Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 2974~2985
e-ISSN: 2548-6861

Medical Named Entity Recognition from Indonesian Health-News using BiLSTM-CRF with Static and Contextual Embeddings

Darnell Ignasius [1]*, Ika Novita Dewi [2]*, Maria Bernadette Chayeenee Norman [3]*, Ramadhan Rakhmat Sani [4]*
* Study Program of Information System, Faculty of Computer Science, Universitas Dian Nuswantoro
112202206905@mhs.dinus.ac.id [1], ikadewi@dsn.dinus.ac.id [2], 112202206930@mhs.dinus.ac.id [3], ramadhan_rs@dsn.dinus.ac.id [4]

Article Info

Article history:

Received 2025-10-24
Revised 2025-11-06
Accepted 2025-11-12

Keyword:

Named Entity Recognition,
Medical Entity,
Word2Vec,
IndoBERT,
Indonesian Health News.

ABSTRACT

Named Entity Recognition (NER) is vital for structuring medical texts by identifying entities such as diseases, symptoms, and drugs. However, research on Indonesian medical NER remain limited due to the lack of annotated corpora and linguistic resources. This scarcity often leads to difficulties in learning meaningful word representations, which are crucial for accurate entity identification. This research aims to compare the effectiveness of static and contextual embeddings in enhancing entity recognition on Indonesian biomedical text. The experimental setup involved utilizing both static (Word2Vec) and contextual (IndoBERT) embeddings in conjunction with neural architectures (BiLSTM) along with Conditional Random Fields (CRF). The BiLSTM architecture was selected for its ability to capture bidirectional dependencies in language sequences. Specifically, four models: Word2Vec-BiLSTM, Word2Vec-BiLSTM-CRF, IndoBERT-BiLSTM, and IndoBERT-BiLSTM-CRF were evaluated to assess the impact of contextual representations and structured decoding. The models were trained on a manually annotated DetikHealth corpus, where specific medical entities such as diseases, symptoms, and drugs were labeled with the BIO-tagging scheme. Performance was subsequently evaluated based on standard metrics: precision, recall, and F1-score. Results indicate that IndoBERT's contextual embeddings significantly outperform static Word2Vec features. The IndoBERT-BiLSTM-CRF model achieved the highest performance micro-F1 0.4330, macro-F1 0.3297, with the Disease entity reaching an F1-score of 0.5882. Combining contextual embeddings with CRF-based decoding enhances semantic understanding and boundary consistency, demonstrating superior performance for Indonesian biomedical NER. Future work should explore domain-adaptive pretraining and larger biomedical corpora to further improve contextual accuracy.

I. INTRODUCTION

The rise of digital media has led to an explosion of health-related news articles. Many of these articles are typically written in an unstructured format, meaning that crucial information such as the names of diseases, specific symptoms, or drug mentions is embedded within free-form text without standardized tags, categories, or database-friendly fields. This lack of structure makes automated extraction and analysis particularly challenging. Extracting medical entities such as diseases, drugs, and symptoms from these unstructured texts is a crucial step in transforming narrative health information into structured data. Once structured, this information can be systematically analyzed to support clinical research, enable disease surveillance, and strengthen evidence-based decision-making in public health. Through such extraction, it becomes possible to detect outbreak trends automatically, identify frequently mentioned medications, and understand symptom progression in real-world discourse. Consequently, automatically identifying relevant entities from textual health data is critical for tasks

such as information retrieval, knowledge graph construction, and large-scale public health monitoring.

Named Entity Recognition (NER) directly addresses this challenge by identifying and categorizing specific spans of text, such as diseases, symptoms, or medical procedures, enabling the conversion of unstructured content into a structured format. Historically, NER systems were built upon symbolic rules or statistical models. However, these conventional approaches demonstrate limitations when faced with the dynamic terminology and long-range dependencies prevalent in real-world textual data, particularly within specialized domains like medicine [1], [2]. Recent researches have demonstrated that modern deep learning architectures can achieve substantial performance gains in NER, effectively addressing the limitations of earlier approaches and motivating the present investigation [3].

Conventional machine learning approaches to NER depend heavily on handcrafted features and domain specific lexicons. Rule based systems are interpretable and precise but require extensive manual effort and lack generality statistical models such as Hidden Markov Models or Conditional Random Fields (CRF) improved scalability, yet they still depend on feature engineering and large annotated datasets [4]. The limitations of these approaches are particularly evident in specialized domains such as medicine, where new terminologies emerge rapidly and labelled data are scarce. While the integration of symbolic knowledge with deep learning in hybrid methods has been explored to mitigate existing challenges, the imperative for robust Medical NER systems that demonstrate superior generalization across domains and precisely capture complex contextual nuances persists [5].

The emergence of deep neural has been revolutionary. Specifically, Bidirectional Long Short-Term Memory (BiLSTM) networks facilitate the assimilation of extensive contextual data by processing input sequences in both forward and backward directions, thereby enriching each token's representation with both antecedent and subsequent information [6]. To optimize the output sequence, the incorporation of a Conditional Random Field (CRF) layer is highly beneficial. This layer allows the model to leverage global label dependencies and ensure the generation of valid tag sequences, thereby enhancing the overall consistency and reliability of detected entity boundaries. However, simple stacking of BiLSTM and CRF may not fully exploit the representational capacity of transformer based embeddings [7]. For example, dynamic attention network (Dyn-Att Net) proposed by Hou et al. [8] for traditional Chinese medical NER. Their model, which rearranges the BERT-BiLSTM-CRF architecture to better capture semantic and sequential relations, achieved an F1-score of 81.91% and an accuracy of 92.06% on benchmark data. This outcome, along with similar enhancements observed in other specialized domains, highlights that careful architectural design is crucial for significantly boosting NER performance.

Pre-trained language models have emerged as the dominant foundational technology for modern NER. Specifically, models like BERT generate highly contextualized word embeddings, which are instrumental in capturing rich semantic nuances derived from their surrounding text [9]. In medical NER, hybrid architectures such as BERT-BiLSTM-CRF and BERT-BiGRU-CRF have demonstrated superior performance compared to earlier neural models. This advantage stems from their capacity to effectively leverage pre-training on extensive corpora, thereby capturing nuanced contextual representations crucial for the domain [10]. Furthermore, hybrid models that integrate attention mechanisms and trigger matching have demonstrated the capacity to further reduce data requirements. For instance, Tu et al. [11] showed that their attention-based NER model outperformed a traditional BiLSTM-CRF baseline using merely 20% of the training data. These advances collectively underscore the critical importance of contextual embeddings and prominently highlight the substantial potential of transfer learning, particularly in low-resource settings.

