Jurnal Teknik Informatika (JUTIF)
Vol. 3, No. 3, Juni 2022, hlm. 495-500

DOI: https://doi.org/10.20884/1.jutif.2022.3.3.160
p-ISSN: 2723-3863
e-ISSN: 2723-3871

C4.5 ALGORITHM FOR DISASTER IDENTIFIER SYSTEM

Ade Sutedi[*1], Hilmi Aulawi[2], Eko Walujodjati[3], Dini Destiani Siti Fatimah[4]

[1,4]Teknik Informatika, Institut Teknologi Garut, Indonesia
[2]Teknik Industri, Institut Teknologi Garut, Indonesia
[3]Teknik Sipil, Institut Teknologi Garut, Indonesia
Email: [1]adesutedi@itg.ac.id, [2]hilmi_aulawi@itg.ac.id, [3]eko.walujodjati@itg.ac.id, [4]dini.dsf@itg.ac.id

(Naskah masuk: 11 Februari 2022, Revisi: 2 Maret 2022, diterbitkan: 28 Juni 2022)

Abstract

Disaster management is a strategic issue that has been widely studied as a form of mutual responsibility in reducing victims and losses due to disasters. Today, one of the sources of disaster information is spread on Twitter social media. This research explains the implementation of the C4.5 algorithm to classify and map the disaster information that delivers using social media Twitter from the official BNPB_Indonesia account. The disaster data was retrieved and then processed to display in the geographic information system form. The words combination of disaster events, victims, and disaster locations process is carried out using n-gram by divided into unigram, bigram, and trigram to obtain the vocabulary accordance with the database. The C4.5 algorithm in this research was used to classify the disaster information with several categories. The results shows that the C4.5 algorithm can be used to classified the category of disaster and could identified the disaster informasion such as type of disaster, victims, and locations. The result can provide real time information on the distribution of disaster events and their locations using geographical information system. However, for location such as name of provinces only which has many geo-position possibilities (district or sub-district). The determination of the disaster location could be difficult. In addition, to determine the information obtained from post-disaster conditions such as the number of victims, damage, and losses. The comparison of n-gram with predetermined keywords is still constrained by noise of data.

Keywords*: c4.5 algorithm, disaster management, mapping, n-gram, social media.*

1. INTRODUCTION

A disaster is an event that cannot be predicted and often causes many victims. Disaster events are triggered by natural events or human actions which can damage the environment, buildings, property, and lives. This is a challenge for human survival, especially in disaster-prone areas. For this reason, the disaster management process becomes important to overcome the victims and losses that can be detected early. The disaster management process has been developed in many previous studies with the application of frameworks and technologies that can provide solutions in dealing with disasters.

Along with the development of big data, disaster information become the main resource to facilitate the disaster management process [1]. The process widely developed by combining several algorithm approaches [1]-[3], such as design artificial intelligence [2] or internet of thing [3], which can be integrated into a framework [4], [5], [9] supported by social media [6], [9], [10]. These steps can be used for Disaster Risk Reduction before, during the occurrence, and post-disaster to provide relevant information about disasters and accessible at all levels by all stakeholders.

This study focuses on the response process in the event of a disaster and post-disaster recovery by applying the application of Disaster Identification and Disaster Safe Locations [9], [10]. With utilizing the short of messages from social media [8]-[12] Twitter [2], [7]-[11], combined with classification algorithms [13]-[15] to make decisions [7], [10]-[12] in choosing accurate disaster information as a form of response disaster emergency.

In general, disasters are divided into three categories, namely natural disasters, non-natural disasters, and social disasters [16]. Meanwhile, according to the United Nations International Strategy of Disaster Reduction (UN-ISDR) in [17], it is stated that disasters are divided into five categories, namely Geological Aspects of Hazards, Hydro-meteorological Aspects of Hazards, Environmental Aspects of Hazards, Biological Aspects of Hazards, and Technological Aspects of Hazards.

