Published online on the page : https://ejournal.uinbukittinggi.ac.id/index.php/ijokid/

Knowbase : International Journal of Knowledge in Database

| ISSN (Print) 2798-0758 | ISSN (Online) 2797-7501 |

Named Entity Recognition for Uncovering Clinical and Emotional Entities from Breast Cancer Patient Interviews

Norma Alias[1,*], *Agus Sundari*[2]

[1] *University Teknologi Malaysia, Skudai, Malaysia*
[2] *Universitas Islam Negeri Sjech M. Djamil Djambek Bukittinggi, Indonesia*

Article Information

Article History:

Submitted:April 29, 2025
Revision: May 15, 2025
Accepted: June 11, 2025
Published: June 30, 2025

Keywords

Named Entity Recognition (NER), Breast Cancer, Clinical Entities, Emotional Entities, *Rule-Based*, Natural Language Processing

Correspondence
E-mail: its.norma.alias@gmail.com

ABSTRACT

This study aims to develop a Named Entity Recognition (NER) system capable of identifying clinical and emotional entities within interview transcripts of breast cancer patients. The corpus was manually annotated using the BIO scheme across seven main entity categories: Social Support (Dukungan Sosial), Medical Actions (Tindakan Medis), Diagnosis, Negative Emotions (Emosi Negatif), Positive Emotions (Emosi Positif), Symptoms (Gejala), and Spiritual. The annotation process was followed by the implementation of a rule-based method supported by entity dictionaries and word normalization, and the model was evaluated using precision, recall, and F1-score metrics. The analysis results revealed that Dukungan Sosial was the most dominant entity with 347 occurrences, followed by Tindakan Medis and Diagnosis. The rule-based NER model achieved an F1-score of 0.50 for the Diagnosis entity, although its performance on emotional and social entities remained low due to data imbalance. These findings highlight the importance of integrating clinical and emotional aspects in natural language processing to gain a more comprehensive understanding of patient narratives. The proposed approach has potential applications in healthcare text mining for detecting emotional experiences and medical contexts, and it can be further enhanced through the integration of transformer-based models such as IndoBERT to improve entity recognition accuracy.

1. Introduction

Breast cancer is one of the most prevalent types of cancer among women and remains a leading cause of cancer related mortality each year, particularly in developing countries such as Indonesia [1]. Nevertheless, breast cancer patients have the potential to survive and recover by adapting to the conditions they face. This disease affects multiple aspects of patients' lives, particularly the psychological dimension, where psychological disturbances can influence both physical health and overall quality of life. In addition to its physical impact, the process of breast cancer diagnosis and treatment imposes significant emotional stress on patients. The perceptions, experiences, and emotions encountered throughout the treatment process are often reflected in narratives or interviews, which contain valuable information related to patients' clinical conditions as well as their psychological aspects [2]. However, the interview data is generally unstructured, making manual analysis difficult, especially when the data volume is very large, requiring a significant amount of time and being susceptible to researcher subjectivity. Therefore, an artificial intelligence-based approach is needed that can automatically and structurally extract clinical and emotional information.[3].

To cope with these challenges, breast cancer patients require both internal and external support to help them manage and treat their illness [4]. The management of breast cancer can be carried out through pharmacological and non-pharmacological approaches, aiming to assist patients in addressing various issues they encounter, including physical, psychological, and social concerns [5].

Extracting valuable information from unstructured text has become a crucial aspect of various applications in the era of big data, ranging from text mining to information retrieval [6]. The advancement of Natural Language Processing (NLP), particularly through Named Entity Recognition (NER) techniques, has opened significant opportunities for extracting essential information from unstructured textual data. For NLP systems to operate optimally, it is necessary to leverage not only information related to named entities but also nominal concepts, as both are closely interrelated [7]. NER plays a key role in identifying and classifying specific entities within text, such as disease names, symptoms, medications, and even emotional expressions [8]. Although deep learning–based NER approaches have demonstrated promising performance, their effectiveness largely depends on the availability of large annotated corpora. This poses a major challenge, as sufficient training data are often difficult to obtain particularly in specialized domains such as threat intelligence, where annotated data are generally scarce [9].

In recent years, several pre-trained biomedical language models have gained significant attention as effective approaches for Biomedical Named Entity Recognition (BioNER) tasks. Based on the encoder mechanism of the Transformer architecture, these models can be broadly categorized into three types: (1) encoder-based models, such as BioBERT, BlueBERT, and PubMedBERT [10][11]. (2) decoder-based models, such as BioGPT and BioMedLM [12]. and (3) encoder–decoder-based models, which incorporate both encoders and decoders, such as BioBART and SciFive [13]. Although large language models (LLMs) have recently emerged, domain-specific biomedical language models particularly those employing encoder-based architectures—continue to represent the state-of-the-art (SOTA) in biomedical text analysis and research [14][15][16]. Within the healthcare domain, NER has been extensively utilized to extract clinical entities from a variety of textual sources, including electronic medical records, physician narratives, and scientific publications [17][18][19]. Nevertheless, the application of NER to patient interview data—particularly for the simultaneous extraction of both clinical and emotional entities—remains relatively underexplored and poses unique methodological challenges [20].

