ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 12 (2022) No. 5, pp. 213-229
DOI:10.33168/JSMS.2022.0513

Flood Mapping based on Online News using Named Entity Recognition

Abba Suganda Girsang, Ardivo Virsa Siswanto, Bima Krisna Noveta

Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480

agirsang@binus.edu, ardivo.virsa@binus.ac.id, bima.noveta@binus.ac.id

Abstract. Mapping the flood area in the big regions like Indonesia is an interesting challenge of the problem in the wide region like Indonesia is the flood. This paper aims to map the location of floods based on online news media using a mix of methodologies, including Named Entity Recognition (NER) and Leaflet Map. This research uses data from document media online which is accessed freely. This study uses six specific classes to improve the accuracy of the NER model based on the regional level of the Indonesian region using StanfordNER tools to generate the location prediction model from online news. The model is built by training data in six specific classes based on the region of Indonesia. Data media online is extracted from six classes of location using NER, then compiled data using based on the previous model built and Leaflet map to get the prediction spatial location. The results show that the proposed method can map flood locations based on 5 online news to 34 provinces in Indonesia. Moreover, the position accuracy geocoding of the proposed method using Leaflet Map is also quite good in mapping the flood location. Theoretically, this result shows that several flood points are happening in Indonesia using online news and geolocation to get the coordinates so that the location of the flood can be seen. Based on this research, it can be used by the authorities or community to find out about the flood disasters that are around them and to take advantage of flood data.

Keywords: geolocation, flood detection, named entity recognition, natural language processing, StanfordNER, online news.

1. Introduction

Indonesia has geographical, geological, hydrological, and demographic conditions that allow disasters to occur, whether caused by natural or non-natural factors. Following National Disaster Management Agency in 2020, Indonesia experienced 4,650 disasters, but the most dominant natural disaster is flooding. Based on (Yuniarti, 2018), floods can occur due to 2 factors, namely natural and human effects. Rainfall, physiography, erosion and sedimentation, river capacity, drainage capacity, and the influence of tides can all contribute to natural floods. Non-natural factors are caused by changes in watershed conditions, residential areas around the banks, damaged drainage, damaged forests, and improper planning of flood control programs.

Based on (Ramdhan, 2018), almost every municipal authority is responsible for dealing with floods. In tackling the problem of flooding, the capital city of Indonesia or the Government of DKI Jakarta implements short-term, medium-term, and long-term flood prevention policies. The implementation of this policy has inhibiting factors, namely communication factors, resources, bureaucratic or executor attitudes, and organizational structure including bureaucratic workflow arrangements. In the aspect of communication, the government still difficult to communicate with stakeholders. Certainly, communication in the regions is not as good as it is in the city, resulting in the uneven distribution of flood information. The public can find flood information from news on television and radio, but it has a weakness. The weakness is it cannot directly know the flood news that is around the public directly. On the contrary, the public needs to look for each of the news sources to find out.

Based on (Nadeau & Sekine, 2007), The goal of Named Entity Recognition (NER) is to recognize mentions of rigid designators from the text that belongs to predefined semantic types such as a person, location, organization, and so on. NER not only serves as a standalone tool for information extraction (IE), but it also plays an important role in a variety of natural language processing (NLP) applications such as text understanding (Cheng & Erk, 2019; Zhang et al., 2019), Information Retrieval (Guo et al., 2009; Petkova & Croft, 2007), automatic text summarization (Okurowski et al., 2000).

This research proposes building a geocoding flood platform based on online media news with the main contribution as follows.

(1) NER extraction.

NER extraction recognizes 4 classes named (PERSON, LOCATION, ORGANIZATION, and MISC) by default. Some literature uses 4 classes and 6 classes. As (Finkel et al., 2005) do use the person (PER), location (LOC), organization (ORG), and miscellaneous (MISC) classes, then (Syaifudin & Nurwidyantoro, 2016) uses the PERSON, ORGANIZATION, LOCATION, TIME, QUANTITY, and OTHER classes. Because this research was conducted in Indonesia, and the territory of Indonesia is very wide Indonesia has many regional levels.

Because of that, this research uses six specific classes to improve the accuracy of the NER model based on the regional level of the Indonesian region using StanfordNER tools to generate the location prediction model from online news. However, these six specific classes are PROP (province), KAB (city), KEC (sub-district), KEL (village), STREET, and POI (Place of Interest).

(2) Geocoding using Leaflet Maps.

This study also helps to estimate the spatial data of news locations from online news. The Application Programming Interface (API) is not used in this study to obtain the searched spatial location. The location of each news item, on the other hand, is derived through NER extraction and is adjusted to the geographical level in Indonesia. Each region/country has a unique model for displaying a location's position.

2. Literature Review

Big Data is the concept of formatting, storing, and analyzing very large data sets (Reyes et al., 2019). From the big data concept, data on flood news can be stored in big data and the data can be obtained using text mining. Text mining is a field located in computer information science, mathematics, and linguistics computing that can be used to analyze large texts efficiently, transparently, and reproducibly (Antons et al., 2020).

After getting the online news dataset, the text can be processed. According to (Goyal et al., 2018; Miranda-Escalada et al., 2020; Yu et al., 2020), a Named Entity is a marker of a word form that has similar properties from a collection of elements or classes. NER is an established Natural Language Processing (NLP) task that can be used to identify and classify a word from various types of entities. ENAMEX (person, place, organization) and NUMEX (time, money, and percentage) entities can be extracted from structured data using Named Entity. NLP is a set of approaches for making the human language accessible to computers, according to (Eisenstein, 2019). According to (Kogkitsidou & Gambette, 2020), Geographical Name Entity Recognition (GeoNER) was developed to extract and evaluate geographically. Tools that can be used such as CasEN, CoreNLP, Perdido, SEM, and spaCy tools. CoreNLP is a set of Neural Language Processing (NLP) analysis tools in Java and has been provided by StanfordNLP which contains Stanford NER. Stanford NER is a machine learning named entity recognition program based on linear-chain CRF sequence models.

