

# Implementing deep learning-based named entity recognition for obtaining narcotics abuse data in Indonesia

Daris Azhar<sup>1</sup>, Robert Kurniawan<sup>1</sup>, Waris Marsisno<sup>1</sup>, Budi Yuniarto<sup>1</sup>, Sukim<sup>2</sup>, Sugiarto<sup>2</sup>

<sup>1</sup>Faculty of Statistical Computing, Politeknik Statistika STIS, Jakarta, Indonesia

<sup>2</sup>Faculty of Statistics, Politeknik Statistika STIS, Jakarta, Indonesia

## Article Info

### Article history:

Received Aug 21, 2022

Revised Jan 13, 2023

Accepted Mar 10, 2023

### Keywords:

Application of named entity recognition  
Bidirectional-long short-term memory-conditional random field named entity recognition  
Convolutional neural network-long short-term memory named entity recognition  
Named entity recognition  
Narcotics abuse

## ABSTRACT

The availability of drug abuse data from the official website of the National Narcotics Board of Indonesia is not up-to-date. Besides, the drug reports from Indonesian National Narcotics Board are only published once a year. This study aims to utilize online news sites as a data source for collecting information about drug abuse in Indonesia. In addition, this study also builds a named entity recognition (NER) model to extract information from news texts. The primary NER model in this study uses the convolutional neural network-long short-term memory (CNNs-LSTM) architecture because it can produce a good performance and only requires a relatively short computation time. Meanwhile, the baseline NER model uses the bidirectional long short-term memory-conditional random field (Bi-LSTMs-CRF) architecture because it is easy to implement using the Flair framework. The primary model that has been built results in a performance (F1 score) of 82.54%. Meanwhile, the baseline model only results in a performance (F1 score) of 69.67%. Then, the raw data extracted by NER is processed to produce the number of drug suspects in Indonesia from 2018-2020. However, the data that has been produced is not as complete as similar data sourced from Indonesian National Narcotics Board publications.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Robert Kurniawan

Faculty of Statistical Computing, Politeknik Statistika STIS

Jln. Otto Iskandardinata No. 64C, Jatinegara, Jakarta Timur, DKI Jakarta, Indonesia

Email: robertk@stis.ac.id

## 1. INTRODUCTION

Named entity recognition (NER), which is a part of natural language processing (NLP), has been widely used in various studies and is still under improvement until now. Several researchers have developed NER model architectures ranging from rule-based NER [1]–[3] to deep learning NER [4], [5] on the corpus of certain languages and in the general domain and certain domains. In general, NER research tends to focus more on building various types of NER model architecture in a particular language corpus to produce the best performing NER model architecture [6]–[8].

NER research with the Indonesian language corpus is generally an implementation of the existing NER model architecture and then applied to the Indonesian language corpus with a general domain coverage [5], [9], [10]. Research related to NER with a general domain scope is usually characterized by the use of general entities such as name of person (PER), name of place/location (LOC), and name of organization (ORG) without any specific purpose for what needs these entities are chosen. Research related to the application of Indonesian language NER in certain specific domains, such as [11]–[14], is quite rare. In fact, many applications of NER can be carried out in certain specific domains, such as the application of NER to extract

information on dietary recommendations [15], the application of NER in the biomedical field [16], the application of NER for cyber security [17], to the application of NER in agriculture [18].

On the other hand, in the worrisome threat of drug danger in Indonesia, there are still problems related to the availability of drug data by the National Narcotics Board that is not up-to-date and the relatively long period of publication of drug case reports (one year). The Indonesian National Narcotics Board should be able to provide up-to-date information that is easily accessible to the public as it is one of the state institutions that plays important role for preventing drug abuse in Indonesia. Based on the problems described, online news can be an alternative to obtain various information related to drug cases in Indonesia. Moreover, the information related to drug cases in online news is relatively easy to find. In addition, the information contained in online news is up-to-date and actual information. These can be a basis for making online news a source of data in obtaining information related to drug cases in Indonesia.

NER is part of NLP that usually uses unstructured sources such as text as data sources [19]. In this study, various information related to drug cases in Indonesia is not manually collected directly from an online news text. There needs to be a tool that can extract entities that contain information related to drug cases from the news text. Therefore, the NER model can be used to extract the required entities in this study.