Research in Indonesian NLP has been accelerated by the release of IndoBERT [12] pre-trained model. IndoBERT is trained on a large corpus of Indonesian news and social media and has achieved state of the art results across a range of language understanding tasks. However, Indonesian presents unique challenges stemming from its linguistic diversity, which includes more than 700 regional languages, pervasive code-switching, and under-resourced dialects [13]. For example, NusaBERT extends IndoBERT by expanding the vocabulary and pre training on a multilingual corpus, this model demonstrates improved performance on tasks covering multiple Indonesian languages [14]. IndoBERTweet adapts IndoBERT for Twitter data by introducing domain specific vocabulary, the authors show that initializing new embeddings with averages of IndoBERT subwords yields better efficiency than projecting Word2Vec vectors [15].

Despite these advances, the majority of Indonesian NER researches continue to focus on general news or social media. While progress has been made, domain-specific efforts remain relatively limited. Notable recent contributions include the IPerFEX dataset, which targets personal financial entities and effectively demonstrates the utility of IndoBERT-BiLSTM-CRF models [15]. Another significant development is the TWCAM model, which integrates Transformers, Word2Vec, convolutional layers, and attention mechanisms to enhance NER in general Indonesian news. This model achieved an F1-score of 0.8178, a considerable improvement over a BiLSTM baseline score of 0.72 [16], [17].

While significant progress has been achieved in general-domain Indonesian NER, prior works have primarily focused on generic or social media texts and have rarely explored domain-specific contexts such as health journalism. Existing studies have demonstrated the effectiveness of hybrid deep learning architectures (e.g., BERT-BiLSTM-CRF) and multilingual biomedical models such as BioBERT and

ClinicalBERT, however, none have systematically examined their applicability to Indonesian medical texts. Furthermore, no previous research has conducted a comparative analysis between static and contextual embeddings for Indonesian biomedical NER, leaving an open question regarding which representation strategy is most effective for this morphologically rich and low-resource language. Another gap concerns the lack of a publicly available, manually annotated corpus that specifically captures medically relevant entities disease, symptom, and drug within the Indonesian health-news domain. These gaps collectively highlight the need for a domain-specific benchmark and an empirical comparison across embedding paradigms to establish a foundation for future biomedical NLP research in Indonesian. To address these issues, the present research introduces a new manually annotated DetikHealth corpus and evaluates four model configurations (Word2Vec-BiLSTM, Word2Vec-BiLSTM-CRF, IndoBERT-BiLSTM, and IndoBERT-BiLSTM-CRF) to investigate the relative contributions of contextual embeddings and structured decoding in Indonesian medical NER.

II. METHOD

This research adopts a supervised comparative framework to evaluate Word2Vec and IndoBERT based BiLSTM, and BiLSTM-CRF architectures for Indonesian medical NER. The research workflow, illustrated in Figure I, includes data preparation, embedding integration, model training, and performance evaluation.

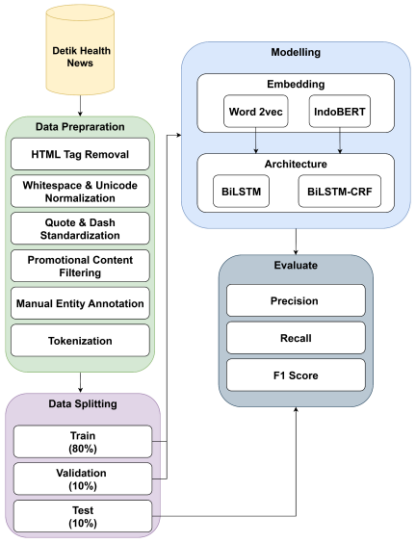

Figure I workflow of the proposed Indonesian health-news NER pipeline

A. Dataset

The dataset used in this research was developed to evaluate Named Entity Recognition (NER) systems on Indonesian health-news texts. The corpus was collected from DetikHealth, one of Indonesia's major online health news portals that regularly publishes articles on disease prevention, treatment updates, healthy lifestyles, and public health policies. Data collection was conducted over a two-week period preceding August 13, 2025, resulting in a total of 272 articles.

Each sentence was annotated with three medical entity types Disease, Symptom, and Drug using Label Studio, followed by tokenization with a rule-based Indonesian tokenizer. As illustrated in Figure II, the annotation workflow began with the definition of medical entity categories and the preparation of detailed annotation guidelines based on the BIO (Begin–Inside–Outside) tagging convention. The process continued with span-based tagging in Label Studio, performed by two trained undergraduate annotators from the Information Systems program. Prior to full annotation, both annotators received instruction and practice using guidelines derived from the ICD-10 (Ministry of Health, Indonesia), the Kamus Kesehatan Indonesia, and the BPOM RI national drug registry to ensure consistency in identifying medical terms. To evaluate reliability, a random 10% subset of sentences was double-annotated and inter-annotator agreement was measured using Cohen's Kappa ($\kappa = 0.81$), indicating substantial agreement. Discrepancies were resolved through discussion until consensus was reached [18]. The finalized annotations were then exported to BIO format for model training and evaluation, ensuring consistent labeling and linguistic coherence of medically relevant entities within the health-news context [13], [17].

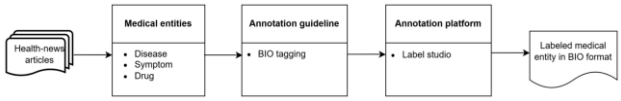

Figure II Entity annotation workflow

To ensure consistency and alignment with established medical terminology, the annotation guidelines for each entity category referred to authoritative national health sources. The Disease category followed the ICD-10 Indonesian adaptation issued by the Ministry of Health (Kementerian Kesehatan RI), which includes recognized pathological and infectious conditions such as *demam berdarah dengue*, *diabetes melitus*, and *tuberkulosis paru*. The Symptom category was defined based on terminology from the Pusat Data dan Informasi Kesehatan (Infodatin) and the Kamus Kesehatan Indonesia, encompassing clinical indicators such as batuk kering, sesak napas, and demam tinggi. Meanwhile, the Drug category was based on the Badan Pengawas Obat dan Makanan (BPOM RI) registry (Daftar Obat dan Bahan Aktif Terdaftar 2024), covering pharmaceutical substances such as parasetamol, amoksisilin, and ibuprofen.

Each entity label adhered to the BIO tagging standard, where "B-" denotes the beginning of an entity, "I-" indicates continuation within the same entity span, and "O" refers to non-entity tokens, providing a structured and linguistically coherent annotation framework.

B. Data Preprocessing

To prepare the raw news articles for modelling, a series of cleaning and normalization steps were applied. HTML tags and other nontextual elements were removed from the scraped pages, and the remaining content was converted to lowercase. Regular expressions were then used to normalize whitespace and unicode characters, and quotation marks and dash characters were standardized. Promotional or unrelated content (e.g., advertisements) was filtered out to retain only informational text [19]. Careful preprocessing is essential for NER because high quality, consistent input improves the ability of neural models to learn contextual patterns and is considered a prerequisite for reliable sequence labelling pipelines in biomedical NLP [20].

