Application of Named Entity Recognition via Twitter on SpaCy in Indonesian (Case Study: Power Failure in the Special Region of Yogyakarta)

R M Yanti[1], I Santoso[2], L H Suadaa[3]

[1,2,3]Politeknik Statistika STIS, Jakarta, Indonesia

E-mail: 221709980@stis.ac.id[1], ibnu@stis.ac.id[2], lya@stis.ac.id[3]

Submitted: 15 July 2021, revised: 24 July 2021, accepted: 29 July 2021

Abstrak. SpaCy merupakan *tools* yang dapat menangani masalah *Natural Language Processing* (NLP) dengan efisien, salah satunya adalah *Named Entity Recognition* (NER). NER digunakan untuk mengekstrak dan mengidentifikasi entitas bernama pada suatu teks. Namun, sejauh ini SpaCy belum merilis *pre-train* model NER untuk Bahasa Indonesia secara resmi, sehingga perlu dilakukan pembuatan model terlebih dahulu. Model yang dibangun menggunakan data *training* yang diambil dari data keluhan masyakarat di Twitter terkait gangguan listrik di D.I. Yogyakarta. Sebab, berdasarkan laporan statistik PLN tahun 2019, Provinsi D.I. Yogyakarta merupakan provinsi yang sering mengalami gangguan listrik. Adapun penelitian ini dibuat dengan tujuan membangun model NER berbahasa indonesia terkait gangguan listrik di Provinsi D. I. Yogyakarta dengan SpaCy, mengetahui hasil *evaluation metric* dari model yang telah dibangun, dan memetakan persebaran serta mengetahui perbandingan lokasi yang tersebutkan di *tweet* terkait gangguan listrik di Provinsi D. I. Yogyakarta pada tahun 2020. Penelitian ini menghasilkan performa hasil yang baik dengan hasil perhitungan *precision* 95.52%, *recall* 93.27%, dan *f1-score* 94.38%. Kemudian, dilakukan pemetaan berdasarkan entitas lokasi yang terdapat dalam *tweet* terkait gangguan listrik. Dari proses tersebut didapat bahwa jumlah lokasi yang tersebutkan di *tweet* terkait gangguan listrik tertinggi berasal dari Kabupaten Sleman, sedangkan jumlah terendah berasal dari Kabupaten Gunung Kidul. Lalu, bulan yang paling banyak mengalami gangguan listrik adalah bulan Maret 2020, sedangkan yang paling sedikit adalah bulan Juli 2020.

Kata kunci: information extraction; NER; spacy; twitter; gangguan listrik

Abstract. SpaCy is a tool that can efficiently handle Natural Language Processing (NLP) problems, one of which is Named Entity Recognition (NER). NER is used to extract and identify named entities in a text. However, so far SpaCy has not officially released the NER model pre-train for Indonesian, so it is necessary to build a model first. The model was built using training data taken from community complaints data on Twitter related to power failure in D.I. Yogyakarta. Because, based on the 2019 PLN statistical report, the Province of D.I. Yogyakarta is a province that often experiences power failure. This research was made with the aim of building an Indonesian-language NER model related to power failure in the DI Yogyakarta Province with SpaCy, knowing the results of the evaluation metric of the model that has been built, and mapping

the distribution and knowing the comparison of the locations mentioned in the tweet related to power failure in the DI Yogyakarta Province on 2020. This research produces good results with the calculation results of precision 95.52%, recall 93.27%, and f1-score 94.38%. Then, mapping is carried out based on the location entities contained in tweets related to power failure. From this process, it was found that the highest number of locations mentioned in the tweet related to power failure came from Sleman Regency, while the lowest number came from Gunung Kidul Regency. Then, the month that experienced the most power failure was March 2020, while the month that experienced the least was July 2020.

