Named Entity Recognition Using Conditional Random Fields for Flood Detection In Gerbang Kertosusila Based Twitter Data

[1]Ikrimatul Ulumiyyah, [2]Dwi Rolliawati, [3]Andik Izzuddin,
[4]Khalid, [5]Anang Khunaefi, [6]Mujib Ridwan
[123456]Department of Information System, Universitas Islam Negeri Sunan Ampel Surabaya, Indonesia
Email: [1]Ikrimatul@gmail.com, [2] dwi_roll@uinsa.ac.id, [3] andik@uinsa.ac.id,
[4] khalid@uinsa.ac.id, [5] kunaefi@uinsa.ac.id, [6] mujibrw@uinsa.ac.id

Article Info

Article history:
Received Dec 6th, 2023
Revised Mar 7th, 2024
Accepted May 9th, 2024

Keyword:
Conditional Random Fields
Flood Detection
Gerbang Ketosusila
Natural Language Processing
Named Entity Recognition

ABSTRACT

The national strategic area Gerbang Kertosusila East Java should be aware of floods. One of the existing efforts is to place flood sensors at several flood-prone points. However, that way is constrained by the need for more equipment to handle the many needy areas. So it is necessary to develop technology for the dissemination of flood information. Dissemination of flood information was quickly obtained from social media Twitter. One way is to use Twitter's text data source for a Named Entity Recognition model to help detect flood events and their locations. The Named Entity Recognition (NER) model was constructed using the Conditional Random Fields (CRFs) method to achieve research objectives. This research adds slang word handling at the preprocessing stage to improve model performance and the use of the BIO format in the labeling process and POS Tagging in the Feature Extraction process. Evaluation results with five Kfolds, 80% training data, and 20% test data show that the NER CRFs model performs excellently with a Precision of 0.981, Recall of 0.926, and f-measure of 0.950 so that these results can help the community and government regarding the information on the distribution of floods.

Corresponding Author:
Ikrimatul Ulumiyyah,
Department of Information System,
Universitas Islam Negeri Sunan Ampel Surabaya,
Ahmad Yani Street No.117, Jemur Wonosari, Wonocolo Subdistrict, Surabaya, East Java, Indonesia
Email: Ikrimatul@gmail.com

DOI: http://dx.doi.org/10.24014/ijaidm.v7i2. 27062

1. INTRODUCTION

Floods are hydrometeorological natural disasters that often occur in Indonesia and globally. Usually, rainfall and negligence in the process of disposing of garbage are the causes of flooding. Due to intense rainfall and long duration, several Indonesian regions like the national strategic areas Gerbang Kertosusila (Gresik, Bangkalan, Mojokerto, Surabaya, Sidoarjo, and Lamongan) in East Java Province urged to be aware of the hydrometeorological disaster. Gerbang Kertosusila has a tropical climate with an average temperature of 28.5%, humidity average of 75%, and rainfall low with an average of 1,290.50 mm per year and the trends patterns of extreme rainfall on Java Island, Indonesia, from 2009 to 2021, the rainy season from November to May, with January and February being the wettest months [1]. However, the BNBP East Java stated that East Java recorded as many as 65 disaster events throughout the early years of January 2021 via CNN Indonesia reports. Forty-nine flood events dominated the disasters that occurred. Floods negatively impact the environment, leading to a decline in the economic sector [2]. In addition, the flood affects public health, including increased diseases and human psychology. The impact of flooding, if not treated immediately, will increase the loss of society and government. Many efforts have been made for the community to make it easier to detect floods. The other uses the conventional method, like placing a flood detection sensor. However, their devices could be more robust and sufficient to handle many disadvantaged areas [3]. Besides, the way tends to be very expensive. So a breakthrough is needed for new information so that flood information and handling can be faster. That is with using Twitter social media.

Twitter deserves popular microblogging. The fact of Twitter has 353 million users worldwide, with more than 500 million daily tweets. Another fact is that the We are social and Hootsuite reports reveal that the number of Twitter users in Indonesia reached 14.05 million. Twitter is used not only for exchanging information but also for vent events, protests, educational media, and campaigns [4]. That makes many user influencers choose Twitter to convey news of an incident. Such as events, music concerts, traffic, earthquakes, and floods. Related research allows Twitter as a source of research such as search technology, analyzing sentiment topics, measurement of user validity, information traffic mapping, and disaster observation [4], [5].

The Text Mining approach uses Natural Language Processing (NLP) for extracting and filtering information relevant to flood. An essential process of extracting information in the NLP technique is Named Entity Recognition (NER). NER's purpose is to identify entities' names into structured label groups [6]. Several research approaches to studying NER are categorized into rule-based, machine learning, and deep learning[7], [8]. Previous research studies NER implemented for extracting traffic incident location information rule-based approach [9], [10]. Conditional Random Fields (CRFs) are a probabilistic approach for segmentation, label data sequences, such as sequences, trees, or grids. CRFs used to calculate the conditional probability value at the output node designed with a designated input node. Then the sequence probability model is used to detect entities automatically. Like research that has been done in the introduction of named entities in Indonesian texts by Jaariyah[11]. Jaariyah modeled NER using CRFs by providing the best accuracy rate of 87.06%. Continued on Research on introducing Indonesian-language Twitter data entities by Munarko[12]. Monarko performs model testing on three test data, formal, informal, and mixed tweets, and finds the high precision value for testing on all test data. Nevertheless, the testing process depends on selecting the suitable model appropriate to the test data to get high precision. From some of these studies, it can be seen that advantages of the CRFs method. CRFs can determine the number of features required to build a model. In contrast, the HiddenMarkov Model is local, and each word only depends on the current label with every previous label [13]. With these capabilities, the CRFs method can overcome the high assumptions in the Hidden Markov Model method. Most of the research that successfully implemented the implementation of NER using CRFs requires features vector (to simplify the feature extraction process) such as Language features, POS Tagging [14], Morphological Analyze, Gazetteers, and NE annotated corpus[12].

