.1016/j.dajour.2024.100426.

[6] N. Manoonwong, A. Muresan, A. R. Raavi, and X. Zhu, "An Empirical Study of Named Entity Recognition for Electronic Health Records," in *2024 IEEE International Conference on Knowledge Graph (ICKG)*, IEEE, Dec. 2024, pp. 227–234. doi: 10.1109/ICKG63256.2024.00036.

[7] Y. Li *et al.*, "Artificial intelligence-powered pharmacovigilance: A review of machine and deep learning in clinical text-based adverse drug event detection for benchmark datasets," *J. Biomed. Inform.*, vol. 152, p. 104621, Apr. 2024, doi: 10.1016/j.jbi.2024.104621.

[8] J. Kong, L. Zhang, M. Jiang, and T. Liu, "Incorporating multi-level CNN and attention mechanism for Chinese clinical named entity recognition," *J. Biomed. Inform.*, vol. 116, p. 103737, Apr. 2021, doi: 10.1016/j.jbi.2021.103737.

[9] Tedyyana, A., Ratnawati, F., & Kurniati, R. (2019). Rancangan sistem informasi penelitian dan pengabdian masyarakat Politeknik Negeri Bengkalis menggunakan metode UML (Unified Modeling Language). *Sistemasi: Jurnal Sistem Informasi*, *8*(3), 413-423.

[10] R. A. A. Jonker, T. Almeida, R. Antunes, J. R. Almeida, and S. Matos, "Multi-head CRF classifier for biomedical multi-class named entity recognition on Spanish clinical notes," *Database*, vol. 2024, Jul. 2024, doi: 10.1093/database/baae068.

[11] X. Li, H. Zhang, and X.-H. Zhou, "Chinese clinical named entity recognition with variant neural structures based on BERT methods," *J. Biomed. Inform.*, vol. 107, p. 103422, Jul. 2020, doi: 10.1016/j.jbi.2020.103422.

[12] A. Swamy and Srinath S., "POS Tagging and NER System for Kannada Using Conditional Random Fields," *International Journal of Information Retrieval Research*, vol. 11, no. 4, pp. 1–13, Oct. 2021, doi: 10.4018/IJIRR.2021100101.

[13] H. U. Rahman, T. Hanh, and R. Segall, "Disease Named Entity Recognition Using Conditional Random Fields," in *Proceedings of the 7th International Symposium on Semantic Mining in Biomedicine*, Potsdam, Germany, 2016, pp. 37–41.

[14] W. Khan, A. Daud, K. Shahzad, T. Amjad, A. Banjar, and H. Fasihuddin, "Named Entity Recognition Using Conditional Random Fields," *Applied Sciences*, vol. 12, no. 13, p. 6391, Jun. 2022, doi: 10.3390/app12136391.

[15] A. C. Gunawan, E. Elfitriani, P. H. Khotimah, and D. Kurniasari, "Multi-Class Named Entity Recognition for Health Information Extraction Using Indonesian Online News Headlines," Institute of Electrical and Electronics Engineers (IEEE), Jan. 2026, pp. 55–60. doi: 10.1109/ic3ina68387.2025.11325187.

[16] R. P. Kusumawardani and K. N. Kusumawati, "Named entity recognition in the medical domain for Indonesian language health consultation services using bidirectional-lstm-crf algorithm," *Procedia Comput. Sci.*, vol. 245, pp. 1146–1156, 2024, doi: 10.1016/j.procs.2024.10.344.

[17] F. Profesio Putra, G. Agung, P. Mahendra, A. Tedyyana, and M. Noor, "IMPROVING FAQ RETRIEVAL FOR ACADEMIC REGULATIONS USING SEMANTIC EMBEDDINGS AND LLM QUESTION AUGMENTATION."

[18] S. Zhang and N. Elhadad, "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts," *J. Biomed. Inform.*, vol. 46, no. 6, pp. 1088–1098, Dec. 2013, doi: 10.1016/j.jbi.2013.08.004.

[19] A. Elwing Torres, E. Silva de Moura, A. Soares da Silva, M. A. Nascimento, and F. Mesquita, "An experimental study on data augmentation techniques for named entity recognition on low-resource domains," *Natural Language Processing*, vol. 32, no. 2, pp. 184–203, Mar. 2026, doi: 10.1017/nlp.2026.10019.

[20] I. Montani, M. Honnibal, A. Boyd, S. Van Landeghem, and H. Peters, "spaCy: Industrial-strength Natural Language Processing in Python. Zenodo."

[21] L. A. Ramshaw and M. P. Marcus, "Text Chunking Using Transformation-Based Learning," 1999, pp. 157–176. doi: 10.1007/978-94-017-2390-9_10.

[22] Y. Wibisono, "POS Tagger Bahasa Indonesia dengan Python.," 2018.

[23] F. Rashel, A. Luthfi, A. Dinakaramani, and R. Manurung, "Building an Indonesian rule-based part-of-speech tagger," in *2014 International Conference on Asian Language Processing (IALP)*, 2014, pp. 70–73. doi: 10.1109/IALP.2014.6973521.

[24] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," 2003. [Online]. Available: https://api.semanticscholar.org/CorpusID:15019738

[25] N. Limsopatham and N. Collier, "Learning Orthographic Features in Bi-directional LSTM for Biomedical Named Entity Recognition," 2016.

[26] A. P. Ajees, K. Manju, and M. I. Sumam, "An Improved Word Representation for Deep Learning Based NER in Indian Languages," *Information*, vol. 10, no. 6, p. 186, May 2019, doi: 10.3390/info10060186.

[27] M. C. Hinojosa Lee, J. Braet, and J. Springael, "Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores," *Applied Sciences*, vol. 14, no. 21, pp. 9863–9883, Oct. 2024, doi: 10.3390/app14219863.

[28] S. Nemoto, S. Kitada, and H. Iyatomi, "Majority or Minority: Data Imbalance Learning Method for Named Entity Recognition," *IEEE Access*, vol. 13, pp. 9902–9909, 2025, doi: 10.1109/ACCESS.2024.3522972.

[29] Open Resources for Nursing (Open RN), "Medical Terminology, 2nd edition," 2024. Accessed: Dec. 18, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK607454/?report=printable