Keywords: information extraction; NER; spacy; twitter; power failure

1. Introduction

Language is an important communication tool. Communication is the process of exchanging information between humans with the help of language. The rapid development of information technology gives more freedom to the public to express their opinions or complaints on social media. There is billions of information distributed through various technologies and some of it uses natural language. Natural language is the language that humans use to communicate with each other. The language accepted by the computer must be processed and understood before the computer can properly understand the user's intent. Artificial Intelligence (AI) is growing rapidly in the field of research and its application in the real world. For example, one of the rapid developments in the field of language computing is Natural Language Processing (NLP). NLP is a branch of AI that focuses on natural language processing with a computerized approach used to analyze data, text, speech, etc. Therefore, NLP can perform natural language processing like human language [1].

Along with the development of the internet, the availability of online data also develops, especially textual data. The text data contains a lot of information. However, the information in the text is often not visible due to the unstructured form of the text. Thus, information extraction is needed, which is a system to find specific data from Natural Language Text [2]. An example is public complaints related to power failure. The news contained some information such as the duration of the power outage, the date, and the location of the email. However, all the information is hidden in the form of sentences. Therefore, to obtain this important information, it is necessary to conduct an analysis of the entire text. If the complaints are in very large numbers, of course it will be difficult to analyze them. From the abundance of these data and the information they contain, there is a need to extract information automatically from text data [3].

Text data that is also experiencing development is the data contained on Twitter. Twitter is one of the social media that is currently growing rapidly, because it makes it easier for users to communicate. Indonesia is one of the countries with the 3rd largest number of active Twitter users in the world [4]. The increase in Twitter users every year, shows that Twitter is a very popular social media. Therefore, we need a method that can extract the information contained in the tweets contained on Twitter.

So in this study, researchers will use one of the NLP tasks, namely Information Extraction (IE). IE has one sub-task that can assist the process of identifying and extracting information in the form of entities, or commonly called Named Entity Recognition (NER). NER is used to extract and identify named entities (people, locations, organizations) in a text. In this study, NER will be developed through a Machine Learning approach using the Supervised Learning method. Where the model will learn to recognize entity patterns from previously labeled data. So, later NER can recognize entities from a text that has been entered into the NER model. The model development in this research will be assisted by the SpaCy library. SpaCy is an open source library that uses the Python programming language and is useful for handling NLP problems in an efficient way, one of which is NER. However, unfortunately so far SpaCy has not officially released the NER model pre-train for Indonesian.

In addition, it is still rare to find research on NER that uses Indonesian SpaCy as a tool to create NER models. This makes researchers interested in conducting this research. Although SpaCy has not officially released the pre-train model for Indonesian, it is still possible to train the model with the data it has. Where, the training data is adjusted in advance to the format of SpaCy, which will then be trained with several iterations before it can be used. The data used is tweets related to power failure in the Province of D.I. Yogyakarta. Researchers are interested in using these data because electricity is one of the basic human needs and it can be seen from the pattern of human life that always depends on electricity. Without electricity, all basic human needs cannot be met. In addition, commercial activities, public services and telephone networks will be disrupted. Thus, the presence of electrical disturbances will be very detrimental to the community. In this study, the author will use SAIDI (System Average Interruption Duration Index) and SAIFI (System Interruption Frequency Index) as indicators to see the provinces experiencing the highest power failure in Indonesia. Province of D.I. Yogyakarta is a province with the 2nd highest SAIDI and SAIFI on the island of Java after West Java [5]. SAIDI is the average length of outage, while SAIFI is the average outage frequency. SAIDI and SAIFI are indices that show the level of reliability of a distribution system in serving consumers. The smaller the SAIDI and SAIFI values, the better the reliability [6] and vice versa.

2. Theoritical Framework

2.1. Text Mining

Text mining is the application of data concepts and techniques mining to find patterns in process text teks text analysis to find information that useful for certain purposes [7]. In analyzing part or all of the unstructured text, text mining tries to match one text to another according to certain rules [8]. Text Mining or Knowledge Discovery from Text belongs to the sub-field of Data Mining, commonly known as Knowledge Discovery in Databases (KDD). The goal of KDD is to discover knowledge from a variety of data sources, including text data, relational databases, web data, and user log data [9]. Text mining uses the same procedure as the KDD process. However, in text mining, the focus of analysis is not on datasets, but on text documents in the form of unstructured data sets.

