Eduvest – Journal of Universal Studies
Volume 4 Number 06, June, 2024
p- ISSN 2775-3735- e-ISSN 2775-3727

NAMED ENTITY RECOGNITION IN ELECTRONIC MEDICAL RECORDS BASED ON HYBRID NEURAL NETWORK AND TRANSFORMER

Muhammad Sumarudin[1], **Mohammad Syafrullah**[2]
[1,2] Program Studi Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur, Indonesia
Email: 2111601742@student.budiluhur.ac.id, mohammad.syafrullah@budiluhur.ac.id

ABSTRACT

The development of artificial intelligence in the field of health encourages the use of electronic medical records in all health facilities to record health services provided to patients. For hospitals, extracting information from electronic medical records can make it easier for management to make clinical decisions and for researchers to obtain data for research in the medical and nursing fields. Information extraction is part of natural language processing, which requires a deep learning model in the medical and nursing fields. Information extraction is from electronic medical records. The research builds a model of named entity recognition in electronic medical records based on hybrid neural networks, bidirectional encoder representations from transformers, and setting hyperparameters to get the highest accuracy. The research data set is processed from an initial examination form of an adult patient in a hospital onto an electronic medical record in 2022, and the data is pre-processed. Next, perform the entity-tagging phase on the text and divide 70% of the training datasets by 30% of the testing datasets. Training and evaluation of models built using the confusion matrix method. The results of this study show that the entity identification model called bidirectional encoder representations from transformers consistently outperforms the neural network-based entity recognition model in any evaluation metric. The abbreviation of the bi-directional encoder representation of transformer has very high precision, recall, and f1-score values, demonstrating its ability to recognise entities very well. In this study, although the model named Entity Recognition based on neural networks also has high accuracy, the low precision and recall values indicate that this model may have difficulty recognising entities accurately.

KEYWORDS *Electronic Medical Records, Named Entity Recognition, Information Extraction, Natural Language Processing, Hybrid Neural Networks*

How to cite:	Muhammad Sumarudin, Mohammad Syafrullah (2024). Named Entity Recognition In Electronic Medical Records Based On Hybrid Neural Network And Transformer. *Journal Eduvest*. *4* (6): 5263-5279
E-ISSN:	2775-3727
Published by:	https://greenpublisher.id/

INTRODUCTION

Every health care facility must have and implement electronic medical records from the time the patient arrives at the hospital until the patient is discharged, whether the patient is referred, healthy, or dead. The Minister of Health Regulation regulates this further on Medical Records (Regulation of the Minister of Health of the Republic of Indonesia Number 24 of 2022 concerning Medical Records, 2022).

An electronic medical record is a text-based record of a patient's diagnosis through the hospitalization process. It is usually stored in a structured and unstructured format which includes information such as the patient's health status and symptoms, medications, diseases, and various examination indicators. (Liu et al., 2019).

According to Parsaoran and Sitompul (2023) electronic medical record is any statement, record, or explanation made by a health worker for a patient and stored in electronic storage.

For example, the medical record reads "the patient was injured 15 minutes ago when his motorcycle collided with another motorcycle, causing pain in the left wrist and right shoulder, localized swelling and limited movement, no loss of consciousness, no dizziness, and no nausea. The local residents took the patient immediately to the emergency department for CT examination, no abnormalities were seen". Some information that can be obtained [left wrist] is a body part, [limited movement and local swelling] is a symptom, and [CT examination] is an examination. Such electronic medical records make it easier for hospitals and healthcare professionals to analyze information about the patient's condition and provide appropriate treatment recommendations. (Liu et al., 2019).

The development of electronic medical records at the National Central General Hospital Dr. Cipto Mangunkusumo was carried out in stages by transferring forms into electronic form, of course adapted to the characteristics of the form's contents. The form templates are in the form of one choice, more than one choice, and even narrative text.

The implementation of electronic medical records is a hope for researchers in the medical and nursing fields to facilitate obtaining primary and secondary data. On the side for hospital management, the application of electronic medical records can reduce the use of illegible or identified writing. Recording in electronic medical records, especially at the National Central General Hospital Dr. Cipto Mangunkusumo, is accompanied by monitoring activities for filling in patient health records in order to immediately determine the completeness and accuracy of electronic medical records. Therefore, electronic medical records are included in the examination material during the open / closed medical record review conducted by medical records officers and the quality committee. During the examination of records on electronic medical records by medical record officers and the quality committee, abbreviations outside the standard hospital abbreviations that have been determined are still found. In addition, hospital management and researchers had difficulty obtaining electronic medical record data sourced from narrative text.

One method to extract sentences in electronic medical records, identify named entities in the text and classify them into defined categories is named entity recognition. Ambiguity in words is one of the challenges in identifying named entities in narrative text, namely to recognize words that have multiple meanings in several sentences. Some approaches that can be done to detect entities are using rule-base, machine learning, and deep learning. Currently, the deep learning approach has gained much success in the field of natural language processing including named entity recognition. According to Susanty and Sukardi (2021) Recognition of named entities is part of research on natural language processing which is included in the field of artificial intelligence. The following are examples of entity categories in named entity recognition such as location (loc), organization (org), person name (per), time (date), quantity (mg), and others. (Kale and Govilkar, 2017).

In recent years, many studies have concentrated on named entity recognition using deep learning methods because it is a very dominant approach model in terms of named entity recognition research. This is because models built using deep learning based have better accuracy results. (Li et al., 2022).

Named entity recognition has been widely implemented in various fields based on Indonesian, such as in the field of online news media or social media there is research on named entity recognition in Indonesian with bidirectional long short term memory convolutional neural networks. (Gunawan et al., 2018), named entity recognition model on Indonesian tweets with conditional random fields classifier (Munarko et al., 2018)., Classification of event types using a combination of neuroNER and recurrent convolutional neural networks on twitter data (Putra and Fatichah, 2018). (Putra and Fatichah, 2018)Comparison of machine learning and deep learning algorithms for named entity recognition: a case study of disaster data. (Giarsyani et al., 2020), Hybrid conditional random field and K-Means named entity recognition in Indonesian language news documents (Santoso et al., 2020), Comparison of transformer models in fake review detection (Awalina et al., 2022), Classification of Indonesian news with named entity recognition approach (Nurchim et al., 2023).. In the field of Geography, named entity recognition is used to identify location-specific named entities. (Berragan et al., 2023).. In the field of tourism, named entity recognition is used to identify entities named locations (Fakhri et al., 2023). (Fakhri et al., 2023).. In the field of health, named entity recognition is used to identify biomedical named entities such as genes, drugs, diseases, and chemicals (Abdillah et al., 2023). (Abdillah et al., 2023)., health insurance entities (AI-Ash et al., 2019), various kinds of language model named entity recognition in health insurance question and answer system (Jati et al., 2020).

