Jurnal ELTIKOM :
Jurnal Teknik Elektro, Teknologi Informasi dan Komputer

Vol. 10, No. 1, June 2026, page. 99-117
ISSN 2598-3245 (Print), ISSN 2598-3288 (Online)
DOI: http://doi.org/10.31961/eltikom.v10i1.2116
Available online at http://eltikom.poliban.ac.id

DOMAIN ADAPTATION OF BERT MODELS FOR BIOMEDICAL ENTITY EXTRACTION FROM INDONESIAN HEALTH NEWS

Maria Bernadette Chayeenee Norman[1], Ika Novita Dewi[1,2*], Darnell Ignasius[2]
[1)] Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia
[2)] Research Center for Intelligent Distributed Surveillance and Security (IDSS), Universitas Dian Nuswantoro, Semarang, Indonesia
e-mail: mariabernadettecn@gmail.com, ikadewi@dsn.dinus.ac.id, ignasiusdarnell@gmail.com

Received: 16 December 2025 – Revised: 8 April 2026 – Accepted: 21 April 2026

ABSTRACT

Health-related news articles play an increasingly important role in public health monitoring. However, their unstructured linguistic style complicates the automatic extraction of biomedical information. Indonesian health news shows high lexical variation by combining medical terms, colloquial expressions, borrowed English words, and culturally specific symptom descriptions. This condition creates challenges for Named Entity Recognition (NER). To address the limited availability of domain-specific resources, this study compares four Transformer-based models, namely BERT, IndoBERT, RoBERTa, and BioBERT, for biomedical NER in Indonesian health news. A new BIO-annotated dataset consisting of 272 manually labeled articles was constructed and validated, achieving strong inter-annotator agreement (Cohen's Kappa = 0.88). To reduce data limitations, an additional 103 articles were automatically annotated using the best-performing model, RoBERTa, through a semi-supervised approach. All models were fine-tuned under identical settings and evaluated at both BIO and entity levels. The results show that RoBERTa achieves the highest weighted F1-score (0.9543). However, its macro F1-score (0.3873) indicates uneven performance across entity classes because of severe label imbalance, with non-entity tokens dominating the dataset. This finding highlights the importance of emphasizing macro-level evaluation to better reflect entity recognition performance. RoBERTa consistently outperforms the other models, which may be explained by its robust architecture and adaptability to diverse linguistic patterns. In contrast, BioBERT underperforms because of cross-lingual and domain mismatch, as it is pretrained on English biomedical corpora and optimized for scientific text rather than journalistic language. The error analysis further identifies boundary inconsistencies and under-detection of low-frequency entities, especially in the drug and symptom categories.

Keywords*: BIO tagging scheme, biomedical named entity recognition, domain adaptation, Indonesian health news, transformer models.*

I. INTRODUCTION

HEALTH-related news articles provide valuable insight into public health conditions by reporting emerging diseases, medication use trends, and symptom patterns across populations. Extracting structured information from these unstructured texts requires robust Named Entity Recognition systems that can accurately identify diseases, drugs, and symptoms [1], [2]. Transformer-based pretrained language models, such as BERT and RoBERTa, have shown significant improvements in NER performance across multiple domains [3], [4]. Public frameworks, including Hugging Face Transformers, have also supported standardized and reproducible fine-tuning pipelines [5]. Despite this progress, biomedical NER in Indonesian remains significantly underexplored, particularly in journalistic health narratives rather than clinical records or social media posts [6].

This gap is largely caused by linguistic variability and limited resources. Indonesian health news often combines formal medical terms with colloquial expressions, transliterated English terms, and culturally

specific symptom descriptions. For instance, influenza may appear interchangeably as *flu*, *pilek*, or *masuk angin*, depending on regional or editorial style [7]. Medication mentions may shift between generic names, brand labels, and abbreviations, which complicates entity boundary decisions. In addition, unlike English, Indonesian often omits tense or plurality markers, requiring models to rely on contextual cues rather than surface morphology. Existing pretrained Indonesian language models are mostly trained on general-domain corpora, such as Wikipedia, news headlines, or social media comments [8]. These domains rarely include clinical phrasing or pharmacological syntax, creating a vocabulary mismatch when the models are applied to biomedical content. As a result, pretrained models may misclassify short or ambiguous symptom terms, such as *pusing* or *sesak*, when they are used figuratively rather than medically [9].

The absence of reliable biomedical NER tools has several downstream implications. Without automated entity extraction, large-scale health news streams cannot be efficiently analyzed for epidemiological surveillance, misinformation detection, or public sentiment tracking [10]. Manual inspection is infeasible given the publication volume of major news portals. In addition, inconsistent entity recognition increases the risk of misinformative claims about drugs or diseases may circulate undetected [11]. Public health agencies, pharmacovigilance institutions, and digital health platforms would benefit from automated systems that can highlight rising mentions of specific diseases or detect inappropriate medication references [12]. Such systems, however, require NER models that are specifically adapted to the linguistic nuances of Indonesian biomedical news reporting.

To address this gap, a systematic evaluation of domain adaptation strategies for pretrained language models is needed. It remains unclear whether general-purpose multilingual models, such as BERT, can be effectively fine-tuned for Indonesian biomedical contexts or whether monolingual models, such as IndoBERT, offer stronger alignment because of their native lexical grounding [13], [14], [15]. In addition, cross-lingual domain transfer from English biomedical models, such as BioBERT, has not been validated in Indonesian health news [16], [17]. A unified comparison across these architectures is needed to determine whether domain adaptation, language specialization, or pretraining scale has the strongest influence on entity recognition performance.

The present study addresses this need by comparing several BERT-based models for biomedical entity extraction in Indonesian health news from platforms such as DetikHealth. A manually annotated dataset is constructed using the BIO scheme to provide consistent entity boundaries across disease, drug, and symptom categories. Four pretrained language models are fine-tuned under identical experimental settings: multilingual BERT as a general-purpose baseline, IndoBERT as a monolingual model trained on native corpora, RoBERTa as a variant with more extensive masked language modeling, and BioBERT as a domain-adapted biomedical model originally trained on English scientific literature [3], [4], [15], [16]. The contributions of this study are threefold. First, it introduces one of the first publicly structured biomedical NER datasets derived from Indonesian health news, enabling reproducible evaluation in future research. Second, it provides the first domain adaptation benchmark that compares multilingual, monolingual, and biomedical pretrained models within the same experimental protocol [17]. Third, it offers a detailed error categorization that highlights model-specific weaknesses, such as confusion between drug brands and generic names or the misclassification of non-medical symptom usage. The findings are expected to guide model selection strategies for Indonesian biomedical information extraction and support the development of real-time surveillance systems for health trend monitoring.

II. Research Method

This research was designed to examine how different BERT-based language models can be adapted to recognize biomedical entities in Indonesian health news. The overall process in-volved four main stages: data collection, entity annotation, model fine-tuning, and performance evaluation. The dataset was obtained from DetikHealth, a major online health news platform in Indonesia, through an automated web scraping process. The collected articles were then cleaned and prepared by removing unwanted elements, such as advertisements, HTML tags, and redundant formatting [18]. Each article was manually annotated using the BIO scheme to label entity types, namely diseases, drugs, and symptoms. The annotated data were split into training, validation, and testing sets, then used to fine-tune four transformer-based models, namely BERT, IndoBERT, RoBERTa, and BioBERT. Finally, the models were evaluated using standard NER performance metrics, including precision, recall, and F1-score [19].

