Natural Disaster Monitoring Information System From Social Media Data Using Naïve Bayes Algorithm

1st Brilliant Friezka Aina
School of Electrical Engineering
Telkom University
Bandung, Indonesia
brilliantfaina@student.telkomuniversity.ac.id

2nd Meta Kallista
School of Electrical Engineering
Telkom University
Bandung, Indonesia
metakallista@telkomuniversity.ac.id

3rd Ig. Prasetya Dwi Wibawa
School of Electrical Engineering
Telkom University
Bandung, Indonesia
prasdwibawa@telkomuniversity.ac.id

Abstract— **In Indonesia, there have been several natural disasters, such as earthquakes, tsunamis, landslides, floods, and others. Because Indonesia is situated where the Eurasian, Pacific, and Indo-Australian plates converge, this potential natural disaster is caused by this location. Social media information is expanding quickly and becoming more useful. Social media helps to alert people of the disaster's location during a disaster like a flood. Twitter is used as a data search engine in this work. Twitter has been utilized effectively to update the public on current events during emergencies. In order to learn more, we can conduct a search using pertinent hashtags to determining for the incident's location. The test's results will show a map of the Indonesian region, and the disaster's epicenter will be determined using the geolocation provided by the tweet data. The Naive Bayes approach will be used for classification. The clustering process occurs in real time across every region of Indonesia. In this investigation, the accuracy value was 75% based on the k-fold cross-validation test, utilizing a fold value of 3.**

Keywords—Natural disasters, Twitter, Naïve Baiyes.

I. INTRODUCTION

Social media particularly Twitter, has transformed data gathering and analysis. It provides valuable insights for disaster management and sentiment analysis. Twitter is vital for real-time updates and eyewitness reports during natural disasters like floods, wildfires, earthquakes, and storms [1]. Its microblogging format allows users to share brief updates within 280 characters, making it ideal for emergency information dissemination. In this paper we will focus on flyers shared on Twitter or commonly called Tweets to find data on where natural disasters occur. Indonesia is a country prone to natural disasters. To get fast and accurate information, Twitter can be the place.

Mitigation is an effort to reduce the risk of disaster. One of them is by increasing to deal with the threat of the disaster using data from Twitter. To make it easier for people to access disaster information, the author will create a disaster mapping information system that can provide real-time information using tweet data. To get the latest disaster information, the author will take data from Twitter and then grouped into several categories such as earthquakes, floods, and landslides. The collected information will then be classified using the Naïve Bayes algorithm and displayed on maps based on geolocation. Our research focuses on utilizing social media and text mining techniques to improve disaster management. By using Twitter as a sensor to detect and monitor natural disasters, we aim to enhance early warning systems, disaster response, and understand public opinion and emotional responses [2][3]. Our goal is to provide a practical solution for accessing real-time disaster information, reducing risks and consequences of natural calamities.

II. RELATED WORK

There are several research talk about classifying tweets from Twitter massage using machine learning, particularly when it comes to categorizing comments regarding natural disasters like floods, landslides, earthquakes, volcanic eruptions, and others. According to Jun Li (2017), remote sensing systems employ technologies and algorithms to gather data about the earth's surface from a distance. While remote sensing data is generally reliable for obtaining information, its accessibility may be limited. To enhance its utility, additional data from alternative sources, like social media or geographic information systems (GIS), can be incorporated [4]. This process will eventually develop into a system that can automatically retrieve Twitter and social media data from the Internet. Mauldy et al. employed Support Vector Machines, Random Forest Classifier, and Multinomial Naive Bayes in their study..

The primary objective is to reduce time spent on numerous web sources while assisting individuals in gathering news and information on disasters that are local to their area. Several studies have examined this kind of catastrophe categorization utilizing Twitter dataset sources or certain websites. Flood-related natural catastrophe tweets were categorized as tweets in Indonesian by authors Delimayanti et al. The authors analyzed the three alternative algorithms that may be used to categorize flood catastrophes using these approaches [5]. A Disaster News Automatic

Identification method was developed by academics Domala et al. Machine Learning (ML) and Natural Language Processing (NLP) are utilized in the processing stage, with NLP specifically employed for this purpose [6].