2.2. Natural Language Processing

Natural Language Processing (NLP) is a branch of artificial intelligence that allows computers to understand, process, and generate language like humans do [10]. NLP can also be regarded as a set of theory-driven computational techniques to analyze and represent human language automatically [11]. Natural language is the language that humans use to communicate or interact with each other. Natural language or natural language can be found in various languages, one of which is Indonesian. NLP is a branch of Artificial Intelligence (AI), because NLP is designed to allow people to do work like humans. NLP aims to complete language processing like human or natural language.

2.3. Information Extraction

Most of the world's knowledge is recorded in natural language texts, but its effective use in this form is a major challenge. Information extraction offers many possibilities for using this knowledge [12]. Information Extraction (IE) is the process of extracting useful structured information from entities, relationships, objects, events, and many other types of unstructured data. Information extracted from unstructured data to prepare data for analysis [13]. Information extraction consists of several more focused sub-fields, each of which has a difficult problem to solve. For example, in this study, Named Entity Recognition (NER) is used to extract information. Therefore, efficient and accurate conversion of unstructured data in the IE process improves data analysis capabilities.

2.4. Named Entity Recognition

Named Entity Recognition (NER) is an important task in NLP. NER is a challenging task which traditionally required large quantities knowledge in the form of feature and lexicon engineering to achieve high performance [14]. NER can also be defined as a method of extracting information by processing structured and unstructured documents and identifying entities which can be people, locations, organizations, or companies [15]. NER is also the process of identifying and grouping entities in text and is the basis and core of natural language processing (NLP) systems. The use of NER is influenced by the type of language that will be used, because different languages will be treated differently, so the algorithms used in NER will also be different. NER has two tasks, the first is to identify the correct entities, and the second is to classify those entities into certain predefined categories of entities. NER can be categorized into three classes, namely rule-based NER, machine learning-based NER and hybrid NER [16]. Rule-based NER focuses on searching for entity names using predefined artificial rules and consisting of a series of patterns that use a combination of grammatical, syntactic, and spelling characteristics [17]. Then, machine learning-based NER uses a method to solve it by changing the identification task into a classification task using a statistical classification model. In the machine learning-based NER method, the system looks for patterns and relationships in the text to create a model using statistical models and machine learning algorithms. Then, hybrid NER uses a method that combines rule-based NER and machine learning-based NER methods, and uses the strengths of each of these methods to create new methods. In this study, researchers carried out the NER process using a machine learning approach.

2.5. SpaCy

SpaCy is an open source handling Natural Language Processing library that may be used to do a range of tasks such as POS Tagging, Named Entity Recognition, Dependency Parsing, and so on [18]. References [19] states that spaCy doesn't do the most accurate in their evaluation, it performs the fastest comparable maintenance accuracy. The SpaCy model is statistical and every “decision” what they make is a prediction. The SpaCy library does not offer a pre-trained model for Indonesian, but provides an opportunity to do training and getting a model on its own. In this study, the library architecture used in SpaCy is nlp.pipeline. References [20] states that the processing pipeline consists of one or more pipeline components that are called on the doc in sequence. Pipeline components can be added using nlp.add_pipe. The pipeline components can contain statistical models and trained weights, or make only rule-based modifications to the document. SpaCy provides various built-in components for different language processing tasks and also allows adding custom components. In this study, the pipeline used is only NER.