The identified research gap highlights the need to adapt Named Entity Recognition (NER) approaches to the linguistic and cultural characteristics of communication in Indonesia. To date, there have been very few studies that specifically integrate clinical and emotional entity analysis within a single framework, particularly using interview data from cancer patients in the Indonesian language. Moreover, patient interviews in regions such as Bukittinggi exhibit distinctive ways of expressing physical conditions and emotions, requiring linguistic models that can more accurately capture contextual meanings. Based on this background and the identified gap, this study aims to develop and implement an NER model capable of identifying two main types of entities clinical entities (such as diagnoses, medical procedures, and symptoms) and emotional entities (such as anxiety, fear, hope, or sadness) from interview transcripts of breast cancer patients in Bukittinggi.

The main objective of this study is to evaluate the performance of the model in recognizing these entities and to explore the relationships between clinical and emotional entities extracted from the interviews. By combining linguistic and artificial intelligence approaches, this research seeks to make a novel contribution to medical text analytics in Indonesia by providing an integrated understanding of the clinical and psychosocial dimensions of patients through an NER-based analysis of natural conversational data. This analysis is expected to offer deeper insights into patients' holistic experiences and conditions, encompassing both medical and emotional aspects. Consequently, the findings may support the development of clinical decision-support systems, enhance healthcare professionals' empathy, and enrich interdisciplinary research bridging artificial intelligence, linguistics, and health psychology.

Furthermore, this research is exploratory in nature, utilizing a limited dataset consisting of 20 interview transcripts from breast cancer patients in the Bukittinggi region. This approach is expected to provide an initial overview of the potential application of Transformer-based NER models (such as BioBERT or IndoBERT) for analyzing natural language data in the healthcare domain. The findings from this study are anticipated to serve as a foundation for future research with larger datasets and the development of models capable of recognizing a broader range of clinical and emotional entities.

2. Method

The research methodology is designed to integrate linguistic and artificial intelligence approaches to extract both clinical and emotional meanings from interviews with breast cancer patients, aiming to provide novel insights into the emerging field of medical text analytics in Indonesia.

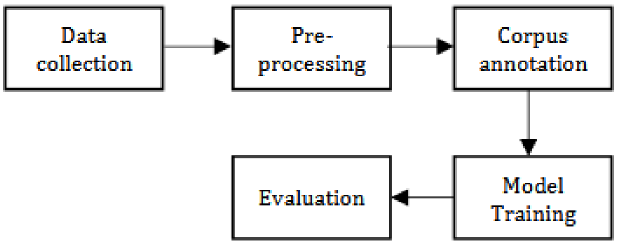


Figure 1. Research Flow Diagram of the Study

As shown in Figure 1, the research flow of this study consists of several main stages, namely data collection, pre-processing, corpus annotation, model training, and evaluation. The process begins with collecting interview transcripts as the primary dataset, followed by text cleaning and tokenization in the pre-processing stage. The annotated corpus is then used to train the Named Entity Recognition (NER) model, and finally, the model performance is evaluated using precision, recall, and F1-score metrics.

2.1. *Research Design*

The dataset consists of 20 interview transcripts from breast cancer patients collected in the Bukittinggi region. Each interview contains a narrative of the patient's experiences throughout the processes of diagnosis, treatment, and recovery. All data were obtained with informed consent from the participants and were anonymized to ensure the confidentiality of their identities. The analyzed data are in the form of unstructured text, with lengths varying between 300 and 800 words per interview. This study employs a computational experimental approach using the Named Entity Recognition (NER) method to extract both clinical and emotional entities from the patient interview data. The research design is exploratory in nature, given the limited amount of data, and aims to assess the initial feasibility of applying NER models within the domain of health and emotion analysis.

