Jurnal Mantik, 6 (1) (2022) 423-430

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik/index

Impact of Text Preprocessing on Named Entity Recognition Based on Conditional Random Field in Indonesian Text

Samuel Indra Gunawan Situmeang

Fakultas Teknik Informatika dan Elektro, Institut Teknologi Del, Toba, 22381, Indonesia

E-mail: samuel.situmeang@del.ac.id

ARTICLE INFO

Article history:
Received: Mar 12, 2022
Revised: Aprl 23, 2022
Accepted:May 30, 2022

Keywords:
Named Entity Recognition,
Text Preprocessing,
Text Analysis,
Conditional Random Field

ABSTRACT

The text preprocessing stage within a natural language processing application framework helps eliminate parts that are not helpful in the text analysis process or particular noise. Despite having a potential impact on the final performance of the application, text preprocessing has not received attention in the text analysis application literature, especially in the named entity recognition application in Indonesian texts. This paper aims to comprehensively examine the impact of text preprocessing in the Indonesian named entity recognition based on a baseline model, namely Conditional Random Field, to find the fittest preprocessing procedures for a NER model compelling performance. Various forms of text preprocessing contribute to the successful recognition of named entities assessed comparatively across three categories: people, places, and organizations. Experimental analysis of the data set reveals that several combinations of preprocessing text forms are useful. Rather than enabling or disabling them all, several combinations can significantly improve the accuracy of Indonesian named entity recognition depending on the entity category.

1. Introduction

Named entity recognition (NER) is an information extraction subtask in finding and classifying named entities mentioned in text into predefined categories such as a person's name, places, and organizations. The term Named Entity (NE) emerged at the sixth Message Understanding Conference (MUC-6) [1], where the detection of NEs was to classify proper nouns defining a Person (pers), Location (loc), or Organization (org). These named entities are grouped in ENAMEX (Entity Name Expression) category. In addition, two categories were used for numerical expressions (NUMEX) and temporal expressions (TIMEX). NER has an important role in natural language processing or other text-based knowledge applications, such as text classification [2], text summarization [3], [4], optimizing information retrieval engine algorithms [5], [6], and text-based recommendation systems [7].

Data-driven modeling or machine learning has become increasingly common, providing practical solutions to many problems, especially NER. In NER, machine learning systems operate by segmenting texts into sequences of W_i words. Each word is assigned a C_i class, generally using the BIO annotation format [8]. NER-based machine learning may use multiple clues and features. Texts may also be segmented into phrases if necessary. Machine learning models in NER includes Naïve Bayes Classifier [9], Conditional Random Field (CRF) [10], [11], [12], [13] Bi-directional Long Short-Term Memory Conditional Random Field (BiLSTM-CRF) [14], and Transformer [15], [16], [17], [18], [19]. Of all these models, CRF is one of the baseline models in NER.

While machine learning models have a substantial impact on the success of a text analysis process, the preprocessing procedures may also influence this performance noticeably. As reported by Hickman *et al.* (2022), several text preprocessing procedures can help improve the validity of subsequent text analysis [20]. Standard procedures in NER are to apply stopword removal [21], lowercase conversion [13], [22], and stemming [23]. Contractions expansion is a commonly used in text analysis [24], [25]. This preprocessing form

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

can be added in preprocessing procedures for NER. In addition, since numbers, commas, hyphens are often seen as clues in identifying a named entity, number to words conversion and hyphen-comma splitting can be considered possible text preprocessing procedures. Despite having a potential impact on the final performance of the application, text preprocessing has not received attention in the text analysis application literature, especially the impact of text preprocessing in the named entity recognition application in Indonesian texts.