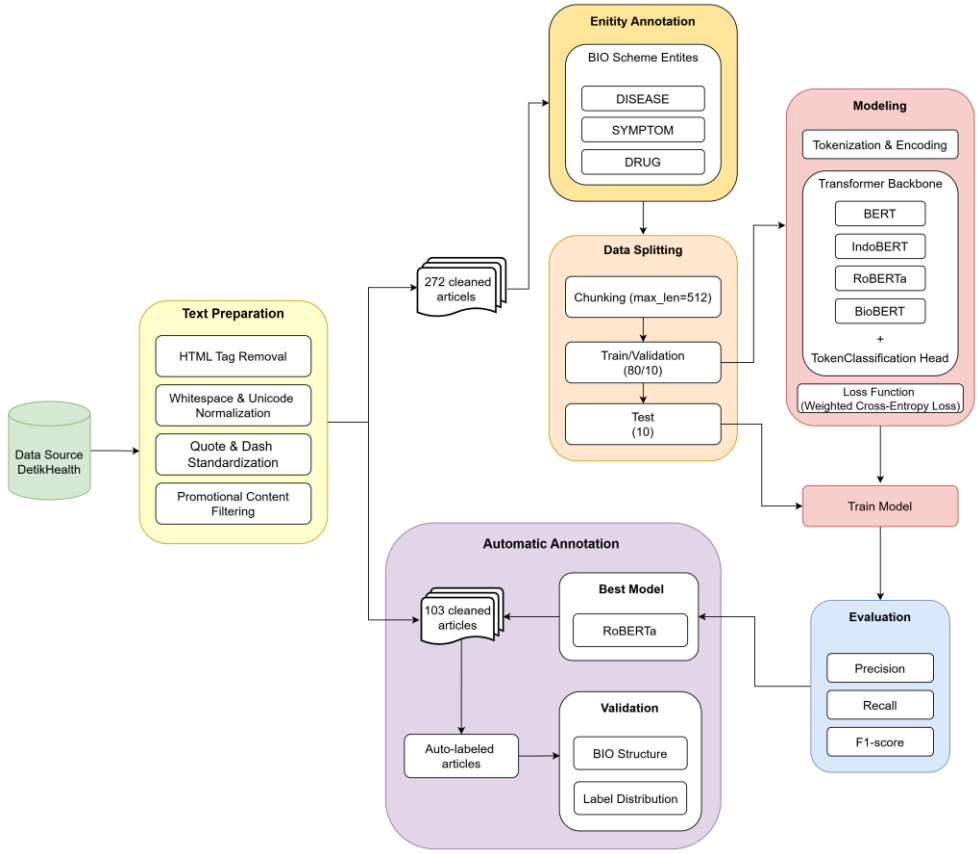

Figure 1. Research Flow

After the evaluation stage, an automatic annotation procedure was conducted using RoBERTa, the best-performing model. This step aimed to expand the corpus by generating automatically labeled health news articles. The resulting annotations were validated through BI-structure consistency checks and label distribution analysis to ensure linguistic coherence and structural reliability. The automatic annotation phase extended the manual dataset, supported scalability, and enabled broader applicability in downstream biomedical NLP tasks. The following subsections describe each step of the research workflow in Figure 1.

A. Data Collection and Text Preprocessing

The dataset used in this research was drawn from Indonesian online health news articles published on DetikHealth. Data collection took place over a fourteen-day period beginning on 13 August 2025, resulting in a total of 272 articles. DetikHealth was chosen as the primary data source because it consistently publishes reliable and up-to-date coverage of medical issues, public health trends, and emerging health phenomena in Indonesia. The platform's range of topics and consistent reporting style made it a representative resource for developing an NER system in the biomedical domain of the Indonesian language [20].

The web scraping process was implemented in Python using several supporting libraries, such as feedparser, BeautifulSoup, and fake_useragent. These tools improved the efficiency of HTML parsing, structured text extraction, and user-agent randomization to reduce request blocking during the automated crawling process [21], [22]. Each retrieved article was stored in JSON format and included essential metadata, such as the article title, publication date, and main text body. To maintain data quality, duplicate and incomplete entries were automatically filtered using string-matching techniques, ensuring that the final dataset contained only unique and complete articles.

After data collection, the corpus underwent a series of preprocessing steps as summarized in Figure 1. The steps included (1) removing HTML tags and special formatting, (2) normalizing whitespace and Unicode characters, (3) standardizing quotation marks and dash symbols, and (4) filtering out promotional or non-editorial content such as advertisements or sponsored materials. Once the text was cleaned, the manual annotation process was carried out using the BIO labeling scheme [23], [24].

TABLE 1
DEFINITION AND EXAMPLES OF BIOMEDICAL ENTITY CATEGORIES USED IN ANNOTATION

Entity Type	Definition	Examples
Disease	Refers to medical conditions, disorders, or illnesses, including formal clinical terminology and colloquial expressions used in Indonesian health news.	*hipertensi*, *diabetes melitus*, *flu*
Drug	Represents names of medications, including generic names, brand names, and commonly used abbreviations appearing in biomedical text.	*paracetamol*, Panadol, amoxicillin
Symptom	Denotes physical or physiological manifestations experienced or reported by individuals, commonly described in medical and layperson contexts.	*demam*, *pusing*, *mual*, *sesak napas*

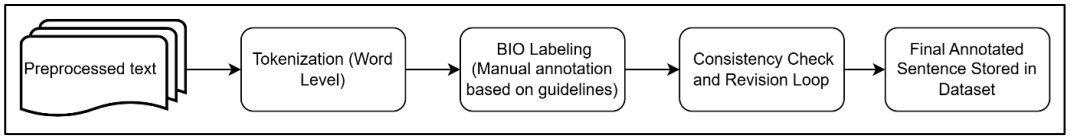

Figure 2. Assigning BIO Tags

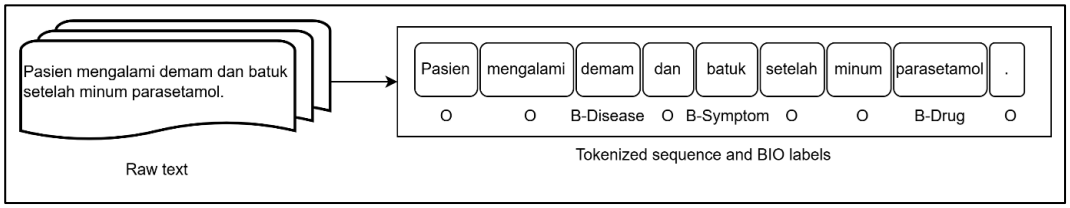

Figure 3. Example of manual BIO annotation workflow for Indonesian biomedical entities

B. Annotation Guidelines and BIO Scheme

After the preprocessing stage, manual annotation was conducted to label entities within the DetikHealth corpus. The annotation followed the BIO (Begin–Inside–Outside) tagging scheme, which is widely used in sequence labeling tasks such as NER. Each token in a sentence was as-signed a tag indicating whether it marked the beginning (B) of an entity, inside (I) an entity span, or was outside (O) any entity. This approach allowed the model to distinguish entity boundaries, including in multi-word expressions and contextually complex cases [25].

Three biomedical entity categories were defined in this study: Disease, Drug, and Symptom. Each entity type was annotated according to the contextual meaning of the term within the article, following the definitions summarized in Table 1.

Annotation was performed manually by the researcher using a token-level interface to ensure accuracy and contextual consistency, with BIO tags assigned to each token, as illustrated in Figure 2. Ambiguous or polysemous terms were resolved based on sentence context, following the manual BIO annotation workflow shown in Figure 3. For instance, the term *asam* was labeled as Symptom in *asam lambung* but tagged as O when appearing in *asam amino*. Similarly, *flu* was labeled as Disease when referring to an actual illness but remained O when used idiomatically.

The final annotated corpus comprised 272 DetikHealth articles stored in CoNLL format, where each line contained a token and its corresponding BIO label, and sentences were separated by blank lines. This dataset served as the foundation for model fine-tuning and evaluation in subsequent experiments.

The annotated corpus was stored in CoNLL format, where each token was aligned with its BIO tag. A Python preprocessing script was implemented to parse the corpus, split long sequences exceeding 512 tokens to fit BERT input limits, and calculate entity distribution [26]. The resulting dataset consisted of 272 news articles transformed into approximately 323 input sequences, annotated with BIO tags across three entity types. Label frequency analysis revealed that non-entity tokens dominated the dataset, while disease mentions were the most frequently labeled entities. This pattern confirmed the expected distribution in health-related news.

C. Model Selection and Domain Adaptation Strategy

This research employs a comparative modeling framework that fine-tunes four Transformer-based architectures, namely BERT, IndoBERT, RoBERTa, and BioBERT, to assess the efficacy of multilingual pretraining, language-specific optimization, and cross-lingual domain adaptation in Indonesian biomedical NER. All models are based on the Transformer encoder architecture, which uses stacked self-

TABLE 2
FINE-TUNING HYPERPARAMETERS CONFIGURATION

Parameter	Value
Batch size	16
Optimizer	AdamW
Learning rate	3e-5
Weight decay	0.1
Epsilon (eps)	1e-8
Epoch	20
Scheduler	Linear
Gradient clipping	1.0
Mixed precision (FP16)	True
Seed	42
Gradient accumulation	2 steps
Precision mode	FP16

$$Attention(Q,K,V) = softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \quad (1)$$

$$\widehat{y_t} = softmax(Wh_t + b) \quad (2)$$

$$\mathcal{L}_{MLM} = -\sum_{i\epsilon M} \log P(x_i | x \backslash M) \quad (3)$$

$$\mathcal{L}_{RoBERTa} = -\sum_{i\epsilon M_t} \log P(x_i | x \backslash M_t) \quad (4)$$

$$\mathcal{L}_{BioBERT} = \mathcal{L}_{MLM}^{\text{general}} + \mathcal{L}_{MLM}^{\text{biomedical}} \quad (5)$$

attention layers to represent text and capture long-range dependencies without relying on recurrent computations [27]. The self-attention mechanism formally computes contextual representations, as shown in (1). Each model incorporates a token-classification head implemented as a linear projection of the final hidden state, as presented in (2) where h_t is the contextual embedding of the token t, and y_t is the predicted BIO label. This unified formulation ensures consistent evalution of the effects of linguistic and domain-specific pretraining across model architectures. BERT-base-uncased serves as the baseline model and uses masked language modeling (MLM) and next sentence prediction (NSP) objectives [3]. The formula is presented in (3). IndoBERT-base-p1 builds on this architecture by using large-scale Indonesian text pretraining to evaluate the impact of language-specific corpora on NER performance [15]. RoBERTa-base improves upon BERT by removing the NSP objective and using dynamic masking and larger batch size optimization, resulting in more robust contextual representations [4]. The formula is presented in (4). BioBERT-base-v1.1 incorporates domain adaptation pretraining on distributed biomedical corpora, such as PubMed and PMC, allowing the model to encode specialized lexical and semantic patterns relevant to clinical terminology [16], as presented in (5). Comparing these models allows for a systematic analysis of how language specialization, architectural refinement, and domain-specific pretraining influence biomedical NER in Indonesian health news.