The Tweedr system was developed by other researchers, including Ashktorab, Z., et al. It is an information-extracting tool for disaster management. The classification, clustering, and extraction functions make up the tool's three core components. In the initial stage of the procedure, a range of classification techniques, such as LDA, Support Vector Machine (SVM), and Logistic Regression, are employed to identify tweets that mention damage or fatalities. For the other strategy, methods for clustering are used to join tweets that are similar to one another using filters. The subsequent step involves incorporating specific emphases and idioms associated with different categories of damaged structures, types of damage, and fatalities found within the filtered tweets. By implementing these three steps, valuable information can be extracted from the tweet streams across various scenarios [7].

III. RESEARCH METHOD

For the development of this system, a web interface will be designed to provide information about natural disasters such as earthquakes, floods, and landslides. This website will serve as a convenient platform for users to access up-to-date information on natural disasters. The data will be sourced from Twitter and classified using the Naive Bayes algorithm.

The web formation system is depicted in Figure 1. The initial stage of the system involves gathering Twitter data, as much as 1350 tweets per natural disaster such as earthquakes, floods, and landslides. Furthermore, it is separated into tweet classes and others. The data will then be processed with text preprocessing and classified using the Gaussian Nave Bayes model.

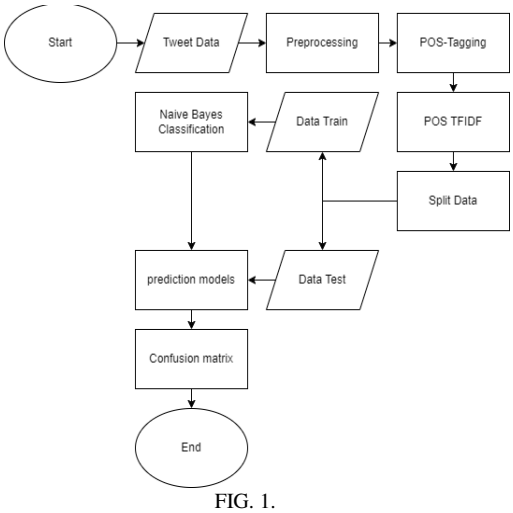

FIG. 1.
System Overview

A. *Functions and Features*

This website will display the map point of the occurrence of the disaster using the Leaflet API. This website makes it easy for users to access information about natural disasters. As for the application-based This website has the following features and functions:

1. Disaster feature to select what disaster to search for. There are three types of disasters, earthquakes, floods, and landslides.
2. The since and until features define start and end dates for disaster.
3. City/district feature to select the city or district where a disaster occurs.
4. The sub-district feature functions to determine the desired sub-district, if any.
5. Generate to run commands.

B. *Data Crawling*

To extract data from Twitter, the Python programming language incorporates the Twitter Intelligence Tool (TWINT) library. The training data comprises tweet data from 2013 to 2021, focusing on keywords such as floods, earthquakes, and landslides. From the selected criteria, a total of 3841 tweets were acquired, which were then categorized into four classes: floods, earthquakes, landslides, and others. The following is an example of the dataset format:

TABLE I.
Data Crawling Result

No	Tweet Data
1	telah terjadi gempa sebanyak 13 kali di selatan Yogyakarta dalam kurun 24 Jam terakhir, selalu waspada dan kita semua berharap semua aman. https://t.co/gzWyzXFI4k
2	Tim Gabungan, Polsek Limau Gotong Royong Jalan Longsor https://t.co/5CHMDvvpUe
3	Ribuan Rumah di Lebak Banten Terendam Banjir https://t.co/84UdHmMHH6

TABLE III.
Data Crawling Result EN ver

No	Tweet Data
1	there have been 13 earthquakes in the south of Yogyakarta in the last 24 hours, always be vigilant and we all hope everyone is safe. https://t.co/gzWyzXFI4k
2	Joint Team, Limau Police Working Together on Landslide Road https://t.co/5CHMDvvpUe
3	Thousands of Houses in Lebak Banten Flooded https://t.co/84UdHmMHH6

C. *Dataset Labelling*

The dataset labelling is done manually, the dataset is divided into 4 label classes, namely Flood, Earthquake, and Landslide, with the provision that data labelling is a tweet of a user who has information or the user is affected by the flood disaster at that time.

In the case of other classes, the dataset includes tweets discussing news or public opinion related to ongoing natural disasters like floods, earthquakes, and landslides. These tweets are posted by users who haven't personally experienced or been affected by these natural disasters. The proportion of data for a flood class is 968, 799 for an earthquake class, 843 for a landslide class, and 1230 for other classes.