This work investigates the impact of several preprocessing procedures for CRF based NER, including contractions expansion, lowercase conversion, stemming, number to words conversion, and hyphen-comma splitting in Indonesian text and on a general text-domain. In this way, this work contributes to extensively assessing the impact of preprocessing tasks on the named entity recognition success in Indonesian text at various feature dimensions and possible interactions among these tasks to find the fittest preprocessing procedures for a CRF based NER model compelling performance. Contractions expansion, lowercase conversion, stemming, number to words conversion, and hyphen-comma splitting are abbreviated as CE, LC, ST, NWC, and HCS, respectively. The experimental settings are briefly described in the following sections.

2. Method

Figure 1 shows the flowchart of experimental methods on text preprocessing in Indonesian NER based on CRF. The process consists of text preprocessing, feature extraction, CRF model training, hyperparameter optimization, CRF model evalution, and performance comparison. The methods are briefly described in the following subsections.

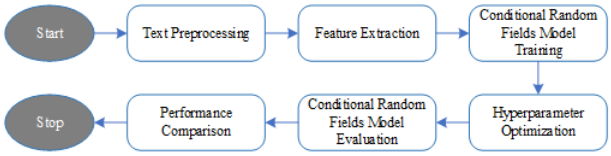

Figure 1. Flowchart of Experimental Methods on text preprocessing in Indonesian NER based on CRF.

2.1 Text Preprocessing

Text preprocessing is usually the first step in the process of building machine learning model, including NER model. There are various of text preprocessing procedures. This research limits the scope to five preprocessing steps to be considered.

1. Contractions Expansion
   Expands contractions, abbreviations, and acronyms to complete sentences, reducing vocabulary size. One technique in doing this step is dictionary mapping. A list of pairs of abbreviations and abbreviations from the official dictionary of the Indonesian language compiled by Language Development and Fostering Agency and published by Balai Pustaka, namely Kamus Besar Bahasa Indonesia, is used.
2. Lowercase Conversion
   Make all text lowercase to reduce vocabulary size.
3. Stemming
   Remove morphological affixes from words to get the word stem. In this study, we use the Indonesian stemmer library, namely Sastrawi. The Sastrawi stemmer is based on Nazief and Adriani algorithm [26], then enhanced with the CS (Confix Stripping) algorithm [27], the ECS Algorithm (Enhanced Confix Stripping) [28], and Modified ECS [29].
4. Number to Words Conversion
   Convert numeric form to the form of words. In this study, we use the num2words library.
5. Hyphen-Comma Splitting
   Split the text by hyphen and comma characters, then keep the characters.

2.2 Feature Extraction

In machine learning, translating raw data into numerical features that can be processed while keeping the information in the original data set is called feature extraction. This research limits the scope to ten types basic features to be considered.

1. The word,
2. The length of the word or number of characters,

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

3. Prefixes and suffixes of the word of varying lengths,
4. The word in lowercase,
5. Stemmed version of the word, which deletes all vowels along with g, y, n from the end of the word, but leaves at least a two long character stem,
6. Punctuation mark clue,
7. Digit clue,
8. Word POS tag,
9. Features number (1) to (8) for the previous word, the following word, and the words two places before and after, and
10. Sentence's beginning (BOS) clue or the sentence's ending (EOS) clue.

2.3 Conditional Random Fields

Conditional Random Fields is a probabilistic statistical model that is commonly utilized for segmenting and labeling sequence data [30]. Given a set of words w_1, w_2, w_3, ..., w_n with named entity tag ne_1, ne_2, ne_3, ..., ne_n in sentence S with length n, then it will be used as features to guess the correct entity label of word w_i. The weighted sum of features determines the likelihood of a labeling result. CRF training aims to assign appropriate weights to all characteristics while minimizing negative log-likelihood or penalizing negative log-likelihood using the training data [31].

The generalized iterative scaling (GIS) technique was initially designed to train maximum entropy models [32]. Later on, it became well known because it is used in the training procedure for CRF [30]. Other approaches to CRF training that use gradient-based numerical optimization algorithms have been proposed because GIS is extremely slow to converge [33]. The limited memory variable metrics (L-BFGS) technique is now a de-facto standard. The L-BFGS method is developed from the second-order Taylor expansion and is a second-order method. L-BFGS greatly surpasses GIS and other gradient-based algorithms in terms of convergence rate. Due to this fact, this research use L-BFGS in the CRF training step.