Recently, deep learning-based NER has outperformed its previous NER methods, such as rule-based NER and machine learning-based NER [20]. The architecture of deep learning-based NER is also quite diverse. However, [8] state that the CNNs-LSTM NER model architecture can produce good performance by requiring relatively shorter computational times. In addition, using a framework in building the NER model will simplify the modeling process. One common NER framework is Flair which already provides the Bi-LSTMs-CRF NER architecture that refers to [7].

Based on the explanation above, this study utilizes the NER model to extract drug case data from online news in Indonesia. The primary NER model architecture used in this study is CNNs-LSTM NER which refers to [8]. In addition, this study also utilizes the flair framework to build the baseline NER using the Bi-LSTMs-CRF architecture that refers to [7]. The baseline NER model is used as a benchmark to see whether or not the performance produced by the primary NER is good.

## **2. METHOD**

### **2.1. Data collection**

The study began with data collection by scraping on one of the online news sites in Indonesia, Okezone. This scraping process gets carried out using the Selenium module with the Python running on Google Collaboratory. The news text data taken is news text related to drug cases. The scraping process uses "drug abuse case" as the search keyword. Elements of news articles collected in the scraping process include article titles, article links, article release dates, and the content of the news text of the article. The articles collected are all searched articles results starting from articles with dates from November 14, 2007, to November 15, 2021. The total number of news articles collected was 3,379 news articles.

### **2.2. Data processing**

This section begins by manually filtering the collected articles. Unrelated articles are not included in the dataset. News articles used for the dataset are articles that provide information related to the arrest of drug abusers and contain information related to other entities, and articles that contain information related to the accumulation of drug cases that occurred in certain areas and at certain times. In addition, the most informative article is chosen if there is a similarity between two or more articles. There are 1,411 articles selected for the dataset of this study. Furthermore, data cleaning is done by removing unnecessary characters and redundant spaces. In addition, all characters in the text are lowercase.

The following process is tokenization which is the process of separating the text into small units called tokens. There are 421,634 tokens in total from the dataset. Each token gets labeled according to its entity category. The entities used in this study include suspect identity (SUS), date (DATE), type of drug evidence (NAR), drug evidence unit (UNIT), location of city or province where the suspect was caught (LOC), number of the case (CASE), miscellaneous (MISC), and other (O). This study uses the begin inside other (BIO) scheme for the labeling process. A total of 45,348 tokens are labeled as primary entities, and the remaining 376,286 tokens are other (O) entities. The number of tokens for each entity category is shown in Table 1. The dataset is split into three parts. Those are training data (80% of the dataset), validating data (10% of the dataset), and testing data (10% of the dataset). The number of articles used in training data, validating data, and testing data are 1,142 articles, 127 articles, and 142 articles, respectively.

Table 1. Number of token for each entity

Entity	Number of Tokens	Entity	Number of Tokens
B-CASE	119	I-DATE	4,912
B-DATE	1,810	I-LOC	1,392
B-LOC	5,007	I-MISC	3,196
B-MISC	3,914	I-NAR	1,119
B-NAR	6,591	I-SUS	1,914
B-SUS	9,110	I-UNIT	3,098
B-UNIT	3,055	O	376,286
I-CASE	111	-	-

### 2.3. Word embedding

Word embedding is a representation of a word in a vector form. Some commonly used word embeddings are Word2Vec [21], GloVe [22], and FastText [23]. This study uses FastText pre-trained word embedding. The NER model that uses pre-trained word embedding has better performance than the model that uses ordinarily trained word embedding [7]. The NER model from [5] resulted in much better performance when using FastText as word embedding compared to when using Word2Vec and GloVe. In addition, FastText has provided word embedding for the Indonesian language corpus.

### 2.4. NER modeling

In this study, NER modeling consists of three parts [8] which are character-level and word-level encoder and tag decoder. Character-level and word-level encoder of the primary NER model uses CNN architecture, while the baseline model uses Bi-LSCRFTM architecture. Meanwhile, the tag decoder part of the primary NER model uses the LSTM architecture, while the baseline model uses CRF. The hyperparameters tuning of the primary NER model is carried out on the type of optimizer, the amount of the learning rate value, and the number of dropout layers in the model. The primary NER model uses the Keras module, while the baseline model uses the Flair framework.