TABLE IIII.
Dataset that already has a label

Tweet	Label
telah terjadi gempa sebanyak 13 kali di selatan Yogyakarta dalam kurun 24 Jam terakhir, selalu waspada dan kita semua berharap semua aman. https://t.co/gzWyzXFI4k	Earthquake

Tim Gabungan, Polsek Limau Gotong Royong Jalan Longsor https://t.co/5CHMDvvpUe	Landslide
Ribuan Rumah di Lebak Banten Terendam Banjir https://t.co/84UdHmMHH6	Flood

D. Pre-Processing

In making this system before it became a web, The dataset will undergo text pre-processing to process and filter the data. This involves cleaning the dataset by removing numbers, punctuation marks, and other elements not relevant to the study. Subsequently, a lemmatization process will be applied to transform words into their base forms. Afterward, words that do not significantly contribute to the overall sentence meaning will be eliminated through the removal of stop words. Lastly, the dataset will undergo word tokenization.

1. Data cleaning

This process is to remove unused text, like remove numbers, punctuation, hashtags, and URL because they have no effect for classification process, after which all sentences are converted to lowercase.

TABLE IVV.
Data Cleaning

Data (Input)	Data (Output)
Telah terjadi gempa sebanyak 13 kali di selatan Yogyakarta dalam kurun 24 Jam terakhir, selalu waspada dan kita semua berharap semua aman. https://t.co/gzWyzXFI4k	telah terjadi gempa sebanyak 13 kali di selatan yogyakarta dalam kurun 24 jam terakhir selalu waspada dan kita berharap semua aman
Tim Gabungan, Polsek Limau Gotong Royong Jalan Longsor https://t.co/5CHMDvvpUe	Tim gabung polsek limau gotong royong jalan longsor
Ribuan Rumah di Lebak Banten Terendam Banjir https://t.co/84UdHmMHH6	Ribu rumah di lebak banten terendam banjir

This process is to remove unused text, like remove numbers, punctuation, hashtags, and URL because they have no effect for classification process, after which all sentences are converted to lowercase.

2. Lemmatization

Lemmatization refers to the process of transforming a word into its base form while preserving its original meaning.

TABLE V.
Lemmatization

Data (Input)	Data (Output)
telah terjadi gempa sebanyak 13 kali di selatan yogyakarta dalam kurun 24 jam terakhir selalu waspada dan kita berharap semua aman	telah jadi gempa banyak 13 kali di selatan yogyakarta dalam kurun 24 jam akhir selalu waspada dan kita harap semua aman
tim gabungan polsek limau gotong royong jalan longsor	tim gabung polsek limau gotong royong jalan longsor
ribuan rumah di lebak banten terendam banjir	ribu rumah lebak banten rendam banjir

The process of lemmatization involves returning each word in a sentence to its root form. This process is intended to minimize the count of distinct words, thus reducing their overall number.

3. Stop Words Removal

Stop Words Removal refers to the process of eliminating insignificant connecting words from a sentence, as their presence or absence does not alter the sentence's meaning. This process is applied to the data that has undergone stop words removal, ensuring that conjunctions with minimal impact are removed.

TABLE VI.
Stop Words Removal

Data (Input)	Data (Output)
telah jadi gempa banyak 13 kali di selatan yogyakarta dalam kurun 24 jam akhir selalu waspada dan kita harap semua aman	jadi gempa banyak selatan yogyakarta dalam kurun jam akhir selalu waspada kita harap semua aman
tim gabung polsek limau gotong royong jalan longsor	tim gabung polsek limau gotong royong jalan longsor
ribuan rumah di lebak banten terendam banjir	ribuan rumah lebak banten rendam banjir

4. Word Tokenization

Word Tokenization is a process that involves splitting a sentence into individual words, resulting in a list of words constituting the sentence. This can be observed in the table below. The output of word tokenization aims to separate the words within a sentence, enabling further processing such as POS tagging, where each word is assigned a corresponding POS tag.