2.2. *Data Preprocessing*

The preprocessing stage was carried out to transform the raw interview transcripts into a structured format suitable for Named Entity Recognition (NER) analysis [21]. The process involved several key steps, including case folding, where all text was converted to lowercase to ensure consistency. Tokenization, which segmented the text into individual words or tokens. Stopword removal, aimed at eliminating commonly used words with little semantic value. Normalization, performed to standardize non-standard spellings and handle mixed-language expressions and lemmatization, which reduced each word to its base or dictionary form. As a result, this stage produced clean and tokenized text data that served as the input for subsequent entity annotation and model training processes.

```
Pseudocode data preprocessing
INPUT mMG, Ed
OUTPUT mMG
    INITIALIZE i, j
    [line, column] ← size(mMG)
    max ← 0
    FOR i ← 1 TO line DO
        FOR j ← 1 TO column DO
            IF max < mMG(i, j) THEN
                max ← mMG(i, j)
            END IF
        END FOR
    END FOR
    FOR i ← 1 TO line DO
        FOR j ← 1 TO column DO
            IF Ed(i, j) = 128 THEN
                mMG(i, j) ← max
            END IF
        END FOR
    END FOR
    // === Tahap 2: Stemming Bahasa Indonesia ===
    INITIALIZE stemmer ← Sastrawi Stemmer
    FUNCTION lemmatize_indonesia(teks):
        IF teks IS EMPTY THEN
            RETURN ""
        ELSE
            RETURN stemmer.stem(teks)
        END IF
    END FUNCTION
    // === Tahap 3: Normalisasi Teks ===
    FUNCTION normalisasi_teks(teks, kamus):
        CONVERT teks TO lowercase
        FOR setiap pasangan (kata_asli, kata_normal) dalam kamus DO
            GANTI semua kemunculan kata_asli dengan kata_normal dalam teks
        END FOR
        RETURN teks
    END FUNCTION
    // === Tahap 4: Deteksi Entitas ===
    FUNCTION deteksi_entitas(teks, kamus_entitas):
        INITIALIZE hasil ← empty list
        FOR setiap (label, daftar_kata) dalam kamus_entitas DO
            FOR setiap kata dalam daftar_kata DO
                IF kata ditemukan dalam teks THEN
                    TAMBAHKAN (kata, label) ke hasil
                END IF
            END FOR
        END FOR
        RETURN hasil
    END FUNCTION
END PROGRAM
```

2.3. *Entity Annotation*

Entity annotation was conducted manually to identify and label both clinical and emotional entities within the interview transcripts. The annotated data were derived from the preprocessed text, which had been cleaned and tokenized in the previous stage. A BIO (Begin, Inside, Outside) tagging scheme was applied to assign entity labels to each token according to its corresponding category. Clinical entities included Diagnosis, Symptom, Medical Action, Medication, and Examination Result, while emotional entities encompassed Sadness, Fear, Anxiety, Hope, Motivation, and other affective expressions. The annotation process was carried out by two trained annotators with expertise in linguistics and health communication to ensure precision and contextual validity . Inter-annotator agreement was assessed using Cohen's Kappa coefficient to evaluate the consistency of labeling, and any discrepancies were resolved through discussion and consensus [22]. The finalized annotated corpus served as the gold-standard dataset for model training and evaluation.

```
Pseudocode Entity Annotation
INITIALIZE hasil ← empty list
```

```
FOR setiap (label, daftar_kata) dalam kamus_entitas DO
        FOR setiap kata dalam daftar_kata DO
            IF kata terdapat dalam teks THEN
                TAMBAHKAN (kata, label) ke hasil
            END IF
        END FOR
    END FOR
RETURN hasil
END FUNCTION
```

2.4. *Model Development (NER)*

The NER model was developed using a Transformer-based architecture, specifically leveraging pre-trained language models such as IndoBERT, which are well-suited for processing biomedical and Indonesian-language text [23] [24]. Creation of an IndoBERT transformer-based NER model that has been optimized to handle the complexities of the legal environment, including mentions of revised statutes and references to tiered articles (verses, letters) [25]. The annotated corpus was divided into training, validation, and testing subsets to optimize and evaluate model performance [26]. The model was fine-tuned on the training set using supervised learning, with hyperparameters (such as learning rate, batch size, and number of epochs) adjusted experimentally. The implementation was carried out using the Hugging Face Transformers library in Python. The goal of this stage was to enable the model to automatically identify clinical and emotional entities from raw interview text.

```
Pseudocode Model Development
    // === Tahap 1: Tokenizer Initialization ===
     LOAD pre-trained tokenizer "indobenchmark/indobert-base-p1"
    // === Tahap 2: Label Encoding ===
    EXTRACT unique labels from dataset df_bio
    CREATE label2id mapping: label → index
    CREATE id2label mapping: index → label
    // === Tahap 3: Model Initialization ===
    LOAD pre-trained model "indobenchmark/indobert-base-p1"
        SET num_labels = total number of unique labels
        ASSIGN label2id and id2label mapping
    // === Tahap 4: Tokenization and Label Alignment ===
    FUNCTION tokenize_and_align_labels(examples):
        tokenized ← tokenizer(examples["tokens"],
                              truncation=True,
                              is_split_into_words=True,
                              padding="max_length",
                              max_length=128)
        INITIALIZE labels ← empty list
        FOR each sentence i in examples["labels"] DO
            word_ids ← tokenized.word_ids(batch_index=i)
            label_ids ← empty list
            previous_word_idx ← None
            FOR each word_idx in word_ids DO
                IF word_idx IS None THEN
                    ADD -100 TO label_ids        // ignore padding tokens
                ELSE IF word_idx ≠ previous_word_idx THEN
                    ADD label2id[label[word_idx]] TO label_ids
                ELSE
                    ADD -100 TO label_ids        // ignore subword tokens
                END IF
                previous_word_idx ← word_idx
            END FOR
            ADD label_ids TO labels
        END FOR
        tokenized["labels"] ← labels
        RETURN tokenized
    END FUNCTION
END INDOBERT_NER_MODEL
```