D. Fine-Tuning Configuration

To ensure fair performance comparisons, all models were trained using a uniform experimental configuration. Each model received tokenized input with a maximum length of 512 tokens. This input was processed by the Transformer layers to generate contextualized representations, which were then classified using a Softmax layer for each token. Training was carried out for 20 epochs using the AdamW optimizer and a linear scheduler with a warm-up ratio of 1% [28]. To maintain training stability, gradients were clipped to a maximum value of 1.0, and training was performed using mixed precision FP16. In addition, gradient accumulation was performed every two steps to balance computational efficiency and convergence stability [29].

To address the class imbalance inherent in NER tasks, where the “O” (Outside) label significantly outnumbers the “B” (Begin) and “I” (Inside) labels, Weighted Cross-Entropy Loss was applied. Class

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (6)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (7)$$

$$F1 - Score = 2\ \times \frac{Precision\ \times Recall}{Precision + Recall} \quad (8)$$

$$P_o = \frac{a\ +\ d}{N} \quad (9)$$

$$P_e = \ p_{A1}p_{A2} + (1 - \ p_{A1})(1 - p_{A2}) \quad (10)$$

$$\kappa = \frac{P_o - \ P_e}{1 - \ P_e} \quad (11)$$

weights were calculated in inverse proportion to the frequency of each class in the training data so that the model paid more attention to minority entity classes [30]. Reproducibility was maintained by setting a fixed seed (42) for all random components in Python, NumPy, and PyTorch, while disabling non-deterministic cuDNN operations. The dataset was divided into 80% training, 10% validation, and 10% testing splits [31]. All experiments were implemented using the Hugging Face Transformers library. Table 2 summarizes the main parameters used during the fine-tuning process.

E. Evaluation Metrics and Reproducibility

The final stage of this research is a quantitative evaluation of the fine-tuned NER models. This evaluation used only the test set to ensure an objective and unbiased assessment of model generalization [32]. Model performance was measured using standard NER metrics, namely precision, recall, and F1-score calculated at the entity level [19]. A prediction was counted as correct only when both the entity boundaries and label type were identified correctly. The mathematical formulations for each metric are presented in (6) - (8).

To provide an aggregated view of performance across entity categories, the macro-averaged F1-score was used because it gives equal weight to all classes, regardless of frequency. Each Transformer model was evaluated separately to allow clear comparisons of its effectiveness in handling biomedical Named Entity Recognition (NER) in Indonesian health news texts.

III. Results and Discussion

A. Results of Entity Annotation and Labeling Process

In this research, the entity annotation process was conducted manually by two annotators, both undergraduate students from the Information System program. They received detailed annotation guidelines adapted from biomedical entity standards and performed the annotation using the Label Studio platform. Entity categories were determined based on authoritative references. Disease entities were identified based on the Indonesian version of the International Statistical Classification of Diseases and Related Health Problems (ICD-10), published by the Ministry of Health of the Republic of Indonesia. Symptom entities were annotated using terminology from the Health Data and Information Center (Infodatin) and the Indonesian Health Dictionary to ensure terminological consistency within the local linguistic context. Drug entities were determined using the official list of registered pharmaceutical products issued by the Indonesian Food and Drug Authority (BPOM RI) to ensure alignment with nationally recognized drug nomenclature.

To ensure the reliability and consistency of the annotation process, we assessed the inter-annotator agreement (IAA) using Cohen's kappa coefficient (κ) [33]. The evaluation was performed on a 10% subset of the 272 total articles, corresponding to 9,096 tokens. Both annotators independently annotated all tokens in this subset following the established annotation guidelines. Cohen's kappa was selected because it accounts for agreement by chance, making it more reliable than simple percent agreement [34].

TABLE 3
CALCULATION OF AGREEMENT SCORES FOR ALL ENTITY CATEGORIES

Entities	N	Support A1	Support A2	Agreement Count	P_o	P_e	κ
Disease	9,096	455	472	9,084	0.9986	0.9624	**0.9650**
Symptom	9,096	312	298	9,088	0.9991	0.9782	**0.9593**
Drug	9,096	126	118	9,090	0.9993	0.9945	**0.8796**

TABLE 4
EXAMPLES OF TOKENIZED ARTICLES WITH BIO LABELS AND TOKEN LENGTHS

Token & Labels	Length
Hong_O Kong_O melaporkan_O kasus_O demam_B-DISEASE chikungunya_I-DISEASE …	427
Antibiotik_B-DRUG adalah_O obat_O yang_O digunakan_O untuk_O …	910
Penyanyi_O Malaysia_O Mohd_O Shah_O Rosli_O terkena_O stroke_B-DISEASE …	204
Kepala_O Badan_O Pengawas_O Obat_O dan_O Makanan_O (_O BPOM_O …	307
Gout_B-DISEASE atau_O penyakit_B-DISEASE asam_I-DISEASE urat_I-DISEASE adalah_O kondisi_O …	453

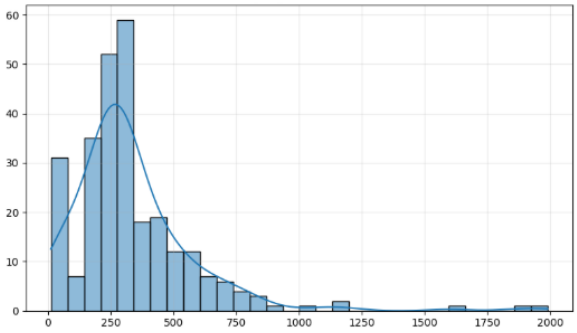

Figure 4. Distribution of Articles Length

The kappa value was computed for each entity type, namely Disease, Symptom, and Drug, using contingency tables and marginal probabilities for both annotators. Observed agreement (P_o) represents the proportion of tokens assigned identical labels by both annotators. Expected agreement (P_e) reflects the probability of random agreement based on the annotator's label distribution [35]. The formulas applied in this calculation are presented in (9) - (11).

Here, a denotes the number of tokens labeled as an entity by both annotators, d denotes the number of tokens labeled as a non-entity by both annotators, and p_{A1} and p_{A2} correspond to the marginal probabilities of annotators 1 and 2, respectively. Table 3 summarizes the complete calculation results for all entity categories.

The resulting kappa values demonstrate a very high level of inter annotator agreement. The Disease entity category achieved a kappa (κ) of 0.965, while the Symptom category obtained a kappa of 0.959. Both values fall within the "almost perfect" agreement range based on the Landis and Koch (1977) classification. This indicates that the annotation guidelines were well understood and applied consistently by both annotators. Although the Drug entity category achieved a slightly lower kappa value of 0.880, it still falls within the substantial agreement category. This minor reduction was likely influenced by the relatively low frequency of drug entities in the dataset. Overall, the average kappa value was 0.935, confirming that the annotated dataset is reliable and suitable for developing Indonesian biomedical NER models.

B. Exploratory Data Analysis

An exploratory data analysis was conducted to examine the basic characteristics of the Indonesian health news corpus used in this research. The dataset comprises 272 articles that underwent tokenization and manual annotation using the Begin-Inside-Outside (BIO) scheme. This scheme produced two primary attributes: tokens with labels and text length. These attributes represent the number of tokens per article, as presented in Table 4.

Analysis of text length revealed substantial variation across articles, with an average of 334.38 tokens per article, a minimum of 12 tokens, a maximum of 1,991 tokens, and a standard deviation of 256.14 tokens. Figure 4 illustrates the distribution of article length and shows a right-skewed pattern. Most articles fall within the range of 200-500 tokens, with the highest frequency around 250 tokens. This

TABLE 5
DETAILED DISTRIBUTION OF TOKEN LABELS IN THE DATASET

Label	Number of Token	%
O	87,965	96.72
B-DISEASE	898	0.99
I-DISEASE	842	0.93
B-SYMPTOM	468	0.51
I-SYMPTOM	530	0.58
B-DRUG	139	0.15
I-DRUG	109	0.12
Total	90,951	100

TABLE 6
ENTITY-LEVEL PERFORMANCE METRICS OF THE BERT MODEL

	Precision	Recall	F1-Score	Support
DISEASE	0.3390	0.8550	0.4855	138.0
SYMPTOM	0.1230	0.4444	0.1927	54.0
DRUG	0.5000	0.3600	0.4186	25.0
micro avg	0.2691	0.6958	0.3881	217.0
macro avg	0.3207	0.5531	0.3656	217.0
weighted avg	0.3038	0.6958	0.4050	217.0

TABLE 7
BIO LABEL-LEVEL PERFORMANCE METRICS OF THE BERT MODEL

	Precision	Recall	F1-Score	Support
O	0.9929	0.9318	0.9614	5590.0
B-DISEASE	0.3311	0.7285	0.4553	70.0
B-SYMPTOM	0.1290	0.4615	0.2016	26.0
B-DRUG	0.4000	0.3333	0.3636	12.0
I-DISEASE	0.2680	0.7647	0.3969	68.0
I-SYMPTOM	0.0784	0.2857	0.1230	28.0
I-DRUG	0.6250	0.3846	0.4761	13.0
accuracy	0.9197	0.9197	0.9197	0.9197
macro avg	0.4035	0.5531	0.3656	5807.0
weighted avg	0.9661	0.9197	0.9389	5807.0

pattern suggests that the corpus contains relevant medical terminology, disease references, and symptom-related expressions, although it is dominated by relatively concise health news articles.