Based on several studies, (Ou et al., 2015) perform sentiment classification based on the Naive Bayesian Classifier and have an F1 score of 100%. (Luoma et al., 2020) Researchers have tried using BERT and succeeded in identifying locations of 93.8%, then (Dashtipour et al., 2017) using Persian Named Entity Recognition get 77.6% result while using a hybrid, with the addition of SVM to get 90.3% results. Afterward, (Chou et al., 2016) used FundanNLP to identify locations by 35.3%, using

StanfordNER to get 21.5% results, and using the program created by Chou at 92.5%. Researchers (Schmitt et al., 2019) compared StanfordNLP, NTLK, Gate, OpenNLP, and SpaCy for accuracy based on the datasets from CoNLL and GMB and the results StanfordNLP had an accuracy of 81.05%, NLTK 48.47%, Gate 53.55%, OpenNLP 42.18%, and SpaCy 54.33%. (Sulaiman & Wahid, 2017) On the other hand, some researchers conducted a study that tried to use StanfordNER to identify 12 newspaper reports. From the twelve newspaper reports, the researcher got the highest accuracy of 58% and the researcher also tried to identify using IllinoisNER with the highest accuracy of 53%. (Kim & Cassidy, n.d.) Other researchers also used StanfordNER to identify the Australian Historical Newspaper. Previously, the researchers did train with 76% accurate results and trained with 600 articles and got 72% location accuracy results. (Neudecker et al., n.d.) Other researchers also used StanfordNER to identify Digital Historic Newspapers for locations in Dutch with an accuracy of 83.8% and in French, which achieved an accuracy of 31%. From (Hu et al., 2018), the researcher did train StanfordNER with 67.2% accurate results, and the researcher trained SpaCy with 66.3% accurate results, the dataset was trained for three and a half months (from Feb 18th, 2017 to May 30th, 2017) from Twitter data.

A geographic Information System (GIS) is a geocoding service that requires people to geocode manually. Real-time geocoding services such as Google Maps are an alternative to these static systems by offering real-time geocoding capabilities using available data and open-source systems. This model can be used for any purpose where people are accustomed to using the Google Maps Geocoding API or other commercial geocoding services (Ozer et al., 2020). The process of acquiring the required address's geographic coordinates (longitude and latitude) is known as geocoding (Cambon et al., 2021). Following the results of (Monir et al., 2021), there is open-source software for geocoding, namely (ArcGIS Online/ESRI, QGIS, and Google Maps), this research uses ESRI because it is faster than Google Maps and QGIS. (Küçük Matci & Avdan, 2018) Other researchers are also trying to experiment using the Google Geocoding API and ArcGIS. Then the research shows that ArcGIS is superior to the Google Geocoding API at coordinates with a certain distance. The results are for ArcGIS to produce an accuracy of 97% and Google Geocoding to produce an accuracy of 94.9%.

There are several APIs for geocoding such as DeGAUSS, ArcGis, Google, Leaflet, Mapbox GL JS, OpenLayers, PostGIS, GeoServer, and SAS that can be used (Brokamp et al., 2018; Duarte et al., 2021; Zunino et al., 2020). Previous research researched by (Girsang et al., 2020) used homemade geocode and google geo-mapping. This research used Geocode and Geomapping Leaflet. By implementing geocoding using the information or online news, it is hoped that flood information can be reached by everyone. Therefore, all application users can see the flooded area and can act as soon as possible.

3. Research Methodology

This research method consists of some steps which are shown in Figure 1 namely data collection, training data, and data testing. Each step is described as follows.

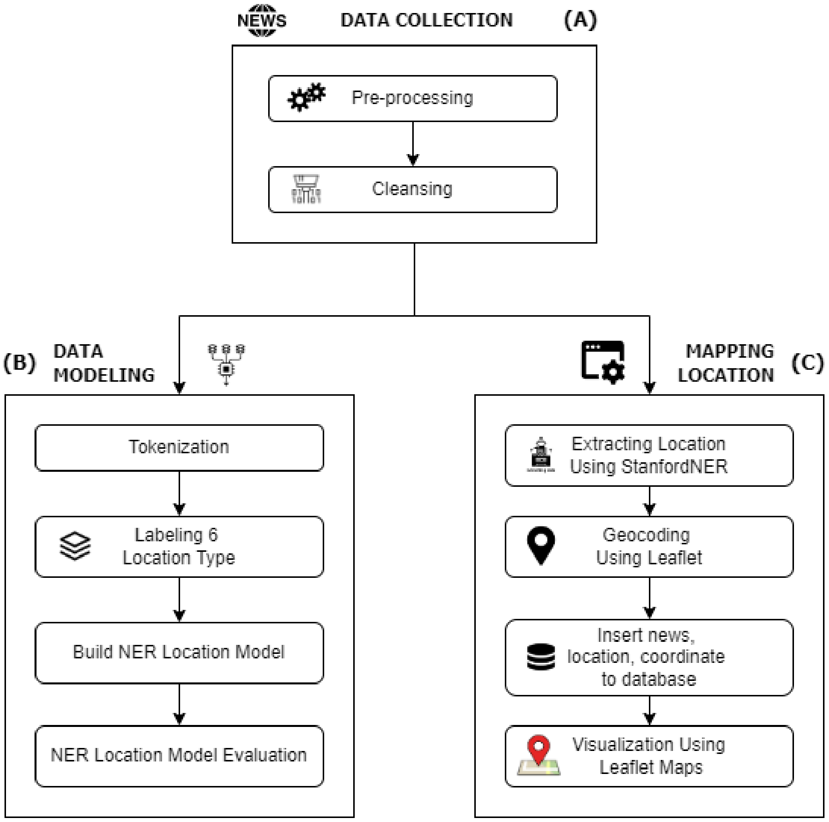


Fig. 1: Flowchart and methods of the experiment.