RESEARCH METHOD

The research method applied in this study crucially chooses a hybrid neural network approach and bidirectional encoder representations from transformers to develop and evaluate a named entity recognition model on Indonesian electronic medical record text. Named entity recognition aims to recognize entities such as medical diagnoses, medications, and other important information in medical record narratives. A hybrid neural network approach combines the advantages of recurrent neural network and long short-term memory to understand temporal context, while transformers are used to overcome long-range dependencies in text.

Methodological steps include data collection from the SIMRS database, model architecture development, and evaluation using metrics such as precision, recall, and f1-score to assess model performance. This research is expected to improve the effectiveness of named entity recognition technology in the context of medical records, support more accurate clinical decision-making, and contribute to the integration of more advanced health information systems.

Design Techniques

The data design technique in this research with several stages including data processing, *preprocessing*, and *tagging* processes and the process of building a *named entity recognition* model.

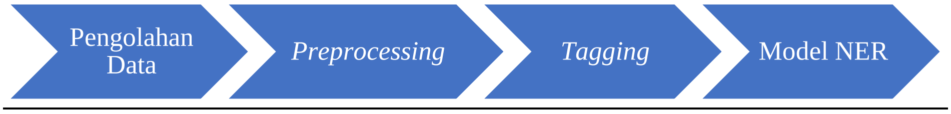


Figure 3.1 Design technique

This research describes the stages of the design technique consisting of data processing, preprocessing, entity tagging, building a neural network-based named entity recognition (NER) model and transformers. The testing techniques used include precision to measure the accuracy of recognizing entities, recall to measure the ability to recognize relevant entities, and F1-score as a harmonious value between precision and recall. The research stages include data collection and analysis, data preprocessing with techniques such as stopword removal, lemmatization, and tokenization, as well as building NER models with hybrid neural networks and transformers. Evaluation was conducted using precision, recall, and F1-score metrics to assess the performance of the model, which will later be implemented in the hospital management information system. The research period lasted for five months, with a detailed schedule for each stage from ETL processing to model deployment.

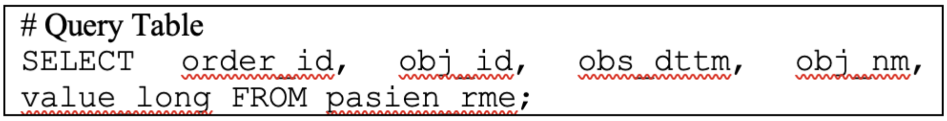
RESULT AND DISCUSSION

Data *Extract-Transfer-Load* (ETL) Process

This research begins with the stage of managing data from electronic medical records and is used as a research dataset. The inpatient adult patient assessment form is the focus of the data area used as a dataset because the form has been implemented at the Dr. Cipto Mangunkusumo National Central General Hospital from 2021, as can be seen in Figure 4.1 which shows the information technology infrastructure that has been built and integrated into the electronic medical record data architecture of this study.

Data Collection

The query command in Figure 4.2 is to view the columns in the table that has been collected:



```
# Query Table
SELECT   order_id,   obj_id,   obs_dttm,   obj_nm,
value_long FROM pasien_rme;
```

Figure 4.2 Query command into a table

Seen in table 4.1, some examples of data that have been collected from the *extract-transform-load* process are accommodated in the research database in one table. The data stored in the table that will be managed to become a data set is value_log as a sentence from an electronic medical record. The value_long field will also be used as a value that is managed or processed in the data preprocessing step so that the final process can be carried out before becoming a dataset, namely tagging.

Table 4.1 Example of data that has been collected

order_id	obj_id	obs_dttm	obj_nm	value_long
135766047	30389	2022-01-01 06:58:25	MAIN COMPLAINTS	A 43-year-old female patient came from the emergency room with complaints of recurrent vomiting since 3 months ago.
135766047	30391	2022-01-01 06:58:25	HISTORY OF CURRENT ILLNESS	Previously the patient complained of recurrent heartburn, the patient took gastric medications, but over time, the patient found it difficult to swallow.
135766047	30394	2022-01-01 06:58:25	FAMILY HISTORY OF DISEASE	Family history of HT, DM, malignancy is denied.
....	....	....	....	....
135766047	42328	2022-01-01 06:58:25	LIST OF PROBLEMS OR DIAGNOSIS	Endometrial Cyst Surgery History (2009 and 2015)

Data *Preprocessing*

The data that has been collected from the *query* results is then carried out in the data *preprocessing* stage. Through the *library* from the *natural languange processing* (nlp-id) collection and this research utilizes various functions for Indonesian such as *stopwords*, *lammatizers*, and *tokenizing*.

Defining Entities

The basic design for defining *class* entities in the electronic medical record domain in this study can be seen in table 4.5 as follows:

Table 4.2 Named entities used

No.	Entity Class	BIO Tagging	Description
1.	Symptoms	B-sym and I-sym	showing symptoms or signs
2.	Body	B-bod and I-bod	showing body parts
3.	Treatment	B-tre and I-tre	show inspection
4.	Nursing	B-nur and I-nur	demonstrate maintenance actions
5.	Disease	B-dis and I-dis	indicates a disease or diagnosis
6.	Abbreviation	B-abb and I-abb	shows the abbreviation
7.	Other	O	indicates a word that has no meaning