Disaster management is a study that continues to be developed to find solutions to reduce the impact and victims of disasters through the process of Mitigation [6]-[8], Preparedness, Response, and Recovery [7], [8] including Rehabilitation and Reconstruction [6]. The study was conducted through a geographic approach with quantitative social media messages based on authoritative data to improve the identification of information in

managing disasters [11] and damage assessment models using social media databases [8]-[12] to increase awareness of disaster situations and rescue operations [9], [10], [12]. Real-time disaster information [3]-[5], integrating artificial intelligence and human awareness [6], sensors in disaster management [7]. Figure 1 shows the stages of disaster management which include a series of pre-disaster activities, during disaster emergency response, and post-disaster.

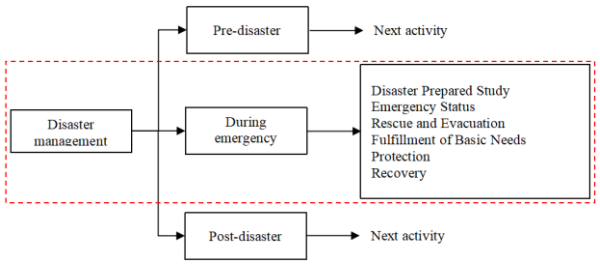

Figure 1. Disaster management [17], [19] and research limitation

Based on [8], information such as evacuation points in disaster areas and safe points for disaster victims as well as parties who will provide assistance in the recovery and evacuation of disaster victims can be mapped. As described in [9], the mapping can be applied in disaster identifier and disaster safe location applications. The mapping is done based on data submitted by application users which is compared with data from Twitter so that the location identification process becomes more valid.

2. RESEARCH METHOD

2.1. Dataset

In this study, the data was retrieved from social media Twitter with the official account of the Indonesian National Disaster Management Agency (BNPB) totaling 128 twits data from March to May 2021. As the example shown in table 1.

Table 1. Tweet Containing Disaster Information Sample

tweets	disaster info
Kepala BNPB Letjen TNI Doni Monardo memberikan piagam peng;hargaan kepada para tenaga kesehatan (Nakes) yang telah berjuang dan berpartisipasi selama satu tahun penanganan Covid-19 di RS Darurat Wisma Atlet, Kemayoran (23/3). #BersatuLawanCovid19 https://t.co/DHgBNPmOCi	No
Kepala BNPB Doni Monardo menghadiri acara peringatan satu tahun Rumah Sakit Darurat Covid-19 (RSDC) Wisma Atlet pada Selasa (23/3) di Kemayoran, Jakarta. Doni juga mengapresiasi partisipasi dan dukungan semua pihak dalam penanganan COVID-19. Selengkapnya: https://t.co/YD11IEoEux https://t.co/eYlpKKf2Zh	No
Hujan dengan intensitas tinggi menyebabkan banjir yang menggenangi pemukiman warga di wilayah Kabupaten Gorontalo Utara, Provinsi Gorontalo. Bencana ini berlangsung sejak hari Selasa (23/3) pada pukul 13.15 WITA. Selengkapnya : https://t.co/Q9sdTTVSTB #InfoBencanaBNPB https://t.co/Q6wkrW7Ciq	**Yes**
Update Infografis percepatan penanganan COVID-19 di Indonesia per tanggal 24 Maret 2021 Pukul 12.00 WIB. #BersatuLawanCovid19 https://t.co/qBlbzdYYA0	No
Acara Malam Puncak peringatan satu tahun RSDC Wisma Atlet diadakan di Tower 3, Selasa (23/3) di Kemayoran, Jakarta. Rangkaian acara dibuka dengan permainan angklung dari para pasien Covid-19 & para tenaga kesehatan. Selengkapnya : https://t.co/jwLTMaC7tD #BersatuLawanCOVID19 https://t.co/P6asChglsB	No
BPBD Kota Batu memonitor kondisi pascalongsor di wilayah Kecamatan Bumiaji, Kota Batu, Provinsi Jawa Timur. Upaya ini menyusul kejadian tanah longsor yang terjadi di Dusun Kajar pada Selasa lalu (23/3). Selengkapnya : https://t.co/KZEf3u2z4z #InfoBencanaBNPB #BNPBIndonesia https://t.co/2cOQnF3sQI	**Yes**
Tanah longsor yang terjadi di Kampung Cibitung, Desa Gunung Malang, Kecamatan Tenjolaya, Kabupaten Bogor, Jawa Barat, pada Minggu (9/5). Tidak ada korban jiwa akibat insiden tanah longsor di kampung tersebut. Selengkapnya: https://t.co/jVVtUX8HcL #InfoBencanaBNPB #BNPBIndonesia https://t.co/B5NjRuskea	**Yes**
BNPB bersama Badan Informasi Geospasial (BIG) berkolaborasi dalam penyusunan Standar Informasi (SI) Geospasial Risiko Banjir di Indonesia melalui rapat persiapan implementasi standar Standar Informasi Risiko Banjir pada Selasa (11/5). Selengkapnya : https://t.co/Jwll7aAE9w https://t.co/V4r9xKVgzh	No
BMKG melaporkan adanya gempabumi susulan (aftershock) dengan parameter magnitudo (M) 5.2 dari gempa sebelumnya M 7.2 yang kemudian dimutakhirkan menjadi M 6.7 di lepas pantai sebelah barat Kabupaten Nias Barat, Sumatera Utara, Jumat (14/5). Selengkapnya: https://t.co/8MfYLJY1uV https://t.co/TbBKcGkjKf	**Yes**

