Vol. 10, No. 3, 2024, pp. 903-913
DOI: https://doi.org/10.29210/020244467

Contents lists available at Journal IICET

JPPI (Jurnal Penelitian Pendidikan Indonesia)
ISSN: 2502-8103 (Print) ISSN: 2477-8524 (Electronic)

Journal homepage: https://jurnal.iicet.org/index.php/jppi

Utilization of geocoding for mapping infrastructure impacts and mobility due to floods in indonesia based on twitter analytics

Muhammad Imam Taufiq[*], Abba Suganda Girsang
Computer Science Department, Faculty of Computer and Information Technology, Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received Jul 20th, 2024
Revised Aug 25th, 2024
Accepted Sept 30th, 2024

Keywords:

Flood impacts,
Geocoding visualization,
Indonesia,
Named Entity Recognition,
Twitter-based data extraction

ABSTRACT

Flooding, a frequent natural disaster in Indonesia, is caused by several factors such as high-intensity rainfall, climate change, inadequate drainage and urban infrastructure challenges, impacting communities, infrastructure and economic activities. The lack of accurate and centralized data hinders government efforts to identify affected areas and respond effectively. Named Entity Recognition (NER), a machine learning-based information extraction tool, offers the potential for geocoding flood-related data from social media, such as Twitter. The purpose of this research is to develop a Named Entity Recognition (NER)-based model to extract location information from Twitter and visualize flood impacts through geocoding. The method used is a combination of Qualitative Analysis with Machine Learning and Geospatial Analysis to assess flooding impacts using Twitter data. Initially, a qualitative analysis of tweets extracts flood-related keywords to identify patterns. Then, Named Entity Recognition (NER) identifies locations, which are converted into geographic coordinates through geocoding for map visualization. The results show that location extraction from flood-related tweets using the Named Entity Recognition (NER) model and geocoding produces very useful and accurate data. About 50% of the flood-related tweets included location tokens, which shows the importance of geographic information in understanding the impact of disasters. The location extraction process using the NER model proved to be effective, although there were some discrepancies between the extracted location tokens and the actual geographic data, especially at the more detailed location level. However, the evaluation results show that 99.5% of the extracted locations correspond to valid locations, especially in the Indonesian region. This shows that the use of the NER model and geocoding is highly effective in analyzing flood impacts and provides significant benefits in disaster management and geospatial analysis based on social media data.

Corresponding Author:

Muhammad Imam Taufiq,
Bina Nusantara University
Email: muhammad.taufiq@binus.ac.id

Introduction

Flooding is a frequent and severe natural disaster in Indonesia, driven by high-intensity rainfall, climate change, inadequate drainage systems, and unplanned urban infrastructure. These floods disrupt daily life, causing damage to public facilities, traffic congestion, and road closures (Dartanto., 2022). The effects are felt most acutely in urban areas, where infrastructure and mobility are highly interconnected (Plyushteva & Schawanen.,

JPPI (Jurnal Penelitian Pendidikan Indonesia)
Vol. 10, No. 3, 2024, pp. 903-913

2023). Despite the recurring nature of floods, government agencies often face challenges in identifying affected locations promptly due to a lack of centralized and accurate data sources, relying instead on fragmented reports from television news or social media platforms. This delay hampers effective disaster response and infrastructure management (Mitchell & Pizzi., 2021).

Social media platforms, particularly Twitter, have become valuable sources of real-time information during disasters. Tweets often contain firsthand reports, including the location and extent of flooding impacts. However, extracting and structuring this information for practical use remains a challenge (Havas & Resch., 2021). Geocoding, the process of converting descriptive location data into geographic coordinates, offers a promising solution. By integrating geocoded data with spatial visualization tools, it becomes possible to map flood impacts, aiding in disaster management and infrastructure planning (Splendiani & Capriello., 2022).

Main section of this study is the usage of NER model such as StandordNER which has been trained using tweets in Indonesian as data in location's format of Indonesia. Named Entity Recognition (NER), a machine learning technique, facilitates the extraction of specific entities such as locations from unstructured text data (Govindarajan et al., 2023). There are four location tokens in this study which are Provinsi (PROP) 'province', Kabupaten (KAB) 'regency', Kecamatan (KEC) 'district', and Kelurahan (KEL) 'rural'. Each location token will be appraised to sort from specific (Kelurahan) to general (Provinsi). Token extraction and appraisal process produces address in Indonesian format; thus, spatial data can be queried using open-source API such as ArcGIS. Spatial data query results include location information and map coordinate that can be visualized as map (Ye et al., 2021).

A previous study was conducted in geocoding from social media to conduct environment planning and monitoring by using StanfordNER. The model compares evaluation result of before and after in re-training process. Evaluation result shows that model which have gone through re-training with ALTA dataset showed decline in precision score, from 94.51% to 86.68% (Lawu et al., 2021). However, the dataset showed increase in recall score from 32.07% to 69.72%; and increase in f1-score from 47.88% to 77.20%.