3. Methodology

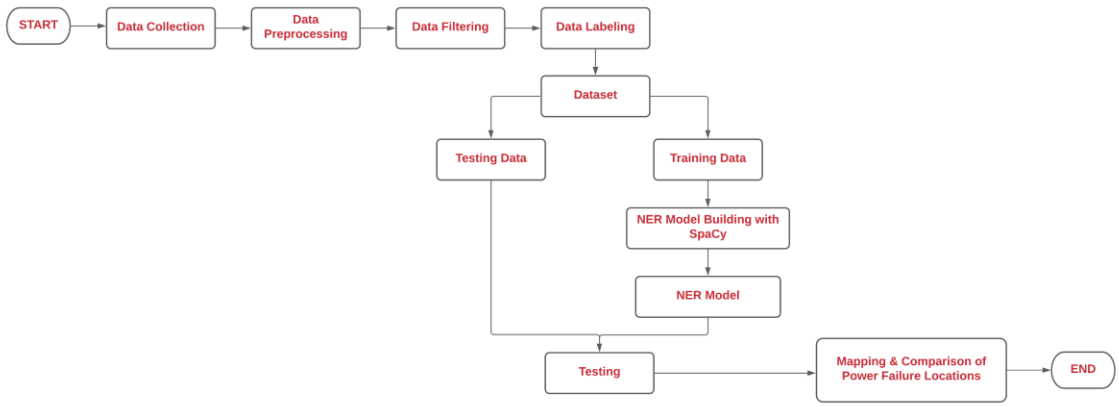

Figure 1. Research flow

3.1. Data Collection

The data used in this study is data sourced from Twitter. Twitter is one of the social media where users can post various things and can communicate with other users. The tweet data collected was taken from the complaints of Twitter users regarding the electrical disturbances that occurred in the D.I. Province. Yogyakarta. Data retrieval in this study refers to the keywords "electricity disturbance", "electricity outage", "power outage", and the @pln_123 account (official account from PT.PLN), the @infolistrikdjty account (official account from PLN Central Java Distribution Main Unit). and DI Yogyakarta), and the @pln_jogja account. In addition, data collection was also obtained from certain hashtags. Hashtags that are commonly used to categorize tweets related to power outages are #InfoPLN. These hashtags are usually used to inform the schedule of power outages, electricity maintenance, electrical repairs, etc.

Data retrieval on Twitter is done using the web scraping method. The purpose of the web scraping process is to obtain information from a website and turn it into a structure that is easy to understand, store and analyze in a database or CSV file. In this study, web scraping for data retrieval from Twitter was carried out using a user interface in the form of Jupyter Lab with the Python programming language. Web scraping in the python programming language can be done using the libraries available in python. The python library used to support the web scraping process in this research is the Twint library. Twint is a tool that allows you to scrape tweets from an account without using the Twitter API.

In the process of web scraping with the python programming language, the twint library extracts all tweets that have certain keywords and at a certain time period. In addition, you can also extract tweets from a specific account by entering the name of the account as a parameter. Then, the tweet data that has been obtained is converted into a Microsoft Excel file with csv format. In this study, the researchers wrote down keywords in the form of a combination of the names of the sub-districts in the province of D. I. Yogyakarta with the previously mentioned keywords. The use of the name of the sub-district is intended so that more data can be obtained than if only using the name of the city.

3.2. Data Preprocessing

The preprocessing process carried out in this study is as follows:

3.2.1. Cleaning

This process is carried out to eliminate user accounts and clean words from characters that have no effect on the classification results. The omitted characters are numbers, punctuation marks, hashtags (#), symbols, "RT" text and the url of the website.

3.2.2. Stopword

This process is done to remove words that have low information from a text. For example "which", "and", "at", "from", and so on.

3.2.3. Tokenization

Tokenization is the stage of separating the text in the tweet into pieces of words called tokens. The space character in the text of the power failure tweet is used as a word separator at this stage. At this stage there are 3 parameters used, namely preserve_case (changes text to lower case), strip_handles (removes mentions), and reduce_len (reduces repeated characters to 3, for example: minnnnnnn becomes minnn).

3.2.4. Remove Duplicates

This stage aims to filter the same tweet data, this is because Twitter has a re-tweet filter that causes the same number of tweets to be retweeted by other users and causes repeated texts with the same topic. Furthermore, tweets that have gone through several preprocessing stages will be converted back into text form, not in token form anymore. It aims to speed up the data labeling process, because the data labeling results must comply with the SpaCy labeling format.

3.3. Data Filtering

The data filtering process is carried out after the data collection process, during the data labeling process, and after the data labeling process. This process is done manually by checking the tweets one by one, with the aim of selecting tweet data that is not in accordance with what is desired, such as tweets containing advertisements related to power failure repair services, tweets that do not include the location of electrical disturbances, tweets related to electrical disturbances. which doesn't happen in DI Yogyakarta and duplicate tweets, one of which contains incorrect location information.