2.5. *Model Evaluation*

Model evaluation was conducted to assess the performance of the fine-tuned NER model [27]. Three standard evaluation metrics precision, recall, and F1-score were used to measure the accuracy of entity recognition for each category. Precision indicates the proportion of correctly identified entities among all predicted entities, while recall measures the proportion of correctly identified entities among all actual entities. The F1-score, which combines both precision and recall, was used as the primary performance indicator. Evaluation results were computed for both clinical and emotional entity types separately, allowing for a comparative analysis of the model's ability to capture factual (clinical) and affective (emotional) information from the text [28].

```
Pseudocode Model Evaluation
    // === Tahap 1: Training Model IndoBERT ===
    INITIALIZE Trainer with parameters:
        - model = IndoBERT for Token Classification
        - args = training_args
        - train_dataset = training dataset
        - eval_dataset = evaluation dataset
        - tokenizer = IndoBERT tokenizer
    CALL trainer.train()        // Melatih model pada data anotasi BIO
    // === Tahap 2: Rule-Based NER Function ===
    FUNCTION ner_rule_based(tokens, kamus_entitas, normalisasi_kamus):
        tokens_norm ← [normalisasi_kamus[t.lower()] IF t.lower() in kamus ELSE t.lower() FOR t in
tokens]
        labels ← list of "O" with length equal to tokens
        FOR each ent_type, ent_list IN kamus_entitas DO
            FOR each ent IN ent_list DO
                ent_tokens ← split(ent)
                FOR i ← 0 TO (len(tokens_norm) - len(ent_tokens)) DO
                    IF tokens_norm[i : i + len(ent_tokens)] == ent_tokens THEN
                        labels[i] ← "B-" + ent_type
                        FOR j ← 1 TO (len(ent_tokens) - 1) DO
                            labels[i + j] ← "I-" + ent_type
                        END FOR
                    END IF
                END FOR
            END FOR
        END FOR
        RETURN list of (tokens_norm, labels)
    END FUNCTION
    // === Tahap 3: Model Prediction ===
    preds_output ← trainer.predict(eval_dataset)
    pred_logits ← preds_output.predictions
    pred_labels ← argmax(pred_logits, axis = -1)
    // === Tahap 4: Cleaning Padding Tokens ===
    true_labels ← []
    pred_labels_clean ← []
    FOR each label_sequence IN eval_dataset["labels"]:
        REMOVE all -100 values from label_sequence
        CONVERT label ID to label name using id2label
        APPEND to true_labels
    END FOR
    FOR each (pred_seq, label_seq) IN zip(pred_labels, eval_dataset["labels"]):
        REMOVE -100 values based on corresponding label_seq
        CONVERT predicted IDs to label names
        APPEND to pred_labels_clean
    END FOR
    // === Tahap 5: Flatten Lists for Evaluation ===
    true_labels_flat ← concatenate all sublists in true_labels
    pred_labels_flat ← concatenate all sublists in pred_labels_clean
    // === Tahap 6: Classification Report ===
    CALCULATE precision, recall, and F1-score
        USING classification_report(true_labels_flat, pred_labels_flat)
    DISPLAY evaluation results
END NER_MODEL_TRAINING_AND_EVALUATION
```

2.6. *Result Analysis*

Result analysis focused on interpreting the patterns and insights derived from the recognized entities. The analysis examined the distribution and co-occurrence of clinical and emotional entities to explore how patients' emotional expressions relate to their clinical experiences [29]. For instance, associations between certain medical procedures and emotional responses such as fear or hope were identified and qualitatively discussed [30]. Visualization tools such as word clouds and co-occurrence graphs were employed to present these relationships. The findings were then analyzed in relation to existing literature to highlight the model's strengths, limitations, and potential implications for future research in medical text analytics and health communication.

3. Results and Discussion

3.1. Entity Annotation Statistics

A total of 1,328 entities were manually annotated using the BIO tagging scheme. These were classified into two major categories**:** Clinical Entities (Diagnosis, Symptom, Medication, Medical Action, Examination Result) and Emotional Entities (Sadness, Fear, Anxiety, Hope, Motivation). Figure 2 illustrates the frequency distribution of all annotated entity categories.