TABLE VII.
Word Tokenization

Data (Input)	Data (Output)
jadi gempa banyak selatan yogyakarta dalam kurun jam akhir selalu waspada kita harap semua aman	[‘jadi’,’gempa’,’banyak’,’selatan’,‘ yogyakarta’,’dalam’,’kurun’,’jam’,‘ akhir’,’selalu’,’waspada’,’kita’,’har ap’,‘semua’,’aman’]
tim gabung polsek limau gotong royong jalan longsor	[‘tim’,’gabungan’,’polsek’,’limau’,‘ gotong’,’royong’,’jalan’,’longsor’]
ribuan rumah lebak banten rendam banjir	[‘ribu’,’rumah’,’lebak’,’banten’,‘ren dam’,’banjir’][‘ribu’,’rumah’,’lebak ’,’banten’,‘rendam’,’banjir’]

E. Text Mining

Text mining, a discipline within the realm of information extraction from natural text, involves extracting latent and undiscovered information [8]. Its primary purpose is to communicate reliable and pertinent information by automatically extracting significant word components. Text mining encompasses various terms, such as text data mining, knowledge discovery from databases, and intelligent text analysis, all of which are focused on extracting information from unstructured sentences [8]. The subsequent section outlines the steps involved in text mining:

1. Collecting unstructured sentences.
2. Transforming the resulting sentences into structured words.
3. Identifying patterns from already structured words
4. Analyze patterns
5. Extract and store important information that is stored and visualized with the classification results [9].

Text preprocessing is the first stage of text mining that is used to modify the data.

The process performed in text preprocessing [10].

1. Case folding: converting the document to lowercase. Characters that are changed "From A to Z".

2. Tokenization: the process of breaking sentences into words. Can be done by removing punctuation marks dot (.), comma (,), spaces, and characters.
3. Filtering: the step of removing words that have no meaning, such as "what, di, di, and" which are in the middle of the sentence.
4. Stemming: is a process that aims to find basic words consisting of prefixes, inserts and suffixes.

F. Naïve Bayes Classifier

The Naive Bayes Classifier (NBC) is a straightforward algorithm employed in data mining techniques, leveraging Bayes theory and classification [11][12]. One of the advantages of NBC is that it doesn't demand a substantial volume of data for parameter determination [13]..
Equation of Naive Bayes Classifier

$$\mathrm{P(H|X)} = \frac{P(X|H)x\ P(H)}{P(X)}$$

Description:

X: Data with unknown class

H: Hypothesis that data X is a specific class

P (H | X): Probability of hypothesis H based on condition X

P(H): Prior probality of hypothesis H

P (X | H): Probability of X based on the condition in hypothesis H

P(X): Probability of X

G. POS TF-IDF

TF-IDF (Inverse Document Frequency Term Frequency) is an extraction technique used in text mining and information retrieval. Inverse Document Frequency to determine the usefulness of the word is needed or not [14]. In classifying the number of sentences, TF-IDF is used. The system generates keywords entered by the user, the use of TF-IDF to extract words taken from the subject of the sentence based on the keywords [15]. This method calculates the weight of the feature term which consists of two parts: TF and IDF. TF focuses more on the number of words that appear in the dataset while IDF is a measure of the general importance of the available sentences [16].
The TF-IDF formula is as follows:

$$TFij = \frac{n_{i.j}}{\sum k^n k.j}$$

$n_{i,j}$ is the number of words that appear in the file, while $\sum k^n k.j$ is the number of words that appear in all files [13].

$$idf_{i=log\frac{|D|}{|\{j:t_i \in d_j\}|}}$$

|D| is the total number of files in the data, $|\{j: ti \in d\ j\}|$ is the number of documents that contain the word. If the word is not in the data, it will cause the dividend to be zero.

$$tfidf_{i,j} = \frac{tf_{i,j\ x\ idf_i}}{\sqrt{\sum t_i \in d_j [tf_{i,j} x idf_i]^2}}$$

$tfidf_{i,j}$ is the weight of the word ti. It can be seen that the high frequency of words in a particular file and the frequency of the data set file can produce a high weight. TF-IDF.
Where:
TF: number of words/terms in the document.
IDF: the number of terms in all documents in the data.
d: document.

H. POS-Tagging

In this process, words in a sentence will be categorized based on their roles and functions. The pre-trained POS-Tagger dataset is used to be implemented in Indonesian.