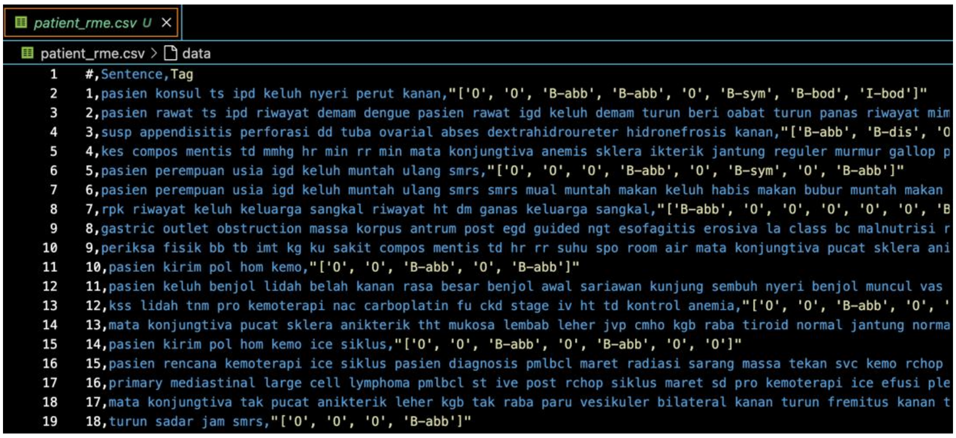
Manual annotation of named entities was carried out for almost two months by researchers together with the hospital medical record committee consisting of internal medicine specialists as electronic medical record evaluators, psychiatric specialists as chairman of the medical record committee, general practitioners as medical service managers, clinical nurses and medical recorders as electronic medical record trainers. The goal was to obtain better entity data and improve the quality of the data. Manual annotation of named entities seen in table 4.6 is used to determine the word class of each word, where this word class will be used to determine words that have entities. This research uses the BIO format which stands for *Beginning* (B), *Inside* (I), and *Outside* (O). This Named entity manual annotation can classify words into 7 Entity *Class* and 13 *Tagging*.

Table 4.3 Example of tagging named entities

Tokenization	Annotated Named Entity Manual
patient	patient, 'O'
Consul	consul, 'O'
ts	ts, 'B-abb'
igd	igd, 'B-abb'
complain	complain, 'O'
pain	pain, 'B-sym'
stomach	belly, 'B-bod'
right	right, 'I-bod'
....	....
patient	patient, 'O'
pro	pro, 'O'
egd	egd,'B-abb'
ligase	ligation, 'B-tre'
ts	ts,'B-abb'
gastro	gastro, 'O'

Research Dataset

A series of extract-transfom-load and data preprocessing processes that have been carried out will obtain data in the database. Figure 4.3 dataset in the form of CSV is the result of the next step from the database and made into a CSV format file.



```
#,Sentence,Tag
1,pasien konsul ts ipd keluh nyeri perut kanan,"['O', 'O', 'B-abb', 'B-abb', 'O', 'B-sym', 'B-bod', 'I-bod']"
2,pasien rawat ts ipd riwayat demam dengue pasien rawat igd keluh demam turun beri oabat turun panas riwayat mim
3,susp appendisitis perforasi dd tuba ovarial abses dextrahidroureter hidronefrosis kanan,"['B-abb', 'B-dis', 'O
4,kes compos mentis td mmhg hr min rr min mata konjungtiva anemis sklera ikterik jantung reguler murmur gallop p
5,pasien perempuan usia igd keluh muntah ulang smrs,"['O', 'O', 'O', 'B-abb', 'O', 'B-sym', 'O', 'B-abb']"
6,pasien perempuan usia igd keluh muntah ulang smrs smrs mual muntah makan keluh habis makan bubur muntah makan
7,rpk riwayat keluh keluarga sangkal riwayat ht dm ganas keluarga sangkal,"['B-abb', 'O', 'O', 'O', 'O', 'O', 'B
8,gastric outlet obstruction massa korpus antrum post egd guided ngt esofagitis erosiva la class bc malnutrisi r
9,periksa fisik bb tb imt kg ku sakit compos mentis td hr rr suhu spo room air mata konjungtiva pucat sklera ani
10,pasien kirim pol hom kemo,"['O', 'O', 'B-abb', 'O', 'B-abb']"
11,pasien keluh benjol lidah belah kanan rasa besar benjol awal sariawan kunjung sembuh nyeri benjol muncul vas
12,kss lidah tnm pro kemoterapi nac carboplatin fu ckd stage iv ht td kontrol anemia,"['O', 'O', 'B-abb', 'O', '
13,mata konjungtiva pucat sklera anikterik tht mukosa lembab leher jvp cmho kgb raba tiroid normal jantung norma
14,pasien kirim pol hom kemo ice siklus,"['O', 'O', 'B-abb', 'O', 'B-abb', 'O', 'O']"
15,pasien rencana kemoterapi ice siklus pasien diagnosis pmlbcl maret radiasi sarang massa tekan svc kemo rchop
16,primary mediastinal large cell lymphoma pmlbcl st ive post rchop siklus maret sd pro kemoterapi ice efusi ple
17,mata konjungtiva tak pucat anikterik leher kgb tak raba paru vesikuler bilateral kanan turun fremitus kanan t
18,turun sadar jam smrs,"['O', 'O', 'O', 'B-abb']"
```

Figure 4.3 CSV dataset

Named Entity Recognition Model

The named entity recognition model built in the experimental environment requires software and hardware for implementation and has certain specifications. The experimental environment specifications in table 4.8 aim to support the performance of the designed system so that it can run properly and have the desired results.

Table 4.4 Research environment specifications

Software	OSx Sonoma 14.0 Visual Studio Code 1.85.1 Python 3.9.2 Xampp for OS X 7.4.33 PHP version 7.4.33 MySQL 5.5.5-10.4.17-MariaDB
Hardware	Processor: 1.4 Ghz Quad Core Intel Core i5 Graphics: Intel Iris Plus Graphics 645 1536 MB Memory: 8 GB 2133 MHz LPDDR3

In addition to some of the software mentioned in table 4.8, the designed model also utilizes several libraries or libraries from python, with details in table 4.9 as follows:

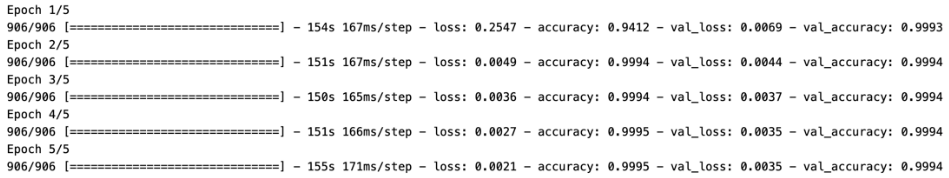
Table 4.5 Library used

Library	Function
import pandas	Manipulate and analyze data, especially tabular data.
import numpy	Perform numerical computations and array operations efficiently.
import sklearn.model_selection	Provides tools to split the dataset into training and testing sets.
import ast	Parses Python expressions from strings, used here to interpret data stored as strings.

import keras.preprocessing.text	Provides tools for text preprocessing, including tokenization.
import tensorflow.keras.preprocessing.sequence	Provides a function to pad the rows of a sequence to a uniform length.
import keras.layers	Provides various layers of neural networks to build the model.
import keras.models	Build and define the neural network model.
import hard.losses	Provide loss function for training optimization.
import keras.callbacks	Provide callbacks to monitor and manage the training process.
import tensorflow	Scientific computing and machine learning.
import sklearn	Provides various machine learning algorithms, including classification, regression, and clustering.
import evaluate	Provides functions for calculating evaluation metrics.
import transformers	Transformers are used for various machine learning tasks, including natural language processing (NLP) tasks such as language translation, language recognition and text classification.