The difference in this study is the use of post tagging as a feature vector in the feature extraction process of NER models using CRFs. In addition, this study also specifically added the stage of handling kata slang. This is different from previous studies that have examined the topic of Introducing Named Entities related to flooding or related to Twitter. One of the reasons for Twitter's use of data sources in flood event detection applications is to save costs compared to conventional methods. This is supported by some references[2], [9]. Then, the performance of using Twitter-based Indonesian NER is still a challenge in the development of Natural Language Processing research. Its performance also still depends on the algorithms and training data used, as found in some references. [9], [11], [12].

In addition, the use of CRFs algorithms in this study requires feature vectors for the feature extraction process, which has been discussed in references [10], [11], [12]. Based on this information, this study examines the results of using post tagging as a feature vector in the feature extraction process of NER models using CRFs. In addition, this study also added a specific stage of handling kata slang. A full description of the other steps will be given in the next section.

Based on the problems and approach above, the research aims to explore how to build a model NER using Conditional Random Fields Method for the detection of flood incident at Gerbang Kertosusila with Twitter data. Then it offers a solution to make it easier for people to know about flood incidents in the Gerbang Kertosusila. NER techniques assist the entity detection process named in the form of the event's occurrence and the location where the event occurred in Tweets. The method used is CRFs and Feature Vector Post Tagging. The previous use of Conditional Random Fields was conducted by Khodra [8] who used POS Tagging on word class labeling as a feature vector to improve the performance of entity recognition model.

2. RESEARCH METHOD

This section presents the steps adopted in pursuing this research. It starts by describing why the existing natural disasters are devoted to researching tweets about floods, then the reasons why social media is widely used as a source of data, especially Twitter. Then data collection methods will be presented, naming the data collection process adopted in this work. Later, the various techniques of analyzing data will be discussed, highlighting the sources of this study relied on in this procedure. Figure 1 will highlight how the research flows.

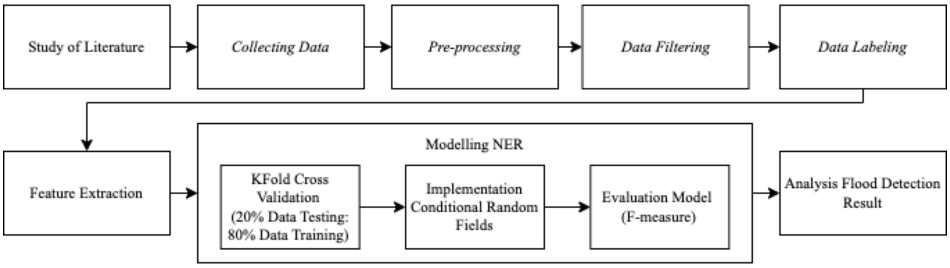
Study of Literature
Collecting Data
Pre-processing
Data Filtering
Data Labeling
Feature Extraction
Modelling NER
KFold Cross Validation (20% Data Testing: 80% Data Training)
Implementation Conditional Random Fields
Evaluation Model (F-measure)
Analysis Flood Detection Result

Figure 1. Flow Research

The Detection of flood tweets with NER using Conditional Random Fields contains the following steps: study literature, data collecting, preprocessing, filtering, labeling, feature extraction, build model then analysis of the result of flood detection. The references in study of literature used meet the topic of Named entity recognition which is related to flooding or related to Twitter. of course these references come from books, NLP journals including text mining, ner and CRFs. From these references the following information can be obtained:

1. One of the reasons for using Twitter data sources in flood event detection applications is to be more economical compared to conventional methods.
2. The performance of using Indonesian-language NER, especially Twitter-based, continues to be a challenge in the development of Natural Language Processing research and still depends on the algorithms and training data used.
3. Using the CRFs algorithm requires a feature vector for the feature extraction process [8], [12], [14].

Based on this information, this study reviews the research results to use post tagging as a feature vector in the feature extraction process on the model NER uses CRFs. and specifically, in this study added the slang word handle stage. A full explanation of other steps will be explained in the next section.

2.1. Material

The material data collected is sourced from Twitter. Twitter data has been collected as tweets related to information on flood events at the Gerbang Kertosusila. Data retrieval on Twitter is done by a web scraping technique with Python programming language. This study uses the available tools, namely Twint. Twint can extract all tweets with a particular keyword in a specific period.

Data has been collected from 2016 to May 2022. Examples of tweets that have been collected are presented in Table 1. To get data like Table 3.1. This study uses the keyword "flood," aka "banjir," and combines city, district, and sub-district in the Gerbang Kertosusila. Such as “banjir surabaya”, "banjir wonocolo" “banjir lamongan”, “banjir gresik”, “banjir bangkalan", etc.