2.2. C4.5 Algorithm

C4.5 is an evolution of ID3 introduced by Quinlan (1993), using the gain ratio as the splitting criterion. The separation stops if the number of instances to be separated is below a certain threshold [20]. The advantages of C4.5 include:

1. C4.5 uses a pruning procedure that removes branches that do not contribute to accuracy and replaces them with leaf nodes.

2. C4.5 allows missing attribute values (missing value).
3. C4.5 handles continuous attributes by dividing the range of attribute values into two subsets (binary separation). In particular, it will look for the best threshold value that maximizes the profit ratio criteria. All values above the threshold are the first subset and all other values are the second subset.

The C4.5 algorithm has been successfully used to model natural disaster information classification on social media [13], landslides disaster [14], and forest fires [15]. In this study, the implementation of the C4.5 algorithm refers to the research [13] in determining disaster attributes, victims, and disaster locations with the following rules:

IF *(disaster = No)* ***THEN*** *Twit_Info_Bencana is No*
IF *(disaster = Yes)* ***THEN*** *Twit_Info_Bencana is Yes*
IF *(disaster = Yes* ***AND*** *victim = No)* ***THEN*** *Twit_Info_Bencana is Yes*
IF *(disaster = Yes* ***AND*** *victim = Yes)* ***THEN*** *Twit_Info_Bencana is Yes*
IF *(disaster = Yes* ***AND*** *victim = No* ***AND*** *location = No)* *THEN Twit_Info_Bencana is Yes*
IF *(disaster = Yes* ***AND*** *victim = Yes* ***AND*** *location = No)* *THEN Twit_Info_Bencana is Yes*
IF *(disaster = Yes* ***AND*** *victim = Yes* ***AND*** *location = Yes)* ***THEN*** *Twit_Info_Bencana is Yes*

Furthermore, these rules are applied to map a disaster location on a web-based application using the PHP language with the steps shown in Figure 2 below.