Of the 90,951 annotated tokens, 96.72% are non-entity tokens ("O"), while biomedical entities collectively contributed only 3.28%. The most frequent entity category was Disease at 1.91%, followed by Symptom at 1.09% and Drug at 0.27%. Table 5 shows the detailed distribution of token labels in the dataset. This distribution confirms a severe class imbalance problem, where the strong dominance of the "O" label significantly affects model learning. As a result, the model tends to prioritize non-entity predictions, producing high weighted F1-scores but relatively low macro F1-scores. This imbalance particularly affects low-frequency classes, such as Drug and Symptom, reducing the model's ability to generalize across all entity categories.

C. Results of BERT Implementation

The BERT model was implemented as the baseline system for biomedical NER using Indonesian health news articles. We evaluated the model using precision, recall, and the F1-score at two analytical levels: the BIO label level and the entity level. This evaluation assessed its ability to recognize linguistic patterns in biomedical terminology appearing in news text.

At the entity level, as shown in Table 6, the model demonstrated uneven performance across entity categories. The DISEASE entity achieved the best results with a recall of 0.8550 and an F1-score of 0.4855. This indicates that BERT was relatively effective in identifying disease-related terms, despite its low precision of 0.3390. This low precision reflects a high number of false positive predictions. Conversely, the SYMPTOM entity showed the weakest performance, with an F1-score of 0.1927 because of its low precision of 0.1230. This suggests that the model struggled to capture the diverse expressions of symptoms commonly found in health news articles. The DRUG entity showed relatively high precision of 0.5000 but limited recall of 0.3600, indicating that BERT was cautious when labeling drug-related expressions but still missed many occurrences. The micro-average score of 0.3881 and the macro-average score of 0.3656 indicate that the model's overall performance was affected by the prevalence of high-frequency classes, especially the DISEASE entity. These findings suggest that, although

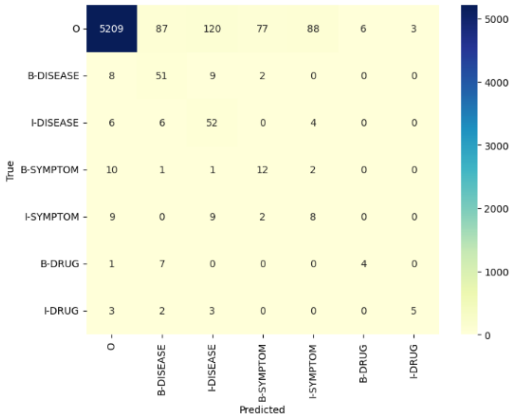

Figure 5. Confusion Metrix of the BERT Model

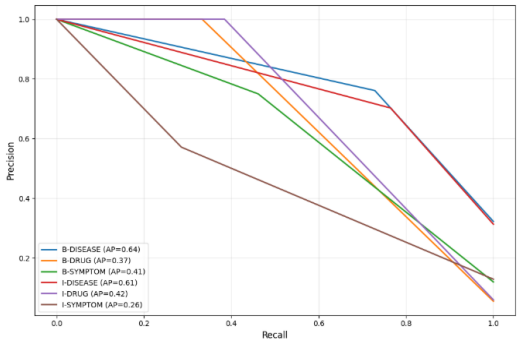

Figure 6. Precision-Recall Curve of the BERT Model

BERT could identify major biomedical entities with reasonable sensitivity, improvements are needed to enhance prediction precision and consistency across categories.

As shown in Table 7, evaluation at the BIO label level indicates that the I-DISEASE label has the highest recall of 0.7647, while I-DRUG has the highest F1-score of 0.4761. This implies that BERT can detect disease mentions with relatively strong sensitivity, although boundary-related errors still reduce label precision. In contrast, B-DRUG shows higher precision (0.4000) but lower recall (0.3333). This suggests that the model is conservative in identifying drug mentions and misses many true positive instances. B-SYMPTOM and I-SYMPTOM both show weaker performance than disease-related labels, with F1-scores of 0.2016 and 0.1230, respectively. This may be due to the limited variety of symptom expressions in the training corpus and the semantic overlap between symptom descriptions and disease-related contexts commonly found in health news texts. At the aggregate level, the macro-average F1-score of 0.3656 and the weighted average F1-score of 0.9389 indicate that the model's overall performance is strongly influenced by the dominance of the "O" (non-entity) label. This imbalance inflates the weighted F1-score, which may overestimate the model's actual capability in recognizing biomedical entities. Therefore, the macro F1-score is considered a more reliable metric for evaluating entity-level performance in this study.

Figure 5 shows a confusion matrix that provides deeper insight into BERT's error distribution at the token level. The "O" class clearly dominates, with 5,209 tokens correctly classified as non-entities. This confirms the influence of label imbalance in the training data. However, the main source of precision degradation is the misclassification of non-entity tokens as entity labels. This includes 87 "O" tokens mislabeled as "B-DISEASE", 120 as "I-DISEASE", and 77 as "B-SYMPTOM". These patterns suggest that the model tends to be overly aggressive in predicting biomedical entities and has difficulty differentiating relevant clinical terminology from general vocabulary, likely because of its limited contextual understanding of the domain.

Further analysis of the DISEASE entity reveals boundary detection issues. Although 51 "B-DISEASE" and 52 "I-DISEASE" tokens were correctly predicted, several transition errors occurred, such as 9 "B-DISEASE" tokens mislabeled as "I-DISEASE" and 6 "I-DISEASE" tokens mislabeled as "B-DISEASE". These inconsistencies show that the model has not fully captured the internal structure of multi-token disease entities. Meanwhile, symptom entities show more dispersed misclassification

TABLE 8
ENTITY-LEVEL PERFORMANCE METRICS OF THE INDOBERT MODEL

	Precision	Recall	F1-Score	Support
DISEASE	0.3390	0.8550	0.4855	138.0
SYMPTOM	0.1077	0.3333	0.1628	54.0
DRUG	0.5000	0.3200	0.3902	25.0
micro avg	0.2711	0.6635	0.3850	217.0
macro avg	0.3156	0.5028	0.3462	217.0
weighted avg	0.3000	0.6635	0.3943	217.0

TABLE 9
BIO LABEL-LEVEL PERFORMANCE METRICS OF THE INDOBERT MODEL

	Precision	Recall	F1-Score	Support
O	0.9910	0.9354	0.9624	5590.0
B-DISEASE	0.3184	0.8142	0.4578	70.0
B-SYMPTOM	0.1153	0.3461	0.1730	26.0
B-DRUG	0.4545	0.4166	0.4347	12.0
I-DISEASE	0.3136	0.7794	0.4472	68.0
I-SYMPTOM	0.0786	0.2500	0.1196	28.0
I-DRUG	0.6000	0.2307	0.3333	13.0
accuracy	0.9235	0.9235	0.9235	0.9235
macro avg	0.4102	0.5389	0.4183	5807.0
weighted avg	0.9647	0.9235	0.9402	5807.0

patterns. For example, 10 “B-SYMPTOM” and 9 “I-SYMPTOM” tokens were mislabeled as “O,” while 9 “I-SYMPTOM” tokens were incorrectly classified as “I-DISEASE”. These findings highlight the semantic overlap between symptom descriptions and disease expressions that the model has not yet differentiated effectively.

The DRUG entity had the weakest performance because of its low frequency in the training corpus. Only 4 “B-DRUG” and 5 “I-DRUG” tokens were correctly identified, while many others were misclassified as disease entities. Notably, 7 “B-DRUG” tokens were mislabeled as “B-DISEASE”. These errors suggest that the model relies more on general contextual cues than on deeper semantic representations of pharmacological terminology. Taken together, the misclassification patterns in the confusion matrix highlight the primary weaknesses of the baseline BERT model: difficulty distinguishing entities from non-entities and substantial semantic overlap across related biomedical entity categories.

The precision-recall curve in Figure 6 reinforces these findings. The B-DISEASE and I-DISEASE curves achieved the highest Average Precision (AP) scores of 0.64 and 0.61, respectively, indicating relatively stable performance in disease recognition. In contrast, the I-SYMPTOM curve yielded the lowest AP score of 0.26, showing persistent difficulty in distinguishing symptom expressions from non-entity descriptive phrases. The gradual decline across the curves further highlights the clear precision-recall trade-offs, particularly for low-frequency entity classes. Overall, these findings suggest that the baseline BERT model reliably identifies disease-related entities, although it still struggles to capture low-frequency biomedical expressions, especially symptoms and drug-related terms.