Table 1. Tweets data

Tweets (Indonesian Version)	Tweets (English Translation Version)
07.48: Banjir di Raya Juanda. Sebaiknya HINDARI jalur ini	07.48: Flooding in Raya Juanda. It is best to AVOID this line
#SSinfo Banjir di Kecamatan Kalitengah Lamongan ini melanda delapan desa dengan ketinggian air di jalan raya paling tinggi 50 cm	#SSinfo The flood in Kalitengah Lamongan District hit eight villages with the highest water level on the road being 50 cm
09.39: Kondisi banjir di Jalan Raya Morowudi Cerme Gresik, Senin (15/3/2021) pagi tadi. Havid pendengar SS via WhatsApp SS melaporkan, ketinggian air sekitar 50-80 cm. Beberapa sepeda motor yg nekat melintas mogok. Untuk mobil, diarahkan lewat Metatu Benjeng-Balongpanggang. (hm) https://t.co/AvLLwc5jXM"	09.39: Flood conditions on Jalan Raya Morowudi Cerme Gresik, Monday (15/3/2021) this morning. SS listener Havid via WhatsApp SS reported that the water height was around 50-80 cm. Some desperate motorbikes cross the strike. For cars, directed via Metatu Benjeng-Balongpanggang. (hm) https://t.co/AvLLwc5jXM"

The study flow is carried out in the data preprocessing, data filtering, data labeling, feature extraction, dividing the data into 2 (training and testing data), and Modelling NER (which includes the implementation process of the Conditional Random Fields method and model evaluation) than analyzing the results of flood detection.

2.1. Data Preprocessing

Preprocessing in this study is carried out to change the text data obtained from the previous process into structural data that is ready for training and testing. Data processing is done with the help of a library based on the Python programming language, namely NLTK. Preprocessing is carried out on the data as follows:

2.1.1. Cleaning

This cleaning process cleans words from characters that do not affect the entity recognition results. The omitted characters are numbers, punctuation marks, usernames, mentions, hashtags, symbols, “RT” text, “FAV” text, and website URLs.

2.1.2. Case folding

Case folding is a process that is applied to character sequences. Where the identified character is not capitalized, it is replaced with an uppercase equivalent of the character or vice versa, not lowercase, is replaced with a lowercase equivalent. So that all character letters have the same characteristics. In this study, we will change all uppercase characters to lowercase to make it easier to process words.

2.1.3. Tokenization

Tokenization is the process of separating text in the form of documents, paragraphs, and sentences into certain parts (tokens). The word separation limit depends on the characteristics of the data used. In word-based tokenizing, all words can be separated based on punctuation marks, space delimiters, and others. This research splits sentences in tweets into tokens. The space character in flooded tweets is used as a word delimiter at this stage.

2.1.4. Stopword

Stop words are commonly used and deliberately avoided to save space and data processing time. In addition, the considered terms contain little information in the text. This study uses Indonesian stop words collected by Owen. For example, “is”, “how,” “ie”, “with,” “which”, “and,” “in,” and etc. [15]. Where the target text tweet is a token, then compares the target with a list of stop words. If it is the same, the stop word is removed, and the process is repeated until the target item is reached.

2.1.5. Handle Slang Word

The cleaning process has been carried out to stopwords on the tweets data, but many words still use informal language, both slang and short forms of other words. One way to decipher short words and slang is by using external resources. That is why there are many slang or slang dictionaries. The Colloquial Indonesian Lexicon [16] and IndoCollex [17] bodies were used in this study.

2.1.6. Stemming

The stemming process is to find root words and root words. In contrast to English, the stemming process in Indonesian texts is done by removing suffixes and prefixes. The process uses the python literary stemming library, which has implemented the Nazief and Adriani algorithm [18].

2.2. Data Filtering

In this process, two stages are carried out: first, filtering by deleting the same tweet because Twitter has a retweet feature that causes the same number of tweets to be retweeted by other users and causes repeated text with the same topic. Then filter tweets manually to select and delete data tweets that do not match what is desired after data collection, during data 340abelling, and after data 340abelling.

Where filtering is done on tweets with the following characteristics: tweets that contain advertisements related to the use of the word flood, for example, “banjir hadiah” or “banjir gol,” tweets that do not include the location of the flood disaster, flood tweets that did not occur in the Gerbang Kertosusila Area and multiple tweets one of which contained incorrect location information.

2.3. Data Labelling

Tweets are labeled using the notation BIO (Begin, Inside, and Other), a labeling scenario that shows the order, which is then classified into three class categories: Begin-Tag, Inside-Tag, and Other. The process of labeling tweets is done by building a location and event dictionary. The dictionary is used for the mapping process between tokens with the location and event of the words. Then the process of determining the beginning and inside is carried out with the rules of B-event, I-event, B-location, I-location, and Other. Furthermore, the data will be used to perform NER modeling.

2.4. Feature Extraction

Feature extraction also performs the task of changing the data originally in text format to be numeric. Then the data will be processed in calculations using conditional random fields. To do this process requires a feature vector or function. The feature vectors used in this study are word parts, post tags, titles, and upper, lower, and nearby words.

For Pos Tag, This study uses the Indonesian Tag Post corpus, the research results by [19], [20]. With tab-separated value format. The Indonesian POS Tagging corpus consists of ten thousand Indonesian language sentences built from 256,683 lexical tokens and is designed to consist of 23-word classes. With POS tagging, conditional random fields will model the behavior of sequential input data and consider the previous context when making predictions.