C. Data Splitting

To ensure a reliable evaluation process, the annotated corpus was divided into training (80 %), validation (10 %) and testing (10 %) subsets. To prevent information leakage across splits, a stratified anti leakage splitting strategy was applied. Sentences were first grouped by their source article, and a multilabel stratification algorithm was used to balance the distribution of the Disease, Symptom and Drug entities across the three subsets.

Long sentences were divided into overlapping chunks with a maximum length of 384 tokens so that entity boundaries were not fragmented. The splitting procedure was iterated over multiple random seeds and the configuration with the lowest variance in entity ratios across splits was selected. Such careful splitting ensures that the evaluation reflects model generalization rather than memorization and adheres to best practices for biomedical and health related NER task [21].

D. Word Embedding Method

Two embedding methods were investigated to assess the impact of static versus contextual word representations. The first approach used Word2Vec, a static embedding model trained on the training portion of the health news corpus using the skip gram architecture. Word2Vec captures semantic similarity by predicting neighboring words and produces dense vector representations that are invariant across contexts [22].

The second approach employed IndoBERT, a pre trained monolingual transformer model for the Indonesian language. IndoBERT provides contextual embeddings in which each word representation depends on its surrounding context, offering superior modelling of polysemy and long-range dependencies. For contextual embedding, IndoBERT's WordPiece tokenizer was applied to segment text into sub-word units before feeding the input into the model this approach preserves morphological information and handles out of vocabulary words effectively, as recommended in transformer based NER studies [23].

E. Model Architecture

NER models were constructed using Bidirectional Long Short-Term Memory (BiLSTM) networks, optionally combined with a Conditional Random Field (CRF) layer. BiLSTMs process sequences in both forward and backward directions, allowing the model to capture long distance contextual dependencies across tokens [24].When a CRF layer is added on top of the BiLSTM output, it models the joint probability of the entire label sequence and enforces valid tag transitions, providing a globally optimal decoding for sequence labelling tasks.

For reproducibility and comparability, the network hyperparameters were fixed across all experiments. Table I lists the values used for the embedding dimension, hidden size, number of BiLSTM layers, dropout rate, batch size, maximum epochs, learning rate, weight decay, gradient clipping threshold and early stopping patience. Setting fixed hyperparameters across models is a common practice in sequence labelling research, as it isolates the effects of embedding representations and architectural components. Models were trained using the AdamW optimizer on a single NVIDIA P100 GPU on Kaggle

TABLE I
PARAMETER SETTING OF BILSTM MODEL

Parameter Setting	Value
embedding dimension	300
hidden size	256
Layers	2
batch	32
learning rate	2×10^{-3}
early-stopping patience	6
Optimizer	AdamW
Max Epoch	12

Four model configurations were evaluated Word2Vec and BiLSTM, Word2Vec and BiLSTM-CRF, IndoBERT and BiLSTM, and IndoBERT and BiLSTM-CRF. The first two settings use static embeddings to examine the effect of adding a CRF decoding layer, while the latter two employ contextual embeddings to assess the impact of IndoBERT.

F. Evaluation

Model performance was evaluated using three standard NER metrics: precision, recall, and F1-score. These metrics were computed at both the token level and the entity level. At the entity level, a prediction is considered correct only if the predicted span exactly matches the ground-truth span in both boundaries and label token-level evaluation, by contrast, assesses each token independently and may reward partial matches [25].

To account for class imbalance among entity types such as frequent symptoms versus rare diseases we report both macro-averaged and micro-averaged F1-scores. Micro-averaging aggregates true positive, false positive, and false

negative counts across all entity classes before computing precision and recall, thereby weighting each prediction equally [26]. Macro-averaging computes the metrics separately for each class and then takes the unweighted mean, giving equal importance to both frequent and rare categories.

The formal definitions of these metrics are given in Equations (1)-(3), where TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively

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

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

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

These evaluation protocols are widely adopted in biomedical and clinical NER research and provide a robust and fair basis for comparing models with different embedding strategies and architectural designs [27].

III. Result and Discussion

A. Dataset Exploration

The final corpus comprises 272 health-related articles collected from DetikHealth, containing approximately 90,951 tokens written entirely in Bahasa Indonesia. Each sentence was manually annotated and resulted in three medical entity categories Disease, Symptom, and Drug.

As illustrated in Figure III, the dataset shows a clear imbalance in entity distribution, with Disease entities occurring most frequently, followed by Symptom and Drug. The corpus contains 87,965 (O) non-entity tokens, 898 (B-DISEASE) and 842 (I-DISEASE) disease tokens, 468 (B-SYMPTOM) and 530 (I-SYMPTOM) symptom tokens, and 139 (B-DRUG) and 109 (I-DRUG) drug tokens. This imbalance reflects the linguistic characteristics of Indonesian health journalism, where diseases are mentioned more often than symptoms or medications, and thus introduces a realistic challenge for model generalization in medical NER.

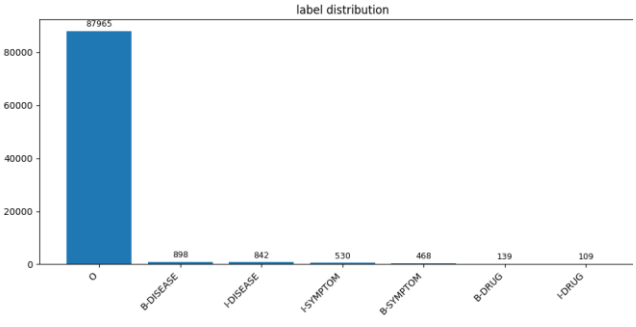

Figure III Label distribution across the DetikHealth corpus

To illustrate the annotation style and linguistic diversity of the corpus, Table II presents representative examples of annotated sentences using the BIO tagging scheme. These examples demonstrate how Indonesian medical terminology appears in compound noun phrases and colloquial phrasing, highlighting the morphological richness of the language and the contextual complexity encountered in automatic entity recognition.

Table II
Examples of Annotated Sentences from the DetikHealth Corpus.

Tokens	Labels	length
Hong Kong melaporkan kasus demam chikungunya ...	O O O O B-DISEASE I-DISEASE O...	427
Badan Pengawas Obat dan Makanan (BPOM RI) ...	O O O O O O O O O…	504
Belakangan game Roblox menjadi perbincangan ...	O O O O O…	279

Overall, the dataset provides a linguistically diverse and domain-specific benchmark for evaluating Indonesian NER systems in the healthcare domain. This analysis also establishes a quantitative foundation for the comparative evaluation of model architectures discussed in the subsequent sections

B. Performance result of Word2Vec-BiLSTM

The baseline configuration utilizing Word2Vec embeddings combined with a BiLSTM network was implemented to evaluate the fundamental capability of static word representations in identifying medical entities within Indonesian health-related texts.