### 2.5. Model evaluation

Evaluation of the NER model is generally done by calculating the value of precision, recall, and F1 score [24]. These three values are obtained by first calculating the number of false positives (FP), false negatives (FN), and true positives (TP). False positive is an outcome where the model incorrectly predicts the primary entity (other than the O entity), whereas true positive is an outcome where the model correctly predicts the primary entity. A false negative is an outcome where the model incorrectly predicts the O entity.

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

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

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall} \quad (3)$$

### 2.6. Further processing of extracted data

This study also shows an overview in the form of data obtained from the extraction by the primary NER model. The data used in this section are news articles from the dataset (the same dataset that is also used for NER modeling) released only from 2018 to 2020 that are then entered into the primary NER model to perform the extraction stage. The number of articles used for each year is 99 for 2018, 98 for 2019, and 112 for 2020. The data extracted by the NER model is still raw. Thus, further processing is needed so that the resulting data will be unique for each case and each entity (no duplication of entities for each occurrence).

In this study, further processing of the extracted entity only processes the SUS entity. The purpose is to obtain the number of drug suspect data. Further processing is carried out using simple rules and has considered the uniqueness of the entity (eliminating redundant entity). The description of the extracted data displayed is data related to the number of suspected drug cases in Indonesia from 2018 to 2020. Then, the data get compared with similar data sourced from the annual publication of the National Narcotics Board.

## 3. RESULTS AND DISCUSSION

### 3.1. Architecture of CNNs-LSTM NER (Primary NER)

The first step is to download the FastText model for the Indonesian language corpus. The dimension of the word vector used is 300. This number is chosen based on [5] which utilizes FastText for the word

embedding and uses a word vector dimension of 300. Another thing to be prepared before carrying out the training process on the model is the input data for the modeling. The input data include word vectors, character indexes for each token, and label indexes.

The NER modeling carried out on the primary model consists of three parts (character-level encoder, word-level encoder, and tag decoder) [8]. The primary NER model uses the CNN architecture for the character-level encoder. This CNN architecture is almost similar to the architecture used in [8]. Figure 1(a) shows the architecture of the primary NER model for the character-level encoder section.

The character-level encoder section begins with embedding all character indexes using a uniform distribution with the interval -0.5 to 0.5 [6]. After that, the results are entered into two convolutional layers. Each convolutional layer uses 50 filters, kernel widths of 3, 1 stride, and a rectified linear unit (ReLU) activation function [8]. There is a dropout layer between the two convolutional layers with a 0.5 dropout rate value. After going through the convolutional layer, the resulting vectors are entered into the max pooling layer. The chosen vector will be concatenated with the word vector from the previous word embedding.

The following part is the word-level encoder. The architecture is shown in Figure 1(b) and is almost similar to the previous part. However, this section does not have a max pooling layer and has a different concatenating vector stage at the end. This part consists of two convolutional layers and a dropout layer between the two convolutional layers. Each convolutional layer uses 800 filters, kernel widths of 5, 1 stride, and a ReLU activation function [8]. The resulting vectors are concatenated with the input vectors from this section.

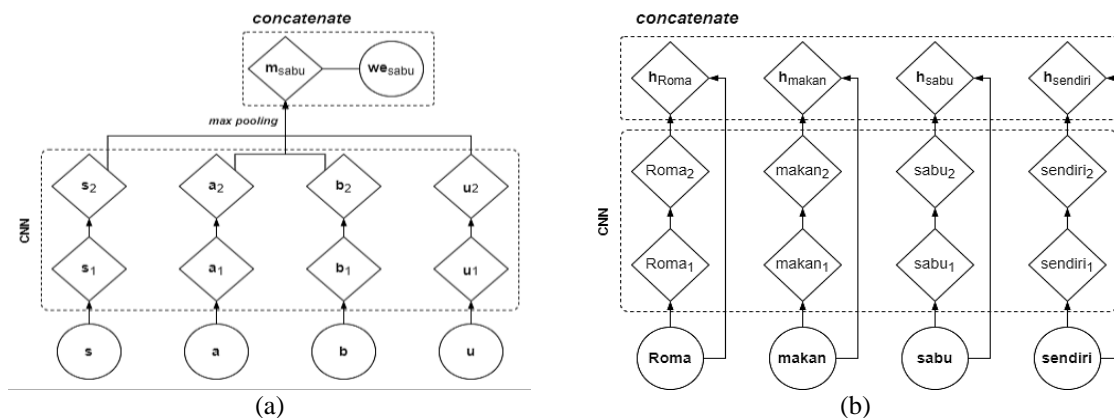


Figure 1. The architecture of, (a) character-level encoder and (b) word-level encoder of CNNs-LSTM NER