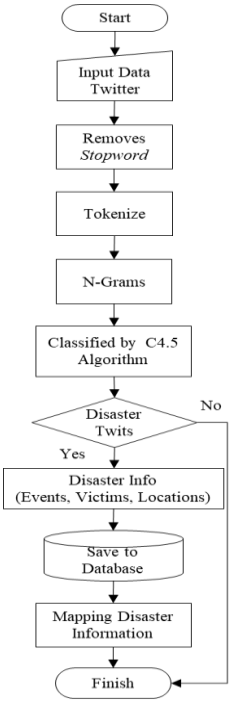

Figure 2. Disaster information selection process from Twitter

2.3. N-gram

N-gram is a gram order probability model of a value of n (n = 1,2,3, ...) words in a sentence [20]. n-gram are widely used in text mining processing as in previous studies [12], [21] which applied the concepts of uni-gram, bi-gram, and tri-gram in disaster studies. In this study, n-gram were used to determine the main keywords of disaster events such as unigram (*banjir*, *gempa*), bigram (*banjir bandang*, *gempa bumi*), and trigram (*angin puting beliung*, *erupsi gunung berapi*). In addition, n-gram is used for keywords in determining disaster locations such as uni-gram (Jakarta, Bandung), bi-gram (Bangka Belitung, Bandar Lampung), and tri-gram (Bandar Pasir Mandoge, Nusa Tenggara Timur). For this reason, the search for disaster locations using n-gram can be more specific to determine the geo-position to be used in disaster mapping.

3. RESULT AND DISCUSSION

After the experiment in this research, we conduct several result for the C4.5 algorithm with n-gram methods implementation i.e.:

3.1. Text Cleaning

The data obtained from Twitter will be selected using a stopword removal technique. At this stage, the words and strings that are not needed will be removed then what is left are the main words which will be evaluated to the next stage. The following text is an example data that using stopword removal on one of the input data taken from Twitter.

> "*Gempabumi dengan parameter magnitudo 7.2 dirasakan kuat oleh masyarakat di Kabupaten Nias Barat, Sumatera Utara, Jumat (14/5) pukul 13.33 WIB. Selengkapnya: https://t.co/jv8ZJkkfHY #InfoBencanaBNPB #BNPBIndonesia https://t.co/e7yTlXhh9d*"

Hasil yang diperoleh setelah proses stopword removal yaitu sebagai berikut.

> "*gempabumi parameter magnitudo 7.2 dirasakan kuat masyarakat kabupaten nias barat sumatera utara jumat selengkapnya infobencanabnpb bnpbindonesia*"

From the results of the stopword removal processing, it can be seen that some words and characters that are not needed are lost so that what is left are words that have more relevant information.

3.2. N-gram Feature

The process of selecting disaster-related information is the most important thing in this research. For this reason, the word combination using n-gram aim to obtain the vocabularies that relevan with the database data in determining disaster events, victims, and disaster locations.

The combination of uni-gram, bi-gram, and tri-gram was used for words arrangement that consisting of one, two, or three words in order to obtain the opportunity for words containing disaster information. In the classification process, the use of disaster information keywords is the main key to separate disaster categories. In this study, the classification process refers to the research [13] in which the keyword selection process was carried out based on general disaster information such as floods, earthquakes, landslides, etc. Next, choose the keywords with information related to the existence of victims in the disaster event such as death, damage, and loss. For the last category, we will choose the information of disaster location such as Jakarta, Bandung, Bangka Belitung, Bandar Lampung, Nusa Tenggara Timur, etc. So, the disaster keywords in this research used come from Indonesian words that listed in table 2.

Table 2. Keyword List Related to Disaster Information

natural disasters (geology and hydrometeorology)	non-natural and biological disasters	social disaster
Gempa Bumi, Tsunami, Gunung Meletus, Banjir, Kekeringan, Angin Topan, Tanah Longsor, Angin Puting Beliung, Gelombang Pasang, Kebakaran Hutan, Kerusakan Lingkungan.	*Gagal Teknologi, Gagal Modernisasi, Kecelakaan Transportasi, Kecelakaan Industri, Epidemi, Wabah Penyakit, Hama dan Penyakit Tanaman, Pencemaran Limbah.*	*Konflik Sosial, Teror Tawuran, Perebutan Sumberdaya, Pencemaran.*

In Figure 3 presented, an example of a combination of words taken from a tweets related to flood disaster information.