2.4 Hyperparameter Optimization

A hyperparameter configuration for machine learning models directly impacts the model's performance [34]. However, it is unclear how to best set a hyperparameter for a given dataset. Furthermore, many machine learning models have several hyperparameters that can interact nonlinearly. It is common to select a set of hyperparameters that provide a model with the best performance on a dataset. This selection process is called hyper-parameter optimization, hyperparameter tuning, or hyperparameter search. A hyperparameter optimization procedure involves defining a search space that can be thought of as an n-dimensional volume. Each hyperparameter represents a different dimension. The dimension is the hyperparameter values that may take on, such as categorical, real-valued, or integer-valued. A point in the search space is a vector with a specific value for each hyperparameter value. The goal of the optimization procedure is to find a vector that results in the best performance of the model after learning, such as maximum accuracy or minimum error. This research use one of the simplest and most common optimization methods, Random Search. Random Search defines a search space as a bounded domain of hyperparameter values and randomly sample points in that domain [35].

2.5 Model Evaluation

The CRF model performance will be evaluated using Precision, Recall, F1-Score used in CoNLL 2003 [36], which can be seen in equation (1), (2), and (3), respectively.

$$Precision = \frac{total\ correct\ named\ entity\ from\ model}{total\ named\ entity\ from\ model} \quad (1)$$

$$Recall = \frac{total\ correct\ named\ entity\ from\ model}{total\ gold\ named\ entity\ in\ dataset} \quad (2)$$

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

2.6 Experiment Settings

The model was built using Singgalang dataset with a total of 48,977 sentences [37]. There are three types of entities: person, place, and organization. In this research, the data set is prepared in BIO annotation format. The experiments will be divided into eight scenarios which are:

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

1. CRF: The CRF model training and testing is carried out without text preprocessing.
2. CRF and CE: The CRF model training and testing is carried out with one text preprocessing procedure, namely Contractions Expansion.
3. CRF and LC: The CRF model training and testing is carried out with one text preprocessing procedure, namely Lowercase Conversion.
4. CRF and ST: The CRF model training and testing is carried out with one text preprocessing procedure, namely Stemming.
5. CRF and NWC: The CRF model training and testing is carried out with one text preprocessing procedure, namely Number to Words Conversion.
6. CRF and HCS: The CRF model training and testing is carried out with one text preprocessing procedure, namely Hyphen-Comma Splitting.
7. CRF and CE+LC+ST+NWC+HCS: The CRF model training and testing is carried out with five text preprocessing procedure, namely Contractions Expansion, Lowercase Conversion, Stemming, Number to Words Conversion, and Hyphen-Comma Splitting.
8. CRF and CE+NWC+HCS: The CRF model training and testing is carried out with three text preprocessing procedure, namely Contractions Expansion, Number to Words Conversion, and Hyphen-Comma Splitting.

The experiment will be carried out by dividing the training and testing data using percentage splitting, with 33% for testing and 77 % for training. There are three types of entities: person, place, and organization.

3. Result and Discussion

The results of the experiments obtained can be seen in Table 1. The table shows that the highest macro-average precision is in the first scenario, namely CRF without text preprocessing with 0.858. This relatively high precision means that CRF model without preprocessing return more relevant named entities than irrelevant ones. From the aspect of macro-average recall, it can be seen that the highest is in the sixth scenario, namely CRF and HCS, with 0.798. This relatively high recall means that the CRF and HCS model return the most relevant named entities. From the aspect of macro-average F1-score, it can be seen that the highest is in CRF without text preprocessing, CRF and NWC, and CRF and CE+NWC+HCS with 0.825. However, the difference between one scenario and another is not significantly different. It can be seen that without the proposed text preprocessing procedure, CRF could still obtain relatively good results.