NER Model Training and Evaluation

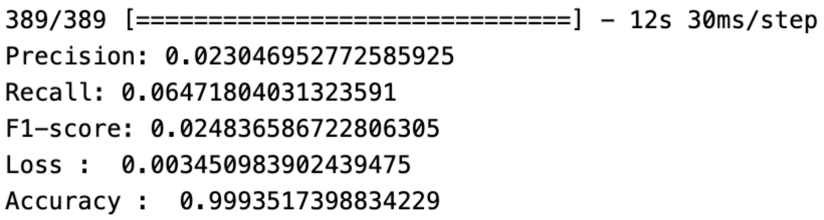
The neural networks used are GRU, bidirectional GRU, LSTM, bidirectional LSTM. The training and evaluation of the neural network-based model whose results are shown in Figure 4.10 were set hyperparameters at the time of model compilation with adam selected as the optimizer, accuracy selected as the metrics and 5 epochs as the number of rounds of model training.



```
Epoch 1/5
906/906 [==============================] - 154s 167ms/step - loss: 0.2547 - accuracy: 0.9412 - val_loss: 0.0069 - val_accuracy: 0.9993
Epoch 2/5
906/906 [==============================] - 151s 167ms/step - loss: 0.0049 - accuracy: 0.9994 - val_loss: 0.0044 - val_accuracy: 0.9994
Epoch 3/5
906/906 [==============================] - 150s 165ms/step - loss: 0.0036 - accuracy: 0.9994 - val_loss: 0.0037 - val_accuracy: 0.9994
Epoch 4/5
906/906 [==============================] - 151s 166ms/step - loss: 0.0027 - accuracy: 0.9995 - val_loss: 0.0035 - val_accuracy: 0.9994
Epoch 5/5
906/906 [==============================] - 155s 171ms/step - loss: 0.0021 - accuracy: 0.9995 - val_loss: 0.0035 - val_accuracy: 0.9994
```

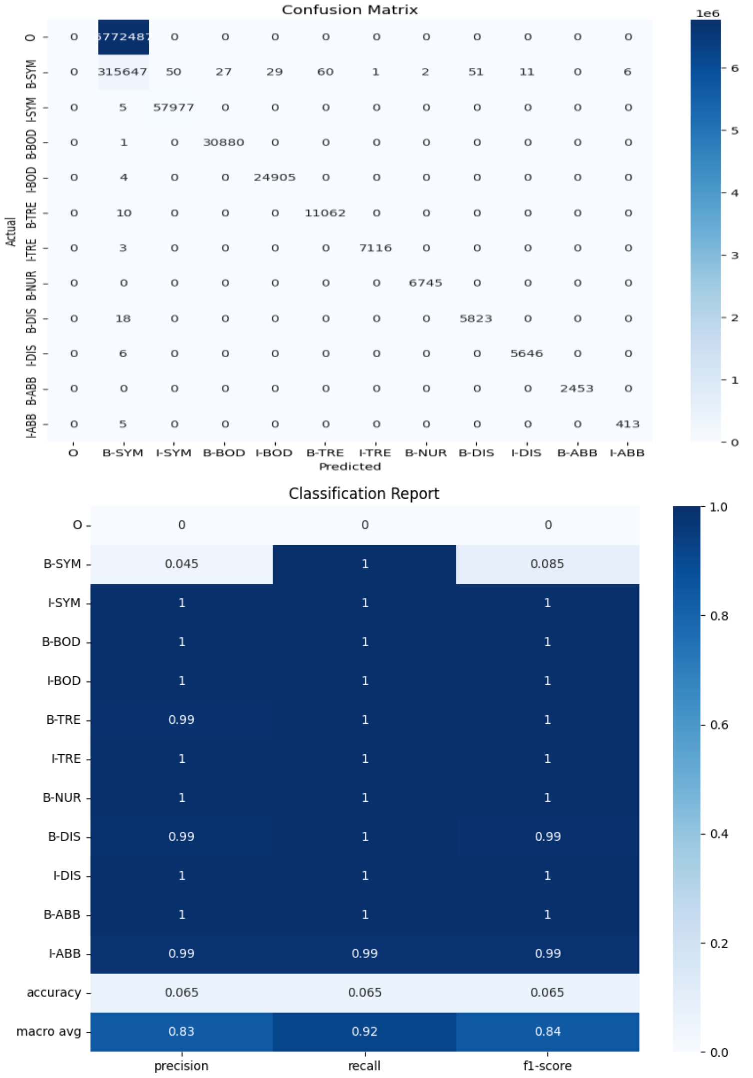
Figure 4.4 Training of neural network-based NER model

Evaluation of the neural network-based model in Figure 4.11 processed with the test dataset to obtain the accuracy, precision, recall, and f1-score metric values.



```
389/389 [==============================] - 12s 30ms/step
Precision: 0.023046952772585925
Recall: 0.06471804031323591
F1-score: 0.024836586722806305
Loss :  0.003450983902439475
Accuracy :  0.9993517398834229
```

Figure 4.5 Evaluation of neural network-based NER model



Confusion Matrix

Actual \ Predicted	O	B-SYM	I-SYM	B-BOD	I-BOD	B-TRE	I-TRE	B-NUR	B-DIS	I-DIS	B-ABB	I-ABB
O	0	6772487	0	0	0	0	0	0	0	0	0	0
B-SYM	0	315647	50	27	29	60	1	2	51	11	0	6
I-SYM	0	5	57977	0	0	0	0	0	0	0	0	0
B-BOD	0	1	0	30880	0	0	0	0	0	0	0	0
I-BOD	0	4	0	0	24905	0	0	0	0	0	0	0
B-TRE	0	10	0	0	0	11062	0	0	0	0	0	0
I-TRE	0	3	0	0	0	0	7116	0	0	0	0	0
B-NUR	0	0	0	0	0	0	0	6745	0	0	0	0
B-DIS	0	18	0	0	0	0	0	0	5823	0	0	0
I-DIS	0	6	0	0	0	0	0	0	0	5646	0	0
B-ABB	0	0	0	0	0	0	0	0	0	0	2453	0
I-ABB	0	5	0	0	0	0	0	0	0	0	0	413