2.6 Modeling

2.6.1. Preparation KFold

Distribution of Proportions of Training and Testing Datasets. This stage divides the proportions of training and testing datasets NER modeling. Distribution configuration is done using the Kfold Cross-Validation technique Field[21]. details of the Kfold Cross validation scenario in this study were carried out with the following process:

1. Divide the amount of data by five folds so that each fold consists of 20% of the dataset.
2. Each fold, for example, is K1 – K5. And set 1 proportion. (K1) as testing, and the rest as training data (K2 – K5). This means there is a proportion of 20% for testing data and 80% for training data.
3. Iterate processes a and b until all partition folds are tested.

So that it produces five folds - each partition consists of 392 data. The research uses a ratio of 80:20, so the training stage is carried out with 1568 data, and the testing stage is 392 data. So that iterations are carried out with the possibility of different testing data for each fold used.

2.6.2. Conditional Random Fields

CRF is used in many applications, including Natural Language Processing, Computer Vision, and bioinformatics. One form of CRF is the linear chain CRF shown in Figure 1.

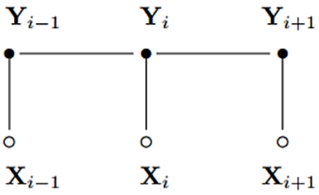

Figure 2. linier Chain CRF

Probabilistic for segmentation and labeling data sequences, such as sequences, trees or graphs. When we condition the graph on X globally, i.e., when the values of random variables in X are fixed or given, all the random variables in set Y follow the Markov property $p(Y_u/X, Y_v, u\neq v) = p(Y_u/X,Y_x, Y_u\sim Y_x)$, where $Y_u\sim Y_x$ signifies that Y_u and Y_x are neighbors in the graph. So it is written in equations (1) and (2) [22], [23], [24].

$$p(y|x) = \frac{1}{Z(x)}\prod_C \Psi_C(y_c, x) \quad (1)$$

$$Z(x) = \sum_c \prod_c \Psi_C(y_c, x) \quad (2)$$

Functional potential refers to the number of data labels that bind to features simultaneously. Features are usually referred to as data patterns in some CRF literature. In the simple case, the potential function is the exponential of the sum of the weights of all the feature functions. The expressions for the derivatives as well as the log-likelihood [25]. Actually, it has all that we need to code a CRF model. We can code the likelihood using the abovementioned equations, use belief propagation to calculate the marginals, work out the derivatives, and optimize the likelihood using off-the-shelf optimization algorithms like L-BFGS. Then the existing CRFSuite library with Skicit-Learn Wrapper for the research because of limited time [26].

2.6.3. Evaluation

Evaluation by measuring the accuracy of Named Entity Recognition performance goes well or not well. This research uses f-measure performance measurement. f-measure is generated from the recall and precision results in equation (5).

$$precision = \frac{true\ positif}{true\ positif + false\ positif} \quad (3)$$

$$recall\ = \frac{true\ positif}{true\ positif\ +\ false\ negatif} \tag{4}$$

$$f - measure\ = \frac{true\ positif\ +\ true\ negatif}{True\ Positif + False\ Positif + True\ Negatif + False\ Negatif} \tag{5}$$

For location and event entities that are correctly identified are counted as true positives, while those incorrectly are counted as false positives [27]. Meanwhile, other entities correctly identified are counted as true negatives, while those incorrectly identified are counted as false negatives. Then the precision calculation is performed using formula (3), and the recall value is calculated using formula (4). With the f-measure measurement(5), this research will determine whether the model's performance is going well.

2.7. Analysis of Flood Detection

After the model is made, the model detects the entity location of the flood event and describes the distribution of the location of the flood events in the Kertosusila Gate Area. They then presented the results of the evaluation and testing of the NER model from the model that has been made and the conclusions drawn from the flood detection results.

3. RESULTS AND ANALYSIS

3.1. Data Collection

It aims to ease researchers' work when filtering data, 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 tweet's content with Indonesian language characteristics and object data type. From the results of scraping tweets of floods in Gerbang Kertosusila on Twitter, 2408 tweets related to floods were obtained.

3.2. Preprocessing

As seen in table 1. Data that is still dirty and cannot be used as training and testing data for building the NER model. Therefore, before going to the next step, the data must be cleaned first, commonly known as data preprocessing. The actions include cleaning, case folding, tokenizing, stopwords, handling slang words, and stemming and filtering that shown at table 2. Starting from 2408, reduced to 1960 data ready to be labeled.

Table 2. Preprocessing Result

Tweet (Indonesia Languange)	Tweet (After Preprocessing)
Buduran Sidoarjo menuju Surabaya banjir plus macet	['buduran', 'sidoarjo', 'menuju', 'surabaya', 'banjir', 'plus', 'macet']
Radio ANDIKA 11.52 HINDARI jalur Trosobo arah Surabaya MACET. Imbas dari banjir setelah jembatan layang Trosobo....	['radio', 'andika', 'hindari', 'jalur', 'trosobo', 'arah', 'surabaya', 'macet', 'imbas', 'banjir', 'jembatan', 'layang', 'trosobo']
09.39: Kondisi banjir di Jalan Raya Morowudi Cerme Gresik, Senin (15/3/2021) pagi tadi. Havid pendengar SS via WhatsApp SS melaporkan, ketinggian air sekitar 50-80 cm. Beberapa sepeda motor yg nekat melintas mogok. Untuk mobil, diarahkan lewat Metatu-Benjeng-Balongpanggang. (hm)	['banjir', 'jalan', 'raya', 'morowudi', 'cerme', 'gresik', 'pagi', 'havid', 'dengar', 'suara surabaya', 'dari', 'whatsapp', 'suara surabaya', 'lapor', 'tinggi', 'air', 'cuma', 'sepeda', 'motor', 'nekat', 'lintas', 'mogok', 'mobil', 'arah', 'metatu', 'benjeng', 'balongpanggang']

3.2. Labelling

Tweets that have been processed will then be labeled using the BIO notation (Begin, Inside, and Other) as a labeling scheme. That indicates the sequence, which is then classified into five classes: B-event, I-event, B-location, and I-Location and Other. In this study, the researcher deliberately uses the O notation to complement sentences because it allows it to influence post-tags in feature extraction. Table 3 example of labeling results.