As summarized in Table III, the model achieved a micro-average F1-score of 0.2777, with individual class performances of 0.3971 for Disease, 0.1503 for Symptom, and 0.0426 for Drug. The substantial difference between Disease and the two other categories (approximately 0.25-0.35 points) illustrates a pronounced class imbalance effect, where frequent and lexically distinct entities are captured more effectively. Precision and recall for the Disease class reached 0.3793 and 0.4167, respectively, indicating relatively stable detection, while both Symptom and Drug entities exhibited weaker precision-recall trade-offs due to their sparsity and semantic overlap with common non-entity expressions.

Table III
Entity Level Performance Metrics Of The Word2Vec-Bilstm Model On The Test Set.

	Precision	Recall	F1-Score	Support
Disease	0.3793	0.4167	0.3971	132
Drug	0.0370	0.0500	0.0426	20
Symptom	0.1238	0.1912	0.1503	68
micro avg	0.2491	0.3136	0.2777	220
macro avg	0.1801	0.2193	0.1967	220
weighted avg	0.2692	0.3136	0.2886	220

To further interpret these results, Figure IV presents the confusion matrix that visualizes token-level prediction distributions across the seven output classes, including entity boundaries (B-, I-) and the non-entity category (O). The matrix highlights a strong dominance of the non-entity (O) class, which accounts for 13,013 correctly predicted tokens, confirming that the model's learning dynamics are largely governed by the overrepresentation of neutral text. This heavy skew toward non-entity tokens causes the model to favor conservative predictions, where uncertain or contextually ambiguous tokens are defaulted to the O class. The highest confusion occurs between B-DISEASE and I-DISEASE, with 67 and 48 correctly predicted instances but 57 and 55 false transitions respectively, indicating that while the model identifies the presence of disease-related information, it fails to consistently recognize the start and continuation of multi-token spans. This irregular boundary detection is typical of BiLSTM architectures trained on unbalanced corpora, as the recurrent context window is insufficient to differentiate entity onset patterns from internal entity tokens, especially when both occur in similar syntactic positions.

Further analysis of Figure IV shows that Symptom entities exhibit scattered misclassifications, with 35 B-SYMPTOM and 38 I-SYMPTOM tokens predicted incorrectly as O. The model tends to under-recognize symptom terms that share overlapping linguistic structures with general descriptive expressions, leading to uncertainty in entity assignment. Additionally, 22 B-SYMPTOM tokens were confused with B-DISEASE, suggesting a systematic overlap between disease and symptom mentions where boundary distinctions are not clearly learned. This cross-entity confusion demonstrates that the static Word2Vec embeddings fail to encode contextual distinctions between medically related terms. The Drug entity class shows the weakest structural integrity: 14 B-DRUG and 13 I-DRUG tokens were misclassified as O, and only 2 instances of B-DRUG were correctly identified. This near-absence of correct Drug detection indicates that the BiLSTM relies primarily on surface-level frequency patterns rather than semantic cues, resulting in limited discrimination of low-frequency entities. The misclassification patterns in Figure 3 reveal that while the model captures partial semantic grouping for frequent classes, it lacks contextual precision and boundary consistency across minority medical entities.

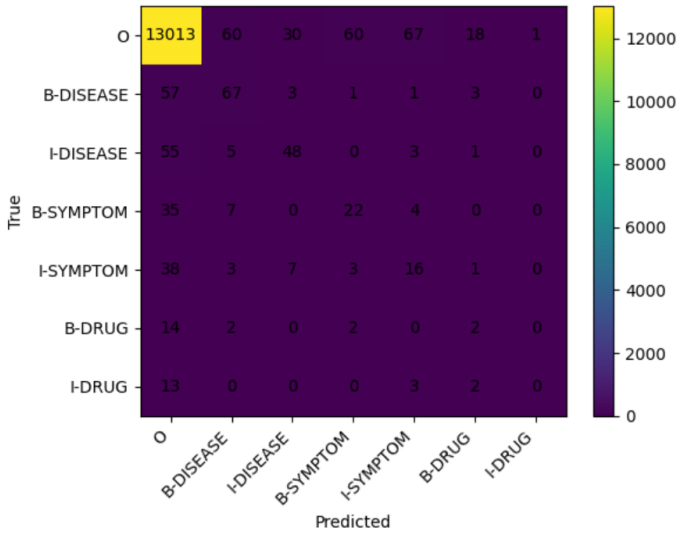

Figure IV Confusion matrix of the Word2Vec-BiLSTM model

A broader view of model behavior is provided by the Precision Recall (PR) curves illustrated in Figure V, which depict how the model's confidence changes across varying recall thresholds. The B-DISEASE and I-DISEASE curves occupy the largest area under the curve (AUC), maintaining precision above 0.8 up to recall levels of approximately 0.6. The B-SYMPTOM and I-SYMPTOM curves drop steeply beyond recall 0.4, demonstrating that the model rapidly loses discriminative ability as it attempts to generalize symptom terms. The B-DRUG and I-DRUG curves cluster near the lower left corner, representing near-random precision and confirming that low-frequency entities are effectively neglected during training. This visual trend reinforces the numerical results in Table IV, showing that static Word2Vec embeddings lack contextual adaptability, and that entity recognition is primarily driven by token frequency rather than semantic structure.

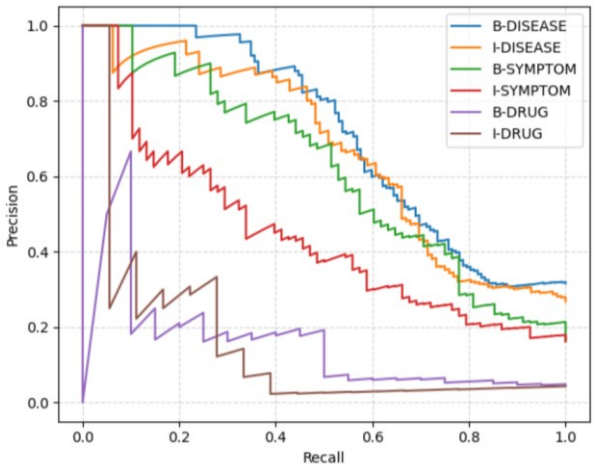

Figure V Precision-Recall curves for each entity category produced by the Word2Vec-BiLSTM model.

C. *Effect of the CRF Layer on the Word2Vec-BiLSTM Model*

The integration of a CRF layer atop the BiLSTM architecture was intended to assess whether structured decoding could enhance label sequence consistency and improve entity boundary detection. The performance outcomes presented in Table IV show that the Word2Vec-BiLSTM-CRF configuration achieved a micro-average F1-score of 0.2357 and a macro-average F1-score of 0.2387. At the entity level, the F1-scores were 0.3538 for Disease, 0.2909 for Symptom, and 0.0714 for Drug.

TABLE IV ENTITY LEVEL PERFORMANCE METRICS OF THE WORD2VEC-BILSTM-CRF MODEL ON THE TEST SET.