Classification Report

	precision	recall	f1-score
O	0	0	0
B-SYM	0.045	1	0.085
I-SYM	1	1	1
B-BOD	1	1	1
I-BOD	1	1	1
B-TRE	0.99	1	1
I-TRE	1	1	1
B-NUR	1	1	1
B-DIS	0.99	1	0.99
I-DIS	1	1	1
B-ABB	1	1	1
I-ABB	0.99	0.99	0.99
accuracy	0.065	0.065	0.065
macro avg	0.83	0.92	0.84

Figure 4.6 Confusion matrix based on neural network

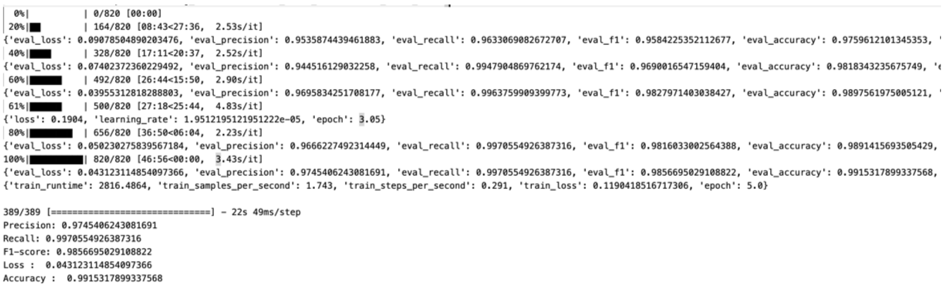
Confusion matrix and classification report on the neural network-based named entity recognition model can be seen in Figure 4.12 and summarized evaluation comparison in Table 4.12 shows all neural network-based named entity recognition models have varying values.

Table 4.6 Evaluation of neural network model

Model	Precision	Recall	F1-Score	Loss	Acuracy
LSTM	0.00494	0.06471	0.00901	0.00354	0.99934
RNN-biLSTM	0.02304	0.06471	0.02483	0.00345	0.99935
GRU	0.00494	0.06471	0.00902	0.00396	0.99937
RNN-biGRU	0.02304	0.06471	0.02483	0.00341	0.99932

The evaluation results of all neural network-based named entity recognition models in table 4.12 can be summarized as follows: precision of 0.02304 shows that the model only correctly predicts 2.30% of labels, recall of 0.06471 shows that the model only detects 6.47% of correct labels, f1-score of 0.0248 shows that the model has low precision and recall, loss of 0.00345 shows that the model has low error, accuracy of 0.99935 shows that the model has very high accuracy.

In general and as can be seen in Figure 4.13, the neural network-based named entity recognition model has low precision, recall, and f1-score, but has very high accuracy. This shows that the neural network-based named entity recognition model is very good at predicting correct labels, but often misidentifies correct labels as false labels.



```
 0%|          | 0/820 [00:00]
 20%|█         | 164/820 [08:43<27:36,  2.53s/it]
{'eval_loss': 0.09078504890203476, 'eval_precision': 0.9535874439461883, 'eval_recall': 0.9633069082672707, 'eval_f1': 0.9584225352112677, 'eval_accuracy': 0.9759612101345353, '
 40%|██        | 328/820 [17:11<20:37,  2.52s/it]
{'eval_loss': 0.07402372360229492, 'eval_precision': 0.944516129032258, 'eval_recall': 0.9947904869762174, 'eval_f1': 0.9690016547159404, 'eval_accuracy': 0.9818343235675749, 'e
 60%|███       | 492/820 [26:44<15:50,  2.90s/it]
{'eval_loss': 0.03955312818288803, 'eval_precision': 0.9695834251708177, 'eval_recall': 0.9963759909399773, 'eval_f1': 0.9827971403038427, 'eval_accuracy': 0.9897561975005121, '
 61%|███       | 500/820 [27:18<25:44,  4.83s/it]
{'loss': 0.1904, 'learning_rate': 1.9512195121951222e-05, 'epoch': 3.05}
 80%|████      | 656/820 [36:50<06:04,  2.23s/it]
{'eval_loss': 0.050230275839567184, 'eval_precision': 0.9666227492314449, 'eval_recall': 0.9970554926387316, 'eval_f1': 0.9816033002564388, 'eval_accuracy': 0.9891415693505429,
100%|█████| 820/820 [46:56<00:00,  3.43s/it]
{'eval_loss': 0.043123114854097366, 'eval_precision': 0.9745406243081691, 'eval_recall': 0.9970554926387316, 'eval_f1': 0.9856695029108822, 'eval_accuracy': 0.9915317899337568,
{'train_runtime': 2816.4864, 'train_samples_per_second': 1.743, 'train_steps_per_second': 0.291, 'train_loss': 0.1190418516717306, 'epoch': 5.0}

389/389 [==============================] - 22s 49ms/step
Precision: 0.9745406243081691
Recall: 0.9970554926387316
F1-score: 0.9856695029108822
Loss :  0.043123114854097366
Accuracy :  0.9915317899337568
```

Figure 4.7 Example of BERT-based model evaluation results

The evaluation of the named entity recognition model based on bidirectional encoder representations from transformers can be seen in Figure 4.13 with an accuracy value of 0.99153 or 99% in recognizing named entities, of course with a loss value of 0.04312 or 0.4% indicating a low error.

Table 4.7 BERT-based NER model evaluation tabulation

Hyperparameters Pretrained indolem/IndoBERT					
No.	**Precision**	**Recall**	**F1**	**Loss**	**Accuracy**
1.	0.94982	0.99048	0.96973	0.07344	0.98210
2.	0.96022	0.99524	0.97742	0.06107	0.98634
3.	0.97171	0.99592	0.98366	0.05057	0.99043
4.	0.94748	0.98890	0.96774	0.06859	0.98108
5	**0.97454**	**0.99705**	**0.98566**	**0.04312**	**0.99153**

Training and evaluation of named entity recognition models based on bidirectional encoder representations from transformers with limited device capabilities, hyperparameters are set which can be seen in table 4.11 then the results in table 4.13 are the best in model 5 can be concluded as follows: precision of

0.97454 shows that the model only misidentifies 0.03% of labels, recall of 0.99705 shows that the model only misses 0.01% of correct labels, f1-score of 0.98566 shows that the model has very high precision and recall, loss of 0.04312 shows that the model has low error, accuracy of 0.99153 shows that the model has very high accuracy.