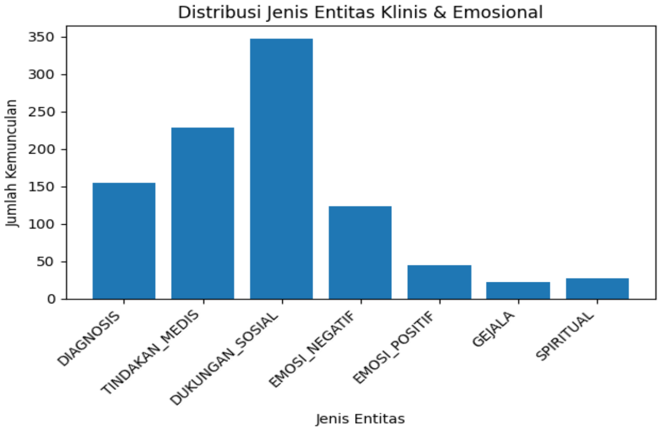


Figure 2. Distribution of Clinical and Emotional Entity Types

Based on Figure 1, the Social Support (Dukungan Sosial) entity represents the most frequently occurring category, with a total of 347 instances. This finding indicates that social interactions particularly support from family members, friends, and healthcare professionals play a crucial role in patients' experiences throughout their treatment process. Furthermore, the Medical Actions (Tindakan Medis) entity appeared 228 times, while Diagnosis occurred 155 times. These frequencies suggest that the majority of patient narratives focus on medical aspects, including examination procedures, treatment processes, and the results of disease diagnosis that they experienced.

The Negative Emotions (Emosi Negatif) entity also appeared prominently, totaling 123 occurrences, reflecting emotional responses such as fear, sadness, worry, and anxiety that frequently arise during patients' journeys of confronting cancer. On the other hand, the Positive Emotions (Emosi Positif) entity, with 45 occurrences, represents expressions such as enthusiasm, gratitude, and optimism, which although less frequent remain an important component of patients' emotional adaptation. Meanwhile, the Spiritual and Symptoms (Gejala) entities were less prevalent, with 27 and 22 occurrences, respectively. This indicates that spiritual aspects and specific physical complaints were not always explicitly expressed by patients during the interviews, yet they still form an integral part of the narrative context. Overall, these annotation results reveal that the interview corpus contains a rich combination of both clinical and emotional information. Such variation is essential for the development of the NER model, as it enables the system to recognize not only the medical but also the psychological contexts embedded in the patients' natural language expressions.

In addition to the frequency distribution, a word cloud visualization was generated to provide an overview of the most frequently occurring terms within the annotated corpus. The size of each word in the visualization corresponds to its frequency, allowing a quick visual interpretation of dominant expressions used by patients during interviews As shown in Figure 3**.**

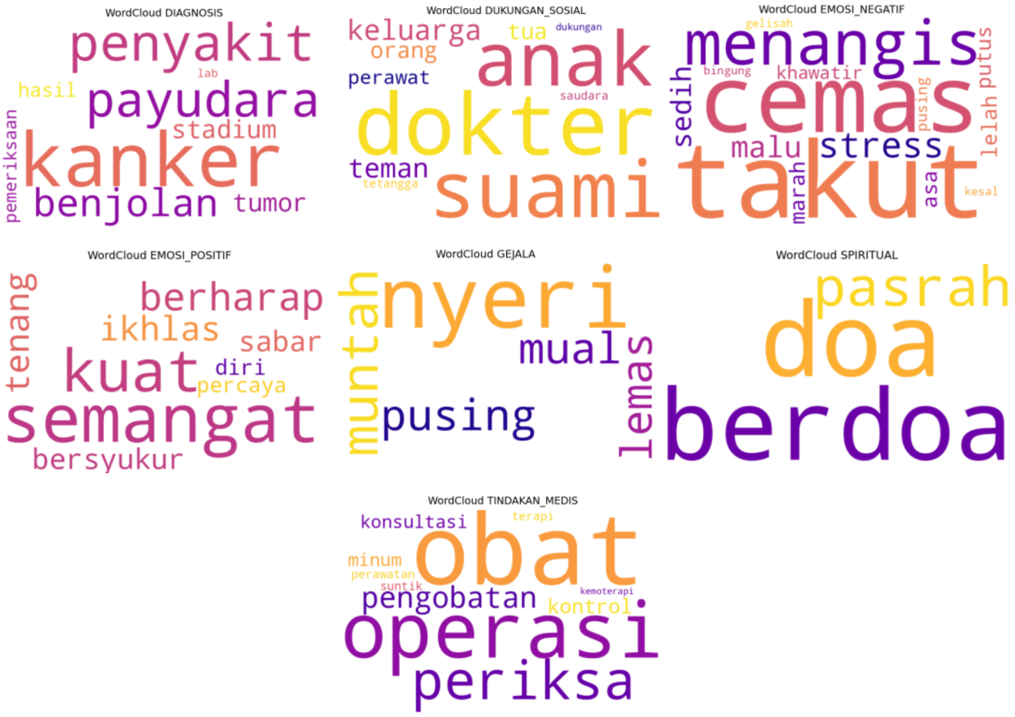
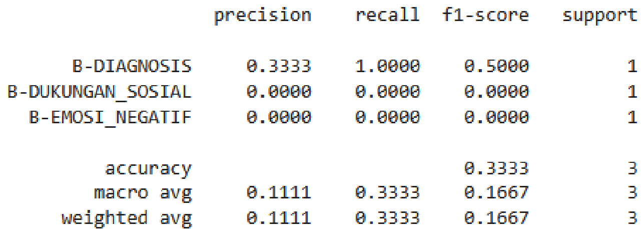