Table 3. Labelling Result

Token	Label
Banjir	B-event
Jalan	O
Morowudi	O
Kulon	O
Cerme	B-location
Gresik	I-location

Based on the sample results of the labeling of the BIO notation. Goes well according to the rules made in the dictionary and can recognize entities event in the word flood aka ”banjir”, the location entity for the subdistrict Cerme as B-location and I-location in Gresik. However, it cannot recognize street names where the dictionary is still limited, not yet covering the street location entity. So that it is known that 26679 words have been assigned the name of the entity of five classes. When The tweet labeling process is done, Then the data that has been labeled will be broken down into training data and testing data.

3.1. Modelling

3.1.1 Feature Extraction

In the post tag feature, there are 26679 words with the POS NN (Noun) having a total of 17438. Then followed by POS VB (Verb) with 3350. POS JJ (Adjective) occupies the third position with 1549, the last position in POS UH (Interjection) 3 words. With this information, it is known that the dataset has an unbalanced proportion in the class of noun words. Even so, there is no special treatment to overcome the unbalanced proportions in the NN word class. Because in Table 4, important keywords such as flood, rain, and Gresik are defined as POS NN and NNP. A total of 2160 words for flood, 389 for Gresik etc.

Table 4. Total Token in POS

Token		Total
Indonesian	English	
Banjir	Floods	2160
Gresik	Gresik	389
Jalan	Road/Street	324
Hujan	Rainfall	313
Desa	Village	278

Next is the feature extraction process to extract the features used, namely the word part, post tag, title, upper, lower, and nearby word. Then these features are converted into the scikit-learn library format. Where the Extraction Feature gives real number output even if it's only 0 or 1.

3.1.2. Model Training

The Conditional Random Fields training model uses kfold = 5. The scenario is carried out with a dataset proportion of 80% for the training model, namely 1568 data and 20% testing model with 392 data. In the sklearn crf wrapper, it will first conduct training data and send the results of the training data with the appropriate label. Research has built a CRFs model with the lbfgs algorithm configuration, parameters c1 and c2 each of 0.1. Then there is a maximum small dataset iteration is done 100 times with possible transition boolean true and verbose boolean true. With this configuration, the training model is completed with a kfold of 5 repetitions. So that it can complete the training model in a few seconds.

3.1.3. Model Testing

Evaluate the new model with precision testing scenarios, recall and f-measure on the data testing results of Kfold as much as 5 folds and consists of 392 tests data. matrix of test results based on named entity per fold partition. The report shows the precision of the precision, recall, and f-measure classification metrics on an entity basis The metric is calculated using positive is true and false, negative is true and false. Positive and negative in this case are the entity names for the class label being predicted. the average precision, recall, and f-measure values are shown in Table 5. Based on the average, the model has a Precision value of 0.981, Recall of 0.926, and f-measure of 0.950.

Table 5. Test Report

Fold	Precision	Recall	F-Measure
K1	0.969	0.894	0.925
K2	0.975	0.922	0.946
K3	0.990	0.956	0.972
K4	0.985	0.969	0.977
K5	0.983	0.899	0.929
Avg	0.981	0.926	0.950

3.1.4. Analysis Result of NER Flood Detection

With the results of these tests, the model can detect entity events with B-events and I-events and location entities labeled B-location and I-location of flood events. the distribution Of the location of the flood incident in the Gerbang Kertosusila.

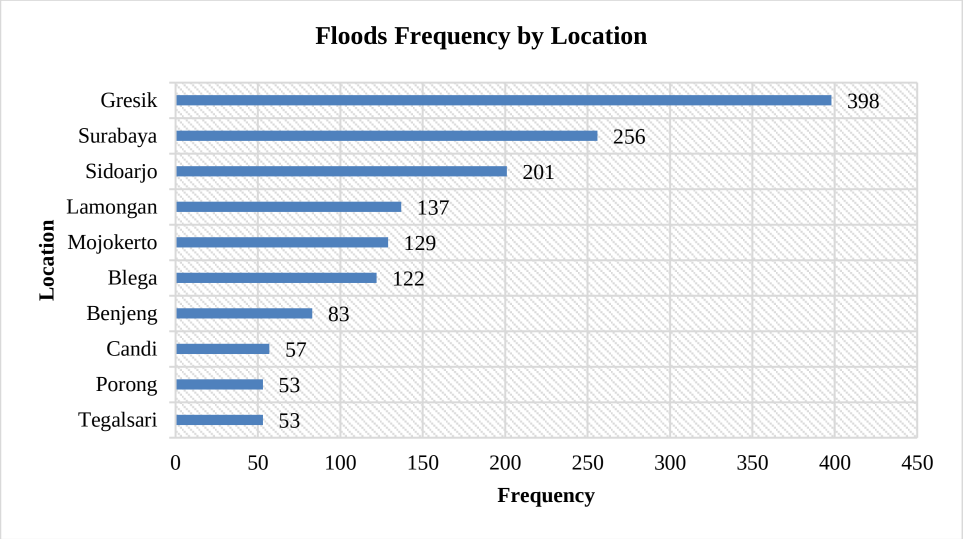

Figure 3. Location Distribution

Results in Figure 1 were obtained by calculating the frequency of flood events based on filtering tweets labeled B-location and I-location. Flood frequency distribution It is known that the location of Gresik is the location of the highest flood of 398, then Surabaya with as many as 262 and Sidoarjo with as many as 201. So this model can be applied to flood tweets.