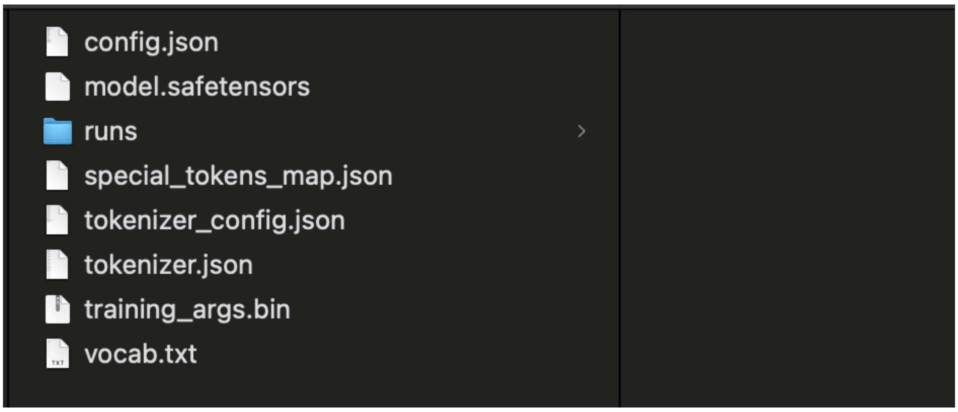
In general, the named entity recognition model based on bidirectional encoder representations from transformers has very high precision, recall, and f1-score, and has low loss. This shows that the named entity recognition model based on bidirectional encoder representations from transformers is very good at predicting correct labels and does not misidentify correct labels as false labels.

From the comparison of the two approaches in building the named entity recognition model, it can be concluded that the bidirectional encoder representations from transformers-based approach consistently outperforms the neural network-based approach in every evaluation metric. The bidirectional encoder representations from transformers approach in named entity recognition mode has very high precision, recall, and f1-score values, indicating its ability to recognize entities very well, Although the neural network-based named entity recognition model has high accuracy, the low precision and recall values indicate that this model may have difficulty in recognizing entities accurately, the significant difference between the evaluation results of the two shows that the bidirectional encoder representations from transformers approach has an advantage in the named entity recognition task compared to the neural network.

Application of the NER Model

In this research, the bidirectional encoder representations from transformers approach in the named entity recognition model is selected for implementation or deployment. The named entity recognition model based on bidirectional encoder representations from transformers is set to provide maximum results and then enter the implementation or deployment stage.

The named entity recognition model based on bidirectional encoder representations from transformers that have been compiled in Figure 4.14 will form a model.safetensors, tokenizer.json file that can be called in other files or inference API (Application Programming Interface) to be run as new learning.



config.json
model.safetensors
runs
special_tokens_map.json
tokenizer_config.json
tokenizer.json
training_args.bin
vocab.txt

Figure 4.8 Compiled BERT-based model

The named entity recognition model based on bidirectional encoder representations from transformers will be embedded in a web-based information system in the electronic medical record module, where users will enter sentences then the inference API (Application Programming Interface) from huggingface.co as a back-end will provide named entity output from the process of the named entity recognition model based on bidirectional encoder representations from transformers that have previously been built.

Hyperparameters Result Comparison

From table 4.13, we can see that the loss value decreases from 0.07344 in model 1 to 0.06107 in model 2 and 0.05057 in model 3 as shown in figure 4.16. This shows that the model becomes more accurate in recognizing named entities. The precision, recall, and f1-score values also show the same trend. The precision and recall values increase from model 1 to model 3, which shows that the model becomes better at recognizing the correct named entities. The f1-score value also increased, indicating that the model became better overall at recognizing named entities. The accuracy value also increased from model 1 to model 3, indicating that the model became better at correct named entity recognition. Overall, it can be concluded that the IndoBERT pretrained model becomes more accurate in named entity recognition as the loss value decreases and the precision, recall, and f1-score values increase.

Model 1 has a fairly high loss value, which is 0.07344. This shows that the model still has errors in recognizing named entities. The precision and recall values of this model are also quite high, which are 0.94982 and 0.99048 respectively. However, the f1-score value of this model is lower than its precision and recall values, which are 0.96973. This shows that the model still has errors in recognizing correctly named entities. The accuracy value of this model is also quite high, which is 0.98210.

Model 2 has a lower loss value than model 1, which is 0.06107. This shows that the model becomes more accurate in recognizing named entities. The precision and recall values of this model also increased from model 1, which are 0.96022 and 0.99524 respectively. The f1-score value of this model also increased from model 1, which is 0.97742. This shows that the model has become better overall at recognizing named entities. The accuracy value of this model also increased from model 1, which is 0.98634.

Model 3 has the lowest loss value of the three models, which is 0.05057. This shows that the model is the most accurate in recognizing named entities. The precision and recall values of this model are also the highest of the three models, which are 0.97171 and 0.99592 respectively. The f1-score value of this model is also the highest of the three models, which is 0.98366. This shows that the model is the best overall at recognizing named entities. The accuracy value of this model is also the highest of the three models, which is 0.99043.

Model 4 has a higher loss value than the third model, which is 0.06859. This shows that the third model is still the most accurate in named entity recognition. The precision and recall values of this model are also lower than the third model, which are 0.94748 and 0.98890 respectively. The f1-score value of this model is lower than the third model, which is 0.96774. This shows that the model is not as good overall as the third model in named entity recognition. The accuracy value of this model is also lower than the third model, which is 0.98108.

Model 5 has the lowest loss value of the five models, which is 0.04312. This shows that the model is the most accurate in recognizing named entities. The precision and recall values of this model are also the highest of the five models, which are 0.97454 and 0.99705 respectively. The f1-score value of this model is also the highest of the five models, which is 0.98566. This shows that the model is the best overall at recognizing named entities. The accuracy value of this model is also the highest of the three models, which is 0.99153.