D. Results of IndoBERT Implementation

IndoBERT, a model designed specifically for the Indonesian language, was expected to overcome the limitations of multilingual BERT by offering richer linguistic representations that more accurately capture local morphological and syntactic structures. However, the experimental results indicate that these linguistic advantages do not automatically lead to better performance in biomedical NER. According to the entity-level evaluation in Table 8, IndoBERT achieved a macro-average F1-score of 0.3462, which is slightly lower than that of the baseline BERT model. This suggests that enhanced language understanding alone is insufficient to bridge the semantic gap between general domain Indonesian text and biomedical terminology.

Entity-level analysis shows that the DISEASE category remains the best performer, with a high recall of 0.8550 and an F1-score of 0.4855. However, its low precision of 0.3390 indicates a tendency toward over-detection, resulting in a substantial number of false positives. A similar pattern was observed for the SYMPTOM entity, which achieved a recall of 0.3333 and a precision of only 0.1077. This demonstrates IndoBERT’s difficulty in distinguishing symptom expressions from general descriptive phrases with similar lexical structures. Meanwhile, the DRUG entity achieved a moderate F1-score of 0.3902, likely because of the small number of training samples and weak semantic associations in pharmacological contexts.

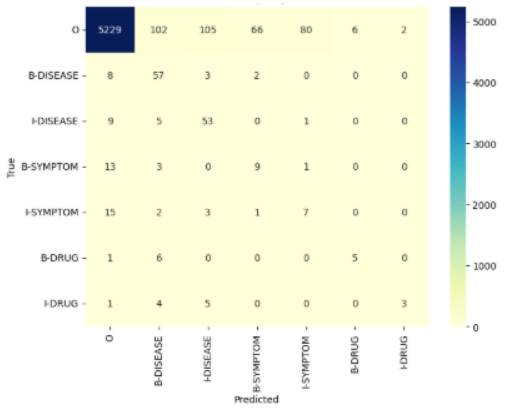

Figure 7. Confusion Metrix of the IndoBERT Model

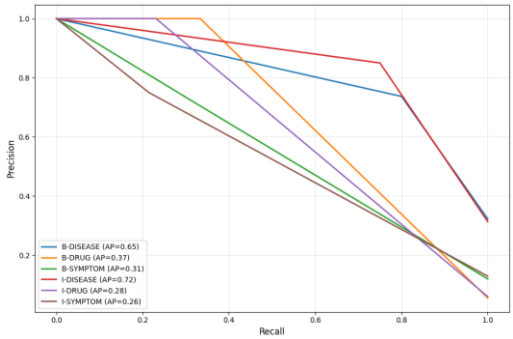

Figure 8. Precision-Recall Curve of the IndoBERT Model

The BIO-level evaluation in Table 9 provides a more granular view of IndoBERT's ability to detect entity boundaries. Although the B-DISEASE and I-DISEASE labels have high recall values of 0.8142 and 0.7794, respectively, their precision values remained low at 0.3184 and 0.3136. This resulted in moderate F1-scores. These results suggest that, while IndoBERT can identify disease mentions with reasonable sensitivity, it still struggles with accurate boundary detection. Conversely, the symptom-related labels, B-SYMPTOM and I-SYMPTOM, showed substantially lower performance, with F1-scores of 0.1730 and 0.1196. This reflects the linguistic complexity of symptom expressions in health news texts. Drug-related labels have somewhat better F1-scores, ranging from 0.3333 to 0.4347, but misclassifications remained common because of data imbalance and limitations in domain-specific vocabulary.

These findings are supported by the confusion matrix, as shown in Figure 7, which reveals error patterns similar to those observed in the BERT baseline model. The non-entity class "O" dominates the dataset, with 5,590 tokens in the test set, and also contributes the most misclassifications. IndoBERT mislabeled 102 "O" tokens as "B-DISEASE", 105 as "I-DISEASE", and 66 as "B-SYMPTOM", indicating a bias toward entity prediction driven by unresolved domain mismatch. Conversely, IndoBERT misclassified 15 "I-SYMPTOM" and 13 "B-SYMPTOM" tokens as "O", reflecting its difficulty in recognizing symptom-related expressions in ambiguous contexts.

As shown in Figure 8, the precision-recall curves reveal that IndoBERT achieves higher AP scores than the baseline BERT model for several entity labels, including B-DISEASE (0.65), I-DISEASE (0.72), and B-SYMPTOM (0.33). This indicates stronger class separability. However, these improvements are not fully reflected in the F1-score because of suboptimal confidence calibration in IndoBERT's predictions. Overall, although IndoBERT shows improved linguistic understanding of Indonesian text, its limited biomedical vocabulary and insufficient domain-specific representation continue to restrict precision improvement.

E. Results of RoBERTa Implementation

The evaluation of the RoBERTa model was conducted to determine the extent to which its extensive pretraining strategy and optimized architecture improve NER performance in the biomedical domain. The results indicate that RoBERTa consistently outperforms preceding models, especially in its ability to generalize across diverse entity categories.

TABLE 10
ENTITY-LEVEL PERFORMANCE METRICS OF THE ROBERTA MODEL

	Precision	Recall	F1-Score	Support
DISEASE	0.3924	0.8333	0.5336	138.0
SYMPTOM	0.2162	0.2962	0.2500	54.0
DRUG	0.5833	0.2800	0.3783	25.0
micro avg	0.3641	0.6359	0.4630	217.0
macro avg	0.3973	0.4698	0.3873	217.0
weighted avg	0.3706	0.6359	0.4451	217.0

TABLE 11
BIO LABEL-LEVEL PERFORMANCE METRICS OF THE ROBERTA MODEL

	Precision	Recall	F1-Score	Support
O	0.9898	0.9611	0.9753	5590.0
B-DISEASE	0.3684	0.8000	0.5045	70.0
B-SYMPTOM	0.1951	0.3076	0.2388	26.0
B-DRUG	0.5714	0.3333	0.4210	12.0
I-DISEASE	0.3617	0.7500	0.4880	68.0
I-SYMPTOM	0.1818	0.2142	0.1967	28.0
I-DRUG	0.6000	0.2307	0.3333	13.0
accuracy	0.9473	0.9473	0.9473	0.9473
macro avg	0.4669	0.5138	0.4511	5807.0
weighted avg	0.9658	0.9473	0.9542	5807.0

TABLE 12
ENTITY-LEVEL PERFORMANCE METRICS OF THE BIOBERT MODEL

	Precision	Recall	F1-Score	Support
DISEASE	0.3674	0.8333	0.5099	138.0
SYMPTOM	0.1192	0.3333	0.1756	54.0
DRUG	0.5000	0.2400	0.3243	25.0
micro avg	0.2920	0.6405	0.4011	217.0
macro avg	0.3288	0.4688	0.3366	217.0
weighted avg	0.3209	0.6405	0.4053	217.0

TABLE 13
BIO LABEL-LEVEL PERFORMANCE METRICS OF THE BIOBERT MODEL

	Precision	Recall	F1-Score	Support
O	0.9902	0.9443	0.9667	5590.0
B-DISEASE	0.3594	0.7857	0.4932	70.0
B-SYMPTOM	0.1111	0.3461	0.1682	26.0
B-DRUG	0.4444	0.3333	0.3809	12.0
I-DISEASE	0.3250	0.7647	0.4561	68.0
I-SYMPTOM	0.1000	0.2500	0.1428	28.0
I-DRUG	0.6666	0.1538	0.2500	13.0
accuracy	0.9312	0.9312	0.9312	0.9312
macro avg	0.4281	0.5111	0.4083	5807.0
weighted avg	0.9647	0.9312	0.9447	5807.0

At the entity level, as shown in Table 10, RoBERTa achieved a macro-average F1-score of 0.3873. The highest performance was observed for the DISEASE entity (F1 = 0.5336). Notable improvements were also recorded for the SYMPTOM (F1 = 0.2500) and DRUG (F1 = 0.3783) categories. These results show that RoBERTa can maintain a more balanced trade-off between precision and recall across entity types. Notably, the DISEASE entity showed high recall (0.8333) and improved precision (0.3924), suggesting a clear reduction in the false positives that dominated the predictions of BERT and IndoBERT.

Further analysis at the BIO level, as shown in Table 11, reveals clear structural improvements in boundary detection. The B-DISEASE and I-DISEASE labels achieved high recall values of 0.8000 and 0.7500, along with stable precision values of 0.3684 and 0.3617. These results produced F1-scores ranging from 0.4880 to 0.5045. This balance suggests that RoBERTa effectively preserves boundary consistency for the onset and continuation of disease entities. However, labels such as B-SYMPTOM and I-DRUG still show noticeable precision-recall gaps. This is likely due to uneven data distribution and semantic overlap among entities, which make minority entities more difficult for the model to capture.

The confusion matrix in Figure 9 further supports these findings by illustrating a substantial reduction in misclassification from the non-entity class ("O") to entity classes. Only 76 "O" tokens were incorrectly predicted as B-DISEASE, and 33 "O" tokens as B-SYMPTOM. This highlights the improved robustness of RoBERTa's contextual representation in distinguishing entity boundaries.

