# Identification of Symptoms Based on Natural Language Processing (NLP) for Disease Diagnosis Based on International Classification of Diseases and Related Health Problems (ICD-11)

[*1]Fariz Bramasta Putra, [*2]Alviansyah Arman Yusuf, [*3]Heri Yulianus, [*4]Yogi Putra Pratama, [*5]Dzakiyah Salma Humairra, [*6]Urfiyatul Erifani, [*7]Dwi Kurnia Basuki, [*8]Sritrusta Sukaridhoto, [*9]Rizqi Putri Nourma Budiarti

[*1-8]*Department of Informatics and Computer Engineering, Politeknik Elektronika Negeri Surabaya*, Surabaya, Indonesia

[*9]*Department of Engineering, Nahdlatul Ulama University of Surabaya*, Surabaya, Indonesia

E-mail: [*1]farisbram@gmail.com, [*2]alviansyah.1997@gmail.com, [*3]Heri4n@gmail.com, [*4]yogiputrapra@gmail.com, [*5]dzakiyahsalma@ce.student.pens.ac.id, [*6]fifierifani@it.student.pens.ac.id, [*7]dwiki@pens.ac.id, [*8]dhoto@pens.ac.id, [*9]rizqi.putri.nb@unusa.ac.id

***Abstract*— The digestive system is a vital organ, considering its function in processing food and drinks that are consumed every day so it is very important to be maintained. But sometimes people have a lack of awareness and knowledge of the initial symptoms of digestive diseases so that being a factor causing the disease to be serious can even cause death. Identification of symptoms as early as possible is very important for the diagnosis process so that immediate control measures can be taken to prevent and overcome the spread of disease. Anamnesis process is needed to get the symptoms of the disease, question and answer process between the patient and medical personnel whose results are stored in the Electronic Medical Record (EMR) in the form of narration to assist in the process of Clinical Decision Support (CDS). EMR is often difficult to do computing processing due to inappropriate grammar. For computers to process natural languages, Natural Language Processing (NLP) is used. In this study, an NLP system was created that can identify symptoms of the digestive disease by using to optimize the CDS process. The method used to identify symptoms of the disease is Named Entity Recognition (NER), which determines which tokens are included in the symptoms of the disease. The model trained with 800 epochs produces f1-score accuracy of 0.79. Experimental results show that the NER process supported by stemming and removing stopwords in pre-processes can improve system accuracy.**

***Keywords—Natural Language Processing, Named Entity Recognition, ICD-11, Electronic Medical Record***

## I. INTRODUCTION

The disease is an abnormal condition of parts of the body or mind of humans. To find out, identify a type of disease or health problem experienced by the patient is to make a diagnosis. The diagnosis process aims to find out or identify a type of disease so that efforts can be made to immediately control the diseases, it is also an effort to prevent and overcome the spread of diseases.

One way to diagnose the disease is the Anamnesis technique, which is to do the question and answer directly or indirectly between patients and health workers (who can diagnose disease) [1]. Anamnesis is divided into two types, the first is Auto History, which is a question and answer process that is aimed directly at the patient or who has the disease. To be able to make Auto History, the patient is conscious, mature and communicative (ability to communicate well). The second type is Allo anamnesis, which is a question and answer process that is carried out between health workers with family or relatives of patients, such as parents, siblings, and friends. Usually, this process is carried out when the conditions for performing anamnesis cannot be fulfilled, such as patients or patients who are still children, the patient is unconscious, the patient is not communicative, and the patient experiences impaired memory so it is not possible to do a question and answer process. After the diagnosis process is done, the disease can be classified based on symptoms using an ICD.

The International Classification of Diseases and Related Health Problems (ICD-11) is a list of medical classifications made by the World Health Organization (WHO) which contains disease codes, signs and symptoms, and various details about other diseases. Disease classification in the ICD list uses numerical codes, letters, or numeric letter combinations, this aims to homogenize names and classes of diseases, injuries, symptoms and factors that affect health. ICD is an international standard in the medical world in terms of diagnostic classification of diseases. The 10th ICD has been made to the tenth revision (ICD-11) which began in 1983 and was completed in 1992. More than 70,000 codes are found on the ICD-11, which is far from the 14,000 ICD-9 codes. These codes represent each classification of the type of asset along with the index and other detailed information. [1]