The hyperparameters set in model 4 of model 3 are differentiated in train_batch_size and eval_batch_size with higher values, but have smaller confusion matrix results ranging from precision, recall, f1-score, loss and accuracy values. While it is interesting when the hyperparameter set in model 5 of model 3 is differentiated in train_bacth_size and eval_batch_size with smaller values, but has better confusion matrix results than model 3.

CONCLUSION

This research shows important contributions in the development of deep learning models for the task of named entity recognition in Indonesian electronic medical records. The named entity recognition model based on bidirectional encoder representations from transformers achieved 99% evaluation accuracy with optimal hyperparameters settings. The model effectively identifies entities such as symptoms, diseases, abbreviations, body parts, examinations, and treatments in medical record text, with the ability to accurately predict labels and reduce misidentification. Experiments showed that the use of data preprocessing and entity tagging facilitated the automatic labeling process, and training using pre-trained IndoBERT contributed to superior results compared to previous neural network-based models. In addition, this research produced a prototype system for detecting named entities in electronic medical records, which can serve as a basis for further development. However, this study also identified that to achieve evaluation standards close to human capabilities, future developments need to expand the scope of entities and consider the use of specialized pre-trained models in Indonesian-language biomedical or medical domains, as well as the addition of categories such as radiology, clinical pathology, anatomical pathology, and microbiology in the classification of named entities.

REFERENCES

Abdillah, A.F., Purwitasari, D., Junita, S., Purnomo, M.H., 2023. Pengenalan Entitas Biomedis dalam Teks Konsultasi Kesehatan Online Berbahasa Indonesia Berbasis Arsitektur Transofmers. Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) 10, 131–140. https://doi.org/10.25126/jtiik.2023106337

AI-Ash, H.S., Fanany, I., Bustamam, A., 2019. Indonesian Protected Health Information Removal using Named-Entity Recognition, in: International Conference on Information & Communication Technology and System.

Alsaaran, N., Alrabiah, M., 2021. Classical Arabic Named-Entity Recognition Using Variant Deep Neural Network Architectures and BERT. IEEE Access 9, 91537–91547. https://doi.org/10.1109/ACCESS.2021.3092261

Awalina, A., Bachtiar, F.A., Utaminingrum, F., Korespondensi, P., 2022. Perbandingan Model Transformer pada Deteksi Ulasan Palsu. Jurnal Teknologi Informasi dan Ilmu Komputer 9, 597–604. https://doi.org/10.25126/jtiik.202295696

Berragan, C., Singleton, A., Calafiore, A., Morley, J., 2023. Transformer Based Named-Entity Recognition for Place Name Extraction from Unstructured Text. International Journal of Geographical Information Science 37, 747–766. https://doi.org/10.1080/13658816.2022.2133125

Chen, P., Zhang, M., Yu, X., Li, S., 2022. Named-Entity Recognition of Chinese Electronic Medical Records Based on A Hybrid Neural Network and Medical MC-BERT. BMC Med Inform Decis Mak 22. https://doi.org/10.1186/s12911-022-02059-2

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation, in: Conference on Empirical Methods in Natural Language Processing (EMNLP).

Devlin, J., Chang, M.-W., Lee, K., Google, K.T., Language, A.I., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of NAACL-HLT. pp. 4171–4186.

Dey, R., Salem, F.M., 2017. Gate-Variants of Gated Recurrent Unit (GRU) Neural Networks, in: IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS).

Fakhri, M., Putra, D.A., Fathan Hidayatullah, A., Wibowo, A.P., Nastiti, K.R., Yogyakarta, T., Siliwangi, J., Utara, R., Id, A.A., 2023. Named-Entity Recognition of Tourist Destinations Reviews in The Indonesian Language. Jurnal Linguistik Komputasional 6.

Giarsyani, N., Hidayatullah, A.F., Rahmadi, R., 2020. Komparasi Algoritma Machine Learning dan Deep Learning untuk Named-Entity Recognition : Studi Kasus Data Kebencanaan. Jurnal Informatika & Rekayasa Elektronika 3.

Gunawan, W., Suhartono, D., Purnomo, F., Ongko, A., 2018. Named-Entity Recognition for Indonesian Language using BiLSTM-CNNs, in: Procedia Computer Science. Elsevier B.V., pp. 425–432. https://doi.org/10.1016/j.procs.2018.08.193

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., Lew, M.S., 2016. Deep Learning for Visual Understanding: A Review. Neurocomputing 187, 27–48. https://doi.org/10.1016/j.neucom.2015.09.116

Haris, M., Pustaka, T., Diponegoro, M.H., Kusumawardani, S., Hidayah, I., 2021. Tinjauan Pustaka Sistematis: Implementasi Metode Deep Learning pada

Prediksi Kinerja Murid. Jurnal Nasional Teknik Elektro dan Teknologi Informasi | 10.

Hatta Fudholi, D., Abida Nayoan, R.N., Fathan Hidayatullah, A., Brahma Arianto, D., 2022. A Hybrid CNN-BiLSTM Model for Drug Named-Entity Recognition. Journal of Engineering Science and Technology 17, 730–0744.

Hutter, F., Lücke, J., Schmidt-Thieme, L., 2015. Beyond Manual Tuning of Hyperparameters. KI - Kunstliche Intelligenz 29, 329–337. https://doi.org/10.1007/s13218-015-0381-0

Jaariyah, N., Rainarli, E., 2017. Conditional Random Fields untuk Pengenalan Entitas Bernama pada Teks Bahasa Indonesia. Jurnal Ilmiah Komputer dan Informatika 6.