This study focuses on development of geocoding model by using Named Entity Recognition (NER) with data from social media as sources. Data usage that coming from social media is based on how fast the information can be extracted so the obtained information is timely relevant with the event. Data used are Tweets in Indonesian on Twitter with impacts of urban flooding, such as traffic jam, damages on facilities, and road closing as the topic (Putra et al., 2022).

The novelty of this research lies in the development of a Named Entity Recognition (NER)-based geocoding model for the extraction and mapping of flood impacts in Indonesia using Twitter data. By utilizing social media as a source of fast and relevant information, this research creates a new approach to identify flood-affected locations in real-time, which are then visualized in the form of maps using ArcGIS and Folium APIs. This study aims to develop a geocoding model based on NER to analyze flood impacts in Indonesia using Twitter data. By leveraging the speed and availability of social media information, the research seeks to create a scalable system for visualizing affected areas and understanding infrastructure and mobility challenges caused by floods. The findings will contribute to improved disaster response and urban planning by offering a data-driven approach for identifying and addressing flood impacts. Future developments include integrating additional data sources, such as news articles, to enhance the accuracy and specificity of location extraction and geocoding.

Method

The method used is Qualitative method with Machine Learning and Geospatial Analysis approach is a combination that combines text analysis and spatial analysis to evaluate the impact of flooding based on Twitter data. The approach starts with a qualitative analysis of the collected tweets, where keywords related to flood impacts are extracted to identify context and patterns in the text (Mollick et al., 2023). Next, using machine learning-based Named Entity Recognition (NER), locations in the tweets were identified and mapped to extract geospatial information. The extracted location information is then converted into geographic coordinates through geocoding, which enables data visualization on a map (Mota et al., 2021). This study consists of several step as shown on figure 1.

Data Collection

This study collects 3,441 Tweets from Twitter that was conducted by using Twitter API. Extracted information are coming from tweets with impacts of flooding in Indonesia as topic during 2021. The data are focusing on tweets with 15 keywords about impact of flooding which includes traffic jam, damages on facility, and road closing.

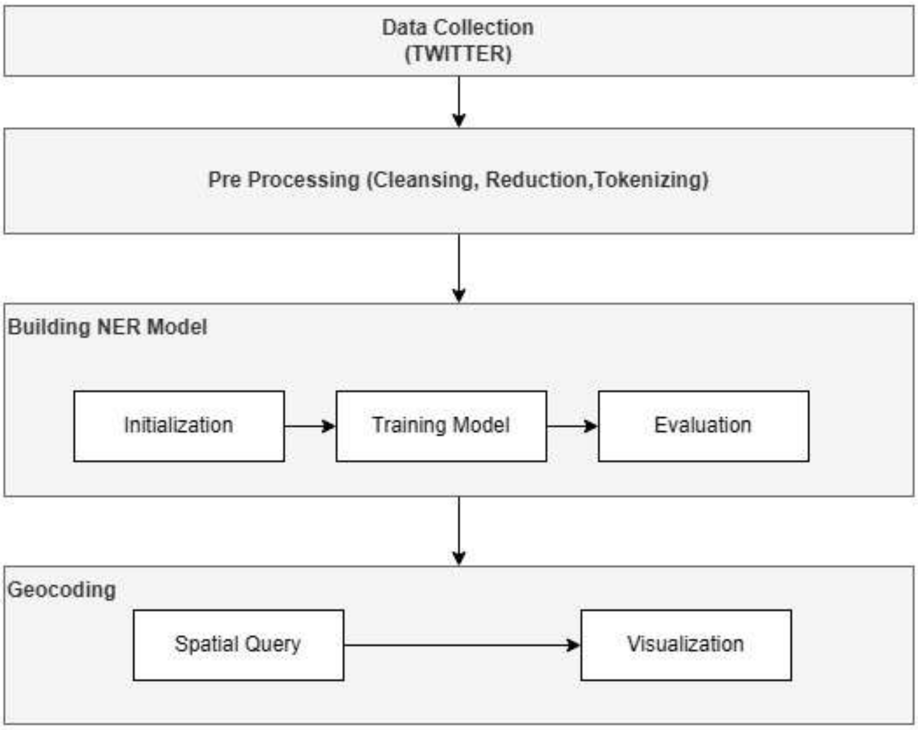

Figure 1 <Research Methodology used in This Study>

Pre-Processing

Pre-Processing step is conducted to ensure the tweets are free from elements that can affect evaluation result such as URL, special characters, and emoji. This step split into two sections of process:

Cleansing

During the cleansing process, texts were cleaned by removing symbols such as hastag, special character, and emoji to ensure the training process becomes more accurate. This process is conducted automatically by using regular expressions as shown on Table 1.