TABLE 1
NER MODEL PERFORMANCE

Model	Macro Avg. Precision	Macro Avg. Recall	Macro Avg. F1-Score
CRF	**0.858**	0.796	**0.825**
CRF and CE	0.854	0.796	0.824
CRF and LC	0.837	0.757	0.795
CRF and ST	0.821	0.757	0.787
CRF and NWC	0.855	0.797	**0.825**
CRF and HCS	0.850	**0.798**	0.823
CRF and CE+LC+ST+NWC+HCS	0.826	0.746	0.784
CRF and CE+NWC+HCS	0.857	0.796	**0.825**

Table 2 shows more comprehensive results for Organization named entities in BIO format. It can be seen that the highest macro-average precision is in the first scenario, namely CRF without text preprocessing, with 0.907 for B-Organization and 0.851 for I-Organization. This relatively high precision means that CRF model without preprocessing return more relevant B-Organization and I-Organization than irrelevant ones. From the aspect of macro-average recall, it can be seen that the highest is in the sixth scenario for B-Organization with 0.831 and the fifth scenario for I-Organization with 0.798. This relatively high recall means that the CRF and HCS models return the most relevant B-Organization. Meanwhile, the CRF and NWC models return the most relevant I-Organization. From the macro-average F1-score aspect, the highest is in the eighth scenario for B-Organization with 0.866 and the first scenario for I-Organization with 0.819.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

TABLE 2
NER MODEL PERFORMANCE ON ORGANIZATION NAMED ENTITY

Model	Named Enity	Macro Avg. Precision	Macro Avg. Recall	Macro Avg. F1-Score
CRF	B-Organization	**0.907**	0.827	0.865
	I-Organization	**0.851**	0.789	**0.819**
CRF and CE	B-Organization	0.903	0.829	0.864
	I-Organization	0.836	0.793	0.814
CRF and LC	B-Organization	0.882	0.795	0.836
	I-Organization	0.830	0.757	0.792
CRF and ST	B-Organization	0.862	0.788	0.824
	I-Organization	0.807	0.760	0.783
CRF and NWC	B-Organization	0.903	0.830	0.865
	I-Organization	0.839	**0.798**	0.818
CRF and HCS	B-Organization	0.894	**0.831**	0.861
	I-Organization	0.830	0.796	0.812
CRF and CE+LC+ST+NWC+HCS	B-Organization	0.868	0.779	0.821
	I-Organization	0.809	0.747	0.777
CRF and CE+NWC+HCS	B-Organization	0.904	0.830	**0.866**
	I-Organization	0.845	0.788	0.815

Table 3 shows more comprehensive results for Person named entities in BIO format. It can be seen that the highest macro-average precision is in the first and eighth scenario for B-Person with 0.850 and the second scenario for I-Person with 0.850. This relatively high precision means that both scenarios return more relevant B-Person than irrelevant ones. Also, CRF and CE return more relevant I-Person than irrelevant ones. From the aspect of macro-average recall, it can be seen that the highest is in the eighth scenario for B-Person with 0.731 and the sixth scenario for I-Person with 0.681. Compared to Organization named entity, this recall is slightly lower. From the macro-average F1-score aspect, the highest is from the eighth scenario for B-Person with 0.786 and the first scenario for I-Person with 0.705.

TABLE 3
NER MODEL PERFORMANCE ON PERSON NAMED ENTITY

Model	Named Enity	Macro Avg. Precision	Macro Avg. Recall	Macro Avg. F1-Score
CRF	B-Person	**0.850**	0.729	0.785
	I- Person	0.734	0.679	**0.705**
CRF and CE	B- Person	0.848	0.729	0.784
	I- Person	0.731	0.677	0.703
CRF and LC	B- Person	0.845	0.713	0.773
	I- Person	**0.745**	0.657	0.698
CRF and ST	B- Person	0.823	0.700	0.757
	I- Person	0.722	0.677	0.699
CRF and NWC	B- Person	0.848	0.727	0.783
	I- Person	0.730	0.677	0.702
CRF and HCS	B- Person	0.845	0.730	0.783
	I- Person	0.729	**0.681**	0.704
CRF and CE+LC+ST+NWC+HCS	B- Person	0.828	0.693	0.755
	I- Person	0.733	0.657	0.693
CRF and CE+NWC+HCS	B- Person	**0.850**	**0.731**	**0.786**
	I- Person	0.736	0.674	0.704