In the data filtering process carried out after data collection, researchers checked all tweets in each sub-district file. This aims to make checking easier, compared to checking one by one in the city files. Then, in the data filtering process carried out at the time of data labeling, the researcher checked the tweets while labeling. If there is a tweet that does not match, then the tweet is removed (not labeled). Finally, in the data filtering process, which was carried out after labeling the data, the researcher checked for duplication of data by combining all city files, the result was that there was duplicated data and one of them had incorrect location information. Tweets with incorrect location information are removed.

3.4. Data Labeling

The preprocessed tweets will then be labeled using BIO notation (Begin, Inside and Other) as a labeling scheme that indicates the order which is then classified into two classes, namely B-LOC and I-LOC. In this study, the researcher deliberately did not use the O notation to facilitate and speed up the labeling process. The tweet labeling process is carried out through the website agateteam.org/spaCynerannotate/. Because the training data must be in spacy format and the website already supports the data format. Then the data that has been labeled will be split into training data and testing data with a ratio of 90:10. Furthermore, the data will be used to perform NER modeling.

3.5. NER Model Building with SpaCy

The NER model development process is carried out using the SpaCy library. The architectural model or algorithm used in SpaCy for building the NER model is a Transition-Based Parser. The process of building the NER model with SpaCy is carried out after the data labeling process. In the training data that has been labeled, a training model is carried out which then the NER model can be used to classify entities on power failure tweets.

3.6. NER Model Testing

In this study, researchers used evaluation metrics such as f1-score, precision and recall which were used to test the model. Testing the NER model was carried out using the SpaCy library on testing data with a total of 596 tweets.

3.7. Mapping and Comparison of Power Failure Locations

The entity mapping of the location of the power outage is done using the shiny library and leaflets in R Studio. The purpose of this mapping is to see the number of power failure locations mentioned in all tweets in each city in the province of D.I. Yogyakarta which occurred in the period of January 1, 2020 to December 31, 2020 in the form of an area. Then a comparison of the entities of the location of the electricity disturbance is made by manually calculating the locations mentioned in all tweets, which are then grouped by month and city. The purpose of this comparison is to see which months have the highest and lowest complaints of electrical disturbances in each city in the province of D.I. Yogyakarta which occurred in the period from January 1, 2020 to December 31, 2020.

4. Result and Discussion

4.1. Data Collection

Data retrieval of power failure tweets from Twitter in this study was carried out using the web scraping method. The electricity disturbance tweet that was taken is a tweet that contains public complaints and notifications from PLN officers regarding power failure that occurred in the period January 2020 to December 2020 in the province of D.I. Yogyakarta. Scraping data on Twitter is done using libraries available in the python programming language. The library available and used for scraping data in this study is the Twint library. To be able to get tweets regarding power outages at D.I. Yogyakarta, the researchers used a combination of sub-district names in D.I. Yogyakarta with "power failure", @pln_123, @infolistrikdjty, @pln_jogja, and #infopln when scraping data. It aims to ease the work of researchers when filtering data and to make it easier to group tweets by city name and to get more data when compared to combining these keywords with city names. then, the content taken in this study is only the date and content of the tweet. From the results of scraping tweets of electrical disturbances on Twitter, 15,374 tweets related to electrical disturbances were obtained.