TABLE VIII.
POS-Tagging

Data (Input)	Data (Output)
jadi gempa banyak selatan yogyakarta dalam kurun jam akhir selalu waspada kita harap semua aman	[(‘jadi‘,‘NN‘),(‘gempa‘,‘JJ‘),(‘banyak‘,‘NN‘),(‘selatan‘,‘NN‘),(‘yogyakarta‘,‘NN‘),(‘kurun‘,‘NN‘),(‘jam‘,‘NN‘),(‘akhir‘,‘JJ‘),(‘selalu‘,‘RB‘)(‘waspada‘,‘JJ‘),(‘kita‘,‘NN‘),(‘harap‘,‘NN‘),(‘semua‘,‘RB‘),(‘aman‘,‘JJ‘)]
tim gabung polsek limau gotong royong jalan longsor	[(‘tim‘,‘NN‘),(‘gabung‘,‘JJ‘),(‘polsek‘,‘NN‘),(‘limau‘,‘NN‘),(‘gotong‘,‘JJ‘),(‘royong‘,‘JJ‘),(‘jalan‘,‘RB‘),(‘longsor‘,‘NN‘)
ribu rumah lebak banten rendam banjir	[(‘ribu‘,‘RB‘),(‘rumah‘,‘NN‘),(‘lebak‘,‘NN‘),(‘banten‘,NN)](‘rendam‘,‘NN‘),(‘banjir‘,‘NN‘)

TABLE IX.
Used Data

Document	Data	Label
D1	so many earthquakes south of yogyakarta in the last hour period always be alert we hope everyone is safe	Earthquake
D2	the team joined the Limau police, working together for the landslide	Flood
D3	Thousands of houses in Lebak Banten were flooded	Landslide

Table above shows the dataset used to calculate POS TF*IDF using the same 3 documents as the documents used in the preprocessing process.

IV. SYSTEM DESIGN & OVERVIEW

A. System overview

information system from social media data is in the following figure below.

FIG. 2.
Application Web Front Page

In the picture above the user is required to input parameters available on the web. It can be seen in the picture that there are disaster options, namely floods, earthquakes,

landslides. The user is required to choose one of the existing disasters, then the user is required to select the location of the city / district to see the location where the disaster occurred, there is a sub-district column which is optional.

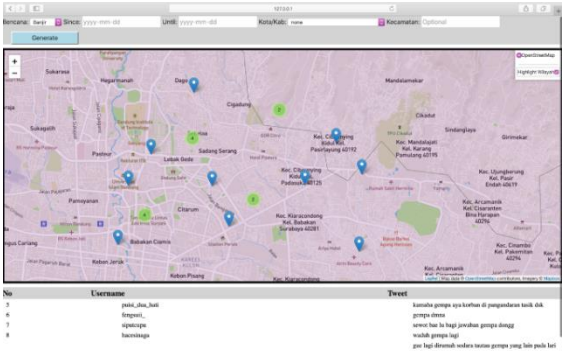

FIG. 3.
Tweet Mapping Results

Note:
1. Username: puisi_dua_hati, tweet: kumha gempa aya korban di pangandaran tasik dsk
2. Username: fengsuii_, tweet: gempa dmna
3. Username: siputcupu, tweet: sewot banget lu bagi jawaban gempa dongg
4. Username: hacesinaga, tweet: waduh gempa lagi gua lagi dirumah sodara tautau gempa yang lain pada lari.

The image above is the result of mapping tweet data based on the geolocation in the tweet. The results of the classification can also be seen in the table below the map.

B. Dataset

The dataset used is from user tweets and is divided into 4 categories, namely earthquakes, floods, landslides, and others. Data with earthquake, flood, landslide categories are tweets that contain information about users being affected, while other classes are user tweets that are not affected and only express their opinions about the disaster. The entire tweet data is taken from twitter and grouped one by one by the author.

C. Weighting

Weighting is done to convert the dataset into numbers that are calculated based on the formula from POS-TF*IDF, which will give weight to each word so that it can proceed to the classification process.

$$POS\ TF(t,d) = \frac{1*5}{\sum((1*5)(1*5)(1*5)(1*3))} \quad (1)$$

$$IDF(t) = 1 + loglog\left(\frac{3}{1}\right) \quad (2)$$

$$POS\ TF * IDF = 0.4106 \quad (3)$$

TABLE X.
Word Frequency Calculation Process

POS TF			
Word	D1	D2	D3
jadiNN	0,278	0	0
gabungJJ	0	0,278	0
ribuRB	0	0	0,167

Table above is the result of calculations that have been carried out using the POS-Term Frequency, which aims to get the frequency value. Words that have a POS-Tag "noun" (NN) and "verb" (VB) will be multiplied by 5, and words that have a POS-Tag "adjective" (JJ) and "adverb" (RB) will be multiplied by 3.

TABLE XI. IDF Calculation Process

Word	IDF
jadiNN	1,477
gabungJJ	1,477
ribuRB	1,477

Table above is the result of calculations carried out using the inverted document frequency, which aims to find out how often the word appears in all documents.