The last part is the tag decoder section. The architecture used in this section is the LSTM architecture shown in Figure 2. There is no architectural difference in the tag decoder part between this study and [8]. This section consists of one LSTM layer and a dense layer at the end. The LSTM layer uses a unit value of 250 (the dimension of the resulting vector is 250) and is close to the unit value used for the LSTM layer in [6], with a value of 275. Other parameter values determined in this part are a sigmoid activation function in the recurrent section, a tanh activation function for the output section, a 0.5 dropout rate value for the input section, and a 0.25 dropout rate value for the recurrent part.

The output of the LSTM layer will become an input for the softmax activation function. Softmax produces a value of probability for each element in its output vector. If all elements in an output vector are summed up, the result will be worth one. The output vector index whose corresponding element is the highest will become the predicted label index by the model.

### 3.2. Hyperparameters tuning

The hyperparameters tuning is carried out on the type of optimizer, the magnitude of the learning rate value, and the number of dropout layers in the model. This study uses adaptive moment estimation (Adam) and Nesterov-accelerated adaptive moment estimation (Nadam) as Adam was also used by [8] in its modeling, and Nadam was able to produce the best NER model performance compared to the use of other optimizers in [4]. This study uses 0.001 for the learning rate as it was also used in [8]. Meanwhile, 0.002 is used for the learning rate as this is the best learning rate value for the Adam or Nadam optimizer based on [25]. This study adds two more dropout layers to reduce overfitting in the model [26]. The two additional layers are each placed in the

section after concatenating at the word-level encoder stage and in the section after passing through the LSTM layer at the tag decoder stage.

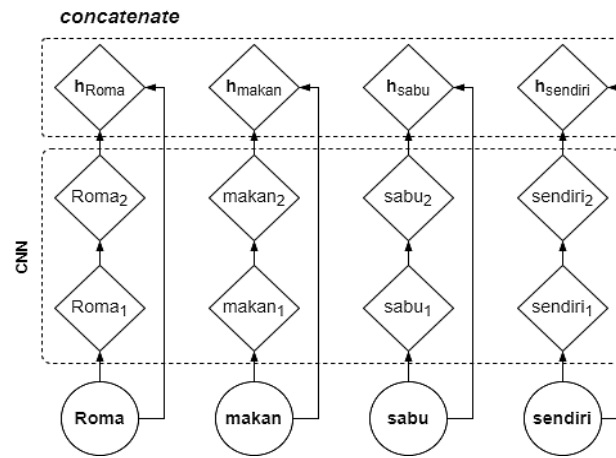


Figure 2. Architecture of tag decoder of CNNs-LSTM NER

**3.3. Training and validation on primary NER model**

The training process is carried out on eight models. Those are all possible models with combinations of various possible hyperparameter options used in these models. The epoch value used in conducting the training process is 25. It is due to [4] that showed the unimproved NER training performance after the 20<sup>th</sup> iteration and above. Following this statement, [6] also stated that the NER model training process carried out using an epoch value of more than 18 caused a decrease in model development performance due to overfitting.

In each iteration, the first thing to do is to divide the training data into several batches based on its length (text length). This study uses sparse categorical cross-entropy for the loss function, as it is commonly used in Keras if the output data type is categorical. In addition, the metrics values such as precision, recall, and F1 score are calculated in the training process. Those metrics are also computed in the validation process to determine the best model in the entire modeling iteration for a particular model. The model that has the highest F1 score at the validation stage will be the best model representing each possible model.

**3.4. CNNs-LSTM NER (Primary NER) modeling results**

Table 2 shows the testing results carried out on the eight models that have been built using different hyperparameter options. It shows that the performance is not much different between all the models. The gap between the model with the best and the lowest F1 score performance is only 0.81%. However, the model chosen as a representation of the primary NER model is a model that uses hyperparameters, which include a learning rate of 0.001 with four dropout layers and an Adam optimizer. It results in 82.54% of F1 score performance.