4.2. Data Preprocessing

The scrapping data of power failure tweets is data that is still dirty and cannot be used as training and testing data for the formation of the NER model. Therefore, before entering the next stage, the data must be cleaned first or commonly referred to as preprocessing data. Data preprocessing is carried out starting from cleaning, stopword, tokenization, and removing duplicates. From 15,374 tweets, preprocessing was carried out and resulted in 6,105 tweets

4.3. Data Filtering

After filtering several times, the final tweet was obtained with a total of 5,955 tweets. The following table will present the number of tweets related to power failure in each city:

Table 1. The number of tweets related to power failure in each city

		Number of Tweets		
		Scraping	Pre-Processing	Final Tweet
City	Bantul	5,543	1,887	1,857
	Gunungkidul	821	237	225
	Kulon Progo	921	539	502
	Sleman	6,969	2,783	2,736
	Yogyakarta	1,120	659	635
Total		15,374	6,105	5,955

4.4. Data Labeling

In this study, the researcher deliberately did not use the O notation to facilitate and speed up the labeling process. Then, the LOC label in this study indicates the location entity. The labeling process is done manually through the agateteam.org/spaCynerannotate website. The following are the results of data labeling in accordance with the spacy format:

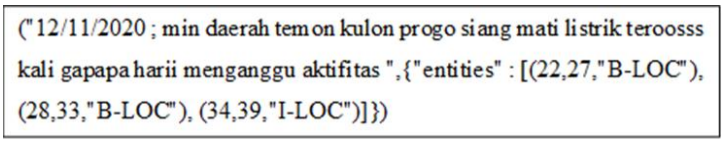

Figure 2. Example of data labeling

4.5. NER Model Testing

Testing of the NER model was carried out using the SpaCy library on testing data with a total of 596 tweets, which were measured by calculating precision, recall, and f1-score. The following are the results of the evaluation metrics for all entities :

Table 2. Test results of the NER model on SpaCy

Evaluation Metric	
Precision (%)	95.52
Recall (%)	93.27
F1-Score (%)	94.38

From the table above, it can be concluded that the f1-score of the NER model that has been built is 93.04%. Then, in addition to obtaining precision, recall, and f1-score values for all entities, the researcher also obtained precision, recall, and f1-score values for each entity, namely B-LOC and I-LOC. The following are the results of the evaluation metrics for each entity :

Table 3. Evaluation metric results for each entity

Evaluation Metric	***Entity***	
	B-LOC	**I-LOC**
Precision (%)	95.93	90
Recall (%)	93.17	94.74
F1-Score (%)	94.53	92.31

From the table above, it can be concluded that the f1-score of the B-LOC entity is 94.53% and the f1-score of the I-LOC entity is 92.31%. So it can be said that the I-LOC entity is an entity that has more errors.

4.6. Mapping and Comparison of Power Failure Locations

The following is a map display of the number of power failure locations that were tweeted in each city/district during 2020:

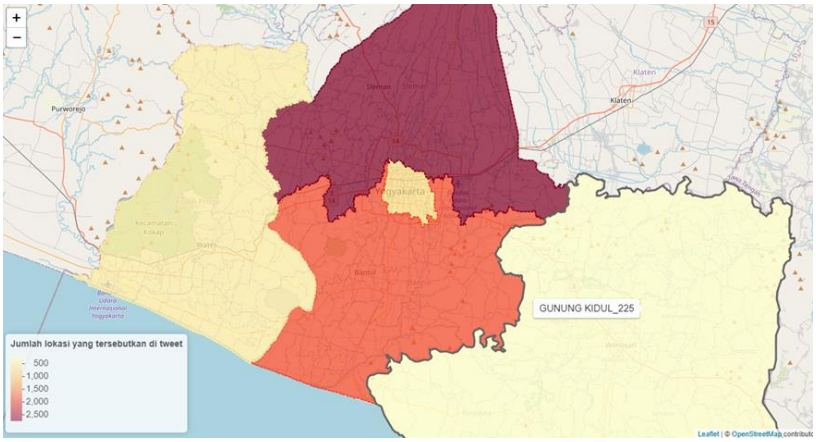

Figure 3. Power failure location mapping

From the picture above, it can be concluded that Sleman Regency is an area that has the highest number of locations mentioned on Twitter, with a total of 2,736 locations, this is indicated by the red area. Meanwhile, Gunungkidul Regency is the area that has the lowest number of locations mentioned on Twitter, with a total of 225 locations, this is indicated by the pale cream colored area. Then, the following is the result of a comparison of the locations mentioned on Twitter for each city in D.I.Yogyakarta Province affected by power failure per month :