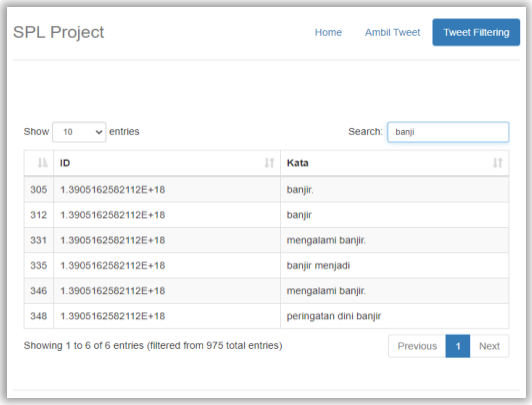

Figure 3. Several n-gram words combination contain information about flood (*Banjir*)

3.3. Classification Process

In this study, the classification process is carried out in two types, namely manual and automatic. The manual classification process is grouped based on disaster information categories by validating the information by the user administrator. While the automatic classification process is carried out by the generated from C4.5 algorithm according to the stages in the previously described section.

The manual method selects the tweets containing information related to events, victims, and disaster locations and saves them into the database. This input process is also for verification, validation, and comparison of the amount of disaster and non-disaster information that will be displayed on a graphic information map. This process also can to select disaster information that does not have a predetermined classification category requirement that the information cannot be classified automatically. For example, in some cases where the location is less specific, such as only knowing the province area, the determination of the disaster location point becomes irrelevant because it has many Geo-position possibilities (district or sub-district). In addition, to determine post-disaster conditions such as the number of victims, damage, and losses due to disasters.

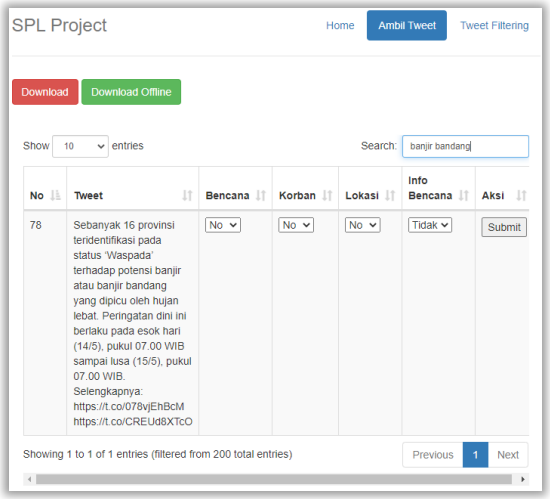

Figure 4. Disaster without location information

3.4. Disaster Identifier System

The execution of the disaster information will be distributed to the mapping of disaster identifier system. This application was made by the PHP and other web language for input and output processes that supported for graphic information system with map library taken from open street map [21]. The following example below is several codes for select the category of disaster event using PHP.

```
if (mysqli_num_rows($result) > 0) {
    while ($row = mysqli_fetch_assoc($result)) {
    $id = $row["user_id"];
    $kata = $row["kata"];
    $cek = mysqli_query($conn,
```

```
    "SELECT * FROM `disaster` WHERE `user_id` = '$id'");
        if (!mysqli_num_rows($cek) > 0) {
        $insert = mysqli_query($conn,
        "INSERT INTO `disaster`(`user_id`, `incident`)
        VALUES ('$id', '$kata')");
        }
        else {
        $insert = mysqli_query($conn, "SELECT * FROM
        `disaster`");
        }
        echo $id . " " . $kata . "<br>";
}
```