This improvement is reflected in the precision-recall curves shown in Figure 10, where RoBERTa maintains stable precision across different recall levels. The strongest results were achieved by the I-

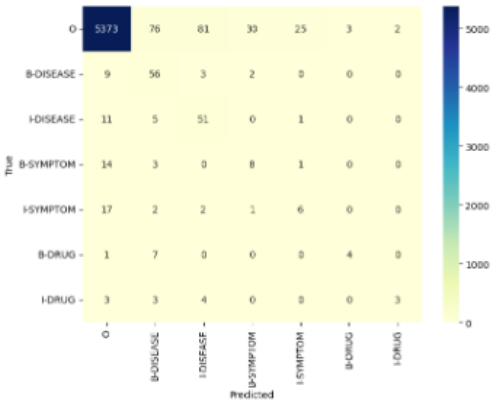

Figure 9. Confusion Metrix of the RoBERTa Model

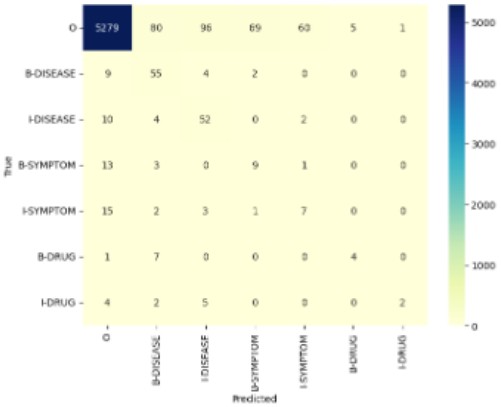

Figure 11. Confusion Metrix of the BioBERT Model

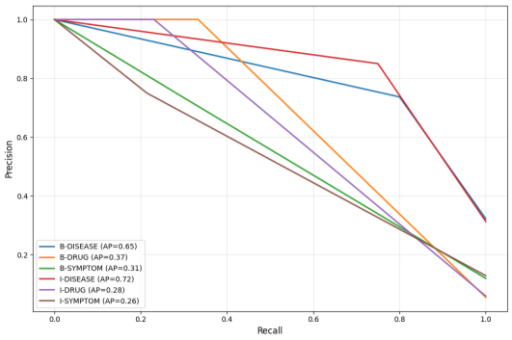

Figure 10. Precision-Recall Curve of the RoBERTa Model

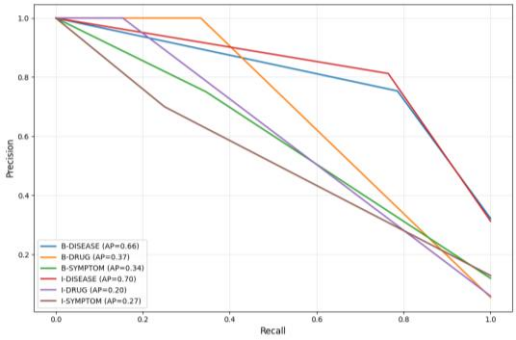

Figure 12. Precision-Recall Curve of the BioBERT Model

DISEASE (AP = 0.72) and B-DISEASE (AP = 0.68) labels. This demonstrates that RoBERTa produces better-ranked probability estimations and uses a more efficient decision-making mechanism to determine optimal classification thresholds.

Overall, the findings confirm that architectural refinements and larger-scale pretraining corpora have a greater influence than language specialization in enabling effective cross-domain adaptation. RoBERTa successfully overcomes several limitations of earlier models by providing more stable, generalized, and contextually adaptive semantic representations for biomedical NER.

F. Results of BioBERT Implementation

The evaluation of the BioBERT model aimed to assess the effectiveness of transferring biomedical domain knowledge from English pretraining corpora to Indonesian biomedical NER. BioBERT demonstrated strong capability in recognizing globally standardized medical concepts. However, its performance deteriorated substantially for entities that required contextual and language-specific understanding. This indicates that domain-specific pretraining can enhance the semantic representation of universal disease terminology.

As presented in Table 12, BioBERT achieved a macro-averaged F1-score of 0.3366 at the entity level, placing it between BERT and IndoBERT, although still below RoBERTa. BioBERT's primary strength was in the DISEASE category, where it attained an F1-score of 0.5099, with a precision of 0.3674 and a recall of 0.8333. These findings suggest that domain-specific pretraining can improve the semantic representation of universal disease terminology, especially terms that remain relatively stable across languages.

In contrast, BioBERT performed poorly on entities that rely on local linguistic variation. The SYMPTOM entity achieved an F1-score of only 0.1756, mainly because of its low precision of 0.1192, while the DRUG entity achieved an F1-score of 0.3243 with limited recall of 0.2400. These findings show that English-based domain transfer is insufficient for colloquial or region-specific Indonesian medical expressions. This creates a substantial lexical and cultural mismatch, restricting cross-language generalization.

A more detailed assessment at the BIO-level, shown in Table 13, reinforces this pattern. BioBERT performs well on disease-related labels, with B-DISEASE and I-DISEASE achieving F1-scores of

TABLE 14
SUMMARY OF AUTOMATIC ANNOTATION RESULTS (ROBERTA)

Metric	Value
Total Tokens	2,330
BIO Inconsistencies	15
Entity Proportion	6.70%
Dominant Entity	B-SYMPTOM (58 tokens)

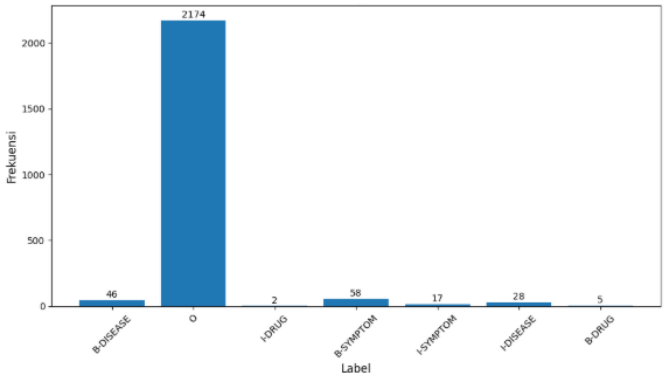

Figure 13. Label Frequency Distribution in Automatically Annotated Data

0.4933 and 0.4561, respectively, along with high recall values of 0.7857 and 0.7647. These results suggest that BioBERT consistently detects token-level patterns associated with disease entities. However, the model's performance drops sharply for labels requiring nuanced contextual comprehension. B-SYMPTOM and I-SYMPTOM both yield low F1-scores of 0.1682 and 0.1429, while I-DRUG records a recall of only 0.1538. This demonstrates the model's difficulty in capturing the structural variability of Indonesian multiword medical expressions.

The confusion matrix in Figure 11 further highlights these limitations through systematic misclassification patterns. A considerable number of false positives originate from "O" tokens, which are non-entities, being misclassified as biomedical entities. In total, 80 such tokens were misclassified as B-DISEASE, 96 as I-DISEASE, and 69 as B-SYMPTOM. These misclassifications reveal BioBERT's difficulty in aligning its domain-specific knowledge with the Indonesian linguistic context. Several false negatives were also found in symptom-related labels, with 13 B-SYMPTOM and 15 I-SYMPTOM tokens remaining unrecognized. Confusion between entity types, such as 7 B-DRUG tokens incorrectly predicted as B-DISEASE, further illustrates that BioBERT recognizes general biomedical semantics but struggles to differentiate finer entity distinctions.

The precision-recall curves in Figure 12 offer a closer look at these trends. Labels such as I-DISEASE (AP = 0.70) and B-DISEASE (AP = 0.66) maintained stable precision across a wide range of recall values. This suggests that the model can predict disease entities with relatively high confidence. Conversely, labels such as I-SYMPTOM (AP = 0.27), I-DRUG (AP = 0.20), and B-SYMPTOM (AP = 0.34) showed significant precision decreases as recall increased. This indicates an unfavorable trade-off, where broader entity retrieval substantially increases classification errors.

Overall, BioBERT's evaluation highlights the complex dynamics of cross-lingual domain adaptation. While the model successfully transfers universal biomedical knowledge, it is not sufficiently adaptive to the linguistic and semantic characteristics of Indonesian. These findings demonstrate that domain-specific pretraining alone does not guarantee strong cross-language performance without an explicit linguistic alignment mechanism. Consequently, BioBERT is an important step in understanding the limitations and potential of transformer-based models for multilingual biomedical NER.