	Precision	**Recall**	**F1-Score**	**Support**
Disease	0.3594	0.3485	0.3538	132
Drug	0.0392	0.4000	0.0714	20
Symptom	0.3810	0.2353	0.2909	68
micro avg	0.1872	0.3182	0.2357	220
macro avg	0.2598	0.3279	0.2387	220
weighted avg	0.3369	0.3182	0.3087	220

Compared with the baseline, the inclusion of CRF slightly lowered the Disease F1 (from 0.3971 to 0.3538) but improved Symptom recognition (from 0.1503 to 0.2909) and notably raised the Drug recall (from 0.05 to 0.40). This suggests that the CRF layer better captures sequential dependencies between entity boundaries, particularly for multi-token expressions, although at the cost of overgeneralization in low-frequency classes. The enhancement in recall for Drug entities is accompanied by reduced precision (0.0392), implying that the layer increased sensitivity to entity cues but also introduced a higher false-positive rate in ambiguous contexts.

To further analyze these outcomes, Figure VI illustrates the confusion matrix visualizing token-level distributions across the seven output labels. The non-entity (O) class remains dominant, accounting for 12,794 correct predictions, which demonstrates that the model still prioritizes majority-class stability over entity differentiation. A clearer pattern of structured labeling can be observed within the Disease and Symptom categories: 55 B-DISEASE and 39 I-DISEASE tokens were correctly identified, indicating improved continuity of multi-token spans. However, 51 I-DISEASE tokens were still misclassified as O, showing that CRF alone could not fully eliminate context dilution. Symptom entities display better boundary consistency, with 19 B-SYMPTOM and 11 I-SYMPTOM predictions aligning correctly, though several false transitions remain. An unusual spike occurs in Drug-related predictions, where 167 non-entity tokens were misclassified as B-DRUG, explaining the rise in recall but sharp drop in precision. This shift demonstrates that while CRF strengthens sequential label coherence, it also amplifies mislabeling for classes with limited contextual examples.

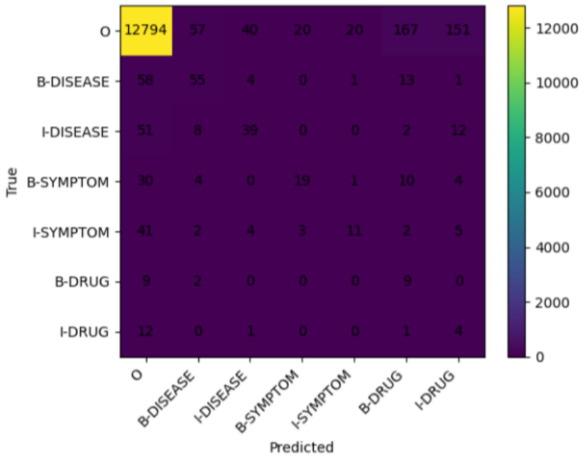

Figure VI Confusion matrix of the Word2Vec-BiLSTM-CRF model

The behavioral dynamics of precision and recall across entity types are further clarified in Figure VII, which presents the Precision-Recall (PR) curves for each tag. The curves for B-DISEASE and I-DISEASE appear smoother and maintain higher precision stability up to a recall of 0.6, signifying more reliable positive predictions compared with the baseline model. The B-SYMPTOM and I-SYMPTOM curves display broader coverage, indicating enhanced detection of symptom sequences with fewer fragmented boundaries. By contrast, the B-DRUG and I-DRUG curves remain unstable and concentrated near the lower-left corner of the plot, revealing erratic detection behavior driven by sparse data. The overall curve patterns suggest that the CRF decoding mechanism introduces structured prediction benefits but cannot fully compensate for the semantic rigidity of static Word2Vec embeddings.

These findings establish a stronger foundation for comparing the contribution of contextualized embeddings in the next configuration. The subsequent analysis therefore evaluates how replacing static representations with contextual embeddings from IndoBERT can further enhance semantic understanding and boundary precision in medical entity recognition

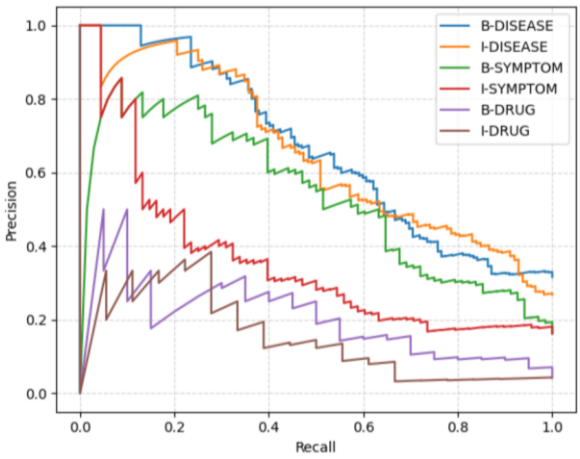

Figure VII Precision-Recall (PR) curves for each entity category produced by the Word2Vec-BiLSTM-CRF model.

D. Performance of the IndoBERT-BiLSTM Model

The adoption of contextualized embeddings through IndoBERT combined with a BiLSTM network marks a substantial shift from static to dynamic semantic representation. As shown in Table V, this configuration achieved a micro-average F1-score of 0.3937 and a macro-average F1-score of 0.3525, representing a clear improvement over both the Word2Vec-BiLSTM and Word2Vec-BiLSTM-CRF models.

Class-level results reveal F1-scores of 0.4632 for Disease, 0.3085 for Symptom, and 0.2857 for Drug, demonstrating consistent gains across all categories. The contextual embeddings enable the model to generalize more effectively to varied medical terms by capturing both lexical and syntactic dependencies. Precision values rose notably for Drug (from 0.0392 to 0.3333) and Symptom (from 0.1238 to 0.2417), while recall increased for Symptom from 0.1912 to 0.4265. These improvements suggest that IndoBERT's bidirectional attention mechanism successfully provides richer contextual cues, enhancing discrimination among overlapping medical entities. The improvement can also be attributed to IndoBERT's subword-level tokenization, which captures morphological and affix variations common in Indonesian medical terminology, allowing the model to better represent nuanced linguistic structures and achieve higher lexical adaptability.

TABLE V
ENTITY LEVEL PERFORMANCE METRICS OF THE INDOBERT-BILSTM MODEL ON THE TEST SET

	Precision	**Recall**	**F1-Score**	**Support**
Disease	0.4314	0.5000	0.4632	132
Drug	0.3333	0.2500	0.2857	20
Symptom	0.2417	0.4265	0.3085	68
micro avg	0.3472	0.4545	0.3937	220
macro avg	0.3355	0.3922	0.3525	220
weighted avg	0.3638	0.4545	0.3992	220

To complement the quantitative metrics, Figure VIII visualizes the distribution of token-level predictions, providing a clearer view of how contextual embeddings influence boundary consistency and inter-entity confusion. The confusion matrix indicates that the IndoBERT-BiLSTM system effectively reduces misclassification at entity boundaries compared with static embedding models. The non-entity (O) label remains dominant, with 13,024 correctly predicted tokens, yet its proportion of false positives against entity classes has decreased relative to earlier architectures. The Disease entity category demonstrates improved consistency, with 77 correctly classified B-DISEASE tokens and 75 I-DISEASE tokens, indicating stronger sequence continuity within disease mentions. Misclassifications between B- and I-boundaries were notably fewer (50 and 29, respectively), evidencing that the contextual encoder allows the BiLSTM to better capture intra-entity token relations.