(A) Data Collection.

Data was collected from trusted online news such as Detik.com, Kompas.com, and Tempo. co, Liputan6.com, and CNNIndonesia.com which can be seen in Table 1. The researchers use online news because they are more dependable. Therefore, the news that the researchers present are protected from a hoax or fake news.

Table 1: Online news.

Online News	Count
Detik.com	100
Kompas.com	100
Tempo.co	100
Liputan6.com	100
CNNIndonesia.com	100

Pre-Processing. During the pre-processing stage, the researchers did the data selection process. The data that is taken has a location. This stage has to be done manually to separate data that does not have a location and discard the data such as in Table 2.

Table 2: Pre-processing.

News	Conclusion
Jakarta - Banjir yang menggenangi permukiman padat penduduk di Kembangan Utara, Jakarta Barat, kini sudah mulai surut. Camat Kembangan, Joko Sukarno, mengungkap penyebab banjir terjadi di wilayahnya.	Yes
Salah seorang warga, Kamaludin menyebut air datang sejak Jumat (15/7) malam kemarin sekitar pukul 22.30 WIB. Air datang cepat dengan ketinggian 100 cm.	No

Table 3: Cleansing.

Online News Before Cleansing	Online News After Cleansing
Jakarta - Banjir yang menggenangi permukiman padat penduduk di Kembangan Utara, Jakarta Barat, kini sudah mulai surut. Camat Kembangan, Joko Sukarno, mengungkap penyebab banjir terjadi di wilayahnya. Baca artikel detiknews, "Camat: Banjir di Kembangan Utara karena Kali Angke Hulu Belum Diturap" selengkapnya https://news.detik.com/berita/d-6183375/camat-banjir-di-kembangan-utara-karena-kali-angke-hulu-belum-diturap. Download Apps Detikcom Sekarang https://apps.detik.com/detik/	Jakarta - Banjir yang menggenangi permukiman padat penduduk di Kembangan Utara, Jakarta Barat, kini sudah mulai surut. Camat Kembangan, Joko Sukarno, mengungkap penyebab banjir terjadi di wilayahnya.

Cleansing. In the cleansing stage, things that are not needed will be removed so that the training process becomes more accurate, and the testing process gets satisfactory results. The processes in this stage are the removal of excess spaces and links will be deleted such as in Table 3.

(B) Data Modelling.

Four stages are conducted to train the model which is tokenization, labeling, training model, and model evaluation.

Table 4: Tokenization.

Before Tokenization	After Tokenization
Jakarta - Banjir yang menggenangi permukiman padat penduduk di Kembangan Utara, Jakarta Barat, kini sudah mulai surut. Camat Kembangan, Joko Sukarno, mengungkap penyebab banjir terjadi di wilayahnya.	Jakarta, Banjir, yang, menggenangi, permukiman, padat, penduduk, di, Kembangan, Utara, Jakarta, Barat, kini, sudah, mulai, surut, Camat, Kembangan, Joko, Sukarno, mengungkap, penyebab, banjir, terjadi, di, wilayahnya.

*Tokenization***.** Tokenization is a technique of breaking text into tokens. Tokens are frequently words since words are the most prevalent semantically significant components of texts (Welbers et al., 2017). Full texts are too specific to execute any meaningful computations with, hence this stage is critical for computational text analysis as in Table 4.

Table 5: Class location.

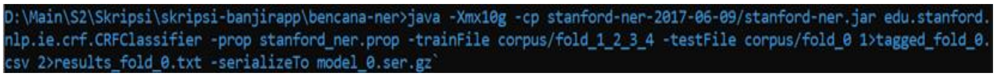
Class	Description
PROP	Provinces
KAB	Districts
KEC	Sub-districts
KEL	Villages
STREET	Roads
POI	Place of Interest (Building, Statue)
O	Other

Labeling 6 class location. Labeling is the process of name tagging the location tokens using Named Entity Recognition (NER) with StanfordNER tools. In labeling, there are six classes, namely six location classes according to the regional level in Indonesia and there is one additional class that is not included as a location. The class that is not included as a location is class (O) as other as in Table 5. The labeling example of news can be seen in Table 6.

Table 6: Example of labelling.