## II. LITERATURE STUDY

### A. Speech Recognition

Speech recognition is a system used to recognize human commands in the form of sound which is then translated into a form of data that can be understood by computers. Speech recognition is included in the introduction of biometrics, which is to analyse physical and human behaviour. Speech recognition is the development of techniques and systems that allow a computer or a device to receive input in the form of human speech. The system can recognize and understand every word spoken by digitizing words and matching the digital signals with a digital signal pattern stored in the database. The voice signal is converted into a digital signal by turning it into a set of numbers which are then adjusted by certain codes to be identified. Identification results can be written directly, or can be used as an order to do a job, for example to open an application on a smartphone which is usually done by operating a touchscreen, with the presence of speech recognition can be done by voice command. [2]

### B. Natural Language Processing (NLP)

NLP is a field of computer science that deals with the interaction between computers and humans, such as Indonesian or English. NLP is used to process writing in order to understand what is said by humans. The main purpose of

978-1-7281-4449-8/19/$31.00 ©2019 IEEE

NLP is to make machines or computers to understand the meaning of human language so that they can provide appropriate responses. NLP application in the medical field is very important as Clinical Decision Support (CDS) which helps health professionals make clinical decisions, deal with medical data about patients or with the knowledge of drugs needed to interpret the data [3].

### C. *Pyhton-Django Framework*

Python is one of the major languages are used for the development of both desktop and web applications that has features that take care of common programming tasks sothat it is easily applied in various fields [4-6]. Django is a web framework that uses python language. Django was createdin 2003 by Adrian Holovaty and Simon Willison. Python has packages or libraries outside of Django so it's easy to combine them for various purposes, such as the web, analytics data, data frames, and more.

Django uses the MVC architecture or stands for Model View Controller, which makes execution faster and neater. MVC is a component or idea that separates entities between views that display data to users, controllers that manage data flow and requests, and models are structures that can be retrieved from a database

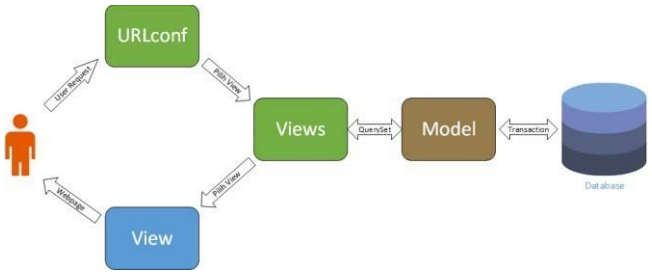


Fig. 1. Django Architecture

MVC on Django is translated into URL config and views as a controller, model, and template as the view. The user requests a URL that URLconf will handle. Then URLconf will select views that match the request from the client. If data from the database is needed, views will request a model to handle the data needed from the database and respond back to views. Furthermore views will choose the appropriate HTML template to be displayed on the web page.

## III. IMPLEMENTATION

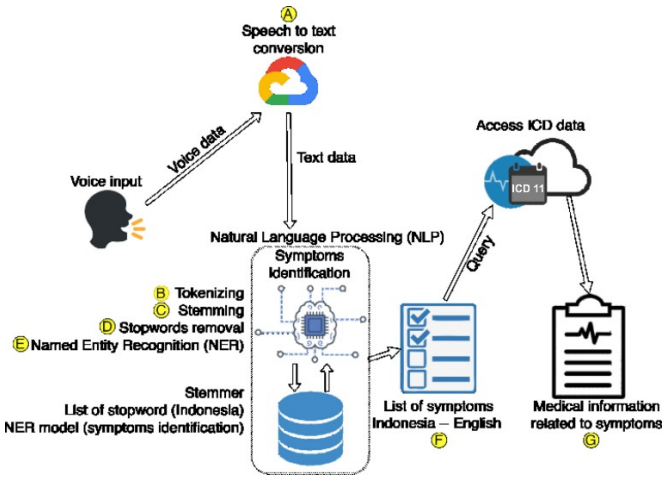


Fig. 2. Workflow Diagram

### A. *Voice Recognition*

The process in this system starts with recording the patient's voice. the recorded voice is the sound of the patient complaining about the symptoms of gastrointestinal disease experienced. the audio file used has a bitrate specification of 16000. As a sample, sound recording is done three times with different complaints shown in fig. 1.