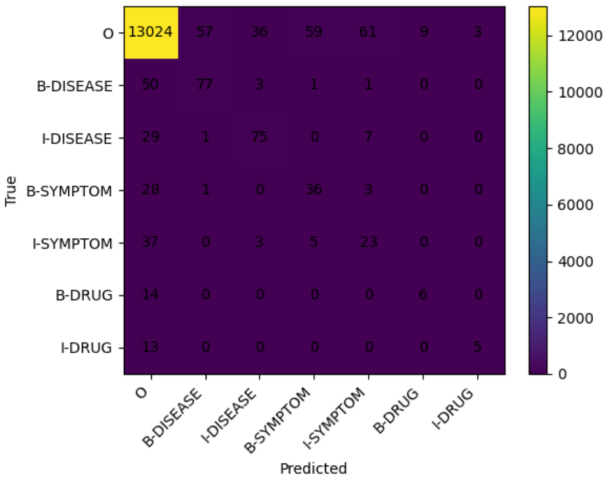

Figure VIII Confusion matrix of the Indobert-BiLSTM model

Similarly, Symptom entities show a marked improvement in boundary detection, with 36 B-SYMPTOM and 23 I-SYMPTOM tokens correctly predicted an indication that IndoBERT effectively distinguishes subtle contextual variations among symptom-related expressions. In contrast, the Drug entity still faces sparsity-related limitations, with only 6 correctly predicted B-DRUG tokens out of 20 samples, suggesting that while contextualization aids recognition, the scarcity of examples constrains precision stability.

A more nuanced interpretation of the model's behavior is captured in Figure IX, which depicts the Precision-Recall (PR) curves across all entity tags. The B-DISEASE and I-DISEASE curves show the largest and smoothest areas under the curve, maintaining precision above 0.8 across a broad recall range up to 0.7, confirming robust and consistent detection of disease terms. The B-SYMPTOM and I-SYMPTOM curves also expand significantly compared with previous models, indicating better trade-offs between recall and precision, especially in cases of overlapping or semantically related tokens.

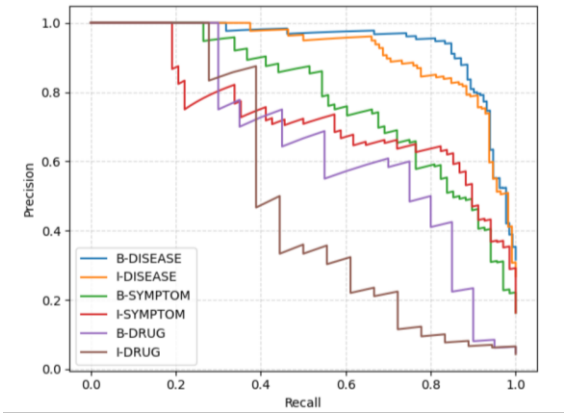

Figure IX Precision-Recall (PR) curves for each entity category produced by the Indobert-BiLSTM.

For the Drug class, the curves have improved curvature compared with the baseline, reflecting IndoBERT's contribution to semantic understanding, although instability persists due to limited training data. The overall PR patterns emphasize that contextual embedding integration enhances the discriminative capacity of the BiLSTM, yielding a more stable and semantically aware tagging process that better represents domain-specific nuances in Indonesian medical text. These consistent improvements suggest that contextual representations substantially enhance the sequential modeling capabilities of BiLSTM networks, motivating further evaluation of how structured decoding through CRF can refine label coherence and entity boundary precision.

E. Integrating CRF with IndoBERT-BiLSTM

The integration of a Conditional Random Field (CRF) layer on top of the IndoBERT-BiLSTM architecture was designed to combine the contextual understanding of transformer embeddings with structured sequence decoding. The quantitative results presented in Table VI show a micro-average F1-score of 0.4330 and a macro-average F1-score of 0.3297, representing a clear improvement compared to the IndoBERT-BiLSTM configuration without CRF (micro-F1 = 0.3937).

TABLE VI
ENTITY LEVEL PERFORMANCE METRICS OF THE INDOBERT-BILSTM-CRF MODEL ON THE TEST SET

	Precision	**Recall**	**F1-Score**	**Support**
Disease	0.5414	0.6439	0.5882	132
Drug	0.1250	0.1500	0.1364	20
Symptom	0.2066	0.3676	0.2646	68
micro avg	0.3742	0.5136	0.4330	220
macro avg	0.2910	0.3872	0.3297	220
weighted avg	0.4001	0.5136	0.4471	220

The Disease category achieved the highest F1-score of 0.5882, followed by Symptom at 0.2646 and Drug at 0.1364. These gains, particularly for Disease entities, demonstrate the positive interaction between contextualized embeddings and structured decoding, which enables the model to enforce valid label transitions within entity spans. Precision and recall also improved for Disease, from 0.4314 and 0.5000 to 0.5414 and 0.6439, respectively, reflecting a more balanced and confident classification process. This enhancement indicates that the CRF layer effectively consolidates IndoBERT's semantic features by learning transition probabilities between BIO tags, thus strengthening boundary coherence across sequences.

Beyond leveraging IndoBERT's WordPiece tokenizer, this research acknowledges the morphological complexity of the Indonesian language, which involves affixation, reduplication, and compounding that often obscure medical entity boundaries. Although explicit morphological preprocessing or customized tokenization rules were not implemented, IndoBERT's subword-level, context-dependent representations proved well suited to these linguistic characteristics. The model effectively handled multiword disease names, affixation patterns (such as *meN-*, *di-*, *-kan*, *-nya*), and colloquial variations commonly found in health news. By encoding each token relative to its surrounding words, IndoBERT successfully disambiguates polysemous terms for instance, distinguishing demam as a symptom from its idiomatic usage and reduces data sparsity through WordPiece segmentation.

This enables more accurate handling of rare or morphologically complex medical terms and spelling variants, yielding embeddings that are more discriminative between entity and non-entity spans. These capabilities explain the consistent improvement across all entity types and the particularly strong performance in recognizing Disease entities, where multi-token phrases occur frequently.

At the same time, the inclusion of the Conditional Random Field (CRF) layer plays a critical role in refining prediction consistency. The BIO tagging scheme imposes sequential constraints for example, I-DISEASE cannot validly follow O and the CRF layer explicitly models such transition regularities. This mechanism penalizes illegal label sequences and rewards coherent spans, addressing typical BiLSTM limitations such as fragmented entities, B/I label inversions, and the excessive prediction of the O class. The result is improved boundary integrity and higher recall for longer medical expressions such as demam berdarah dengue, where maintaining correct span structure is essential for semantic accuracy.