3.5. Discussion

Khodra said research to identify an entity's name, place, time, info and other. The built model adds a rule-based filter module stage and an extraction feature module. The combination of the extraction feature module consists of a multi-token tokenization method, and POS tags feature sets in the CRFs settings [8]. The accuracy results obtained are 75%. Unlike Khodra, Yuda Munarko applies Named Entity research Recognition (NER) Indonesian from the Twitter dataset into three categories: informal, formal and mixed. Preprocessing is used in the research namely lowercase and tokenizing to identify the entities person, location, organization, and others. The research obtains precision and recall values respectively 87% and 62% for formal tweets, 90% and 36% for informal tweets, and 86% and 60% for mixed tweets [12].

This study uses the Kfold = 5 scenario in testing and giving an f-measure value with an average of 0.950. CRFs training is carried out by the lbgsf algorithm, and the combination of c1 = 0.1, and c2 = 0.1. Unlike previous research using standard preprocessing by munarko[12], this research adds slang word handles to address informal language or a slang words on a tweet. In the process of data, labeling is done with the BIO format on named entities. The added feature is the post tag. The model obtained by testing scenario ratio of 80% training data and 20% testing data gives a precision value of 0.981, recall of 0.926, and f-measure of 0.950.

The research findings revealed insights, indicating that a combination of value parameters c1 = 0.1 and c2 = 0.1 consistently delivers stable f-measure values. Additionally, it was observed that the composition selection during the preprocessing and feature extraction stages significantly impacts the precision, recall, and f-measure performance of the CRF model. Notably, the incorporation of slang word handling resulted in a noteworthy increase of 12% in precision and 33% in recall compared to the findings of study [12]. Despite consistently high precision values across all research tests [8], [12] it became evident that selecting the right composition and combination of models for the testing data led to improved performance. Furthermore, employing Kfold for cross-testing proved beneficial in enhancing the model's performance by ensuring equal proportions in the testing dataset.

The findings of this study reveal some interesting insights. First, the combination of parameter values c1 = 0.1 and c2 = 0.1 consistently yields a stable f-measurement value. In addition, the selection of composition during the preprocessing and feature extraction stages also had a significant impact on precision performance, acquisition, and measurement of f CRF models. There was a significant improvement in precision (by 12%) and memory (by 33%) by combining kata slang treatment. Although high precision scores are consistently high across all research tests, it is important to choose the right composition and combination of models for test data to improve performance.

In addition, this study also applied the Conditional Random Fields (CRF) method for flood event detection. The results showed that the built-in model successfully identified the entity's event and location in

BIO (Blocation, B-event, I-location, and I-event) formats. Using this model, it is possible to determine the distribution of flooding in the Ketosusila Gate area. The Gresik region is the most frequently flooded area, followed by Surabaya, Sidoarjo, and other areas.

Although this study has good results, there are some weaknesses that need to be considered. First, this study only focuses on the detection of flood events in the Ketosusila Gate area. More research involving other regions can provide a more comprehensive picture of the spread of floods in the region. In addition, this study also used kfold=5 scenarios for cross-testing. Using different scenarios or combining other testing methods can provide a better understanding of the model's performance.

For further research, there are several areas that can be explored. First, research can involve further analysis of the factors causing flooding in the Ketosusila Gate area as keywords. In addition, research can expand the scope of the area and compare the distribution of floods in different areas. In addition, research may also consider using methods other than CRF for flood event detection. By conducting further research in this case, it can provide more in-depth insight into flood management and mitigation efforts in the region.

In conclusion, the findings of this study provide an important insight into the effect of parameter combinations and the selection of composition in flood event detection using the CRF method. Although this study has drawbacks and there is room for further research, the constructed model shows good results with high accuracy values. By continuing this research, it can make a valuable contribution in understanding and handling floods in the Ketosusila Gate area and other areas. In subsequent studies, there were several areas to explore. First, research may consider the use of alternative algorithms or NLP semantic techniques in addition to CRF for flood event detection. Algorithms such as LSTM (Long Short-Term Memory) or Transformer, BERT (Bidirectional Encoder Representations from Transformers) can be an interesting choice for use in flood event detection models. Combining these algorithms with the CRF method can provide better results and can improve the performance of the detection model.

In addition, it is necessary to pay attention to the potential gap that may occur with the results of this study. Although the constructed model provides good results with a high degree of accuracy, there are still some possible gaps that need to be considered. One of them is the possibility of overfitting, where the model is only able to recognize flood events in the exercise data used in the study. To overcome this, subsequent studies can use more diverse datasets and conduct tests on data that have never been seen before. In addition, another gap that can be considered is the possibility of errors in data annotation. The data annotation process can affect the performance of the detection model, and errors or mismatches in the annotation can produce inaccurate results. Therefore, subsequent studies may consider the use of cross-evaluation testing methods or annotation verification methods to ensure the quality of the data used.