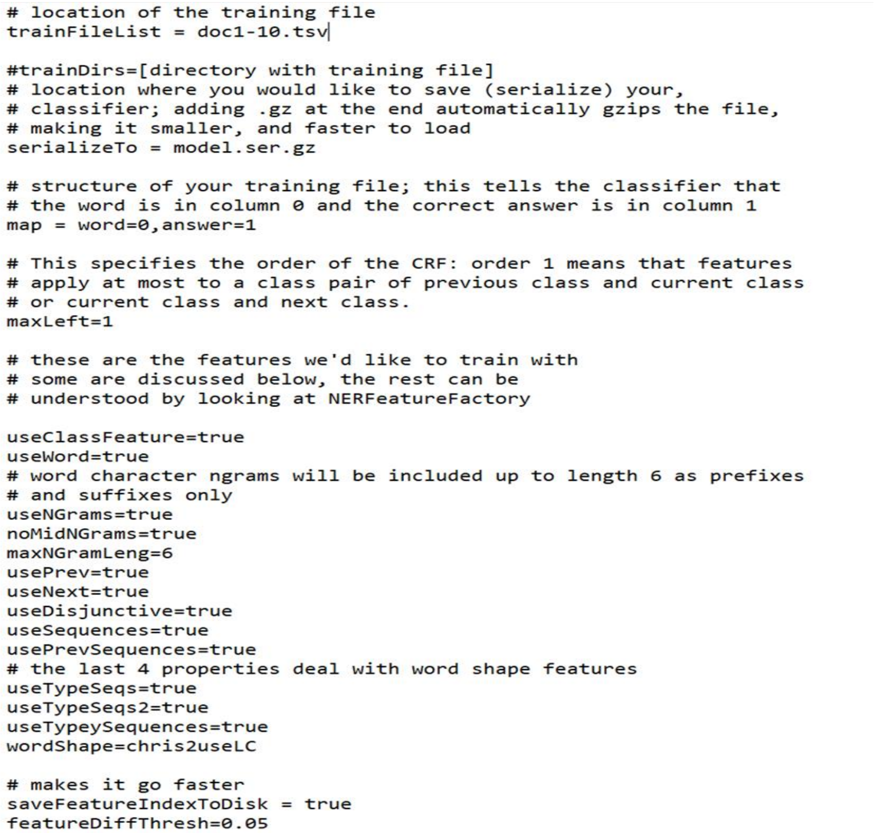
Words	Label
Jakarta	PROP
Banjir	O
Yang	O
Menggenangi	O
Permukiman	O
Padat	O
Penduduk	O
Di	O
Kembangan	KEL
Utara	KEL
Jakarta	KAB
Barat	KAB

Kini	O
Sudah	O
Mulai	O
Surut	O
Camat	O
Kembangan	KEC
Joko	O
Sukarno	O
Mengungkap	O
Penyebab	O
Banjir	O
Terjadi	O
Di	O
Wilayahnya	O



```
D:\Main\S2\Skripsi\skripsi-banjirapp\bencana-ner>java -Xmx10g -cp stanford-ner-2017-06-09/stanford-ner.jar edu.stanford.
nlp.ie.crf.CRFClassifier -prop stanford_ner.prop -trainFile corpus/fold_1_2_3_4 -testFile corpus/fold_0 1>tagged_fold_0.
csv 2>results_fold_0.txt -serializeTo model_0.ser.gz`
```

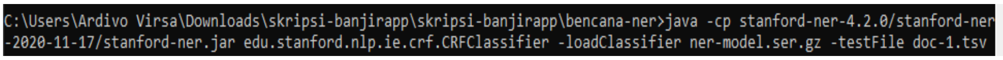
Figure 2: Training model.



```
# location of the training file
trainFileList = doc1-10.tsv

#trainDirs=[directory with training file]
# location where you would like to save (serialize) your,
# classifier; adding .gz at the end automatically gzips the file,
# making it smaller, and faster to load
serializeTo = model.ser.gz

# structure of your training file; this tells the classifier that
# the word is in column 0 and the correct answer is in column 1
map = word=0,answer=1

# This specifies the order of the CRF: order 1 means that features
# apply at most to a class pair of previous class and current class
# or current class and next class.
maxLeft=1

# these are the features we'd like to train with
# some are discussed below, the rest can be
# understood by looking at NERFeatureFactory

useClassFeature=true
useWord=true
# word character ngrams will be included up to length 6 as prefixes
# and suffixes only
useNGrams=true
noMidNGrams=true
maxNGramLeng=6
usePrev=true
useNext=true
useDisjunctive=true
useSequences=true
usePrevSequences=true
# the last 4 properties deal with word shape features
useTypeSeqs=true
useTypeSeqs2=true
useTypeySequences=true
wordShape=chris2useLC

# makes it go faster
saveFeatureIndexToDisk = true
featureDiffThresh=0.05
```

Fig. 3: Properties file.

Build NER Location Model. The model will be trained using the command prompt and the command-line is executed like in Figure 2. The properties file as in Figure 3 that is used to build the model contains the information of the dataset file path that is going to be trained and the destination file path for the model results in the form of a ser.gz extension file.



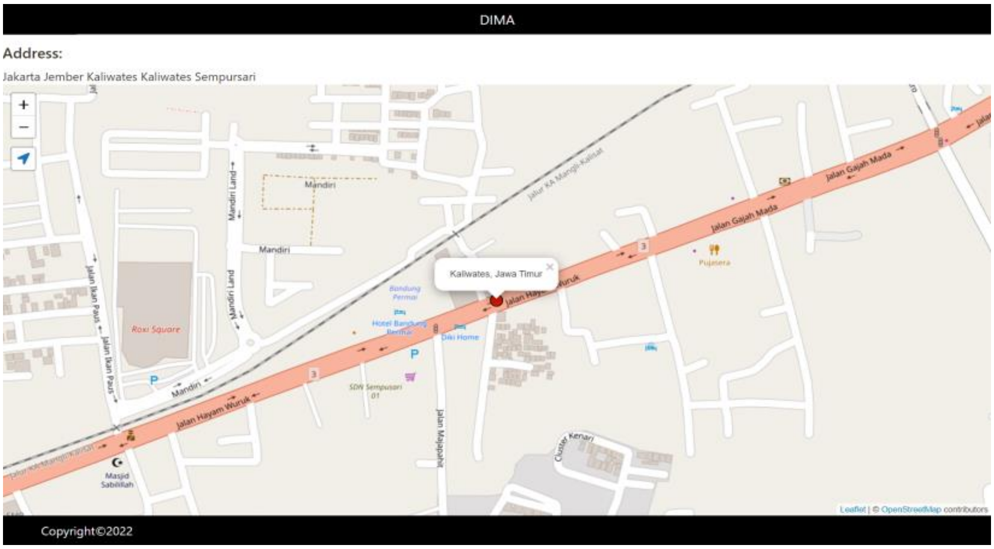
```
C:\Users\Ardivo Virsa\Downloads\skripsi-banjirapp\skripsi-banjirapp\bencana-ner>java -cp stanford-ner-4.2.0/stanford-ner
-2020-11-17/stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier ner-model.ser.gz -testFile doc-1.tsv
```

Fig. 4: Model’s command-line.