Table 4 shows more comprehensive results for Place named entities in BIO format. It can be seen that the highest macro-average precision is in the eighth scenario for B-Place with 0.939 and the third scenario for I-Place with 0.872. From the aspect of macro-average recall, it can be seen that the highest is in the first, sixth, and eighth scenario for B-Place with 0.898. Meanwhile, it can be seen that the highest is in the fifth, sixth, and

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

eighth scenario for I-Place with 0.854. Compared to Organization and Person named entity, this recall is slightly higher. From the macro-average F1-score aspect, the highest is in the sixth and eighth scenarios for B-Place with 0.918 and the fifth scenario for I-Place with 0.863.

TABLE 4
NER MODEL PERFORMANCE ON PLACE NAMED ENTITY

Model	Named Enity	Macro Avg. Precision	Macro Avg. Recall	Macro Avg. F1-Score
CRF	B-Place	0.938	**0.898**	0.917
	I- Place	0.867	0.852	0.859
CRF and CE	B-Place	0.938	0.897	0.917
	I- Place	0.870	0.851	0.860
CRF and LC	B-Place	0.893	0.832	0.862
	I- Place	0.828	0.789	0.808
CRF and ST	B-Place	0.885	0.826	0.854
	I- Place	0.827	0.789	0.808
CRF and NWC	B-Place	0.937	0.896	0.916
	I- Place	**0.872**	**0.854**	**0.863**
CRF and HCS	B-Place	0.938	**0.898**	**0.918**
	I- Place	0.865	**0.854**	0.860
CRF and CE+LC+ST+NWC+HCS	B-Place	0.888	0.823	0.854
	I- Place	0.827	0.779	0.802
CRF and CE+NWC+HCS	B-Place	**0.939**	**0.898**	**0.918**
	I- Place	0.868	**0.854**	0.861

4. Conclussion and Further Research

From the experiments performed, NER model-based CRF without text preprocessing procedure could still obtain relatively good results. However, some indications using one or a combination of text preprocessing procedures such as Contractions Expansion, Number to Words Conversion, and Hyphen-Comma Splitting can improve model performance slightly. Future research must investigate other forms of preprocessing and more sophisticated machine learning models since there are many text preprocessing procedures and machine learning models that are not covered in this research.

It can also be seen that both the CRF model without text preprocessing and the CRF model with text preprocessing are better at recognizing Place named entities, followed by Organization and Person named entities. This fact is because the Organization and Person named entities have a wider value domain than Place named entities. Therefore, in future studies, it is necessary to investigate more representative features with more diverse data sets.

References

[1] R. Grishman dan B. Sundheim, "Design of The MUC-6 Evaluation," in *Proceedings of the 6th conference on Message understanding - MUC6 '95*, 1995, hal. 1.

[2] W. Shishah, "Fake News Detection Using BERT Model with Joint Learning," *Arabian Journal for Science and Engineering*, vol. 46, no. 9, hal. 9115–9127, Sep 2021.

[3] M. E. Khademi dan M. Fakhredanesh, "Persian Automatic Text Summarization Based on Named Entity Recognition," *Iranian Journal of Science and Technology - Transactions of Electrical Engineering*, Jul 2020.

[4] S. I. G. Situmeang, R. K. Lubis, F. J. N. Siregar, dan B. J. D. C. Panjaitan, "Movie Summarization based on Indonesian Subtitles with Restricted Boltzmann Machine," in *Proceedings of 2019 4th International Conference on Sustainable Information Engineering and Technology, SIET 2019*, 2019, hal. 338–342.