TABLE I. VOICE RECORD DATA

| No. | Speech Input |
|---|---|
| 1 | BAB saya lebih dari tiga kali sehari dan fesesnya cair. Saya juga sering merasa mual, muntah, demam, dan sakit perut. |
| 2 | Sakit perut bagian atas, sakit bertambah setelah makan. Saya juga demam selama tiga hari dan terkadang muntah |
| 3 | Sudah dua hari ini saya demam dan sering menggigil, perut juga sakit dan diare. |

NLP is a technique that processes data in the form of text, so that input in the form of sound in this study needs to be converted into text. The scope of language that is processed in this study is Indonesian. This conversion makes the whole input into one sentence.

TABLE II. SPEECH TO TEXT CONVERSION

| No. | Speech to Text Conversion |
|---|---|
| 1 | BAB saya lebih dari 3 kali sehari dan fesesnya cair saya juga sering merasa mual muntah demam dan sakit perut |
| 2 | sakit perut bagian atas sakit bertambah setelah makan saya juga demam selama 3 hari dan terkadang muntah |
| 3 | sudah sudah 2 hari ini saya demam dan sering menggigil perut juga sakit dan diare |

### B. *Tokenizing*

In NLP, tokenizing is the first processes in NLP that identify basic tokens or units for the next process. In simple terms, tokenizing breaks down large text data into smaller shapes to facilitate the analysis process, such as paragraphs being sentences, or sentences being words. In this study, we do not break paragraphs into sentences but sentences into words because the results of speech to text conversion are in the form of a long sentence. [7]

TABLE III. WORD TOKENIZING

| No. | Word Tokenizing |
|---|---|
| 1 | [BAB, 'saya', 'lebih', 'dari', '3', 'kali', 'sehari', 'dan', 'fesesnya', 'cair', 'saya', 'juga', 'sering', 'merasa', 'mual', 'muntah', 'demam', 'dan', 'sakit', 'perut'] |
| 2 | ['sakit', 'perut', 'bagian', 'atas', 'sakit', 'bertambah', 'setelah', 'makan', 'saya', 'juga', 'demam', 'selama', 'tiga', 'hari', 'dan', 'terkadang', 'muntah'] |
| 3 | ['sudah', 'dua', 'hari', 'ini', 'saya', 'demam', 'dan', 'sering', 'menggigil', 'perut', 'juga', 'sakit', 'dan', 'diare'] |

### C. *Stemming*

Stemming works by transforming words into their basic forms. the basic form is not always the same as the root word. in general, the basic word in Indonesian has a combination [8]:

Prefix 1 + Prefix 2 + Basic Word + Suffix 3 + Suffix 2 + Suffix 1 (1)

TABLE IV. STEMMING RESULTS

| No. | Stemming Results |
|---|---|
| 1 | [‘BAB’, 'saya', 'lebih', 'dari', '3', 'kali', 'hari', 'dan', 'feses', 'cair', 'saya', 'juga', 'sering', 'rasa', 'mual', 'muntah', 'demam', 'dan', 'sakit', 'perut'] |
| 2 | ['sakit', 'perut', 'bagi', 'atas', 'sakit', 'tambah', 'telah', 'makan', 'saya', 'juga', 'demam', 'lama', 'tiga', 'hari', 'dan', 'terkadang', 'muntah'] |
| 3 | ['sudah', 'dua', 'hari', 'ini', 'saya', 'demam', 'dan', 'sering', 'gigil', 'perut', 'juga', 'sakit', 'dan', 'diare'] |

The Nazief & Adriani (NA) algorithm used in the stemming process in this research resulted in an overall accuracy score of 95.9% using 30 data, which is included in a more accurate algorithm than other stemming algorithms because it has used a basic word dictionary as a reference main [9]

*D. Stopwords Removal*

Almost all NLP implementations in the Machine Learning field use the stopwords removal method. This method works by removing several conjunctions but does not affect the overall content. Stopwords removal is used to improve system performance in order to effectively process the data needed.