G. Validation of Newly Annotated Data

A total of 103 additional health news articles scraped on November 18, 2025, were processed using RoBERTa, the best-performing model, to generate an automatically annotated biomedical Named Entity Recognition (NER) corpus. The model produced 2,330 labeled tokens with a BIO consistency error rate of only 0.64% (15 inconsistent B/I transitions), indicating stable structural annotation. A detailed overview of these results is presented in Table 14. The label distribution shows that non-entity tokens dominate at 93.3%, while disease and symptom entities are the most frequently detected categories, consistent with the semantic patterns found in Indonesian health articles. These tendencies are reflected in

TABLE 15
WEIGHTED AVERAGE PERFORMANCE COMPARISON ACROSS MODELS

	F1-Score	Precision	Recall
RoBERTa	0.9543	0.9658	0.9473
BioBERT	0.9447	0.9648	0.9313
IndoBERT	0.9402	0.9647	0.9235
BERT	0.9389	0.9662	0.9198

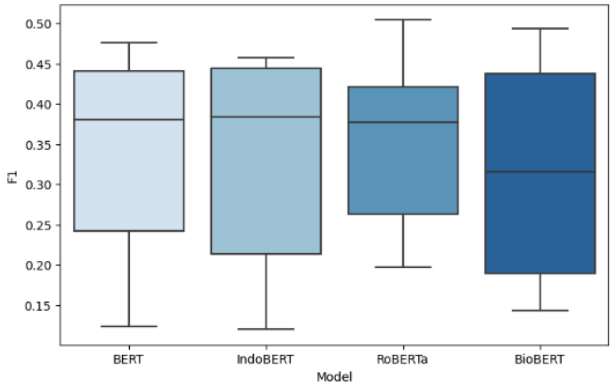

Figure 14. Boxplot of F1-Score Distribution Across Models

the specific token variants detected by the model, such as *deteksi*, *penyakit*, *stroke*, and *obesitas* under B-DISEASE; *pembuluh*, *darah*, and *jantung* under I-DISEASE; *toleransi*, *nyeri*, and *pemindaian* under B-SYMPTOM. Meanwhile, the identification of low-frequency biomedical terms such as Cesium-137 and Cs-137 as B-DRUG demonstrates the model's capacity to capture rare but medically relevant vocabulary.

Overall, the results demonstrate that RoBERTa generalizes effectively to unseen data while preserving the entity distribution characteristics observed in the initial experiment. The label distribution visualization in Figure 13 confirms that the automatically generated entity boundaries are linguistically coherent. However, several anomalies persist, including the mislabeling of biomedical expressions as "Symptom," a pattern that suggests limited Indonesian pretraining in the biomedical domain and leads to semantic ambiguity for medically nuanced terms. Therefore, additional pretraining on Indonesian biomedical corpora is recommended to improve lexical precision. These results show that the automatic annotation stage reliably supplements the manually curated dataset, thereby strengthening the overall availability of Indonesian biomedical NER resources.

H. Discussion

1) Comparison of All Models

A comparative evaluation was conducted across all models to assess the effectiveness of Transformer-based architectures for Indonesian Biomedical NER. Table 15 presents the evaluation results based on precision, recall, and F1-score using the weighted average approach. RoBERTa achieved the highest overall performance with an F1-score of 0.9543, followed by BioBERT (0.9447), IndoBERT (0.9402), and BERT (0.9389). RoBERTa's consistent superiority across all three metrics demonstrates its ability to balance predictive accuracy (precision) and comprehensive entity retrieval (recall). However, aggregate weighted metrics do not fully capture model performance stability across entity types with varying levels of difficulty.

To address the limitations of aggregate metrics, Figure 14 presents a boxplot showing the distribution of F1-scores across entity categories. This analysis reveals model consistency beyond average values. BERT and IndoBERT have similar distribution patterns, with a median F1-score of 0.38, wide interquartile ranges, and lower whiskers that fall at very low values (0.12 to 0.15). These characteristics indicate instability, where the models perform well on entities such as Disease but struggle with more complex categories such as Symptom. This pattern suggests limited robustness and a strong dependence of performance on entity type.

In contrast, RoBERTa exhibits a concentrated distribution with a narrow interquartile range and a much higher lower whisker (0.20), despite its entity-level median of 0.37. This demonstrates that RoBERTa offers both strong aggregate performance and consistent behavior across entity types. Among all

TABLE 16
SUMMARY OF ERROR PATTERNS ON INDONESIAN HEALTH NEWS NER

Error Type	Description	Actual Examples from Experiments
Over-detection	General tokens incorrectly labeled as medical entities.	Words such as "perawatan" and "peningkatan kasus" were predicted as B-SYMPTOM by RoBERTa and BioBERT.
Under-detection	Medical entities missed and labeled as O.	"hipertensi pulmonal" was labeled O by IndoBERT, although the correct labels were B-DISEASE I-DISEASE.
BIO Boundary Error	Incorrect B/I transitions for multi-word entities.	"gagal ginjal akut" was predicted as B-DISEASE, B-DISEASE, I-DISEASE by BERT instead of B-I-I.
Entity-Type Confusion	Entity detected but assigned to the wrong category.	"parasetamol" was predicted as B-DISEASE by BioBERT; "pulmonary edema" was predicted as B-SYMPTOM rather than B-DISEASE.

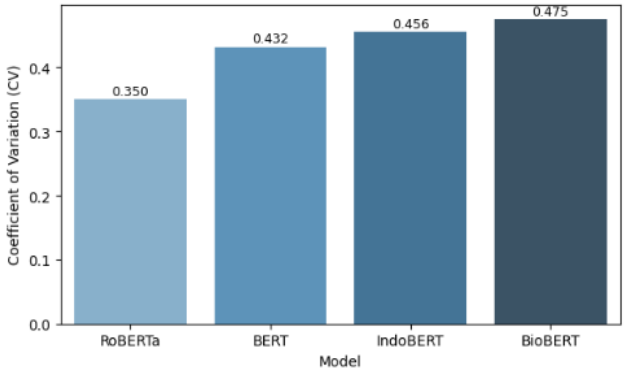

Figure 15. Coefficient of Variation Across Models

models, BioBERT exhibits the most polarized distribution. Its upper whisker extends to the highest value of 0.50, while its lower whisker drops to 0.14. This wide dispersion suggests that BioBERT functions primarily as a domain-specific model. It demonstrates strong performance on entities whose terminology is closely aligned with PubMed-style biomedical corpora, reflecting the effective transfer of domain knowledge from its English biomedical pretraining. However, its performance decreases markedly for Indonesian biomedical expressions that require deeper contextual interpretation and adaptation to linguistic variability. This contrast underscores BioBERT's limited robustness when applied to categories that rely heavily on local language structures. Tokenization differences and stylistic mismatch further contribute to this performance gap, as BioBERT is optimized for English scientific text, whereas the dataset consists of Indonesian semi-formal journalistic language. This indicates that domain-specific pretraining alone is insufficient without proper alignment in language and text style. In other words, effective domain adaptation requires not only domain-specific knowledge but also alignment with the target language and text characteristics.

This observation is further supported by Figure 15, which shows the coefficient of variation (CV) for each model. With a CV of 0.350, RoBERTa demonstrates minimal performance variation across entity types. Meanwhile, BioBERT (0.475), IndoBERT (0.456), and BERT (0.432) exhibit substantially higher variance, reflecting their greater sensitivity to label-specific characteristics. Thus, RoBERTa is the most robust model because it maintains stable performance across diverse entity types without over-reliance on specific categories.

Taken together, these findings suggest that RoBERTa's enhanced architecture, featuring dynamic masking and a more diverse training corpus, enables stronger generalization to syntactic and semantic variability in Indonesian biomedical texts. Although BioBERT benefits from its English biomedical pretraining, particularly for disease entities, it shows weaker adaptability to linguistically localized categories, such as Symptom and Drug. Overall, evidence from both aggregate metrics and distribution-aware evaluations confirms that RoBERTa is the most effective and robust model for biomedical NER in Indonesian health news articles.

2) Error Analysis

An error analysis was conducted to examine the systematic misclassification patterns that emerged from the automatic annotation process and the final model predictions. The errors were found to be linguistically consistent rather than random. They were largely driven by the stylistic characteristics of

Indonesian health news and the limited coverage of medical terminology in pretrained models. During automatic annotation, inconsistencies mainly occurred in BIO boundary assignments for multi-word entities, such as “gagal ginjal akut” and “penyakit jantung koroner”. These inconsistencies introduced noise into the model training process.

Four dominant error categories were identified across all models, as summarized in Table 16: over-detection, under-detection, BIO boundary errors, and entity-type confusion. Over-detection frequently involved journalistic terms, such as "perawatan" or "peningkatan kasus", that were incorrectly labeled as B-SYMPTOM or B-DISEASE. Under-detection commonly occurred with low-frequency or highly technical medical expressions, such as "Remdesivir" or "hipertensi pulmonal," which were often misclassified as O, particularly by IndoBERT and BioBERT. BIO boundary errors occurred in multi-word disease mentions where sequences such as "penyakit jantung bawaan" were incorrectly split into multiple B-DISEASE tags. Finally, entity-type confusion occurred in semantically overlapping categories. For example, "kelelahan kronis" was mislabeled as B-DISEASE instead of SYMPTOM, and "parasetamol" was predicted as B-DISEASE rather than B-DRUG.

3) Limitations of the Research

This research has several limitations. First, the dataset size remains relatively small compared to large-scale biomedical NER corpora, which may affect model robustness. Second, the dataset is derived from a single source, DetikHealth, which limits cross-domain generalization. Third, the dataset shows significant class imbalance, with non-entity tokens dominating the corpus. Future work should address these limitations by expanding the dataset, incorporating multiple sources, and applying more advanced data balancing strategies. Despite these limitations, this research provides a valuable benchmark for biomedical NER in Indonesian health news, particularly in low-resource settings.