[5] P. S. Banerjee, B. Chakraborty, D. Tripathi, H. Gupta, dan S. S. Kumar, "A Information Retrieval Based on Question and Answering and NER for Unstructured Information Without Using SQL," *Wireless Personal Communications*, vol. 108, no. 3, hal. 1909–1931, Okt 2019.

[6] B. Topcu dan I. D. El-Kahlout, "TR-SEQ: Named Entity Recognition Dataset for Turkish Search Engine Queries," in *International Conference Recent Advances in Natural Language Processing, RANLP*, 2021, hal. 1417–1422.
[7] T. P. Sariki dan B. G. Kumar, "A Book Recommendation System Based on Named Entities," *Annals of Library and Information Studies*, vol. 65, no. 1, hal. 77–82, 2018.
[8] L. A. Ramshaw dan M. P. Marcus, "Text Chunking Using Transformation-Based Learning," hal. 157–176, Mei 1999.
[9] R. Rifani, M. A. Bijaksana, dan I. Asror, "Named Entity Recognition for an Indonesian Based Language Tweet using Multinomial Naive Bayes Classifier," *Indonesia Journal on Computing (Indo-JC)*, vol. 4, no. 2, hal. 119–126, Sep 2019.
[10] Q. Zhang, C. Xue, X. Su, P. Zhou, X. Wang, dan J. Zhang, "Named Entity Recognition for Chinese Construction Documents Based on Conditional Random Field," *Frontiers of Engineering Management*, Jan 2022.
[11] G. Georgiev, P. Nakov, K. Ganchev, P. Osenova, dan K. I. Simov, "Feature-Rich Named Entity Recognition for Bulgarian Using Conditional Random Fields," Sep 2021.
[12] N. Patil, A. Patil, dan B. V. Pawar, "Named Entity Recognition using Conditional Random Fields," *Procedia Computer Science*, vol. 167, hal. 1181–1188, 2020.
[13] Y. Munarko, M. S. Sutrisno, W. A. I. Mahardika, I. Nuryasin, dan Y. Azhar, "Named Entity Recognition Model for Indonesian Tweet using CRF Classifier," *IOP Conference Series: Materials Science and Engineering*, vol. 403, hal. 012067, 2018.
[14] Y. An, X. Xia, X. Chen, F.-X. Wu, dan J. Wang, "Chinese Clinical Named Entity Recognition via Multi-Head Self-Attention Based BiLSTM-CRF," *Artificial Intelligence in Medicine*, vol. 127, hal. 102282, Mei 2022.
[15] A. C. Rouhou, M. Dhiaf, Y. Kessentini, dan S. Ben Salem, "Transformer-Based Approach for Joint Handwriting and Named Entity Recognition in Historical Document," *Pattern Recognition Letters*, vol. 155, hal. 128–134, Mar 2022.
[16] O. Litake, M. Sabane, P. Patil, A. Ranade, dan R. Joshi, "Mono vs Multilingual BERT: A Case Study in Hindi and Marathi Named Entity Recognition," Mar 2022.
[17] L. Cui, Y. Wu, J. Liu, S. Yang, dan Y. Zhang, "Template-Based Named Entity Recognition Using BART," *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, hal. 1835–1845, Jun 2021.
[18] Y. Chang, L. Kong, K. Jia, dan Q. Meng, "Chinese Named Entity Recognition Method Based on BERT," in *Proceedings of 2021 IEEE International Conference on Data Science and Computer Application, ICDSCA 2021*, 2021, hal. 294–299.
[19] J. Luoma dan S. Pyysalo, "Exploring Cross-sentence Contexts for Named Entity Recognition with BERT," hal. 904–914, Jun 2021.
[20] L. Hickman, S. Thapa, L. Tay, M. Cao, dan P. Srinivasan, "Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations," *Organizational Research Methods*, vol. 25, no. 1, hal. 114–146, Jan 2022.
[21] R. M. M. A. K. Syachrul, M. A. Bijaksana, dan A. F. Huda, "Person Entity Recognition for The Indonesian Qur'an Translation with The Approach Hidden Markov Model-Viterbi," *Procedia Computer Science*, vol. 157, hal. 214–220, 2019.
[22] N. Ali, "Chatbot: A Conversational Agent Employed with Named Entity Recognition Model Using Artificial Neural Network," Jun 2020.
[23] R. M. M. A. K. Syachrul, M. A. Bijaksana, dan A. F. Huda, "Person Entity Recognition for The Indonesian Qur'an Translation with The Approach Hidden Markov Model-Viterbi," *Procedia Computer Science*, vol. 157, hal. 214–220, 2019.
[24] U. Naseem, I. Razzak, dan P. W. Eklund, "A Survey of Pre-processing Techniques to Improve Short-Text Quality: A Case Study on Hate Speech Detection on Twitter," *Multimedia Tools and Applications*, vol. 80, no. 28–29, hal. 35239–35266, Nov 2021.
[25] M. Novo-Lourés, R. Pavón, R. Laza, D. Ruano-Ordas, dan J. R. Méndez, "Using Natural Language Preprocessing Architecture (NLPA) for Big Data Text Sources," *Scientific Programming*, vol. 2020, hal. 1–13, Agu 2020.
[26] M. Nazief, B. A. A. & Adriani, "Confix- Stripping: Approach to Stemming Algorithm for Bahasa