Table 4. The number of locations mentioned on tweet

Month	Number of locations mentioned in the tweet				
	Bantul	**Gunungkidul**	**Kulon Progo**	**Sleman**	**Yogyakarta**
1	149	17	44	193	97
2	167	18	58	232	**121**
3	**256**	**40**	**81**	299	42
4	167	25	44	206	67
5	127	***8***	45	207	***29***
6	106	12	27	144	48
7	***61***	9	***21***	***88***	33
8	140	16	29	215	59
9	93	10	33	198	61
10	249	28	46	**372**	58
11	186	21	36	361	96
12	156	21	38	221	33
Total	**1,857**	**225**	**502**	**2,736**	**744**

From table 4, it can be concluded that the total number of tweets is 6,064 locations. Then, in the table it is shown that 3 out of 5 regencies/cities have the highest number of power failure locations mentioned in the tweet in March. Where the 3 districts are Bantul, Gunung Kidul, and Kulon Progo. Meanwhile, for Sleman, the highest number of power failure locations mentioned in the tweet was in October, while Yogyakarta City was in February. The table also shows that 3 out of 5 regencies/cities

have the lowest number of power failure locations mentioned in the tweet in July. Where the 3 districts are Bantul, Kulon Progo, and Sleman. Meanwhile, for Gunung Kidul Regency and Yogyakarta City, the lowest number of power failure locations mentioned in the tweet was in May. So it can be concluded that March 2020 is the month that experiences the most power failure. Meanwhile, July 2020 was the month that experienced the least power failure.

5. Conclusion

Based on the results and discussion above, it can be concluded that The Indonesian Name Entity Recognition (NER) model has been successfully built using the SpaCy library. The entity in the power failure-related tweet that has been successfully extracted is the location entity, with the labels B-LOC and I-LOC. The labeling process has been adapted to the SpaCy format. Next, the Name Entity Recognition (NER) model that was built and used to classify tweets related to power failure resulted in good performance with 95.52% precision calculation results, 93.27% recall, and 94.38% f1-score. In addition, the f-1 score of the B-LOC entity was 94.53% and the f1-score of the I-LOC entity was 92.31%. So it can be said that the I-LOC entity is an entity that has more errors. The number of locations mentioned in the tweet related to the highest electricity disturbance originating from Sleman Regency with a total of 2,736 locations. While the number of locations mentioned in the tweet related to electrical disturbances, the lowest came from Gunung Kidul Regency with a total of 225 locations. March 2020 is the month with the most power failure. Meanwhile, July 2020 is the month that experiences the least power failure. The suggestions from researchers for further research are for further research, it can be developed using other Natural Language Processing (NLP) methods in the Named Entity Recognition (NER) process to obtain accuracy for comparison. It is necessary to conduct a study by comparing further data between the data on the number of locations mentioned in the tweet with the data on electricity disturbances generated from PLN, to see which districts/cities experience the most electricity disturbances and to see in what month the electricity disturbances occur most frequently in 2020. For further research, the addition of an automatic tweet data filtering process can be developed to overcome the limitations of the data filtering process in this study which is still done manually. The results of this study can be used by the community as a temporary reference to increase alertness to power failure in their area and can be used as a reference by PLN to improve its services.

References

[1] M. Chaudhari and S. Govilkar, "A Survey of Machine Learning Techniques for Sentiment Classification", *International Journal on Computational Science & Applications*, vol. 5, no. 3, pp. 13-23, 2015. Available: 10.5121/ijcsa.2015.5302

[2] U. Nahm, "Text mining with information extraction", *https://repositories.lib.utexas.edu/*, 2004. [Online]. Available: http://hdl.handle.net/2152/1280. [Accessed: 07- Jul- 2021].

[3] F. Peng and A. McCallum, "Information extraction from research papers using conditional random fields", *Information Processing & Management*, vol. 42, no. 4, pp. 963-979, 2006. Available: 10.1016/j.ipm.2005.09.002

[4] "Indonesia Pengguna Twitter Terbesar Ketiga di Dunia", *Databoks.katadata.co.id*, 2016. [Online]. Available: https://databoks.katadata.co.id/datapublish/2016/11/22/indonesia-pengguna-Twitter-terbesar-ketiga-di-dunia. [Accessed: 05- Nov- 2020].