TABLE V. STOPWORDS REMOVAL RESULTS

| No. | Stopwords Removal |
|---|---|
| 1 | ['BAB', 'lebih', '3', 'kali', 'hari', 'feses', 'cair', 'juga', 'sering', 'rasa', 'mual', 'muntah', 'demam', 'sakit', 'perut'] |
| 2 | ['sakit', 'perut', 'atas', 'sakit', 'tambah', 'makan', 'juga', 'demam', 'lama', 'tiga', 'hari', 'terkadang', 'muntah'] |
| 3 | ['dua', 'hari', 'saya', 'demam', 'sering', 'gigil', 'perut', 'sakit', 'diare'] |

In this research with the aim of identifying symptoms of a disease that is formed from one or several words, there are several symptoms of a disease that requires conjunctions (stopwords), such as in the symptoms of " tidak nafsu makan" if a word loss is done to just " nafsu makan". The word " tidak" in Indonesian is generally included in the stoplist so that the token is filtered so that in this case the system results are considered wrong in processing the desired data. Overall, the stopwords removal process produces an accuracy of 97.2% from 30 sample data.

The stopwords list in this final project is static, which means that the stoplist is obtained from the agency, institution, or developer of the stopword provider for the filtering process used. Stopword static contains common words in the same language so that they can be applied to other NLP projects, but because of their general nature, they make accuracy in the filtering process less. [10]

*E. Named Entity Recognition*

Named Entity Recognition (NER) is part of Information Extraction (IE) in Natural Language Processing (NLP). NER method requires training data, each case study hasa different dataset. NER was applied in this study to identify symptoms of the digestive disease using the BIO dataset format (begin, Inside, Other) [11]. every sentence in the dataset, if there are words that show symptoms, then labelled "B-GEJ". the label "I-GEJ" used when the word after label "B-GEJ". still related. because this method is focused on identifying symptoms of the disease, information that does not show symptoms is labelled "O".

TABLE VI. BIO DATASET FORMAT

| Word | Label |
|---|---|
| BAB | B-GEJ |
| saya | I-GEJ |
| lebih | I-GEJ |
| dari | I-GEJ |
| 3 | I-GEJ |
| kali | I-GEJ |
| sehari | I-GEJ |
| dan | O |
| fesesnya | B-GEJ |
| cair | I-GEJ |
| saya | O |
| juga | O |
| sering | O |
| merasa | O |
| mual | B-GEJ |
| muntah | B-GEJ |
| demam | B-GEJ |
| dan | O |
| sakit | B-GEJ |
| perut | I-GEJ |

The dataset used is 1621 lines consisting of 100 sentences that have been separated by the word and labelled. The training process is carried out with iterations with 800 epochs variables. The training results produced a micro-f1 score of 0.79. To carry out testing of the models that have been carried out by training, an analysis is carried out by providing input on the complaints of the disease.

The results of the overall accuracy of the NER process in identifying disease symptoms using input data as many as 30 were 74.3%. Data is considered wrong or unsuccessful when there is an error in the separation of symptoms, as in the 24th data. "kram perut" should not be separated by commas which means the two words become one symptom, whereas what happens is there is a "kram, perut" separator which means “kram” and “perut” are considered by the system as two different symptoms. There are also model errors in classifying symptoms that do not include symptoms of digestive diseases, such as "flu" and "batuk" symptoms (in the 30th data) so that the data from the identification process is considered wrong.

Several factors that can affect the accuracy of the identification process include the NER method used and also the model generated from the training dataset. The method used in the NER process in this final project is Bidirectional LSTM-CRF. Some NER studies in Indonesian language processing by Wibawa (2016) for 15 classes of entities produced the highest F1-micro score of 0.50, Budi (2015) who used three entity classes to get the highest F1-micro score of 0.67, and Yudi Wibisono (2018) using four entity class that produces a score of 0.73 [12]. The next factor of accuracy is the model produced from the training dataset that can still be improved by increasing the size and variety of the corpus used in the training dataset to produce the model.

*F. Data Translation*

The translation process in this study was carried out with the aim of changing the data containing symptoms from Indonesian to English because the ICD reference data issued

by WHO was not available in Indonesian. However, the results are not entirely appropriate, this is because during the NLP process, namely stemming and removing stopwords change the structure of Indonesian so that if translated into English, it allows the system to produce incorrect data.

### G. Access ICD Data

Access to ICD data is done to search for classification of symptoms related to input of symptoms of the disease. Data parsed in the form of disease symptom strings resulting from the translation process (previous test), from the ICD system provides feedback in the form of a JSON file that contains a classification of symptoms and / or diseases related to input parsed. From the previous test, if the translation data is appropriate, the data obtained is in the form of classification of related symptoms.