NER Location Model Evaluation. Before the researchers test the model that the researchers have made, the researchers have to evaluate the NER model. The model’s evaluation is executed from a command prompt with a command line which can be seen in Figure 4.

(C) Mapping Location

In this study, the testing process is inputting the flood news dataset manually. As shown below in Figure 5, Berita1.txt which is the dataset is inputted using Choose File button.



DIMA

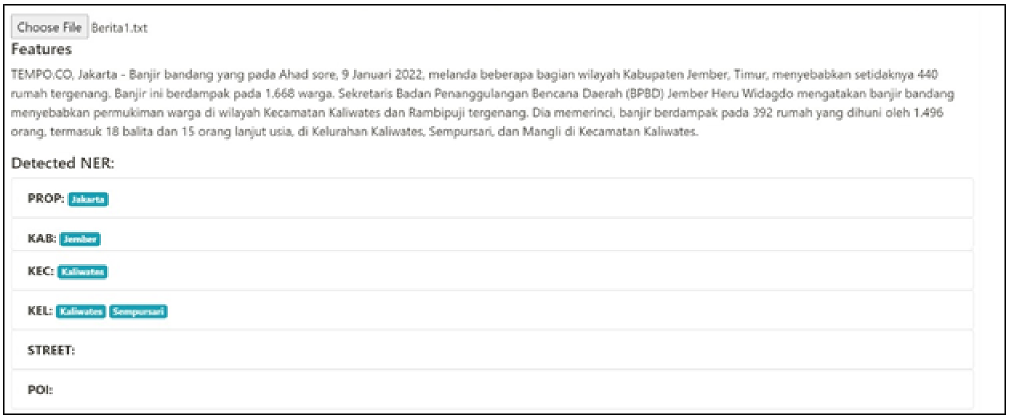
Address:

Jakarta Jember Kaliwates Kaliwates Sempursari

Copyright©2022

Fig. 5: Extracting location using StanfordNER.

Extracting Location Using StanfordNER. After the news has been inserted, the program will display the news content and display the NER results. The results of NER location detection are provinces, districts, sub-districts, villages, roads, and Places of Interest (POI) such as in Figure 5.



Choose File Berita1.txt

Features

TEMPO.CO, Jakarta - Banjir bandang yang pada Ahad sore, 9 Januari 2022, melanda beberapa bagian wilayah Kabupaten Jember, Timur, menyebabkan setidaknya 440 rumah tergenang. Banjir ini berdampak pada 1.668 warga. Sekretaris Badan Penanggulangan Bencana Daerah (BPBD) Jember Heru Widagdo mengatakan banjir bandang menyebabkan permukiman warga di wilayah Kecamatan Kaliwates dan Rambipuji tergenang. Dia memerinci, banjir berdampak pada 392 rumah yang dihuni oleh 1.496 orang, termasuk 18 balita dan 15 orang lanjut usia, di Kelurahan Kaliwates, Sempursari, dan Mangli di Kecamatan Kaliwates.

Detected NER:

PROP: Jakarta

KAB: Jember

KEC: Kaliwates

KEL: Kaliwates Sempursari

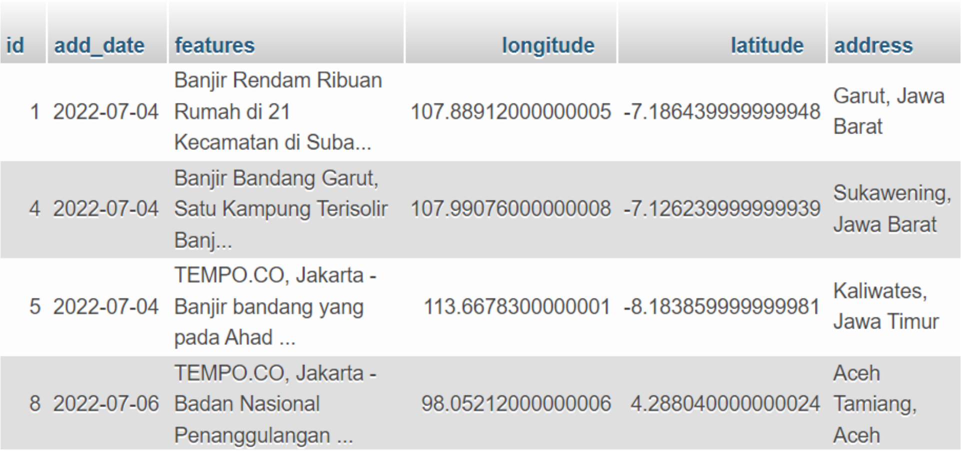
STREET:

POI:

Fig. 6: Geocoding using leaflet.

Geocoding Using Leaflet. After getting the NER tags containing the province, district, sub-district, village, road, and POI from the inserted flood news, the geocoding process can be done to find the coordinates of the news. The program utilizes Leaflet API to get the coordinates of the flood news as shown in Figure 6.