By considering the use of alternative algorithms or NLP semantic techniques and identifying potential gaps that may occur, Further research can contribute more broadly to the development of flood detection and better understanding of flood management in the Ketosusila Gate area and other areas.

4. CONCLUSION

Based on the results of the research and evaluation conducted. Implementing the Conditional Random Fields method for flood event detection has succeeded in recognizing entity events and locations in BIO format (Blocation, B-event, I-location, and I-event). With the model that has been built, the distribution of the occurrence of flooding in the Ketosusila Gate area is known. The Gresik area was the area that experienced the most frequent flooding, with 398, followed by Surabaya with 262, Sidoarjo with 201, and furthermore. Evaluation of the kfold = 5 scenarios, the model has given an average of above 90%, with a ratio of 20% testing and 80% training data. The scenario provides an average precision value of 0.981, recall of 0.926, and f-measure of 0.950. That means the model NER built is very good with a high accuracy value.

With this conclusion, it is important to address the suggestions related to the data used in the research. The current study has limitations due to the narrow scope of data, focusing only on flood events in the Gerbang Kertosusila area based on Twitter data. To enrich the research, it is advisable to expand the study area and incorporate a more comprehensive range of data sources. This can be achieved by combining various social media data sources or by extending the coverage to include the entire country of Indonesia. By doing so, a broader and more diverse dataset can be obtained, enhancing the validity and generalizability of the findings.

In addition to exploring alternative algorithms, there are several potential avenues for future research in the field of flood event detection. One area of focus could be the development of methods or techniques for extracting and recognizing the context or categorization of flood events from text data. This could involve the use of natural language processing techniques, such as semantic analysis or topic modeling, to identify and classify flood-related information in textual sources.

Furthermore, while Conditional Random Fields (CRF) have shown effectiveness in detecting entities, it is worth noting that they have limitations in recognizing the context or categorization of flood events. To overcome this limitation, it is recommended to explore other algorithms, such as BERT, that can provide better contextual understanding. Introducing BERT or similar models can help improve the accuracy and contextual awareness of flood event detection.

REFERENCES

[1] K. Aruna, Dr. M. V. Subramanian (RTD), Dr. B. Jaya sudha, and Bharathidasan university, "Studies on Seasonal variations of rainfall in java island at Indonesia," *J Algebr Stat*, vol. 13, no. 3, pp. 1481–1489, 2022.

[2] D. B. Baranowski *et al.*, "Social-media and newspaper reports reveal large-scale meteorological drivers of floods on Sumatra," *Nat Commun*, vol. 11, no. 1, pp. 1–10, 2020, doi: 10.1038/s41467-020-16171-2.

[3] I. Utami and M. Marzuki, "Analisis sistem informasi banjir berbasis media twitter," *Jurnal Fisika Unand*, vol. 9, no. 1, pp. 67–72, 2020.

[4] M. H. Awalludin, F. Teknik, U. K. Indonesia, and J. D. Bandung, "EVENT DETECTION PADA MICROBLOGGING TWITTER DENGAN METODE DENCLUE UNTUK PEMETAAN LOKASI BENCANA LONGSOR," *JBPTUNIKOMPP*, 2018, [Online]. Available: https://repository.unikom.ac.id/id/eprint/58405

[5] I. Utami and M. Marzuki, "Analisis sistem informasi banjir berbasis media twitter," *Jurnal Fisika Unand*, vol. 9, no. 1, pp. 67–72, 2020, [Online]. Available: http://jfu.fmipa.unand.ac.id/index.php/jfu/article/view/454

[6] E. Kapetanios, D. Tatar, and C. Sacarea, "Named Entity Recognition," *Natural Language Processing*, vol. 8, no. 2, pp. 309–322, 2013, doi: 10.1201/b15472-19.

[7] F. Béchet and B. Mohit, "Named Entity Recognition," *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pp. 257–290, 2011, doi: 10.1002/9781119992691.ch10.

[8] F. Muhammad and M. L. Khodra, "Event information extraction from Indonesian tweets using conditional random field," *ICAICTA 2015 - 2015 International Conference on Advanced Informatics: Concepts, Theory and Applications*, pp. 0–5, 2015, doi: 10.1109/ICAICTA.2015.7335383.

[9] M. Ermawati and J. L. Buliali, "Text Based Approach For Similar Traffic Incident Detection from Twitter," *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, vol. 9, no. 2, p. 63, 2018, doi: 10.24843/lkjiti.2018.v09.i02.p01.

[10] Y. Munarko, U. M. Malang, and Y. Munarko, "Ekstraksi Nama Lokasi Dari Tweets Informasi," *Seminar Teknologi dan Rekayasa (SENTRA)*, pp. 978–979, 2015.

[11] N. Jaariyah and E. Rainarli, "Conditional Random Fields Untuk Pengenalan Entitas Bernama Pada Teks Bahasa Indonesia," *Komputa : Jurnal Ilmiah Komputer dan Informatika*, vol. 6, no. 1, pp. 29–34, 2017, doi: 10.34010/komputa.v6i1.2474.

[12] Y. Munarko, M. S. Sutrisno, W. A. I. Mahardika, I. Nuryasin, and Y. Azhar, "Named entity recognition model for Indonesian tweet using CRF classifier," *IOP Conference Series: Materials Science and Engineering PAPER*, 2018, doi: 10.1088/1757-899X/403/1/012067.