Table 1 <Cleansing Process>

Before	Hujan yang menguyur Kelurahan Kampung Melayu, Jatinegara, Jakarta Timur, DKI Jakarta, sejak Kamis sore kemarin sebabkan ratusan rumah di permukiman padat terendam air. #Banjir https://shorturl.at/IvsFO
After	Hujan yang menguyur Kelurahan Kampung Melayu, Jatinegara, Jakarta Timur, DKI Jakarta, sejak Kamis sore kemarin sebabkan ratusan rumah di permukiman padat terendam air.

Reduction

During the reduction process, redundant elements in tweets such as retweet to reduce the number of data used during training process. Table 2 shows tweets which are considered redundant and not redundant.

Table 2 <Redundant Tweet Sample>

Not Redundant	Hujan yang menguyur Kelurahan Kampung Melayu, Jatinegara, Jakarta Timur, DKI Jakarta, sejak Kamis sore kemarin sebabkan ratusan rumah di permukiman padat terendam air.
Redundant	RT: Hujan yang menguyur Kelurahan Kampung Melayu, Jatinegara, Jakarta Timur, DKI Jakarta, sejak Kamis sore kemarin sebabkan ratusan rumah di permukiman padat terendam air.

Tokenizing

During this process, text is disassembled into words which later called as tokens. Tokens then labelled manually by using location entity that matches with subdivisions of Indonesia as shown on Table 3. Labelling process

JPPI (Jurnal Penelitian Pendidikan Indonesia)
Vol. 10, No. 3, 2024, pp. 903-913

result produces token data which has been labeled correspondently with location entity where any label which doesn't belong to location entity will be marked (O) as others, as shown on Table 4.

Table 3 <Location Entity of Indonesia's Subdivision>

Entity Name	**Information**
Provinsi (PROP)	Province Name
Kabupaten (KAB)	Regency/City Name
Kecamatan (KEC)	District Name
Kelurahan (KEL)	Rural/Village Name

Table 4 <Token Labeling Result>

Text	**Label**
Hujan	O
deras	O
mengakibatkan	O
banjir	O
di	O
sejumlah	O
titik	O
di	O
wilayah	O
Menteng	KEC
Jakarta	KAB
Pusat	KAB

Building NER Model

Model building process is conducting three steps which are as follow:

Initialization

Initiation process is conducted by creating properties file configuration as information regarding directory location of used training data, file name including extension of model, and parameters which are going to be used by the model as shown on Figure 2.

```
trainFile = ./data/dataset.tsv
serializeTo = ner-model-for-tesis.ser.gz
map = word=0,answer=1

useClassFeature=true
useWord=true
useNGrams=true
noMidNGrams=true
maxNGramLeng=6
usePrev=true
useNext=true
useSequences=true
usePrevSequences=true
maxLeft=1
useTypeSeqs=true
useTypeSeqs2=true
useTypeySequences=true
wordShape=chris2useLC
useDisjunctive=true
```

Figure 2 <Properties File Configuration>

Model Training

The training process of model is conducted by running command line to initiate maximum memory usage (mx), CRFClassifier model, and properties file which used as shown on Figure 3. Training result will become.ser.gz format that contains model that have been trained using dataset which have been created before.

Figure 3 <Training Model Command Line>

Evaluation

The model that has been built then proceeds to evaluation process by using testing data as shown on Figure 4 to obtain evaluation result in form of Confusion Matrix on each used location token.

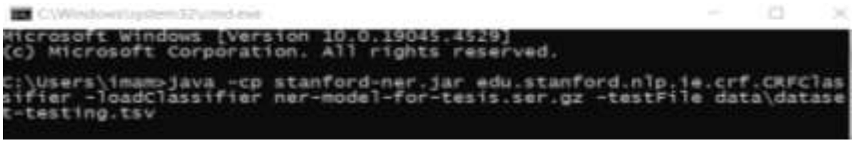

Figure 4 <Evaluation Model Command Line>

Geocoding

Geocoding is conducted to obtain location information in spatial data form which created upon obtained address from extraction process of NER. Geocoding results become a data visualization that taken form of map by using Follum library. Several steps conducted during geocoding process are as follows:

Spatial Query

Obtained address from NER's extraction went through spatial query using ArcGIS' API to obtain spatial data of map coordinate in form of latitude and longitude. Table 5 shows result sample of spatial query that was conducted using ArcGIS API. Obtained data then stored to database to proceed map visualization process as shown on Figure 5.