Table 2. Testing results of primary NER model

No	Hyperparameters	Precision	Recall	F1 score
1	<i>lr</i> =0,001; <i>Adam</i> ; 2 dropout layers	82.50%	82.47%	82.48%
2	<i>lr</i> =0,001; <i>Adam</i> ; 4 dropout layers	80.73%	84.44%	82.54%
3	<i>lr</i> =0,001; <i>Nadam</i> ; 2 dropout layers	83.30%	81.50%	82.39%
4	<i>lr</i> =0,001; <i>Nadam</i> ; 4 dropout layers	78.53%	85.20%	81.73%
5	<i>lr</i> =0,002; <i>Adam</i> ; 2 dropout layers	80.60%	83.76%	82.15%
6	<i>lr</i> =0,002; <i>Adam</i> ; 4 dropout layers	79.54%	84.48%	81.93%
7	<i>lr</i> =0,002; <i>Nadam</i> ; 2 dropout layers	81.00%	84.10%	82.52%
8	<i>lr</i> =0,002; <i>Nadam</i> ; 4 dropout layers	80.23%	84.76%	82.43%

**3.5. Performance evaluation of primary NER and baseline NER**

Based on the results shown in Table 3, the primary NER model exceeds the performance of the baseline NER model. The F1 score for the primary NER model is 82.54%. Meanwhile, the baseline NER model results in 69.67% of the F1 score. Moreover, Table 4 shows the primary NER model performance on its entity level. The model predicts UNIT and CASE entities with the best performance (86% of the F1 score) among all

the entities. Meanwhile, the primary NER model can predict the LOC entity by obtaining an F1 score that only reaches 81%.

Table 3. Performance of baseline and primary NER model

No	Model	Precision	Recall	F1 score
1	Bi-LSTMs-CRF NER (Baseline NER)	76.37%	64.04%	69.67%
2	CNNs-LSTM NER (Primary NER)	80.73%	84.44%	82.54%

Table 4. Entity level performance of CNNs-LSTM NER

No	Entity	Precision	Recall	F1 score
1	SUS	80.21%	89.74%	84.71%
2	DATE	77.94%	90.57%	83.78%
3	LOC	75.21%	87.83%	81.03%
4	NAR	80.86%	91.00%	85.63%
5	UNIT	81.21%	91.53%	86.06%
6	CASE	81.23%	91.38%	86.01%
7	MISC	80.73%	84.45%	82.54%

For the baseline NER, the performance on its entity level is shown in Table 5. It is clear that the baseline NER model failed to predict CASE entities because the performance generated is 0%. In addition, the low prediction result also happens to the MISC entity that only produces 6.39% performance. Meanwhile, the baseline model can predict a NAR entity with the best performance (90.66% of the F1 score).

Table 5. Entity level performance of Bi-LSTM-CRF NER

No	Entity	Precision	Recall	F1 score
1	SUS	74.79%	57.34%	64.91%
2	DATE	73.71%	88.52%	80.43%
3	LOC	59.17%	67.04%	62.86%
4	NAR	95.38%	86.39%	90.66%
5	UNIT	79.09%	83.33%	81.16%
6	CASE	0.00%	0.00%	0.00%
7	MISC	31.82%	3.55%	6.39%

### 3.6. Extracted drug suspect data results

Further processing is only carried out on SUS entities and only focuses on the B-SUS entity. It is divided into two treatments. First, if the token detected by the model consists of letters, it is assumed to contain information regarding the identity of the drug suspect. Then, the number of suspects is counted based on the uniqueness of the B-SUS entity. For the second type of treatment, if the token consists of numbers, it is assumed to contain information regarding the number of perpetrators in an arrest. Lastly, all the number of suspects obtained both from the first and the second treatment will get accumulated for each article.

Based on Table 6, the drug suspects data obtained from the extraction of online news articles are still far below the data from the Indonesian National Narcotics Board publications. It can be caused by the number of online news articles used in this study, which are limited and sourced from the Okezone site only. The completeness of information from online news can also be of particular concern. The related drug abuse cases are not often published online. Various completeness limitations of information on drug cases available on online news will probably affect the completeness of the data produced.