The disaster event was taken from the BNPB_Indonesia Twitter account that processed by splitting words (token) and then separating them into word order combinations in the form of unigram, bigram, and trigram. Furthermore, these words were filtered with the stopword library [23] and then compared with a list of vocabulary related to disasters, victims, and disaster locations based on the generalized rules of the C4.5 algorithm. This automatically process can give better results than previous studies [10]. Figure 5 shows the result of

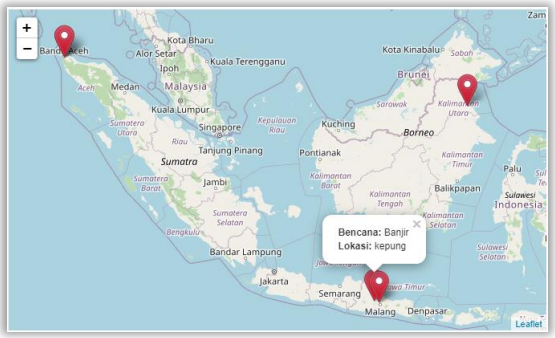

Figure 5. Disaster information mapping

Disaster information is also not only presented in the form of a graphic map, but also equipped with a disaster information table to complete the details of disaster events (number of victims, losses, and conditions) based on information obtained from the BNPB_Indonesia tweet. It also aims to display disaster events that cannot be displayed on a geographical map due to the inaccurate location information which affects the process of classifying disaster information which depicted in figure 6.

Peristiwa	Tanggal
BNPB_Indonesia Adapun tujuh kecamatan yang terdampak banjir menurut laporan Sukasno meliputi Kecamatan Muara Bangkal, Kecamatan Batu Ampar, Kecamatan Muara Ancalong, Kecamatan Long Masengat, Kecamatan Telen, Kecamatan Muara Wahau dan Kecamatan Kombeng. Selengkapnya: https://t.co/QPZKkobXTr https://t.co/ePr65RLIaU	Fri May 21 02:03:57 +0000 2021
BNPB_Indonesia Akibatnya ratusan rumah di kedua Daerah Aliran Sungai (DAS) tersebut terdampak banjir dengan ketinggian bervariasi. Selengkapnya: https://t.co/XRyacIPQVR #InfoBencanaBNPB #BNPBIndonesia https://t.co/jVQzy72Npd	Fri May 14 03:49:00 +0000 2021

Figure 6. Disaster information without specific location

4. CONCLUSION

Based on the research results from the previous section, we conclude that the C4.5 algorithm can be used to classified the category of disaster and could identified the disaster informasion such as type of disaster, victims, and locations. The result can provide real time information on the distribution of disaster events and their locations using geographical information system. However, for location such as name of provinces only which has many geo-position possibilities (district or sub-district). The determination of the disaster location could be difficult. In addition, to determine the information obtained from post-disaster conditions such as the number of victims, damage, and losses. The comparison of n-gram with predetermined keywords is still constrained by noise of data. For this reason, it possible for further research involved a natural language processing (NLP) in determining disasters, evacuation processes, locations, and post-disaster impacts.

REFERENCES

[1] R. J. J. Beerens, H. Tehler, and B. Pelzer, “How Can We Make Disaster Management Evaluations More Useful? An Empirical Study of Dutch Exercise Evaluations,” Int. J. Disaster Risk Sci., vol. 11, no. 5, pp. 578–591, Oct. 2020.

[2] C. Fan, C. Zhang, A. Yahja, and A. Mostafavi, “Disaster City Digital Twin: A vision for integrating artificial and human intelligence for disaster management,” Int. J. Inf. Manage., vol. 56, no. March 2019, p. 102049, Feb. 2021.

[3] P. P. Ray, M. Mukherjee, and L. Shu, “Internet of Things for Disaster Management: State-of-the-Art and Prospects,” IEEE Access, vol. 5, no. i, pp. 18818–18835, 2017.