TABLE XII.
POS TF * IDF Calculation Process

POS TF * IDF			
Word	D1	D2	D3
jadiNN	0,4106	0	0
gabungJJ	0	0,2466	0
ribuRB	0	0	0,2466

Table above is the result of calculations performed using POS-TF * IDF, which is the result of the dataset that has been weighted and is ready to be trained with Naïve Bayes.

D. Evaluation

The evaluation process aims to know the performance of the model formed by using two techniques to test its performance, namely k-fold cross validation and confusion matrix.

E. Classification Evaluation

Evaluation process with a confusion matrix to determine the value of precision, recall, and f1 score. The data is then divided using k-fold cross validation to obtain accurate results, in this process using a value of k = 10. Then, before calculating the confusion matrix parameters, true positive, false positive, true negative, and false negative values are needed.

V. THE RESULT

A. Accuracy Testing

TABLE XIII.
Dataset Accuracy

Testing	Data Training	Data Testing	Results			
			Accuracy	Precision	Recall	F1_Score
1	90%	10%	72%	75%	71%	72%
2	80%	20%	70%	73%	69%	70%
3	70%	30%	70%	74%	69%	71%
4	60%	40%	71%	74%	70%	71%
5	50%	50%	70%	73%	69%	70%
6	45%	55%	69%	72%	68%	69%
7	35%	65%	69%	72%	68%	69%
8	25%	75%	66%	69%	65%	66%
9	15%	85%	64%	67%	62%	64%
10	5%	95%	61%	64%	59%	61%

To measure the performance of the Naïve Bayes algorithm, the following confusion matrix is performed confusion matrix as follows:

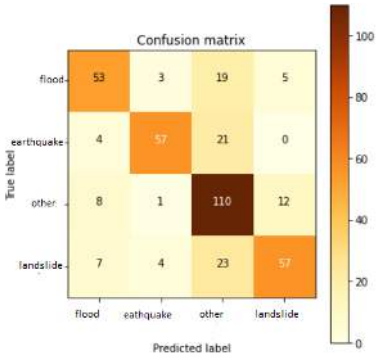

FIG. 4.
Confusion Matrix

Classification performance with the Naïve Bayes method. The confusion matrix results obtained by getting an accuracy value of 0.72 or 72%.

B. Kfold Cross Validation Testing

In this testing process this time the data used amounted to 3840 which contained disasters and tweets. The number of split values used in cross validation is 10 then divided by train data and test data.

After the data is partitioned, the next step is to enter the cross validation experiment. The first experiment is to take fold 1 in the first partition will be used as test data, the rest becomes training data. Then for fold 2 in the partition that was previously the test data for fold 2 becomes the training data and the test data moves to the second partition, and so on up to fold 10. For cross validation testing for experiments up to fold 10.

TABLE XIV.
Partition Score

FOLD	SCORE
FOLD 1	71%
FOLD 2	71%
FOLD 3	75%
FOLD 4	75%
FOLD 5	75%
FOLD 6	73%
FOLD 7	68%
FOLD 8	69%
FOLD 9	75%
FOLD 10	69%

Based on the outcomes from Kfold Cross Falidation testing with 10 trials, the results are as in the table above which results when averaged at 73%. The following is a virtualization image of the bar chart.

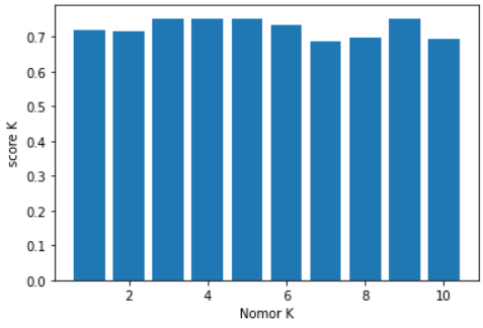

FIG. 5.
Kfold Cross Validation Graphics

Based on the results of the table above, the results obtained in folds 3, 4, 5 and 9 have the same large results, namely 75% as well as being the largest percentage value in the Kfold Cross Validation test.