Table 6. Number of drug suspects data

No	Source	2018	2019	2020
1	Extracted data from online news	404	354	417
2	Publication of National Narcotics Board	59,536	52,709	58,764

## 4. CONCLUSION

This research has utilized online news to obtain information on drug abuse cases in Indonesia. The news site used is Okezone. The CNNs-LSTM NER (Primary NER) model has been successfully built by performing hyperparameters tuning on the value of the learning rate, the number of dropout layers, and the




type of optimizer in the modeling process. The primary NER model was able to significantly exceed the performance of the Bi-LSTMs-CRF NER (Baseline NER) model. The resulted F1 scores of the two models are 82.54% for the primary NER and 69.67% for the baseline NER. This research has also carried out further processing on the SUS entity to produce data on the number of drug suspects that occurred in Indonesia from 2018 to 2020. This study recommends several things for further research: i) Adding more sources of online news sites used, ii) Using the easy data augmentation method, automating filtering processes, and adding more hyperparameter tuning options for further model development, and iii) performing further processing on other entities other than SUS using more complete validation rules and processes.

## REFERENCES




- [1] R. Alfred, L. C. Leong, C. K. On, and P. Anthony, "Malay named entity recognition based on rule-based approach," *International Journal of Machine Learning and Computing*, vol. 4, no. 3, pp. 300–306, 2014, doi: 10.7763/ijmlc.2014.v4.428.
- [2] G. Popovski, S. Kochev, B. K. Seljak, and T. Eftimov, "Foodie: A rule-based named-entity recognition method for food information extraction," *ICPRAM 2019 - Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*, pp. 915–922, 2019, doi: 10.5220/0007686309150922.
- [3] K. Riaz, "Rule-based named entity recognition in Urdu," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 126–135, 2010, [Online]. Available: <https://aclanthology.org/W10-2419>.
- [4] D. Awad, C. Sabty, M. Elmahdy, and S. Abdennadher, "Arabic name entity recognition using deep learning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11171 LNAI, pp. 105–116, 2018, doi: 10.1007/978-3-030-00810-9\_10.
- [5] D. C. Wintaka, M. A. Bijaksana, and I. Asror, "Named-entity recognition on Indonesian tweets using bidirectional LSTM-CRF," *Procedia Computer Science*, vol. 157, pp. 221–228, 2019, doi: 10.1016/j.procs.2019.08.161.
- [6] J. P. C. Chiu and E. Nichols, "Named entity recognition with bidirectional LSTM-CNNs," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 357–370, 2016, doi: 10.1162/tacl\_a\_00104.
- [7] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, pp. 260–270, 2016, doi: 10.18653/v1/n16-1030.
- [8] Y. Shen, H. Yun, Z. C. Lipton, Y. Kronrod, and A. Anandkumar, "Deep active learning for named entity recognition," *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP 2017 at the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*, pp. 252–256, 2017, doi: 10.18653/v1/w17-2630.
- [9] W. Gunawan, D. Suhartono, F. Purnomo, and A. Ongko, "Named-entity recognition for Indonesian language using bidirectional LSTM-CNNs," *Procedia Computer Science*, vol. 135, pp. 425–432, 2018, doi: 10.1016/j.procs.2018.08.193.
- [10] A. S. Wibawa and A. Purwarianti, "Indonesian named-entity recognition for 15 classes using ensemble supervised learning," *Procedia Computer Science*, vol. 81, pp. 221–228, 2016, doi: 10.1016/j.procs.2016.04.053.
- [11] A. Zahra, A. F. Hidayatullah, and S. Rani, "Bidirectional long-short term memory and conditional random field for tourism named entity recognition," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 4, pp. 1270–1277, 2022, doi: 10.11591/ijai.v11.i4.pp1270-1277.
- [12] F. Y. Azalia, M. A. Bijaksana, and A. F. Huda, "Name indexing in Indonesian translation of hadith using named entity recognition with naïve bayes classifier," *Procedia Computer Science*, vol. 157, pp. 142–149, 2019, doi: 10.1016/j.procs.2019.08.151.
- [13] H. S. Al-Ash, I. Fanany, and A. Bustamam, "Indonesian protected health information removal using named entity recognition," *Proceedings of 2019 International Conference on Information and Communication Technology and Systems, ICTS 2019*, pp. 258–263, 2019, doi: 10.1109/ICTS.2019.8850995.
- [14] J. Santoso, E. I. Setiawan, C. N. Purwanto, E. M. Yuniarno, M. Hariadi, and M. H. Purnomo, "Named entity recognition for extracting concept in ontology building on Indonesian language using end-to-end bidirectional long short term memory," *Expert Systems with Applications*, vol. 176, 2021, doi: 10.1016/j.eswa.2021.114856.
- [15] T. Eftimov, B. K. Seljak, and P. Korošec, "A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations," *PLoS ONE*, vol. 12, no. 6, 2017, doi: 10.1371/journal.pone.0179488.
- [16] L. Yao, H. Liu, Y. Liu, X. Li, and M. W. Anwar, "Biomedical named entity recognition based on deep neural network," *International Journal of Hybrid Information Technology*, vol. 8, no. 8, pp. 279–288, 2015, doi: 10.14257/ijhit.2015.8.8.29.
- [17] K. Simran, S. Sriram, R. Vinayakumar, and K. P. Soman, "Deep learning approach for intelligent named entity recognition of cyber security," *Communications in Computer and Information Science*, vol. 1209 CCIS, pp. 163–172, 2020, doi: 10.1007/978-981-15-4828-4\_14.
- [18] B. Drury and M. Roche, "A survey of the applications of text mining for agriculture," *Computers and Electronics in Agriculture*, vol. 163, 2019, doi: 10.1016/j.compag.2019.104864.
- [19] S. Sarawagi, "Information extraction," *Publishers Inc*, pp. 9–31, 2008, doi: 10.1007/978-3-030-12375-8\_2.
- [20] A. Thomas and S. Sangeetha, "Deep learning architectures for named entity recognition: A survey," *Advances in Intelligent Systems and Computing*, vol. 1082, pp. 215–225, 2020, doi: 10.1007/978-981-15-1081-6\_18.
- [21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 2013.
- [22] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1532–1543, 2014, doi: 10.3115/v1/d14-1162.
- [23] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pp. 3483–3487, 2019.
- [24] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 50–70, 2022, doi: 10.1109/TKDE.2020.2981314.
- [25] T. Dozat, "Incorporating nesterov momentum into Adam," *ICLR Workshop*, no. 1, pp. 2013–2016, 2016.
- [26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