IV. Conclusion

This research conducted a systematic, comparative evaluation of four Transformer-based models, namely BERT, IndoBERT, RoBERTa, and BioBERT, for biomedical named entity recognition (NER) using Indonesian health news articles. As shown in Table 15 and Figure 14, the results suggest that performance differences are mainly influenced by the robustness of the models’ pretraining architectures rather than by language specialization or domain adaptation alone. RoBERTa achieved the most consistent performance across models, with a micro-averaged F1-score of 0.463087 and a macro-averaged F1-score of 0.387340. However, the relatively low macro F1-score indicates that overall model performance remains uneven across entity classes, mainly because of severe label imbalance in the dataset, where non-entity tokens dominate.

In contrast, IndoBERT and BioBERT showed clear limitations in bridging linguistic and domain gaps. IndoBERT experienced a decrease in F1-score despite its monolingual advantage, while BioBERT displayed polarized performance by excelling on disease entities but struggling with symptom entities (F1-score of 0.17). This research introduces a manually annotated biomedical NER dataset derived from Indonesian health news, supported by high annotation reliability (Cohen’s Kappa = 0.88). Although the dataset remains relatively small and shows significant class imbalance, it provides an initial benchmark for comparing multilingual, monolingual, and domain-adaptive models in Indonesian biomedical NLP.

The error analysis shown in Table 16 reveals that model errors are systematic and influenced by contextual ambiguity, semantic overlap among entity types, and tendencies toward over-detection and under-detection. Boundary prediction inconsistencies in the BIO scheme further contribute to performance limitations. These findings highlight the need for improved contextual generalization and more effective handling of biomedical terminology in Indonesian texts, which is often not well captured by English-centric pretrained models. Despite its contributions, this study has several limitations. The dataset size remains relatively small, and the use of a single data source, DetikHealth, may limit the generalizability of the model to other Indonesian news platforms with different linguistic styles. In addition, class imbalance within the dataset affects the model’s ability to accurately recognize low-frequency entities. Future work should focus on expanding the dataset with more diverse sources, improving class balance, and exploring domain-adaptive pretraining strategies tailored to Indonesian biomedical texts. The integration of external knowledge resources, such as medical ontologies, may further enhance the recognition of domain-specific entities.

DECLARATION OF AI AND AI ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used ChatGPT (OpenAI) in order to improve language clarity, grammar, and overall readability. After using this tool, the authors carefully reviewed and edited the content as needed and take full responsibility for the content of the publication.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Maria Bernadette Chayeenee Norman: Writing - Original Draft, Methodology, and Visualization. **Ika Novita Dewi**: Conceptualization, Supervision, and Writing - Review & Editing. **Darnell Ignasius**: Data Curation and Validation.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of the Faculty of Computer Science, Universitas Dian Nuswantoro (UDINUS). Appreciation is also extended to the Research Center for Intelligent Distributed Surveillance and Security (IDSS) and the UDINUS Digital Health Laboratory for their valuable support, facilities, and contributions to this research.

REFERENCES

[1] P. Bose, S. Srinivasan, W. C. Sleeman IV, J. Palta, R. Kapoor, and P. Ghosh, "A survey on recent named entity recognition and relationship extraction techniques on clinical texts," *Appl. Sci.*, vol. 11, no. 18, p. 8319, 2021.

[2] J. Ravikumar and P. R. Kumar, "Machine learning model for clinical named entity recognition," *Int J Electr Comput Eng*, vol. 11, no. 2, pp. 1689–1696, 2021.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[4] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv Prepr. arXiv1907.11692*, 2019.

[5] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[6] L. Nemes and A. Kiss, "Information extraction and named entity recognition supported social media sentiment analysis during the COVID-19 pandemic," *Appl. Sci.*, vol. 11, no. 22, p. 11017, 2021.

[7] H. Hernina, Y. Karlina, and D. A. Puspitasari, "The Indonesian terms of disease names: A corpus linguistic study," *OKARA J. Bhs. dan sastra*, vol. 17, no. 1, pp. 14–31, 2023.

[8] F. Ikhwantri, "Cross-lingual transfer for distantly supervised and low-resources Indonesian NER," in *International Conference on Computational Linguistics and Intelligent Text Processing*, Springer, 2019, pp. 391–405.

[9] F. Gallego and F. J. Veredas, "Recognition and normalization of multilingual symptom entities using in-domain-adapted BERT models and classification layers," *Database*, vol. 2024, p. baae087, 2024.

[10] N. Perera, M. Dehmer, and F. Emmert-Streib, "Named entity recognition and relation detection for biomedical information extraction," *Front. cell Dev. Biol.*, vol. 8, p. 673, 2020.

[11] A. Wuehrl, L. Grimminger, and R. Klinger, "An entity-based claim extraction pipeline for real-world biomedical fact-checking," in *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, 2023, pp. 29–37.

[12] Y. Xiong *et al.*, "Improving deep learning method for biomedical named entity recognition by using entity definition information," *BMC Bioinformatics*, vol. 22, no. Suppl 1, p. 600, 2021.

[13] P. Dufter and H. Schütze, "Identifying elements essential for BERT's multilinguality," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4423–4437.

[14] D. Sebastian, H. D. Purnomo, and I. Sembiring, "Bert for natural language processing in bahasa Indonesia," in *2022 2nd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, IEEE, 2022, pp. 204–209.

[15] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 757–770.

[16] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[17] Y. Gu *et al.*, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Trans. Comput. Healthc.*, vol. 3, no. 1, pp. 1–23, 2021.

[18] Z. Xu, Z. Liu, Y. Yan, Z. Liu, G. Yu, and C. Xiong, "Cleaner pretraining corpus curation with neural web scraping," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2024, pp. 802–812.

[19] B. Jehangir, S. Radhakrishnan, and R. Agarwal, "A survey on named entity recognition—datasets, tools, and methodologies," *Nat. Lang. Process. J.*, vol. 3, p. 100017, 2023.

[20] N. Istiqomah and F. Novika, "Comparative Performance of IndoBERT and IndoLEM Baseline Models for Post-Disaster Health Information Extraction from Indonesian Online News," *J. Comput. Sci. Informatics Eng.*, vol. 4, no. 3, pp. 158–174, 2025.

[21] A. Abodayeh, R. Hejazi, W. Najjar, L. Shihadeh, and R. Latif, "Web scraping for data analytics: A beautifulsoup implementation," in *2023 sixth international conference of women in data science at prince Sultan University (WiDS PSU)*, IEEE, 2023, pp. 65–69.

[22] W. A. S. Farias, D. M. A. Melo, L. M. dos Santos, Â. A. S. de Oliveira, R. L. B. A. Medeiros, and Y. K. R. O. Silva, "Web Scraping as a scientific tool for theoretical reference," 2024.

[23] K. Al Sharou, Z. Li, and L. Specia, "Towards a better understanding of noise in natural language processing," in *Proceedings of the International conference on recent advances in natural language processing (RANLP 2021)*, 2021, pp. 53–62.

[24] M. R. A. Rajesh and D. T. Hiwarkar, "Exploring preprocessing techniques for natural languagetext: a comprehensive study using python code," *Int. J. Eng. Technol. Manag. Sci.*, vol. 7, no. 5, pp. 390–399, 2023.

[25] A. Muntakim, F. Sadaf, and K. M. A. Hasan, "BanglaMedNER: A gold standard medical named entity recognition corpus for Bangla text," in *2023 6th International Conference on Electrical Information and Communication Technology (EICT)*, IEEE, 2023, pp. 1–6.

[26] A. Jaiswal and E. Milios, "Breaking the token barrier: Chunking and convolution for efficient long text classification with bert," *arXiv Prepr. arXiv2310.20558*, 2023.

[27] W. Wongso, H. Lucky, and D. Suhartono, "Pre-trained transformer-based language models for Sundanese," *J. Big Data*, vol. 9, no. 1, p. 39, 2022.

[28] L. Guan, "Weight prediction boosts the convergence of adamw," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 2023, pp. 329–340.

[29] R. Tiwari, "Stabilizing the training of deep neural networks using Adam optimization and gradient clipping," *Int. J. Sci. Res. Eng. Manag.*, vol. 7, no. 1, 2023.

[30] M. R. Rezaei-Dastjerdehei, A. Mijani, and E. Fatemizadeh, "Addressing imbalance in multi-label classification using weighted cross entropy loss function," in *2020 27th national and 5th international iranian conference on biomedical engineering (ICBME)*, IEEE, 2020, pp. 333–338.

[31] K. M. Kahloot and P. Ekler, "Algorithmic splitting: A method for dataset preparation," *IEEE access*, vol. 9, pp. 125229–125237, 2021.

[32] D. E. Birba, "A Comparative study of data splitting algorithms for machine learning model selection." 2020.

[33] M. Boguslav and K. B. Cohen, "Inter-annotator agreement and the upper limit on machine performance: evidence from biomedical natural language processing," *Stud. Health Technol. Inform.*, vol. 245, pp. 298–302, 2017.

[34] G. Abercrombie, T. Dinkar, A. C. Curry, V. Rieser, and D. Hovy, "Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement," in *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, 2025, pp. 63–74.

[35] B. Plank, "The 'problem' of human label variation: On ground truth in data, modeling and evaluation," in *Proceedings of the 2022 conference on empirical methods in natural language processing*, 2022, pp. 10671–10682.