TABLE VII. SYMPTOM DATA SEARCH RESULTS IN ICD REFERENCES USING APPROPRIATE DATA

| Hasil Pencarian Data pada ICD |
|---|
| 'constipation':<br>'Constipation',<br>'Constipation disorder (TM1)',<br>'Encopresis with constipation or overflow incontinence',<br>'Encopresis without constipation or overflow incontinence',<br>'Slow transit constipation',<br>'Irritable bowel syndrome, constipation predominant', |

If the translation results are not appropriate, the data parsed to ICD is also in the form of unclear data so that the resulting data is not appropriate. This can occur because there are two different languages processing, namely Indonesian as input while English as reference on ICD.

TABLE VIII. SEARCH RESULTS DATA ON ICD BY USING INPUT DATA THAT DOES NOT MATCH

| Hasil Pencarian Data pada ICD |
|---|
| 'dizzy': [],<br>'no appetite': [] |

### H. Web Interface

The user interface in this research uses web applications so it is necessary to do a web design view. The web display is divided into three pages, namely the first page for recording sound and conversion into written form. The speech recognition system works in streaming so users can see the results of voice conversions directly on the webpage.


Fig. 3. Speech Recognition Page

The second page is used for the process of identifying symptoms of the disease. on this web page an NLP process is carried out, starting from tokenizing, stemming, stopwords removal, and NER. the results of the entire NLP process are displayed on a page containing symptoms of digestive diseases.

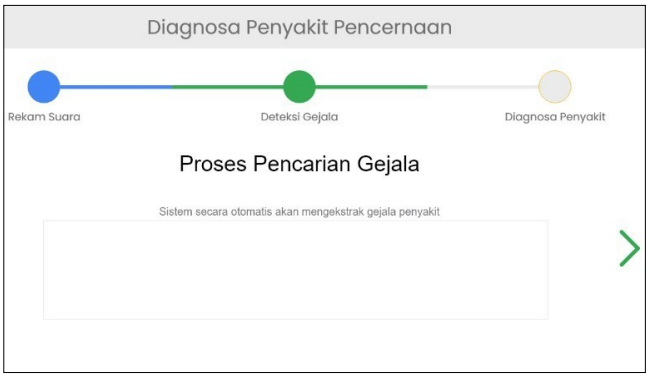


Fig. 4. Web Page Design For The NLP Process

Third page for diagnosis of disease based on ICD-11 references. on this web page the process of parsing data is done to the ICD database. parsed data is a list of symptoms of digestive diseases that have passed the NLP process. The data obtained is in the form of classification of symptoms of related diseases based on ICD references.

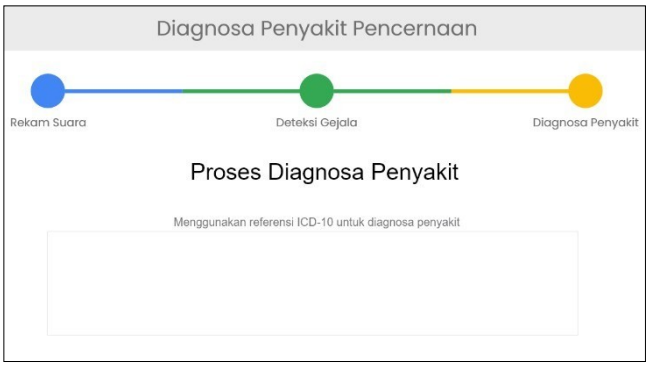


Fig. 5. Web Page Design To Display The Results Of ICD Data Parsing

## IV. EXPERIMENT RESULT

TABLE IX. DATA INPUT

| No. | Input |
|---|---|
| 1 | berat badan saya turun drastis setelah seminggu saya tidak nafsu makan |
| 2 | kepala saya sering pusing saya juga nyeri pada uluh hati terkadang mulas |

The experiment was conducted in two ways, first was the identification of symptoms by only using the NER method without pre-processing to prove deep learning as end-to-end learning. the second way is to use NLP pre-processes before being processed with NER.

TABLE X. NER RESULT

| Input | NER Result | |
|---|---|---|
| | ***Text*** | ***Type*** |
| 1 | berat badan saya turun | GEJ |
| | tidak nafsu makan | GEJ |
| 2 | pusing | GEJ |
| | nyeri pada uluh hati | GEJ |
| | mulas | GEJ |

From the table above it can be seen that the results of identifying symptoms of the disease using only the NER method can detect only words that contain information on symptoms of the disease.