[5] "Laporan Statistik - PT PLN (Persero)", *PT PLN (Persero)*, 2021. [Online]. Available: https://web.pln.co.id/stakeholder/laporan-statistik. [Accessed: 05- Nov- 2020].

[6] S. Hani, G. Santoso, and R. D. Wibowo, "Penempatan Recloser Sebagai Parameter Keandalan Sistem Proteksi Pada Sistem Distribusi", *Simp. Nas. RAPI XVIIII – 2019 FT UMS*, pp. 21–27, 2019

[7] M. Mursyidah and H. T. Hidayat, "Klasifikasi Teks Emosi Bahasa Aceh Menggunakan Metode Term Frekuensi / Invers Dokument Frekuensi," *Jurnal Infomedia*, vol. 2, no. 1, pp. 14–19, 2017, doi: 10.30811/.v2i1.462.

[8] I. Adiwijaya, "Text Mining dan Knowledge Discovery", *Kolokium bersama komunitas datamining Indonesia & soft-computing Indonesia*, pp. 1–9, 2006.

[9] A. Hotho, A. Nürnberger and G. Paaß, "A brief survey of text mining", *In LDV Forum-GLDV Journal for Computational Linguistics and Language Technology*, vol. 20, no. 1, pp. 19-62, 2005.

[10] X. Schmitt, S. Kubler, J. Robert, M. Papadakis, and Y. Letraon, "A Replicable Comparison Study of NER Software : StanfordNLP, NLTK, OpenNLP, SpaCy, and Gate", *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pp. 338–343, 2019.

[11] E. Cambria and B. White, "Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]", *IEEE Computational Intelligence Magazine*, vol. 9, no. 2, pp. 48-57, 2014. Available: 10.1109/mci.2014.2307227.

[12] R. Grishman, "Information Extraction", *IEEE Intelligent Systems*, vol. 30, no. 5, pp. 8–15, 2015, doi: 10.1109/MIS.2015.68.

[13] K. Adnan and R. Akbar, "An analytical study of information extraction from unstructured and multidimensional big data", *Journal of Big Data*, vol. 6, no. 1, 2019. Available: 10.1186/s40537-019-0254-8.

[14] J. P. C. Chiu and E. Nichols, "Named Entity Recognition with Bidirectional LSTM-CNNs", *Transactions of the Association for Computational Linguistics*, vol. 4, no. 2003, pp. 357–370, 2016, doi: 10.1162/tacl_a_00104.

[15] A. Mansouri, L. S. Affendey, and A. Mamat, "Named Entity Recognition Approaches", *International Journal of Computer Science and Network Security*, vol. 8, no. 2, pp. 339–344, 2008.

[16] Y. Wu, T. Fan, Y. Lee and S. Yen, "Extracting Named Entities Using Support Vector Machines", *Knowledge Discovery in Life Science Literature*, pp. 91-103, 2006. Available: 10.1007/11683568_8.

[17] I. Budi and S. Bressan, "Association rules mining for name entity recognition," *Proceedings of the Fourth International Conference on Web Information Systems Engineering, 2003. WISE 2003., 2003*, pp. 325-328, doi: 10.1109/WISE.2003.1254504.

[18] E. Partalidou, E. Spyromitros-Xioufis, S. Doropoulos, S. Vologiannidis, and K. I. Diamantaras, "Design and implementation of an open source Greek POS Tagger and Entity Recognizer using spaCy," *Proc. - 2019 IEEE/WIC/ACM Int. Conf. Web Intell. WI 2019*, pp. 337–341, 2019, doi: 10.1145/3350546.3352543.

[19] D. Gavrilov et al., "Feature Extraction Method from Electronic Health Records in Russia," *Fruct.Org*, pp. 497–500, 2020

[20] "Library Architecture", *Spacy*. [Online]. Available: https://spacy.io/api. [Accessed: 07- Jul- 2021].