Table 5 **<**Spatial Query Process Result Sample**>**

Address	**Coordinate Data**
Menteng, DKI Jakarta	***Latitude*:** -6.199523 ***Longitude*:** 106.833134 **Address:** Menteng, Indonesia

text	address	arcgis_address	arcgis_lat	arcgis_lng
UPDATE! Malam Ini Gedebage Bandung Banjir Lagi, La...	Bandung	Bandung, Jawa Barat	-6.902163	107.619114
Banjir Rendam Ruas Tol Pondok Aren arah Serpong, T...	Pondok Aren	Pondok Aren, Banten	-6.263128	106.714039
cuaca hujan deras sekitar pademangan, lalu lintas ...	Mangga Dua	Mangga Dua, Sawah Besar, DKI Jakarta	-6.137490	106.826791
Kondisi terkini di pintu perlintasan kereta api (k...	Bandung	Bandung, Jawa Barat	-6.902163	107.619114
07.03 - Jalan Bekasi Raya Selepas pertigaan Harapa...	Menteng	Menteng, DKI Jakarta	-6.199523	106.833134
07.10 WibPagi ini jl raya sawangan arus lalu linta...	Depok	Depok, Jawa Barat	-6.397915	106.822081

Figure 5 <Geocoding Process Result Database>

Visualization

Figure 6 <Map Visualization Using Folium>

Stored data in database then went through visualization process by using Folium library. Data to map visualization result can be seen as shown on Figure 6. Generated map point represents information of location that was impacted by flooding in a certain area.

Results and Discussions

Result

Table 6 shows tweet percentage which has location information from whole tweets that were used as data.

Table 6 <Tweets that have Location Token Percentage>

	Data Counts	**Percentage**
Having Location Token	1,739	50,5%
Not-Having Location Token	1,702	49,5%
Total	3,441	100%

Based on the data shown in Table 6, it is clear that 50% of the tweets related to floods contain location tokens. This indicates that location information is commonly embedded in tweets about floods, making them ideal candidates for location extraction using Named Entity Recognition (NER) models. The ability to extract location data from these tweets is crucial, as it helps to map out affected areas and understand the geographical spread of flood-related events. Several studies have emphasized the importance of location-based information in social media data, particularly in the context of natural disasters, where real-time location tracking can significantly aid in disaster management and response (Cui et al., 2021; Khan et al., 2023). NER models have been widely used to extract geospatial information from text, and this study highlights their effectiveness in flood-related tweets (Hou et al., 2024).

However, as illustrated in Table 7, a portion of the tweets with location information does not match the actual location, which suggests discrepancies between the extracted location token and the true geographic data. To validate the extracted locations, the study compares the location tokens from the tweets with the actual location entities derived from the ArcGIS API. This comparison revealed that seven tweets contained location information that did not correspond with the actual location. These mismatches could be attributed to factors such as incorrect geotagging, colloquial language, or the use of location names that are not recognized by the geocoding process. Similar challenges have been identified in other studies on geospatial data extraction from social media, where location names might be ambiguous or incorrectly recognized by algorithms (Nikkanen & Räsänen., 2023; Feng et al., 2022).

In addition to the location token extraction process, the study implements a geocoding step to further refine the location data. This process is limited to the Indonesian region, meaning that any location mentioned outside of Indonesia is considered a mismatch. The geocoding method used helps to ensure that only relevant location data is retained, further enhancing the dataset's accuracy. Despite the 7 mismatches identified, the overall quality of the dataset is considered excellent, as nearly all the locations extracted from the tweets match the actual locations. The success of the geocoding process aligns with findings from previous research that underscores the importance of location validation in the context of disaster-related social media data (Sathianarayaman et al., 2024; Chen & Lim., 2021). Overall, the results suggest that the NER model, in combination with the geocoding process, provides reliable location information, which can be used effectively for disaster response and analysis.

Table 7 <Location Correspondency Percentage>

	Data Counts	**Percentage**
Location Matches	1732/1739	99.5%
Location does not Match	7/1739	0.5%

Table 7 shows the percentage of location correspondence in the dataset used, with 99.5% of the data showing matching locations, while only 0.5% does not match. This indicates that the vast majority of location entities recorded in the dataset are highly accurate. In Named Entity Recognition (NER)-based research, data quality is a crucial factor in determining the model's success. This result suggests that the NER model used was able to identify location entities very accurately, supporting the goal of extracting relevant information from the text.

Table 8, which presents the evaluation results of each location token obtained from the NER model and converted into a Confusion Matrix, provides a more detailed picture of the model’s performance. A Confusion Matrix is a common evaluation tool used to measure the accuracy of classification predictions in machine learning. By measuring true positives, false positives, true negatives, and false negatives, a Confusion Matrix allows researchers to gain deeper insights into how well the model is identifying the correct locations. The evaluation results show that the dataset quality is excellent because nearly all of the data demonstrates a high level of location matching.