TABLE XI. PREPROCESS RESULT TABLE

| Input | Pre-process Result |
|---|---|
| 1 | Berat badan turun drastis minggu tidak nafsu makan |
| 2 | Kepala sering pusing juga nyeri uluh hati terkadang mulas |

Pre-process NLP makes data cleaner and more efficient for the NER process because less data is processed. From the initial input, the pre-processing removes the conjunctions from the entire data using the stopwords removal technique. For the tokenizing process carried out during the NER process that processes each word in one sentence.

TABLE XII. NER RESULT

| Input | NER Result | |
|---|---|---|
| | ***Text*** | ***Type*** |
| 1 | berat badan saya turun drastis | GEJ |
| | tidak nafsu makan | GEJ |
| 2 | pusing | GEJ |
| | nyeri uluh hati | GEJ |
| | mulas | GEJ |

The output produced when using the NLP pre-process is slightly different. The difference lies in the presence and absence of conjunctions in both experiments. data that has no conjunctions is more informative, because identification of the symptoms needed is only the symptoms, there is no need for conjunctions.

## V. CONCLUSION

Based on the experiments that have been conducted, it can be concluded that. This research applies and develops the Named Entity Recognition (NER) model in the process of identifying symptoms of digestive diseases with an accuracy rate of 74.3%. NLP pre-process is still not fully accurate in filtering data on stemming and stopwords removing processes with an accuracy rate of 95.9% and 97.2%. The performance of the NER depends on the variety of datasets and the process of the training carried out. Language differences in data input (Indonesia) with health references on ICD (English) and translate processes can make the parsed data incompatible.

## REFERENCES


[1] Febriyanti, R. I. M., & Sugiarti, I. (2015). Analisis kelengkapan pengisian data formulir anamnesis dan pemeriksaan fisik kasus bedah. Jurnal Manajemen Informasi Kesehatan Indonesia (JMIKI), 3(1).

[2] Wardhany, V. A., Kurnia, M. H., Sukaridhoto, S., Sudarsono, A., & Pramadihanto, D. (2015, September). Smart presentation system using hand gestures and Indonesian speech command. In 2015 International Electronics Symposium (IES) (pp. 68-72). IEEE.

[3] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

[4] Rasyid, A., Yusuf, A. A., Falah, M. F., Panduman, Y., Albaab, M. R. U., Basuki, D. K., ... Yudianto, F. (2019, September). Pothole Visual Detection using Machine Learning Method integrated with Internet of Think Video Streaming Platform. . In 2019 International Electronics Symposium (IES).

[5] Yusuf, A. A., Basuki, D. K., Sukaridhoto, S., Pratama, Y. P., Putra, F. B., Yulianus, H. (2019, September). ArmChain - A Blockchain Based Sensor Data Communication For the Vehicle as a Mobile Sensor Network. In 2019 International Electronics Symposium (IES).

[6] Pratama, Y. P., Basuki, D. K., Sukaridhoto, S., Yusuf, A. A., Yulianus, H., Faruq, Putra, F. B. (2019, September). Designing of a Smart Collar for Dairy Cow Behavior Monitoring with Application Monitoring in Microservices and Internet of Things-Based Systems. In 2019 International Electronics Symposium (IES).

[7] Webster, J. J., & Kit, C. (1992). Tokenization as the initial phase in NLP. In COLING 1992 Volume 4: The 15th International Conference on Computational Linguistics.

[8] Agusta, L. (2009). Perbandingan algoritma stemming Porter dengan algoritma Nazief & Adriani untuk stemming dokumen teks bahasa indonesia. Konferensi Nasional Sistem dan Informatika, 2009, 196-201.

[9] Wahyudi, D., Susyanto, T., & Nugroho, D. (2017). Implementasi Dan Analisis Algoritma Stemming Nazief & Adriani Dan Porter Pada Dokumen Berbahasa Indonesia. Jurnal Ilmiah SINUS, 15(2),49-56.

[10] Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S. M. M., & Williams, H. E. (2007). Stemming Indonesian. ACM Transactions on Asian Language Information Processing, 6(4), 1–33.

[11] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.

[12] Wibisono, Y., & Khodra, M. L. (2018). Pengenalan Entitas Bernama Otomatis untuk Bahasa Indonesia dengan Pendekatan Pembelajaran Mesin.