VI. CONCLUSION

the study successfully developed a natural disaster monitoring information system using social media data and a web-based Naïve Bayes classification method. The system was able to accurately classify tweets related to floods, earthquakes, and landslides based on user information and geolocation data. The evaluation of the system using the Confusion Matrix showed an overall accuracy of 72%, with precision, recall, and f1-score values ranging from 74% to 88% for each disaster type. The system has the potential to provide real-time information to disaster response teams and aid in disaster management efforts. However, further research is needed to improve the accuracy and effectiveness of the system, especially in identifying and classifying emerging disaster types. Overall, this study provides a promising approach to utilizing social media data for natural disaster monitoring and response.

REFERENCES

[1] Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Perez-Meana, H., Portillo-Portillo, J., Sanchez, V., & García Villalba, L. (2019). Using Twitter Data to Monitor Natural Disaster Social Dynamics: A Recurrent Neural Network Approach with Word Embeddings and Kernel Density Estimation. Sensors, 19(7), 1746. https://doi.org/10.3390/s19071746.

[2] Wu, C. - H. (2016). SOCIAL SENSOR: AN ANALYSIS TOOL FOR SOCIAL MEDIA. International Journal of Electronic Commerce Studies, 7(1), 77 – 94. https://doi.org/10.7903/ijecs.1411.

[3] Luqyana, W. A., Cholissodin, I., & Perdana, R. S. (2018). Analisis Sentimen Cyberbullying Pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine. Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya, 2(11), 4704-4713.

[4] J. Li et al., "Social Media: New Perspectives to Improve Remote Sensing for Emergency Response," in Proceedings of the IEEE, vol. 105, no. 10, pp. 1900-1912, Oct. 2017, doi: 10.1109/JPROC.2017.2684460.

[5] M. K. Delimayanti, R. Sari, M. Laya, and M. R. Faisal, "Pemanfaatan Metode Multiclass-SVM pada Model Klasifikasi Pesan Bencana Banjir di Twitter," p. 9, 2021.

[6] J. Domala et al., "Automated Identification of Disaster News for Crisis Management using Machine Learning and Natural Language Processing," in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, Jul. 2020, pp. 503–508. doi: 10.1109/ICESC48915.2020.9156031.

[7] Ashktorab, Z., Brown, C., Nandi, M. and Culotta, A., May. "Tweedr: Mining twitter to inform disaster response". In Proceedings of 11th ISCRAM Conference, USA, 2014.

[8] K. L.Sumathy and M. Chidambaram, "Text Mining: Concepts, Applications, Tools and Issues An Overview", International Journal of Computer Applications, vol. 80, no. 4, pp. 29-32, 2013. Available:10.5120/13851-1685 [Accessed 23 August 2021]

[9] S. Dang, "Text Mining : Techniques and its Application", IJETI International Journal of Engineering & Technology Innovations, vol. 1,p.222014.Available:https://www.researchgate.net/publication/273038150_Text_Mining_Techniques_and_its_Application. [Accessed 23 August 2021]

[10] Kurniawan, B., Fauzi, M., & Widodo, A. Klasifikasi Berita Twitter Menggunakan Metode Improved Na √ Øve Bayes. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 1, no. 10, p. 1193-1200, juli 2017. ISSN 2548-964X. Tersedia pada:

<https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/361>. [Accessed 09 August 2021]

[11] M. Allahyari et al., "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques", arXiv.org, 2021. [Online]. Available: https://arxiv.org/abs/1707.02919v2. [Accessed: 23- Aug- 2021].

[12] A.Indriani.Klasifikasidataforumdenganmenggunakanmetodenaa ÃÉ ÃÑvebayesclassifier.InSeminarNasional Aplikasi Teknologi Informasi (SNATI), volume 1, 2014.

[13] QAULI ADZKIA, Klasifikasi Status Kemacetan Menggunakan Naïve Bayes Classification (Studi Kasus di Persimpangan Buah Batu Kota Bandung), 2020

[14] SRIVIDYA, Kotagiri; SOWJANYA, A. Mary. Aspect based sentiment analysis using pos tagging and tfidf. International Journal of Engineering and Advanced Technology (IJEAT), 2019, 8

[15] S. Kim and J. Gil, "Research paper classification systems based on TF-IDF and LDA schemes", Human-centric Computing and Information Sciences, vol. 9, no. 1, 2019. Available: 10.1186/s13673-019-0192-7 [Accessed 23 August 2021].

[16] C. Liu, Y. Sheng, Z. Wei and Y. Yang, "Research of Text Classification Based on Improved TF-IDF Algorithm," 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE), 2018, pp. 218-222, doi: 10.1109/IRCE.2018.849294.