Based on previous research, several studies also show that NER models trained with high-quality datasets tend to yield more accurate results. For instance, in a study by Deekshith., (2021) discussing the use of digital technology in marketing, data processing accuracy heavily depends on the quality of the data used for training the model. This is also reflected in another study by Hair & Sarstedt., (2021) which states that data selection and cleaning are crucial steps in improving AI system performance. The accuracy of an NER model in recognizing locations can contribute significantly to applications that require automated text processing, such as geographic data analysis or market research.

Overall, the results shown in Table 7 and Table 8 strongly support the conclusion that the dataset used is of very high quality, and the applied NER model has high accuracy in identifying location entities. This shows that the data processing in this research successfully ensured that most of the data was correctly identified, which is crucial for the success of further analysis and practical applications in related fields.

Table 8 <NER Model Evaluation Result>

Token	**Acc**	**Prec**	**Recall**	**F1**	**TP**	**TN**	**FP**	**FN**
PROP	0,87	0,97	0,93	0,95	3483	96	262	268
KAB	0,82	0,96	0.90	0.93	9282	332	981	1031
KEC	0,46	0,66	0,53	0,59	405	202	352	358
KEL	0,41	0,96	0,57	0,71	1208	50	911	911
Total	**0,75**	**0,95**	**0,85**	**0,90**	**14378**	**680**	**2506**	**2568**

Table 8 presents the evaluation results of the NER model, including key metrics such as accuracy (Acc), precision (Prec), recall (Recall), F1-score (F1), true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for different tokens: PROP (Province), KAB (District), KEC (Subdistrict), and KEL (Village). The overall accuracy of the model is 75%, which indicates a relatively strong performance, especially considering the complexity of natural language processing tasks involving location identification.

For the PROP (Province) and KAB (District) tokens, the model performs exceptionally well, achieving accuracy scores of 87% and 82%, respectively. This result suggests that the NER model is particularly effective in identifying higher-level geographic locations such as provinces and districts. The precision and recall for these tokens are also high, with F1-scores of 0.95 and 0.93, respectively. This indicates that the model is very good at both correctly identifying these locations (precision) and ensuring it captures almost all instances where these locations appear (recall).

On the other hand, the KEC (Subdistrict) and KEL (Village) tokens show lower performance, with accuracy scores of 46% and 41%, respectively. These lower scores can be attributed to the lack of sufficient information related to subdistricts and villages in the tweet data used for training the model. As the dataset may have had limited references to these more granular locations, the model struggled to correctly identify them, leading to a higher number of false positives and false negatives.

The confusion matrix results further highlight the challenges with KEC and KEL tokens. The number of false positives (FP) and false negatives (FN) for these tokens is notably high, which suggests that the model either mistakenly identified locations or failed to detect actual location mentions. This is a common issue in NER tasks, especially when the dataset is imbalanced or lacks adequate representation of certain entities.

In terms of practical applications, the NER model can be highly useful in geographic data analysis, such as extracting location-based insights from social media or news articles. However, the model's limitations for identifying subdistricts and villages suggest that additional data or enhanced model training could improve its accuracy in these areas. For instance, incorporating more localized data or applying more advanced techniques such as fine-tuning or augmenting the dataset could boost performance for KEC and KEL tokens.

Furthermore, the model's high performance for broader geographic categories like provinces and districts can be valuable in applications such as regional marketing, disaster response, and public health research, where identifying larger administrative divisions is key. The accuracy for these tokens ensures that the model can still provide meaningful location-based analysis in many practical use cases, even if finer-grained location recognition needs further improvement.

Discussion

This part demonstrates visualization result in a form of map by using Folium library based on latitude and longitude coordinate which were obtained from ArcGIS API.

Figure 7 <Map Visualization Using Folium>

The integration of geographical data and real-time social media updates, such as tweets, offers an innovative approach for mapping the impact of flooding, especially in urban areas like Central Jakarta. As shown in Figure 7, each tweet in the database is linked to specific latitude and longitude coordinates, creating map points that reflect both the location and the intensity of the flood's impact. This method not only offers a dynamic representation of flooding events but also helps authorities understand the real-time effects of natural disasters on urban infrastructure and daily life.

Previous studies have demonstrated the utility of social media in disaster response, particularly in mapping the impacts of flooding. For example, research by Mata et al.., (2023) emphasized the value of customer orientation in understanding how public responses to emergencies, such as floods, could improve service delivery and response times. Similarly, Okonkwo & Awad., (2023) explored how digital tools, including social media, can enhance business marketing strategies, highlighting the importance of real-time data in crisis management. In the case of Jakarta, the classification of flooding impacts into categories like road closures (green), traffic jams (blue), and facility damages (red) adds a layer of clarity to the map, enabling city planners and emergency responders to prioritize actions based on the severity and location of the impacts (Ferreira et al., 2024).