Figure 3. word cloud visualization of Clinical and Emotional Entity Types

Meanwhile, emotionally charged words such as takut (fear), sedih (sadness), and cemas (anxiety) also appear frequently, highlighting the emotional dimension that accompanies the clinical experience.

3.2. *Model Evaluation Results*

The model evaluation was carried out using the classification report that presents performance metrics for each entity type, including precision, recall, and F1-score. As shown in Figure 4, the classification report summarizes the model's performance across different entity categories using standard NER evaluation metrics.



```
                  precision    recall  f1-score   support

      B-DIAGNOSIS     0.3333    1.0000    0.5000         1
 B-DUKUNGAN_SOSIAL     0.0000    0.0000    0.0000         1
   B-EMOSI_NEGATIF     0.0000    0.0000    0.0000         1

         accuracy                         0.3333         3
        macro avg     0.1111    0.3333    0.1667         3
     weighted avg     0.1111    0.3333    0.1667         3
```

Figure 4. Classification Report of the NER Model Performance

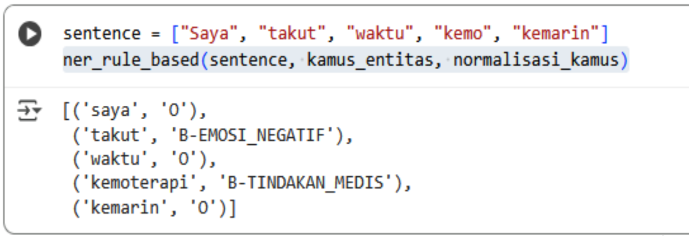
Figure 3 presents the classification report of the Named Entity Recognition (NER) model, showing the evaluation metrics of precision, recall, and F1-score for each entity label. The model was tested on three entity categories: B-DIAGNOSIS, B-DUKUNGAN_SOSIAL, and B-EMOSI_NEGATIF.

The B-DIAGNOSIS entity achieved a precision of 0.333**,** recall of 1.000, and an F1-score of 0.500, indicating that the model successfully recognized all instances of this label, although with limited precision due to a small number of samples. Conversely, the B-DUKUNGAN_SOSIAL and B-EMOSI_NEGATIF entities scored 0.000 across all metrics, suggesting that the model failed to detect these categories in the test set.

The overall model accuracy was 0.333, while both macro and weighted averages of F1-score were 0.1667, which reflects the model's difficulty in generalizing to underrepresented entity types. This result also indicates that the dataset imbalance affected the model's ability to recognize emotional and social support entities compared to diagnostic ones.

3.3. *Rule-Based NER Implementation*

A rule-based Named Entity Recognition (NER) system was implemented using manually constructed dictionaries for clinical and emotional categories. The system successfully identified entities within the tokenized sentences. For example, in the sentence ["Saya", "takut", "waktu", "kemo", "kemarin"], the tokens takut and kemoterapi were recognized as B-EMOSI_NEGATIF and B-TINDAKAN_MEDIS, respectively. This output demonstrates that the rule-based NER module was able to effectively map tokens to their corresponding entity categories using predefined lexical rules. Figure 5 shows that the rule-based system correctly labeled the tokens "takut" and "kemoterapi" as emotional and medical entities.



```
sentence = ["Saya", "takut", "waktu", "kemo", "kemarin"]
ner_rule_based(sentence, kamus_entitas, normalisasi_kamus)
```

```
[('saya', 'O'),
 ('takut', 'B-EMOSI_NEGATIF'),
 ('waktu', 'O'),
 ('kemoterapi', 'B-TINDAKAN_MEDIS'),
 ('kemarin', 'O')]
```

Figure 5. Rule Base NER

4. Conclusion

This study successfully developed a Named Entity Recognition (NER) system for identifying both clinical and emotional entities within the interview transcripts of breast cancer patients. The annotated corpus contained seven main entity categories, consisting of Social Support (Dukungan sosial), Medical Actions (Tindakan Medis), Diagnosis, Negative Emotions (Emosi Negatif), Positive Emotions (Emosi Positif), Symptoms (Gejala), and Spiritual. The analysis results revealed that Dukungan Sosial appeared as the most dominant entity, followed by Tindakan Medis and Diagnosis, indicating that patient narratives were largely centered around medical experiences and social interactions during treatment.