## BIOGRAPHIES OF AUTHORS






**Daris Azhar**    is a fresh graduate and obtained his Bachelor's degree in Computational Statistics in 2022 from Politeknik Statistika STIS, Indonesia. He is currently working in Badan Pusat Statistik (BPS) Kota Makassar, the Central Bureau of Statistics of Makassar Municipality. He can be contacted at email: [daris.azhar@bps.go.id](mailto:daris.azhar@bps.go.id).






**Robert Kurniawan**    is a researcher at Politeknik Statistika STIS Jakarta who focuses on social science, disaster management and the environment, and big data. He is currently pursuing a doctorate in population and environmental education at the State University of Jakarta. He is actively conducting research and writing reference books with ISBNs such as Easy Understanding of Nonparametric Statistics in the Health Sector and Regression Analysis using R. He hopes that big data can provide new public health-related insights. He can be contacted at email: [robertk@stis.ac.id](mailto:robertk@stis.ac.id).






**Dr. Waris Marsisno**    is lecturer at Politeknik Statistika STIS Jakarta. His research interests are mainly on new development in data collection and processing for production of official statistics. He can be contacted at email: [waris@stis.ac.id](mailto:waris@stis.ac.id).






**Budi Yuniarto**    is a lecturer and researcher at Politeknik Statistika STIS. As researcher, he focuses on computational statistics, machine learning and big data. As a lecturer, he teaches courses in Data Mining, Statistical Computing, and Multivariate Analysis. His bachelor's degree was obtained from the Sekolah Tinggi Ilmu Statistik, Jakarta (now Politeknik Statistika STIS), while his master's degree was from the Institut Teknologi 10 November, Surabaya. He can be contacted at email: [byuniarto@stis.ac.id](mailto:byuniarto@stis.ac.id).



**Sukim**    is a lecturer and researcher at the Statistics Polytechnic STIS Jakarta who focuses on social science, demographic study, disaster management, and the environment. As a lecturer, he teaches statistical computing, data exploration and visualization, survey management, survey laboratory, and sampling methodology. He can be contacted at email: [sukim@stis.ac.id](mailto:sukim@stis.ac.id).



**Sugiarto**    is a Lecturer at Politeknik Statistika STIS Jakarta, and focusing on social sciences, economics, official statistics and Management. Currently actively doing research and writing books. He hopes that this paper can provide new insights regarding public health. He can be contacted at email: [soegie@stis.ac.id](mailto:soegie@stis.ac.id).