By color-coding these impacts, the map becomes not only an informative tool for understanding the spread of the flood but also a resource for efficient disaster response. For instance, road closures, represented in green, can be immediately flagged for detours or traffic management interventions, while blue areas signifying traffic jams can direct efforts to clear blocked routes. The red spots, which indicate damaged facilities, highlight where repair crews are most needed. This detailed categorization aligns with findings from Caliskan et al., (2020), who examined digital marketing strategies and noted the effectiveness of using targeted, real-time data to adapt to changing conditions. Thus, the mapping system shown in Figure 7 offers more than just geographical data it serves as a critical tool for understanding and managing the aftermath of floods, contributing to a more resilient and responsive urban environment.

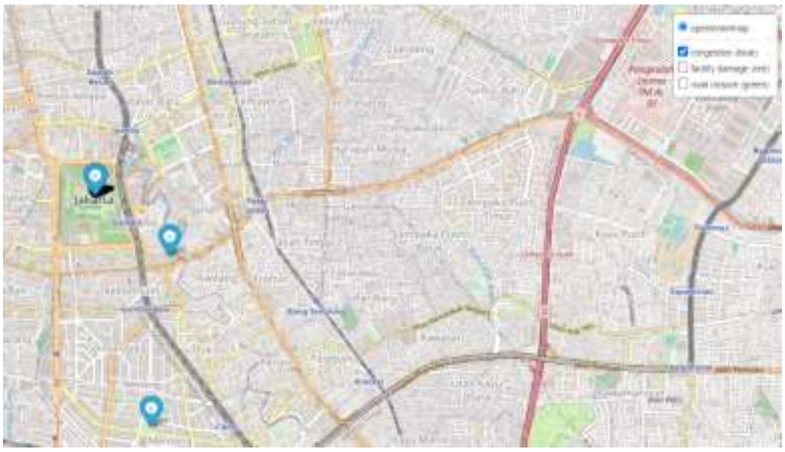

Figure 8 <Filter Based by Category>

Figure 9 <Detailed Information of Map Point>

In the context of displaying geospatial data on maps, particularly when dealing with a large dataset, performance and clarity are critical considerations. For instance, when there are 1,732 tweets with associated map points, showing all of them at once can overwhelm the map's rendering system, leading to performance issues. To address this, filtering methods are commonly employed to reduce the number of points shown at any given time. This not only enhances the map's performance but also improves the overall user experience by preventing clutter. In the case of the tweets related to flooding, one potential solution is to filter the map points by category, based on the impact caused by the flood. This can simplify the visualization by grouping similar events and making it easier for users to understand the data at a glance (Salamkar., 2024).

Research has demonstrated that effective data filtering techniques can significantly improve map interactivity, particularly in scenarios with large volumes of data. For instance, a study by Alam et al., (2022) discusses how spatial data clustering, based on event types or impacts, can streamline the presentation of data on interactive maps. By doing so, users can focus on specific categories without being overwhelmed by the sheer number of data points. In the context of flood-related tweets, categorizing tweets based on their impact such as "damaged infrastructure," "displaced people," or "water levels" can provide more meaningful insights and allow for more targeted responses (Iparraguirre-Villanueva et al., 2023).

Furthermore, providing detailed information for each map point adds value to the visualization, as it allows users to access deeper context on demand. As users interact with the map, they can click on specific points to bring up a pop-up that contains additional details, such as the category of impact and the tweet content. This level of detail facilitates manual validation processes, where human users can cross-check the tweet's content against the corresponding location. According to previous studies, such as the one by (Mansourian & Oucheikh., 2024), combining map-based visualization with detailed textual data helps in verifying the accuracy of geospatially-tagged information, particularly in disaster-related contexts. These verification processes are crucial for ensuring that the data accurately reflects the on-ground situation, enabling better decision-making in disaster management and response efforts (Tekumalla & Banda, 2023).

Conclusions

Based on the research results, it shows that location extraction from flood-related tweets using the Named Entity Recognition (NER) model and geocoding produces very useful and accurate data. About 50% of the flood-related tweets included location tokens, which shows the importance of geographic information in understanding the impact of disasters. The location extraction process with the NER model proved effective, although there were some discrepancies between the extracted location tokens and the actual geographic data, especially at the more detailed location level. However, the evaluation results show that 99.5% of the extracted locations correspond to valid locations, especially in the Indonesian region. This shows that the use of NER and geocoding models is highly effective in analyzing flood impacts and provides significant benefits in disaster management as well as geospatial analysis based on social media data.