The combination of IndoBERT's contextual embeddings and CRF-based decoding thus creates a synergistic effect. IndoBERT contributes deep semantic representations that reduce token-level ambiguity, while the CRF layer enforces globally consistent label sequences over these representations. This complementary interaction stabilizes prediction boundaries, leading to a micro-F1 of 0.4330 and a Disease F1 of 0.5882, along with fewer B–I confusions and reduced misclassification into the non-entity category. In a low-resource setting like the present corpus of 272 articles with class imbalance, such synergy becomes particularly

valuable. IndoBERT distributes contextual information across morphologically related terms, while the CRF acts as a structured prior that enhances label coherence under limited supervision.

To complement these metrics, Figure X presents the confusion matrix, which provides a detailed view of token-level prediction distributions across the seven output labels. The non-entity (O) class remains dominant, with 12,962 correctly predicted tokens; however, the reduced number of false positives demonstrates improved model precision compared with previous configurations. The Disease entity category shows the most substantial refinement, with 96 B-DISEASE and 68 I-DISEASE tokens correctly identified and a marked decrease in misclassifications across adjacent labels.

This confirms that the CRF layer stabilizes boundary consistency, particularly in cases where entity spans are longer or semantically dense. The Symptom class also benefits from the structured decoding mechanism, achieving 35 correct B-SYMPTOM and 21 I-SYMPTOM predictions figures that indicate improved recognition of multi-token phrases related to physiological conditions. Although Drug entities remain the most challenging due to limited data, the model successfully recognized 5 I-DRUG and 5 B-DRUG instances, showing marginal but steady gains over the IndoBERT-BiLSTM model. These observations highlight that CRF's transition modeling alleviates label fragmentation and enhances intra-entity consistency, particularly in high-frequency classes.

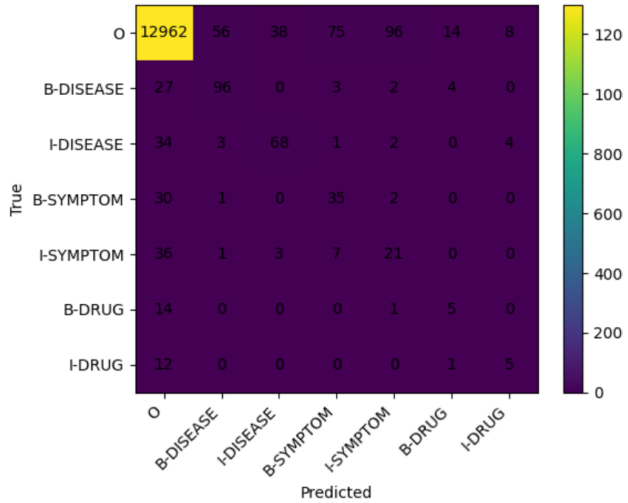

Figure X Confusion matrix of the Indobert-BiLSTM-CRF model

The behavior of the model across precision-recall trade-offs is illustrated in Figure XI, which shows smoother and more stable PR curves for all major entity categories. The B-DISEASE and I-DISEASE curves maintain precision above 0.8 for recall levels up to 0.6, indicating a more reliable and context-aware disease extraction process. The B-SYMPTOM and I-SYMPTOM curves exhibit broader coverage and reduced volatility, suggesting that structured decoding improves detection confidence for symptom-related expressions while mitigating false positives.

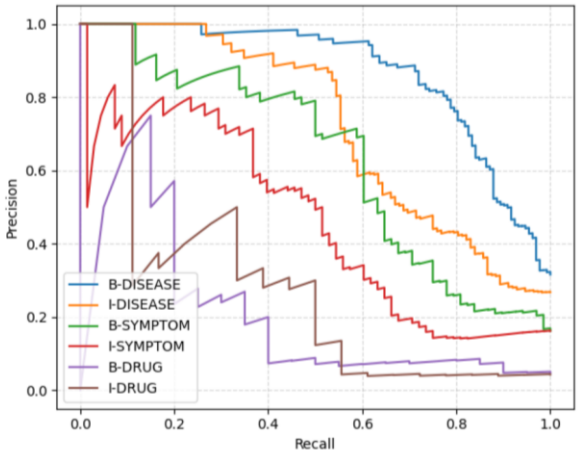

Figure XI Precision-Recall (PR) curves for each entity category produced by the Indobert-BiLSTM-CRF

The Drug curves, while still less stable, show visible improvement in curvature compared with earlier models, implying that the CRF enhances class boundary discrimination even for rare entities. IndoBERT's subword-level contextualization, combined with CRF's transition-level regularization, produces a complementary effect semantic depth from transformer embeddings is grounded by syntactic discipline through sequential decoding.

IV. Conclusion

This research systematically evaluated neural architectures for Indonesian medical named entity recognition, progressing from static embeddings to contextualized models with structured decoding. The results demonstrated that contextual embeddings from IndoBERT substantially outperformed Word2Vec, particularly in capturing complex entities such as diseases and symptoms. Incorporating a Conditional Random Field (CRF) further improved boundary consistency and reduced fragmented predictions. Among all configurations, the IndoBERT-BiLSTM-CRF model achieved the best performance, with a micro-average F1-score of 0.4330 and a macro-average F1-score of 0.3297, confirming the effectiveness of combining contextual semantics with structured decoding. Paired t-tests across model F1-scores indicated statistically significant differences ($p < 0.05$) between contextual and static embeddings, confirming that the observed performance gains were not due to random variation. However, these F1-scores remain relatively modest, largely due to the inherent complexity of medical entities which often appear as multiword or morphologically rich expressions and their low frequency in health-news data, making them more difficult for the model to generalize.

Despite these promising results, several limitations remain. The dataset is relatively small and imbalanced, which may limit generalization across diverse clinical narratives.

IndoBERT's pretraining corpus is not fully biomedical, reducing coverage for rare technical terms. Moreover, the evaluation focused mainly on sequence labeling performance, without testing downstream interpretability or clinical utility. Addressing these issues through domain-adaptive pretraining and larger annotated corpora is an essential direction for improvement.

Overall, the findings affirm that Indonesian biomedical text benefits from subword-level contextualization and sequence regularization, reflecting its linguistic complexity. Integrating semantic depth from transformers with CRF-based structural discipline yields coherent and context-aware entity recognition. Future studies should expand this approach with domain-specific corpora and hierarchical tagging to enhance semantic precision and support practical biomedical NLP applications such as information retrieval and knowledge extraction.

REFERENCES

[1] K. Pakhale, "Comprehensive Overview of Named Entity Recognition: Models, Domain-Specific Applications and Challenges," Sept. 25, 2023, *arXiv*: arXiv:2309.14084. doi: 10.48550/arXiv.2309.14084.