[4] M. Abdel-basset, R. Mohamed, M. Elhoseny, and V. Chang, “Evaluation framework for smart disaster response systems in uncertainty environment,” Mech. Syst. Signal Process., vol. 145, p. 106941, 2020, doi: 10.1016/j.ymssp.2020.106941.

[5] A. M. A. Saja, A. Goonetilleke, M. Teo, and A. M. Ziyath, “A critical review of social resilience assessment frameworks in disaster management,” Int. J. Disaster Risk Reduct., vol. 35, no. February, p. 101096, Apr. 2019.

[6] Z. Li, Q. Huang, and C. T. Emrich, “Introduction to social sensing and big data computing for disaster management,” Int. J. Digit. Earth, vol. 12, no. 11, pp. 1198–1204, Nov. 2019.

[7] M. Erdelj, M. Król, and E. Natalizio, "Wireless Sensor Networks and Multi-UAV systems for natural disaster management," Comput. Networks, vol. 124, pp. 72–86, Sep. 2017.

[8] M. Yu, C. Yang, and Y. Li, "Big Data in Natural Disaster Management: A Review," Geosciences, vol. 8, no. 5, p. 165, May 2018.

[9] C. Slamet, A. Rahman, A. Sutedi, W. Darmalaksana, M. A. Ramdhani, and D. S. Maylawati, "Social Media-Based Identifier for Natural Disaster," IOP Conf. Ser. Mater. Sci. Eng., vol. 288, p. 012039, Jan. 2018.

[10] A. Sutedi, "Rancang Bangun Aplikasi Pengidentifikasi Bencana dan Lokasi Aman Bencana Berbasis Media Sosial," J. Algoritm., vol. 16, no. 2, pp. 239–246, Feb. 2020.

[11] J. P. de Albuquerque, B. Herfort, A. Brenning, and A. Zipf, "A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management," Int. J. Geogr. Inf. Sci., vol. 29, no. 4, pp. 667–689, 2015.

[12] S. Shan, F. Zhao, Y. Wei, and M. Liu, "Disaster management 2.0: A real-time disaster damage assessment model based on mobile social media data—A case study of Weibo (Chinese Twitter)," Saf. Sci., vol. 115, no. February, pp. 393–413, Jun. 2019.

[13] A. Sutedi, S. Rahayu, R. Elsen, and A. D. Supriatna, "Natural disaster topic selection using decision tree classification," J. Phys. Conf. Ser., vol. 1402, no. 7, p. 077034, Dec. 2019.

[14] Y. Handrianto and M. Farhan, "C.45 Algorithm for Classification of Causes of Landslides," SinkrOn, vol. 4, no. 1, p. 120, Oct. 2019.

[15] M. T. Anwar, H. D. Pumomo, S. Y. J. Prasetyo, and K. D. Hartomo, "Decision Tree Learning Approach To Wildfire Modeling on Peat and Non-Peat Land in Riau Province," in 2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2018, pp. 409–415.

[16] Undang-Undang Republik Indonesia, Nomor 24 Tahun 2007, Tentang Penanggulangan Bencana.

[17] D. L. Setyowati, Pendidikan Kebencanaan. 2019.

[18] Sarwidi, "Peran teknik sipil dalam penanggulangan bencana alam," no. September, pp. 1–18, 2012.

[19] L. Rokach and O. Maimon, Data Mining with Decision Trees Theory and Application, 2nd editio. World Scientific Publishing Co. Pte. Ltd. 5.

[20] D. Jurafsky and J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2020.

[21 I. F. Rozi, A. T. Firdausi, and K. Islamiyah, "Analisis Sentimen Pada Twitter Mengenai Pasca Bencana Menggunakan Metode Naïve Bayes Dengan Fitur n-gram," JIP (Jurnal Inform. Polinema), vol. 6, no. 2, pp. 33–39.

[22] "OpenStreetMap." [Online]. Available: https://www.openstreetmap.org/.

[23] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," 2003.