Insert News, Address, Coordinate to Database. After the program display, the result of the flood news's coordinate, the coordinate (latitude and longitude), news, and address are inserted into the database like in Figure 7. This study uses MySQL Database to store the information.



id	add_date	features	longitude	latitude	address
1	2022-07-04	Banjir Rendam Ribuan Rumah di 21 Kecamatan di Suba...	107.88912000000005	-7.186439999999948	Garut, Jawa Barat
4	2022-07-04	Banjir Bandang Garut, Satu Kampung Terisolir Banj...	107.99076000000008	-7.126239999999939	Sukawening, Jawa Barat
5	2022-07-04	TEMPO.CO, Jakarta - Banjir bandang yang pada Ahad ...	113.6678300000001	-8.183859999999981	Kaliwates, Jawa Timur
8	2022-07-06	TEMPO.CO, Jakarta - Badan Nasional Penanggulangan ...	98.05212000000006	4.288040000000024	Aceh Tamiang, Aceh

Fig. 7: Insert news, address, coordinate to database.

4. Results and Discussion

4.1. Results

The results of this study, first evaluate the NER model that has been trained and find the accuracy of the Leaflet Geocode.

Table 7: Model’s evaluation result.

Entity	P	R	F1	TP	FP	FN
KAB	0.9967	0.9952	0.9960	3971	13	19
KEC	0.9981	0.9952	0.9967	2095	4	10
KEL	0.9933	0.9866	0.9899	1477	10	20
POI	0.9946	0.9919	0.9933	3317	18	27
PROP	0.9985	0.9961	0.9973	2062	3	8
STREET	1.0000	0.9976	0.9988	409	0	1
Totals	0.9964	0.9937	0.9950	13331	48	85

The evaluation result of the NER Location Model as shown in Table 7 is quite satisfactory because it can achieve a high level of accuracy for each of the NER classes which are KAB 99,6%, KEC 99,67%, KEL 98.99%, POI 99.33%, PROP 99.73%, and STREET 99.88%.

Table 8: Accuracy leaflet geocode.

Input Address	Result Address	Accuracy
Jakarta Jember Kaliwates Kaliwates Sempursari	Kaliwates, Jawa Timur	81.75
Jakarta Aceh Aceh Tamiang Aceh Utara Langsa	Aceh Tamiang, Aceh	74.69
Jakarta Jakarta Selatan Jalan Kemang Raya Kawasan Kemang Pasca Hujan Deras	Kemang	79.38
Lumajang Candipuro Sumbermujur	Sumbermujur, Candipuro, Jawa Timur	100
Jakarta Jawa Tengah Semarang Tlogosari Kaligawe Semarang Barat Muktiharjo Kidul Tambakrejo Krobokan Jalan Sawojajar BPBD	Tlogo, Tuntang, Semarang, Jawa Tengah	77.4
Jawa Barat Jakarta Subang Bogor Depok Tangerang Bekasi Cianjur Cimanggung Cisarua Puncak Bogor Sehari Puncak Bogor Gunung Mas Institut Pertanian Bogor Gunung Mas Puncak Bogor Puncak	Puncak, Cigugur, Jawa Barat	74.25
Jawa Barat Subang Pada Garut Subang Subang Utara Bandung Pamanukan Pamanukan Bojong Sungai Cipunagara Masjid Al-Hadad Pamanukan Kali Cigadung STTG Garut	Garut, Jawa Barat	74.25
Garut Sukawening Karangtengah Sukawangi Sukaresmi Cintamanik Cinta Caringin Sukawening Sukamukti Sukaresmi Sukalilah Sungai Ciloa BPBD) Garut BPBD Garut Gunung Papandayan Polres Garut	Sukawening, Jawa Barat	74.25
Sulawesi Tengah Sigi Rogo Dolo Selatan	Rogo, Dolo Selatan, Sulawesi	100
Sumatera Barat Solok Lembah Gumanti	Lembah Gumanti, Sumatera Barat	100

As displayed in Table 8, the accuracy results were obtained by using StanfordNER name tagging as an input address and using Leaflet Geocode to find flood locations. The results of this study are quite satisfactory because it can achieve an accuracy of **83.597%** using Leaflet Geocode.

4.2. Discussions

This section will discuss the display of the mapping of online news.

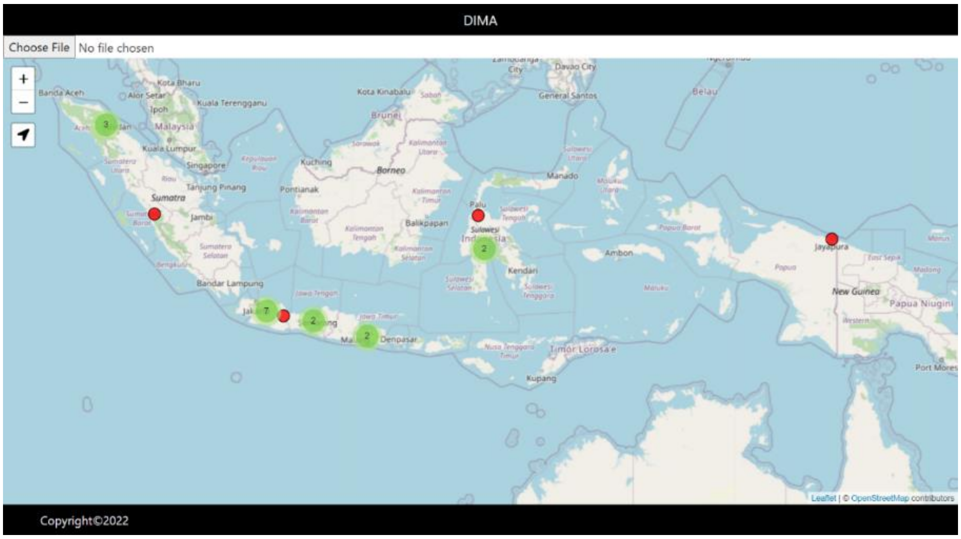


Fig. 8: Retrieve and display the coordinate on the leaflet map.

At this stage, which is shown in Figure 8, the program retrieved the coordinate data from the database containing the latitude and longitude data for the specific flood news. Then the model inserts it into the Leaflet Map and the map will show the flood points in the form of a red dot and regions of Indonesia.