Jati, B.S., Widyawan, S.T., Muhammad Nur Rizal, S.T., 2020. Multilingual Named-Entity Recognition Model for Indonesian Health Insurance Question Answering System, in: International Conference on Information and Communications Technology. Institute of Electrical and Electronics Engineers Inc., pp. 180–184. https://doi.org/10.1109/ICOIACT50329.2020.9332027

Kale, S., Govilkar, S., 2017. Survey of Named-Entity Recognition Techniques for Various Indian Regional Languages. Int J Comput Appl 164, 37–43. https://doi.org/10.5120/ijca2017913621

Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S.F., Forshee, R., Walderhaug, M., Botsis, T., 2017. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. J Biomed Inform. https://doi.org/10.1016/j.jbi.2017.07.012

Li, J., Sun, A., Han, J., Li, C., 2022. A Survey on Deep Learning for Named-Entity Recognition. IEEE Trans Knowl Data Eng 34, 50–70. https://doi.org/10.1109/TKDE.2020.2981314

Li, S., Lin, N., Xiao, L., Jiang, S., 2020. IndoAbbr: A New Benchmark Dataset for Indonesian Abbreviation Identification, in: International Conference on Asian Language Processing. Institute of Electrical and Electronics Engineers Inc., pp. 241–246. https://doi.org/10.1109/IALP51396.2020.9310514

Liu, X., Zhou, Y., Wang, Z., 2019. Recognition and Extraction of Named Entities in Online Medical Diagnosis Data Based on a Deep Neural Network. J Vis Commun Image Represent 60, 1–15. https://doi.org/10.1016/j.jvcir.2019.02.001

Luo, L. xia, 2019. Network Text Sentiment Analysis Method Combining LDA Text Representation and GRU-CNN. Pers Ubiquitous Comput 23, 405–412. https://doi.org/10.1007/s00779-018-1183-9

Luthfi, E.T., Izzah, Z., Yusoh, M., Aboobaider, B.M., 2022. BERT Based Named-Entity Recognition for Automated Hadith Narrator Identification. International Journal of Advanced Computer Science and Applications 13.

Munarko, Y., Sutrisno, M.S., Mahardika, W.A.I., Nuryasin, I., Azhar, Y., 2018. Named-Entity Recognition Model for Indonesian Tweet using CRF Classifier, in: IOP Conference Series: Materials Science and Engineering. Institute of Physics Publishing. https://doi.org/10.1088/1757-899X/403/1/012067

Nurchim, N., Nurmalitasari, N., Long, Z.A., 2023. Indonesian News Classification Application with Named-Entity Recognition Approach. Jurnal Infotel 15. https://doi.org/10.20895/infotel.v15i2.909

Otter, D.W., Medina, J.R., Kalita, J.K., 2021. A Survey of the Usages of Deep Learning for Natural Language Processing. IEEE Trans Neural Netw Learn Syst 32, 604–624. https://doi.org/10.1109/TNNLS.2020.2979670

Parsaoran, A., Sitompul, H., 2023. Penggunaan Rekam Medis Elektronik Untuk Pasien Rawat Jalan Di Fasilitas Kesehatan Indonesia : Literature Review. Jurnal Ilmiah Multidisipline 37, 37–46. https://doi.org/10.5281/zenodo.7837905

Peraturan Menteri Kesehatan Republik Indonesia Nomor 24 Tahun 2022 Tentang Rekam Medis, 2022.

Putra, F.N., Fatichah, C., 2018. Klasifikasi Jenis Kejadian Menggunakan Kombinasi neuroNER dan Recurrent Convolutional Neural Network pada Data Twitter. Jurnal Ilmiah Teknologi Sistem Informasi 4, 81–90. https://doi.org/10.26594/register.v4i2.1242

Rachman, V., Savitri, S., Augustianti, F., Mahendra, R., 2020. Named-Entity Recognition on Indonesian Twitter Posts using LSTM Networks, in: International Conference on Advanced Computer Science and Information Systems.

Rashel, F., Luthfi, A., Dinakaramani, A., Manurung, R., 2014. Building an Indonesian Rule-Based Part-of-Speech Tagger, in: Proceedings of the International Conference on Asian Language Processing 2014 (IALP 2014).

Rochmawati, N., Hidayati, H.B., Yamasari, Y., Peni, H., Tjahyaningtijas, A., Yustanti, W., Prihanto, A., 2021. Analisa Learning Rate dan Batch Size pada Klasifikasi Covid menggunakan Deep Learning dengan Optimizer Adam. Journal Information Engineering and Educational Technology 2.

Rothman, D., 2021. Transformers for Natural Language Processing.

Russell, S., Norvig, P., 2010. Artificial Intelligence A Modern Approach.

Sugiyono., 2019. Metodelogi Penelitian Kuantitatif dan Kualitatif dan R&D. Bandung: CV. Alfabeta.

Salminen, J., Hopf, M., Chowdhury, S.A., Jung, S. gyo, Almerekhi, H., Jansen, B.J., 2020. Developing an Online Hate Classifier for Multiple Social Media Platforms. Human-centric Computing and Information Sciences 10. https://doi.org/10.1186/s13673-019-0205-6

Santoso, J., Setiawan, E.I., Yuniarno, E.M., Hariadi, M., Purnomo, M.H., 2020. Hybrid Conditional Random Fields and K-Means for Named-Entity Recognition on Indonesian News Documents. International Journal of Intelligent Engineering and Systems 13, 233–245. https://doi.org/10.22266/IJIES2020.0630.22

Susanty, M., Sukardi, S., 2021. Perbandingan Pre-Trained Word Embedding dan Embedding Layer untuk Named-Entity Recognition Bahasa Indonesia. Jurnal Pengkajian dan Penerapan Teknik Informatika 14, 247–257. https://doi.org/10.33322/petir.v14i2.1164

Tunstall, L., Werra, L. von, Wolf, T., 2022. Natural Language Processing with Transformers.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention Is All You Need, in: Advances in Neural Information Processing Systems.

Wang, S., Cao, J., 2021. AI and Deep Learning for Urban Computing, in: Urban Book Series. Springer Science and Business Media Deutschland GmbH, pp. 815–844. https://doi.org/10.1007/978-981-15-8983-6_43

Wibawa, M.S., 2016. Pengaruh Fungsi Aktivasi, Optimisasi dan Jumlah Epoch Terhadap Performa Jaringan Saraf Tiruan. Jurnal Sistem dan Informatika 11.

Wintaka, D.C., Bijaksana, M.A., Asror, I., 2019. Named-Entity Recognition on Indonesian Tweets using BiLSTM-CRF, in: Procedia Computer Science. Elsevier B.V., pp. 221–228. https://doi.org/10.1016/j.procs.2019.08.161

Young, T., Hazarika, D., Poria, S., Cambria, E., 2018. Recent Trends in Deep Learning Based Natural Language Processing. IEEE Comput Intell Mag 13, 55–75. https://doi.org/10.1109/MCI.2018.2840738

Yusliani, N., Sufa, M.R.P., Firdaus, A., Abdiansah, Sazaki, Y., 2021. Named-Entity Recognition pada Teks Berbahasa Indonesia menggunakan Metode Hidden Markov Model dan POS-Tagging. Jurnal Linguistik Komputasional 4.