The evaluation results demonstrated that the model was capable of identifying diagnostic entities with a moderate performance level, while other categories such as emotional and social entities remained challenging to detect due to data imbalance and contextual variations in natural language. These findings confirm that combining clinical and emotional components in corpus annotation is crucial for building a more comprehensive understanding of patient narratives.

The proposed NER approach has potential applications in healthcare text mining, particularly for analyzing patient feedback, clinical interviews, and social support patterns in medical contexts. In future research, expanding the dataset and integrating transformer-based models such as IndoBERT or multilingual BERT could further improve entity recognition accuracy, especially for underrepresented emotional and spiritual entities.

References

[1] A. Riandini, U. Safari, N. Riani, F. Khoerunnisa, and D. A. Sulistiani, "Peningkatan Pengetahuan Tentang

Pencegahan Kanker Payudara Melalui 'Sadari' Pada Remaja Di Smk Pelita Alam," *J. Med. Hutama*, vol. 02, pp. 434–440, 2020.

[2] D. Dai, H. Coetzer, S. Zion, and M. Malecki, "Anxiety, Depression, and Stress Reaction/Adjustment Disorders and Their Associations with Healthcare Resource Utilization and Costs Among Newly Diagnosed Patients With Breast Cancer," *J. Heal. Econ. Outcomes Res.*, vol. 10, no. 1, pp. 68–76, 2023, doi: 10.36469/jheor.2023.70238.

[3] B. F. Kalanda and A. J. Cheboi, "Artificial Intelligence in the Analysis of Unstructured Qualitative Data: A Literature Review," *Adv. Soc. Sci. Res. J.*, vol. 12, no. 08, pp. 199–205, 2025, doi: 10.14738/assrj.1208.19286.

[4] Y. R. Putri, Y. Afiyanti, S. Dewi, and A. R. Ma'rifah, "Breast Cancer Patients' Experience of Current Health Services as A Holistic Care: A Qualitative Study," *Malaysian J. Med. Heal. Sci.*, vol. 19, no. 6, pp. 127–135, 2023, doi: 10.47836/mjmhs.19.6.17.

[5] T. Solehati, P. Napisah, A. Rahmawati, I. Nurhidayah, and C. E. Kosasih, "Penatalaksanaan Keperawatan pada Pasien Kanker Payudara; Sistematik Review," *J. Ilm. …*, vol. 10, no. 1, pp. 71–82, 2020, [Online]. Available: http://journal.stikeskendal.ac.id/index.php/PSKM/article/view/672

[6] F. M. Surur *et al.*, "Unlocking the power of machine learning in big data: a scoping survey," *Data Sci. Manag.*, 2025, doi: 10.1016/j.dsm.2025.02.004.

[7] G. Martinelli, F. M. Molfese, S. Tedeschi, A. Fernández-Castro, and R. Navigli, "CNER: Concept and Named Entity Recognition," *Proc. 2024 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. NAACL 2024*, vol. 1, pp. 8329–8344, 2024, doi: 10.18653/v1/2024.naacl-long.461.

[8] M. Theofany Aulia Anwar, S. Hadi Wijoyo, and W. Hayuhardhika Nugraha Putra, "Implementasi Metode TextRank dan Named Entity Recognition Untuk Ekstraksi Kata Kunci Pada Media Online Berita," *J. Sist. Informasi, Teknol. Informasi, dan Edukasi Sist. Inf.*, vol. 5, no. 1, pp. 34–41, 2024, doi: 10.25126/justsi.v5i1.401.

[9] H. Wang, W. Yang, W. Feng, L. Zeng, and Z. Gu, "Threat intelligence named entity recognition techniques based on few-shot learning," *Array*, vol. 23, no. April, p. 100364, 2024, doi: 10.1016/j.array.2024.100364.

[10] 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, doi: 10.1093/bioinformatics/btz682.

[11] P. Lewis, B. Oguz, R. Rinott, S. Riedel, and H. Schwenk, "MLQA: Evaluating cross-lingual extractive question answering," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 7315–7330, 2020, doi: 10.18653/v1/2020.acl-main.653.

[12] E. Bolton *et al.*, "BioMedLM: A 2.7B Parameter Language Model Trained On Biomedical Text," vol. 2015, pp. 1–23, 2024, [Online]. Available: http://arxiv.org/abs/2403.18421

[13] H. Yuan, Z. Yuan, R. Gan, J. Zhang, Y. Xie, and S. Yu, "BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 97–109, 2022, doi: 10.18653/v1/2022.bionlp-1.9.