Indonesia," *Conferences in Research and Practice in Information Technology Series*, vol. 38, no. 4, hal. 307–314, 2005.

[27] J. Asian, H. E. Williams, dan S. M. M. Tahaghoghi, "Stemming Indonesian: A Confix-Stripping Approach," *Conferences in Research and Practice in Information Technology Series*, vol. 38, no. 4, hal. 307–314, Des 2005.

[28] A. Z. Arifin, P. Adhi, K. Mahendra, dan H. T. Ciptaningtyas, "Enhanced Confix Stripping Stemmer and Ants Algorithm for Classifying News Document in Indonesian Language," in *Proceeding of International Conference on Information & Communication Technology and Systems (ICTS)*, 2009.

[29] D. P. Andita Dwiyoga Tahitoe, "Implementasi Modifikasi Enhanced Confix Stripping Stemmer Untuk Bahasa Indonesia Dengan Metode Corpus Based Stemming," Institut Teknologi Sepuluh Nopember, 2010.

[30] J. Lafferty, A. Mccallum, dan F. Pereira, "Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data Abstract," in *Proceedings of the Eighteenth International Conference on Machine Learning*, 1999, vol. 2001, no. June, hal. 282–289.

[31] H. S. Huang, Y. M. Chang, dan C. N. Hsu, "Training Conditional Random Fields by Periodic Step Size Adaptation for Large-Scale Text Mining," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2007, hal. 511–516.

[32] J. N. Darroch dan D. Ratcliff, "Generalized Iterative Scaling for Log-Linear Models," *The Annals of Mathematical Statistics*, vol. 43, no. 5, hal. 1470–1480, 1972.

[33] R. Malouf, "A Comparison of Algorithms for Maximum Entropy Parameter Estimation," in *Proceeding of The 6th Conference on Natural Language Learning - COLING-02*, 2002, vol. 20, hal. 1–7.

[34] L. Yang dan A. Shami, "On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice," *Neurocomputing*, vol. 415, hal. 295–316, Nov 2020.

[35] J. Bergstra dan Y. Bengio, "Random Search for Hyper-Parameter Optimization," *Journal of Machine Learning Research*, vol. 13, no. null, hal. 281–305, 2012.

[36] E. F. Tjong Kim Sang dan F. De Meulder, "Introduction to The CoNLL-2003 Shared Task," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003* -, 2003, vol. 4, hal. 142–147.

[37] I. Alfina, S. Savitri, dan M. I. Fanany, "Modified DBpedia Entities Expansion for Tagging Automatically NER Dataset," in *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2017, hal. 216–221.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.