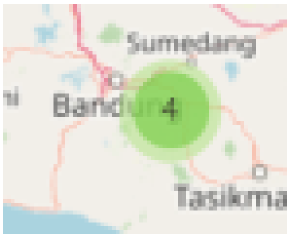


Fig. 9: A group of flood dots.

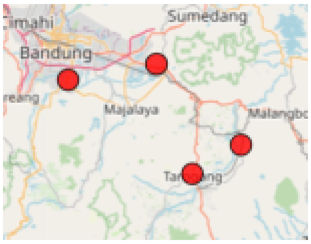


Fig. 10: Multiple red dots to indicate the specific flood points.

Figure 9 shows that if 4 flood points are close, the model will only display one point in the form of a green dot with the flood count. Therefore, the points do not overlap when the user zooms out on the Leaflet Map. Then, when the user zooms in on the Map, the green point will be replaced with multiple red dots for example in Figure 10 to display the precise position for each of the flood points.

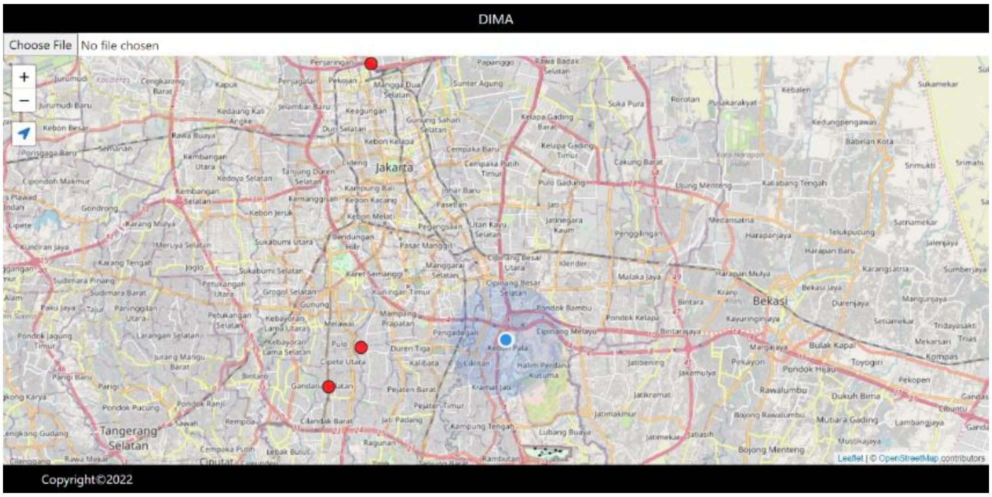


Fig. 11: Flood position and user position

Figure 11 shows the user's position and his/her surroundings. The user's position is marked with a blue dot, and the flood position is marked with a red dot which will help the user to anticipate the flood points. Therefore, the user does not have to worry if they have not watched the news or read the news, because the model can help the user to investigate the floods that occurred in certain places from the Leaflet Map.

5. Conclusions

Floods that are induced by heavy rainfall are still very concerning in Indonesia. Moreover, the flood information that exists is still based on current news and cannot be obtained publicly. The other reason for the flooding problem is some drainages cannot handle heavy rainfall. As a result, based on the online news, this research proposed a helpful model to deliver public flood information by displaying the current flood points. With this strategy, the public can receive accurate information that allows individuals to oversee flood points in specific areas. The online news data are inputted manually into the system (testing phase). Then, the news is processed through the NER Location Model which utilizes StanfordNER tools for name tagging the flood news location. After that, the Leaflet Geocoding will help to identify the flood point's coordinates. Finally, the coordinates (latitude, longitude) are displayed on the Leaflet map, with the red dot representing the flood point and the blue dot representing the application user's location.

The Limitation of this study is If a person's name is the same as the geographical level, such as the name of the location, the name of the street, the name of the village, the name of the sub-district, the name of the district, or the name of the province, the name of the province is used.

This research has a contribution in terms of science, and society. From a scientific point of view, this research uses six specific classes to improve the accuracy of the NER model based on the regional level of the Indonesian region using StanfordNER tools to generate the location prediction model from online news. From the community perspective, this research can be used by the authorities or the community to find out about the flood disasters that are around them and take advantage of the flood data.

For further research on online news-based flood mapping, words that can be equal to six classes can be replaced and added training data with substitute words from certain areas so that the model can extract words, this can be one of the most effective ways to improve classification performance.

References

Antons, D., Grünwald, E., Cichy, P., & Salge, T. O. (2020). The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. *R and D Management*, *50*(3), 329–351. DOI: https://doi.org/10.1111/radm.12408.

Brokamp, C., Wolfe, C., Lingren, T., Harley, J., & Ryan, P. (2018). Decentralized and reproducible geocoding and characterization of community and environmental exposures for multisite studies. *Journal of the American Medical Informatics Association*, *25*(3), 309–314.

Cambon, J., Hernangómez, D., Belanger, C., & Possenriede, D. (2021). tidygeocoder: An R package for geocoding. *Journal of Open Source Software*, *6*(65), 3544.

Cheng, P., & Erk, K. (2019). Attending to Entities for Better Text Understanding. http://arxiv.org/abs/1911.04361.

Chou, C., Chang, C., & Huang, Y. (2016). Boosted web named entity recognition via tri-training. *ACM Trans. Asian Low-Resour. Lang. Inf. Process*, *16*(10). DOI: https://doi.org/10.1145/2963100.

Dashtipour, K., Gogate, M., & Hussain, A. (2017). Persian named entity recognition towards an interactive simulation framework for effective e-learning in university classroom environments. V=view project. https://doi.org/10.1109/ICCI-CC.2017.8109733.