[2] S. Liu, Y. Sun, B. Li, W. Wang, and X. Zhao, "HAMNER: Headword Amplified Multi-Span Distantly Supervised Method for Domain Specific Named Entity Recognition," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 05, pp. 8401–8408, Apr. 2020, doi: 10.1609/aaai.v34i05.6358.

[3] R. Zhang, Y. Shan, and M. Zhen, "Advancing named entity recognition in interprofessional collaboration and education," *Front. Med.*, vol. 12, p. 1578769, June 2025, doi: 10.3389/fmed.2025.1578769.

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

[5] J. Li, A. Sun, J. Han, and C. Li, "A Survey on Deep Learning for Named Entity Recognition," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 50–70, Jan. 2022, doi: 10.1109/TKDE.2020.2981314.

[6] A. Khalid, G. Mustafa, M. R. R. Rana, S. M. Alshahrani, and M. Alymani, "RNN-BiLSTM-CRF based amalgamated deep learning model for electricity theft detection to secure smart grids," *PeerJ Comput. Sci.*, vol. 10, p. e1872, Feb. 2024, doi: 10.7717/peerj-cs.1872.

[7] R. Phukan, N. Baruah, M. Neog, S. KR. Sarma, and D. Konwar, "A Hybrid Neural-CRF Framework for Assamese Part-of-Speech Tagging," *IEEE Access*, vol. 13, pp. 160476–160489, 2025, doi: 10.1109/ACCESS.2025.3609572.

[8] J. Hou, S. Saad, and N. Omar, "Enhancing traditional Chinese medical named entity recognition with Dyn-Att Net: a dynamic attention approach," *PeerJ Comput. Sci.*, vol. 10, p. e2022, May 2024, doi: 10.7717/peerj-cs.2022.

[9] K. Kugler, S. Münker, J. Höhmann, and A. Rettinger, "InvBERT: Reconstructing Text from Contextualized Word Embeddings by inverting the BERT pipeline," Mar. 2024, doi: 10.48694/jcls.3572.

[10] Q. Qin, S. Zhao, and C. Liu, "A BERT-BiGRU-CRF Model for Entity Recognition of Chinese Electronic Medical Records," *Complexity*, vol. 2021, no. 1, p. 6631837, Jan. 2021, doi: 10.1155/2021/6631837.

[11] M. Tu, "Named entity recognition and emotional viewpoint monitoring in online news using artificial intelligence," *PeerJ Comput. Sci.*, vol. 10, p. e1715, 2024, doi: 10.7717/peerj-cs.1715.

[12] M. Kurniawan and B. Juarto, "Unlocking the Potential of IndoBERT for Classification of Indonesian Thesis," in *2024 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, Nov. 2024, pp. 785–791. doi: 10.1109/ICIMCIS63449.2024.10956541.

[13] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," Nov. 02, 2020, *arXiv*: arXiv:2011.00677. doi: 10.48550/arXiv.2011.00677.

[14] W. Wongso, D. S. Setiawan, S. Limcorn, and A. Joyoadikusumo, "NusaBERT: Teaching IndoBERT to be Multilingual and Multicultural," Mar. 04, 2024, *arXiv*: arXiv:2403.01817. doi: 10.48550/arXiv.2403.01817.

[15] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," Sept. 10, 2021, *arXiv*: arXiv:2109.04607. doi: 10.48550/arXiv.2109.04607.

[16] Warto, Muljono, Purwanto, and E. Noersasongko, "Improving Named Entity Recognition in Bahasa Indonesia with Transformer-Word2Vec-CNN-Attention Model.," *Int. J. Intell. Eng. Syst.*, vol. 16, no. 4, p. 655, July 2023, doi: 10.22266/ijies2023.0831.53.

[17] E. Dave and A. Chowanda, "IPerFEX-2023: Indonesian personal financial entity extraction using indoBERT-BiGRU-CRF model," *J. Big Data*, vol. 11, no. 1, p. 139, Sept. 2024, doi: 10.1186/s40537-024-00987-6.

[18] J. R. Dettori and D. C. Norvell, "Kappa and Beyond: Is There Agreement?," *Glob. Spine J.*, vol. 10, no. 4, pp. 499–501, June 2020, doi: 10.1177/2192568220911648.

[19] K. Smelyakov, D. Karachevtsev, D. Kulemza, Y. Samoilenko, O. Patlan, and A. Chupryna, "Effectiveness of Preprocessing Algorithms for Natural Language Processing Applications," in *2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T)*, Kharkiv, Ukraine: IEEE, Oct. 2020, pp. 187–191. doi: 10.1109/PICST51311.2020.9467919.

[20] Nurul Hannah Mohd Yusof, Nurul Adilla Mohd Subha, Nurulaqilla Khamis, and Norikhwan Hamzah, "Named Entity Recognition of an Oversampled and Preprocessed Manufacturing Data Corpus," *J. Adv. Res. Appl. Sci. Eng. Technol.*, vol. 36, no. 1, pp. 203–216, Dec. 2023, doi: 10.37934/araset.36.1.203216.

[21] H. R. Maier *et al.*, "On how data are partitioned in model development and evaluation: Confronting the elephant in the room to enhance model generalization," *Environ. Model. Softw.*, vol. 167, p. 105779, Sept. 2023, doi: 10.1016/j.envsoft.2023.105779.

[22] Q. Chen and M. Sokolova, "Specialists, Scientists, and Sentiments: Word2Vec and Doc2Vec in Analysis of Scientific and Medical Texts," *SN Comput. Sci.*, vol. 2, no. 5, p. 414, Aug. 2021, doi: 10.1007/s42979-021-00807-1.

[23] S. Sun, Q. Hu, F. Xu, F. Hu, Y. Wu, and B. Wang, "Medical named entity recognition based on domain knowledge and position encoding," *BMC Med. Inform. Decis. Mak.*, vol. 25, p. 235, July 2025, doi: 10.1186/s12911-025-03037-0.

[24] X. Zhang *et al.*, "Chinese medical named entity recognition integrating adversarial training and feature enhancement," *Sci. Rep.*, vol. 15, no. 1, p. 14844, Apr. 2025, doi: 10.1038/s41598-025-98465-3.

[25] C. J. Pinard, A. C. Poon, A. Lagree, K.-C. Wu, J. Li, and W. T. Tran, "Precision in Parsing: Evaluation of an Open-Source Named Entity Recognizer (NER) in Veterinary Oncology," *Vet. Comp. Oncol.*, vol. 23, no. 1, pp. 102–108, 2025, doi: 10.1111/vco.13035.

[26] T. Nguyen, D. Nguyen, and P. Rao, "Adaptive Name Entity Recognition under Highly Unbalanced Data," Mar. 10, 2020, *arXiv*: arXiv:2003.10296. doi: 10.48550/arXiv.2003.10296.

[27] A. Srinivasan and S. Vajjala, "A Multilingual Evaluation of NER Robustness to Adversarial Inputs," May 30, 2023, *arXiv*: arXiv:2305.18933. doi: 10.48550/arXiv.2305.18933.