[14] Y. Hu *et al.*, "Improving large language models for clinical named entity recognition via prompt engineering," *J. Am. Med. Informatics Assoc.*, vol. 31, no. 9, pp. 1812–1820, 2024, doi: 10.1093/jamia/ocad259.

[15] Q. Chen *et al.*, "Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations," no. 2, 2023, [Online]. Available: http://arxiv.org/abs/2305.16326

[16] M. Sung, M. Jeong, Y. Choi, D. Kim, J. Lee, and J. Kang, "BERN2: an advanced neural biomedical named entity recognition and normalization tool," *Bioinformatics*, vol. 38, no. 20, pp. 4837–4839, 2022, doi: 10.1093/bioinformatics/btac598.

[17] Y. Yin *et al.*, "Augmenting biomedical named entity recognition with general-domain resources," *J. Biomed. Inform.*, vol. 159, p. 104731, 2024, doi: 10.1016/j.jbi.2024.104731.

[18] H. Zhao and W. Xiong, "A multi-scale embedding network for unified named entity recognition in Chinese Electronic Medical Records," *Alexandria Eng. J.*, vol. 107, no. September, pp. 665–674, 2024, doi: 10.1016/j.aej.2024.09.008.

[19] R. Rhouma *et al.*, "Leveraging mobile NER for real-time capture of symptoms, diagnoses, and treatments from clinical dialogues," *Informatics Med. Unlocked*, vol. 48, no. January, p. 101519, 2024, doi: 10.1016/j.imu.2024.101519.

[20] L. Ramadani, R. A. Nugraha, and Falahah, "Dialectic and Life-cycle of Institutional Logics in IT Governance: Insights from Healthcare Context," *Procedia Comput. Sci.*, vol. 234, pp. 1267–1275, 2024, doi: 10.1016/j.procs.2024.03.124.

[21] J. Mantik, S. Indra, and G. Situmeang, "2022) 423-430 Accredited," *J. Mantik*, vol. 6, no. 1, pp. 423–430, 2021.

[22] R. Permata, Rendika, and L. C. Julianty, “Towards an Automated Essay Evaluation System NLP Based Text Embeddings and Similarity Metrics,” *Digit. Zo. J. Teknol. Inf. dan Komun.*, vol. 16, no. 1, pp. 37–46, 2025, doi: 10.31849/qvjtcn48.

[23] M. T. Manurung, I Gusti Ngurah Lanang Wijayakusuma, and I Putu Winada Gautama, “Named Entity Recognition for Medical Records of Heart Failure Using a Pre-trained BERT Model,” *J. Appl. Informatics Comput.*, vol. 9, no. 2, pp. 341–348, 2025, doi: 10.30871/jaic.v9i2.9170.

[24] Warto *et al.*, *Systematic Literature Review on Named Entity Recognition: Approach, Method, and Application*, vol. 12, no. 4. 2024. doi: 10.19139/soic-2310-5070-1631.

[25] A. Ahmed, A. Abbasi, and C. Eickhoff, “Benchmarking Modern Named Entity Recognition Techniques for Free-text Health Record Deidentification,” *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, vol. 2021, pp. 102–111, 2021.

[26] E. Subowo, I. Bukhori, and Warto, “Corpus Development and NER Model for Identification of Legal Entities (Articles, Laws, and Sanctions) in Corruption Court Decisions in Indonesia,” *Trans. Informatics Data Sci.*, vol. 2, no. 1, pp. 27–39, 2025, doi: 10.24090/tids.v2i1.13592.

[27] M. Sun, S. Xiong, Y. Cai, and B. Zuo, “Positional Attention for Efficient BERT-Based Named Entity Recognition,” *arXiv:2505.01868*, 2025, [Online]. Available: https://arxiv.org/abs/2505.01868

[28] I. Majid, V. Mishra, R. Ravindranath, and S. Y. Wang, “Evaluating the Performance of Large Language Models for Named Entity Recognition in Ophthalmology Clinical Free-Text Notes,” *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, vol. 2024, pp. 778–787, 2024.

[29] A. Rehman, M. Mujahid, A. Elyassih, B. AlGhofaily, and S. A. O. Bahaj, “Comprehensive Review and Analysis on Facial Emotion Recognition: Performance Insights into Deep and Traditional Learning with Current Updates and Challenges,” *Comput. Mater. Contin.*, vol. 82, no. 1, pp. 41–72, 2025, doi: 10.32604/cmc.2024.058036.

[30] F. Gössi *et al.*, “Jo u rn a l P,” *Patient Educ. Couns.*, p. 109386, 2025, doi: 10.1016/j.pec.2025.109386.