Duarte, L., Teodoro, A. C., Lobo, M., Viana, J., Pinheiro, V., & Freitas, A. (2021). An Open Source GIS Application for Spatial Assessment of Health Care Quality Indicators. *ISPRS International Journal of Geo-Information*, *10*(4), 264.

Eisenstein, J. (2019). Introduction to natural language processing. MIT press.

Finkel, J. R., Grenager, T., & Manning, C. D. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 363–370.

Girsang, A. S., Muhamad Isa, S., & Fajar, R. (2020). Implementation of a Geocoding In Journalist Social Media Monitoring System. *International Journal of Engineering Trends and Technology, 68*, 1–4. DOI: https://doi.org/10.14445/22315381/IJETT-VXXXX.

Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review, 29*, 21–43.

Guo, J., Xu, G., Cheng, X., & Li, H. (2009). Named Entity Recognition in Query.

Hu, Y., Mao, H., & McKenzie, G. (2018). *A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements*. DOI: https://doi.org/10.1080/13658816.2018.1458986.

Kim, S. mac, & Cassidy, S. (n.d.). Finding Names in Trove: Named Entity Recognition for Australian Historical Newspapers. Retrieved February 11, 2022, from http://trove.nla.gov.au/ndp/del/article/.

Kogkitsidou, E., & Gambette, P. (2020). Normalisation of 16th and 17th century texts in French and geographical named entity recognition. *Proceedings of the 4th ACM SIGSPATIAL Workshop on Geospatial Humanities*, 28–34.

Küçük Matci, D., & Avdan, U. (2018). Address standardization using the natural language process for improving geocoding results. *Computers, Environment and Urban Systems*, 1–0. DOI: https://doi.org/10.1016/j.compenvurbsys.2018.01.009.

Luoma, J., Oinonen, M., Pyykönen, M., Laippala, V., & Pyysalo, S. (2020). A Broad-coverage Corpus for Finnish Named Entity Recognition. 11–16. http://www.digitoday.fi/.

Miranda-Escalada, A., Farré, E., & Krallinger, M. (2020). Named entity recognition, concept normalization and clinical coding: overview of the cantemist track for cancer text mining in Spanish. *Corpus, Guidelines, Methods and Results. IberLEF@ SEPLN*, 303–323.

Monir, N., Abdul Rasam, A. R., Ghazali, R., Suhandri, H. F., & Cahyono, A. (2021). Address geocoding services in geospatial-based epidemiological analysis: A comparative reliability for domestic disease mapping. *International Journal of Geoinformatics, 17*(5).

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. http://projects.ldc.upenn.edu/gale/.

Neudecker, C., Wilms, L., Faber, W. J., van Veen, T., & Kb, @. (n.d.). *Large-scale refinement of digital historic newspapers with named entity recognition*. Retrieved February 11, 2022, from www.europeana-newspapers.eu.

Okurowski, M. E., Wilson, H., Urbina, J., Taylor, T., Clark, R. C., & Krapcho, F. (2000). Text Summarizer in Use: Lessons Learned from Real World Deployment and Evaluation.

Ou, X. H., Cao, Y., & Mu, X. W. (2015). Classification of Sentiment Sentences Based on Naive Bayesian Classifier. *Journal of Logistics, Informatics and Service Science*, *2*(1), 48–57.

Ozer, M., Zidar, M., Deryol, R., Varlioglu, S., Eldivan, I. S., & Akbas, H. (2020). creating a real-time geocoding system: Implications of open source for the public safety. *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, 1185–1188.

Petkova, D., & Croft, W. B. (2007). Proximity-based Document Representation for Named Entity Retrieval.

Ramdhan, G. (2018). Implementasi kebijakan penanggulangan banjir di DKI Jakarta 2013-2017. *Jurnal Ilmu Administrasi: Media Pengembangan Ilmu Dan Praktek Administrasi*, *15*(1), 78–87.

Reyes, P., Suresh, S., & Renukappa, S. (2019). The adoption of big data concepts for sustainable practices implementation in the construction industry. *Proceedings - 11th IEEE/ACM International Conference on Utility and Cloud Computing Companion, UCC Companion 2018*, 341–348. DOI: https://doi.org/10.1109/UCC-Companion.2018.00079

Schmitt, X., Kubler, S., Robert, J., Papadakis, M., & LeTraon, Y. (2019). A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 338–343.

Sulaiman, S., & Wahid, R. A. (2017). *Using Stanford NER and Illinois NER to Detect Malay Named Entity Recognition Development of Islamic Counselling System View project*. DOI: https://doi.org/10.7763/IJCTE.2017.V9.1128.

Syaifudin, Y., & Nurwidyantoro, A. (2016). Quotations identification from Indonesian online news using rule-based method. *2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 187–194.

Welbers, K., van Atteveldt, W., & Benoit, K. (2017). Text analysis in R. *Communication Methods and Measures*, *11*(4), 245–265. DOI: https://doi.org/10.1080/19312458.2017.1387238.

Yu, J., Bohnet, B., & Poesio, M. (2020). Named entity recognition as dependency parsing. *ArXiv Preprint ArXiv:2005.07150*.

Yuniarti, T. (2018). Kepemimpinan dan pengelolaan modal sosial dalam penanggulangan bencana banjir. *Makna: Jurnal Kajian Komunikasi, Bahasa, Dan Budaya*, *3*(1), 94–128.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). ERNIE: Enhanced Language Representation with Informative Entities.

Zunino, A., Velázquez, G., Celemín, J. P., Mateos, C., Hirsch, M., & Rodriguez, J. M. (2020). Evaluating the performance of three popular web mapping libraries: A case study using argentina’s life quality index. *ISPRS International Journal of Geo-Information*, *9*(10), 563.