[13] W. Ahmed, P. A. Bath, and G. Demartini, "USING TWITTER AS A DATA SOURCE: AN OVERVIEW OF ETHICAL, LEGAL, AND METHODOLOGICAL CHALLENGES," *Emerald Publishing Limited*, vol. 2, pp. 79–107, 2017, doi: https://doi.org/10.1108/S2398-601820180000002004.

[14] N. Patil, A. Patil, and B. V. Pawar, "Named Entity Recognition using Conditional Random Fields," *Procedia Comput Sci*, vol. 167, no. 2019, pp. 1181–1188, 2020, doi: 10.1016/j.procs.2020.03.431.

[15] L. Owen, "Indonesian Stopword Combined." [Online]. Available: https://github.com/louisowen6/NLP_bahasa_resources/blob/master/combined_stop_words.txt

[16] N. A. Salsabila, Y. Ardhito, W. Ali, A. Septiandri, and A. Jamal, "Colloquial Indonesian Lexicon," *2018 International Conference on Asian Language Processing (IALP)*, pp. 226–229, 2018.

[17] D. T. Wijaya, "IndoCollex : A Testbed for Morphological Transformation of Indonesian Colloquial Words," no. 2017, pp. 3170–3183, 2021.

[18] "Sastrawi · GitHub." Accessed: Jun. 22, 2022. [Online]. Available: https://github.com/sastrawi

[19] A. Dinakaramani, F. Rashel, A. Luthfi, and R. Manurung, "Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus," *Proceedings of the International Conference on Asian Language Processing 2014, IALP 2014*, pp. 66–69, 2014, doi: 10.1109/IALP.2014.6973519.

[20] Yudi Wibisono, "POS Tagger Bahasa Indonesia dengan Python – Blog Yudi Wibisono." Accessed: Jun. 22, 2022. [Online]. Available: https://yudiwbs.wordpress.com/2018/02/20/pos-tagger-bahasa-indonesia-dengan-pytho/

[21] L. Mardiana, D. Kusnandar, and N. Satyahadewi, "Analisis Diskriminan Dengan K Fold Cross Validation Untuk Klasifikasi Kualitas Air Di Kota Pontianak," *Bimaster : Buletin Ilmiah Matematika, Statistika dan Terapannya*, vol. 11, no. 1, pp. 97–102, 2022.

[22] R. Klinger, "Classical Probabilistic Models and Conditional Random Fields," *Entropy*, vol. 51, no. December, pp. 282–289, 2007.

[23] C. Sutton and A. McCallum, "An introduction to conditional random fields," *Foundations and Trends in Machine Learning*, vol. 4, no. 4, pp. 267–373, 2011, doi: 10.1561/2200000013.

[24] H. M. Wallach, "ScholarlyCommons Conditional Random Fields : An Introduction Conditional Random Fields : An Introduction," no. February, 2004.

[25] J. Suzuki, E. McDermott, and H. Isozaki, *Training Conditional Random Fields with Multivariate Evaluation Measures.* 2006. doi: 10.3115/1220175.1220203.

[26] N. Okazaki, "a fast implementation of Conditional Random Fields." 2007.

[27] D. J. Hand, P. Christen, and N. Kirielle, "F*: an interpretable transformation of the F-measure," *Mach Learn*, vol. 110, no. 3, pp. 451–456, 2021, doi: 10.1007/s10994-021-05964-1.

BIBLIOGRAPHY OF AUTHORS

Ikrimatul Ulumiyyah received the S.Kom. degree in information systems from UIN Sunan Ampel, Surabaya, Indonesia, in 2022. Recently she research interest has been data analytics, data processing, modeling machine learning, and data science. Besides that, she is also interested in text mining, and natural language processing. she can be contacted at email: ikrimatul@gmail.com

Dwi Rolliawati holds a M.T in Electrical Engineering from ITS, Surabaya, Indonesia. She has been a Lecturer in the Information Systems Study Program at UIN Sunan Ampel from 2014 until now. Besides that, she also serves as head of the information systems study program, where his research concentration is Computer Science, Modeling Simulation, Machine Learning. She can be contacted at email: dwi_roll@uinsa.ac.id

Andik Izzuddin Muhammad Andik Izzuddin is a lecturer at the UIN Sunan Ampel Surabaya in Indonesia. His research interests include computer network, information sytem, information technology and community based research. He can be contacted at email: andik@uinsa.ac.id

Khalid has been a Lecturer in the Information Systems Study Program at UIN Sunan Ampel from 2014 until now, where his research concentration is Data Mining, Natural Language Processing, Data Science, Text Mining, Machine Learning . he can be contacted at email: khalid@uinsa.ac.id

Anang Khunaefi received the B.C. and M.C. degree in informatics engineering from Sepuluh Nopember Institute of Technology, Surabaya, Indonesia, in 2004 and 2013, respectively. He received the Ph.D. degree in computer science and electrical engineering from Kumamoto University, Kumamoto, Japan in 2021. From 2003 to 2010, he was a software engineer focusing on the development of web-based application system using Java programming language, PHP, and Javascript for an IT company based in Jakarta, Indonesia. He wrote two books about the analysis and implementation of information system for mapping students' interest in Indonesia's educational institutions. His research interest includes software engineering, data-driven requirement, business process automation, semantic web service, and the implementation of information system for educational institutions. he can be contacted at email: kunaefi@uinsa.ac.id

Mujib Ridwan has been a Lecturer in the Information Systems Study Program at UIN Sunan Ampel, where his research concentration is technology information, machine learning, intelegent system. he can be contacted at email: mujibrw@uinsa.ac.id