References

Alam, M. M., Torgo, L., & Bifet, A. (2022). A survey on spatio-temporal data analytics systems. *ACM Computing Surveys,* 54(10s), 1–38. https://doi.org/10.1145/3507904

Caliskan, A., Özkan Özen, Y. D., & Ozturkoglu, Y. (2020). Digital transformation of traditional marketing business model in new industry era. *Journal of Enterprise Information Management*, 34(4), 1237-1261. https://doi.org/10.1108/JEIM-01-2020-0033

Chen, Z., & Lim, S. (2021). Social media data-based typhoon disaster assessment. *International Journal of Disaster Risk Reduction,* 64, 102482. https://doi.org/10.1016/j.ijdrr.2021.102482

Cui, P., Peng, J., Shi, P., Tang, H., Ouyang, C., Zou, Q., Liu, L., Li, C., & Lei, Y. (2021). Scientific challenges of research on natural hazards and disaster risk. *Geography and Sustainability,* 2(3), 216–223. https://doi.org/10.1016/j.geosus.2021.09.001

Cui, L., Wu, Y., Liu, J., Yang, S., & Zhang, Y. (2021). Template-based named entity recognition using BART. *ArXiv Preprint ArXiv:2106.01760*. https://doi.org/10.48550/arXiv.2106.01760

Dartanto, T. (2022). Natural disasters, mitigation and household welfare in Indonesia: Evidence from a large-scale longitudinal survey. *Cogent Economics & Finance,* 10(1). https://doi.org/10.1080/23322039.2022.2037250

Deekshith, A. (2021). Data engineering for AI: Optimizing data quality and accessibility for machine learning models. *International Journal of Management Education for Sustainable Development,* 4(4).

Fariz, T. R., Suhardono, S., & Verdiana, S. (2021). Pemanfaatan Data Twitter Dalam Penanggulangan Bencana Banjir dan Longsor. *CogITo Smart Journal*, *7*(1), 135–147. https://doi.org/10.31154/cogito.v7i1.305.135-147

Feng, Y., Huang, X., & Sester, M. (2022). Extraction and analysis of natural disaster-related VGI from social media: review, opportunities and challenges. *International Journal of Geographical Information Science*, 36(7), 1275–1316. https://doi.org/10.1080/13658816.2022.2048835

Ferreira, M. A., Oliveira, C. S., & Francisco, R. (2024). Tsunami risk mitigation: The role of evacuation routes, preparedness, and urban planning. *Natural Hazards.* https://doi.org/10.1007/s11069-024-04917-7

Girsang, A. S., Isa, S. M., & Fajar, R. (2021). Implementation of a geocoding in journalist social media monitoring system. *International Journal of Engineering Trends and Technology*, *69*(12), 103–113.

Govindarajan, S., Mustafa, M. A., Kiyosov, S., Duong, N. D., Raju, M. N., & Gola, K. K. (2023). RETRACTED: An optimization based feature extraction and machine learning techniques for named entity identification. *Optik,* 272, 170348. https://doi.org/10.1016/j.ijleo.2022.170348

Hair, J. F., Jr., & Sarstedt, M. (2021). Data, measurement, and causal inferences in machine learning: Opportunities and challenges for marketing. *Journal of Marketing Theory and Practice,* 29(1), 65-77. https://doi.org/10.1080/10696679.2020.1860683

Havas, C., & Resch, B. (2021). Portability of semantic and spatial–temporal machine learning methods to analyse social media for near-real-time disaster monitoring. *Natural Hazards,* 108, 2939–2969. https://doi.org/10.1007/s11069-021-04679-9

Hou, H., Shen, L., Jia, J., & Xu, Z. (2024). An integrated framework for flood disaster information extraction and analysis leveraging social media data: A case study of the Shouguang flood in China. *Science of The Total Environment,* 949, 174948. https://doi.org/10.1016/j.scitotenv.2024.174948

Iparraguirre-Villanueva, O., Melgarejo-Graciano, M., Castro-Leon, G., Olaya-Cotera, S., Ruiz-Alvarado, J., Epifanía-Huerta, Á., Cabanillas-Carbonell, M., & Zapata-Paulini, J. (2023). Classification of tweets related to natural disasters using machine learning algorithms. *International Journal of Interactive Mobile Technologies (iJIM),* 17(14). https://doi.org/10.3991/ijim.v17i14.39907

Khan, M. T. I., Anwar, S., Sarkodie, S. A., Yaseen, M. R., Nadeem, A. M., & Ali, Q. (2023). Natural disasters, resilience-building, and risk: Achieving sustainable cities and human settlements. *Natural Hazards*, 118, 611–640. https://doi.org/10.1007/s11069-023-05839-1

Kocerka, J., Krześlak, M., & Gałuszka, A. (2022). Ontology extraction from software requirements using named-entity recognition. *Advances in Science and Technology. Research Journal*, *16*(3). https://doi.org/10.12913/22998624/149941

Lai, K., Porter, J. R., Amodeo, M., Miller, D., Marston, M., & Armal, S. (2022). A natural language processing approach to understanding context in the extraction and geocoding of historical floods, storms, and adaptation measures. *Information Processing & Management*, *59*(1), 102735. https://doi.org/10.1016/j.ipm.2021.102735

Lawu, B. L., Lim, F., Susilo, A., & Surantha, N. (2021). Social media data crowdsourcing as a new stream for environmental planning & monitoring: A review. *IOP Conference Series: Earth and Environmental Science*, *729*(1), 012013. https://doi.org/10.1088/1755-1315/729/1/012013

Mansourian, A., & Oucheikh, R. (2024). ChatGeoAI: Enabling geospatial analysis for public through natural language, with large language models. *ISPRS International Journal of Geo-Information*, 13(10), 348. https://doi.org/10.3390/ijgi13100348

Mata, P., Cullano, R. A., Tiu, A. M., Gonzales, G., Selerio Jr., E., Maturan, F., Evangelista, S. S., Burdeos, A., Yamagishi, K., & Ocampo, L. (2023). Public satisfaction with the government's disaster response during Typhoon Odette (Rai). *International Journal of Disaster Risk Reduction,* 84, 103483. https://doi.org/10.1016/j.ijdrr.2022.103483

Mitchell, S. M., & Pizzi, E. (2021). Natural disasters, forced migration, and conflict: The importance of government policy responses. *International Studies Review,* 23(3), 580–604. https://doi.org/10.1093/isr/viaa058

Mollick, T., Azam, M. G., & Karim, S. (2023). Geospatial-based machine learning techniques for land use and land cover mapping using a high-resolution unmanned aerial vehicle image. *Remote Sensing Applications: Society and Environment,* 29, 100859. https://doi.org/10.1016/j.rsase.2022.100859

Motta, M., de Castro Neto, M., & Sarmento, P. (2021). A mixed approach for urban flood prediction using Machine Learning and GIS. *International Journal of Disaster Risk Reduction,* 56, 102154. https://doi.org/10.1016/j.ijdrr.2021.102154

Nikkanen, M., & Räsänen, A. (2023). Spatial data, methods, and mismatches for adaptive governance research. In Political Science and Public Policy 2023 (pp. 99–114). *Edward Elgar Publishing*. https://doi.org/10.4337/9781800888241.00016

Okonkwo, I., & Awad, H. A. (2023). The role of social media in enhancing communication and collaboration in business. *Journal of Digital Marketing and Communication,* 3(1), 19-27. https://doi.org/10.53623/jdmc.v3i1.247

Plyushteva, A., & Schwanen, T. (2022). “We usually have a bit of flood once a week”: conceptualising the infrastructural rhythms of urban floods in Malate, Manila. *Urban Geography*, 44(8), 1565–1583. https://doi.org/10.1080/02723638.2022.2105003

Putra, P. K., Mahendra, R., & Budi, I. (2022). Traffic and road conditions monitoring system using extracted information from Twitter. *Journal of Big Data*, *9*(1), 65. https://doi.org/10.1186/s40537-022-00621-3

Salamkar, M. A. (2024). Data visualization: AI-enhanced visualization tools to better interpret complex data patterns. *Journal of Bioinformatics and Artificial Intelligence,* 4(1), 204–226.

Sathianarayanan, M., Hsu, P.-H., & Chang, C.-C. (2024). Extracting disaster location identification from social media images using deep learning. *International Journal of Disaster Risk Reduction*, 104, 104352. https://doi.org/10.1016/j.ijdrr.2024.104352

Sitorus, M. E. J., Nababan, D., & Bangun, H. A. (2023). Dampak Bencana Banjir Terhadap Kesehatan Masyarakat Siatas Barita. *Tour Abdimas Journal*, *2*(2), 54–59. https://tourjurnal.akupuntour.com/index.php/tourabdimasjournal/article/view/78

Splendiani, S., & Capriello, A. (2022). Crisis communication, social media and natural disasters – the use of Twitter by local governments during the 2016 Italian earthquake. *Corporate Communications: An International Journal*. https://doi.org/10.1108/CCIJ-11-2021-0201

Tekumalla, R., & Banda, J. M. (2022). TweetDIS: A large Twitter dataset for natural disasters built using weak supervision. *In Proceedings of the 2022 IEEE International Conference on Big Data (Big Data)* (pp. 4816–4823). IEEE. https://doi.org/10.1109/BigData55660.2022.10020214

Ye, X., Du, J., Gong, X., Na, S., Li, W., & Kudva, S. (2021). Geospatial and semantic mapping platform for massive COVID-19 scientific publication search. *Journal of Geovisualization and Spatial Analysis*, *5*, 1–12. https://doi.org/10.1007/s41651-021-00073-y