Natural disaster topic selection using decision tree classification

A Sutedi*, S Rahayu, R Elsen and A D Supriatna

Department of Informatics, Sekolah Tinggi Teknologi Garut, Jalan Mayor Syamsu 1, Garut 44151, Indonesia

*adesutedi@sttgarut.ac.id

Abstract. Social media like Twitter used as a reference for popular and trusted news today. This opportunity could be useful manner to inform the natural disasters that occur in certain condition related to their locations. This study aims to proposed model to select the topic related to the natural disasters that occur in Indonesia based on Twitter user timeline. The selected topic will use as a reference to ascertain events, conditions of victims, and position of disaster locations. In this study, the data taken from the official Twitter account of @BNPB_Indonesia. The data extracted into two main classes, and grouped based on types of disasters, victims, and disaster locations or safe point position using decision tree algorithm.

1. Introduction
Natural Disasters are events that occur unexpectedly, instantly cause damage, and victims. In Indonesia, the disaster mostly occurs every year such as earthquake, flood, eruption, tsunami, and many others. This condition is driven by Indonesia's geographical location which is between two continents and two oceans which called Ring of Fire [1]. For the last decade, the incidence of disasters in Indonesia has increased significantly. During this period Indonesia was hit by 11,274 disasters that claimed 193,240 lives and resulted in a total loss of at least Rp.420 trillion [1]. This condition increased every year as presented in table 1 which shows the number of natural disasters that have occurred over the past two years. Related to this condition, the government has planned a strategy that focuses on the seven main priorities [1], include increasing the effectiveness of disaster prevention and mitigation and increasing disaster recovery capacity. In order to implement these priorities, the government has been working on the process of disaster management and recovery by applying the existing technology. But, the effects of natural disasters causing environmental damage which results in the use of existing technology is not optimal. In this case, the appropriate technology is needed because of the impact of natural disasters cannot be predicted correctly.

Today, the presence of social media such as Twitter become important in the sharing process of the information related to disaster events. People can shared any information about disaster condition when the event occur such as earthquake, typhoon as describe in [2], eruption [3], flash flood in [4], based on their geo location [2-4] then inform to the people, non-governmental organization (NGO), government [5], about disaster site and safe point/ Secure Place Locator (SPL) [4], in order to minimize the victims and/or damage and losses, and saving a recovery fund to support decision making in the future [3].
Table 1. The natural disaster event in Indonesia for past two years.

| Disaster      | 2017 | 2018 | Total  | Average/ Year |
|---------------|------|------|--------|---------------|
| Typhoon       | 648  | 913  | 1561   | 780.5         |
| Flood         | 729  | 713  | 1442   | 721           |
| Landslide     | 576  | 613  | 1189   | 594.5         |
| Fire          | 96   | 375  | 471    | 235.5         |
| Earthquake    | 18   | 29   | 47     | 23.5          |
| Abrasion      | 8    | 33   | 41     | 20.5          |
| Eruption      | 2    | 32   | 34     | 17            |
| Dry           | 19   | 93   | 112    | 56            |
| Tsunami       | 0    | 1    | 1      | 0.5           |
| Total         | 2096 | 2802 | 4898   | 2449          |

In order to select the Information that related and unrelated contents of natural disasters. In this paper, the topics selected from twitter user timeline which classified into two main class using decision tree algorithm. The information of events limited to the content based on keyword [2] of Indonesian language such as “Gempa” for earthquake, “Banjir” for flood or flash flood, “Longsor” for landslide, etc. the victims, and the location of natural disaster. This algorithm select the keywords and use as a source of topic that related to the natural disaster for mapping disaster preparedness [6], emergency response [3–5], impact of disaster [2–4], and recovery process [4,7].

2. Methodology

In order to reduce the suffering of disaster caused, Rodríguez-Espíndola et al proposed a method for disaster preparedness system based on collaborative approach and Geographic Information System (GIS) [6]. Combining the social media data and semi-structured to determine the disaster preparedness using information technology and social media events [8]. Another research proposed the prototype of Spatial Data Infrastructure (SDI) that could be used as a platform to access the analysis of spatial data in emergency response and GIS for disaster management [3]. The information of disaster events taken from social media by modelled the automatic classification for uninformative and informative tweets class using Naïve Bayes and SVM [9]. A self-learning algorithm that construct Bag of Words (BoW) to detect the document from news of natural disaster to locate, organize and use valuable information [10]. Along with Basu et al that proposed a decision support framework to collects situational information through interactive crowd-sourcing using SMS from the disaster site, and summarizes such responses to have situational awareness and appropriate decision-making regarding damage or need assessment [7]. Then, the information could mapping the condition of post disasters for secure place locator to handle the victim on recovery process [4].

This research focused on the twitter topic classification related to natural disaster tweets content in order to support preparedness, response, impacted and recovery process based on social media. In this research, We adopted the keyword feature that applied in Sakaki et al [2] the information from crowdsourcing people could be source of topic that related to disaster events [2,3,11]. The process for topic classification from twitter data showed by Figure 1.
Based on [1], disaster risk has several categories that could be potential to the environmental damage. Disaster risk could be Earthquake (Gempa Bumi), Tsunami, Eruption (Gunung Api), Flood/Flash Flood (Banjir), Land Slide (Tanah Longsor), dryness (Kekeringan), Fire (Kebakaran), Extreme Weather (Cuaca Ekstrim) and Abration (Abrasi). In this research, decision tree [11] method used for establishing classification and prediction by C4.5 algorithm which divide two main classes related to disaster topic or not then for the related topic it will select based on the keyword that related to three subclass i.e. keyword from disaster event (Bencana), victims of disaster (Korban), and location of disaster or safe point position (Lokasi bencana/ tempat aman bencana) [12].

3. Result and discussion
In this research, the data taken from the official Twitter account of @BNPB_Indonesia which retrieve 396 tweets data. To determine the related and unrelated of natural disaster data, the algorithm take several steps for topic selection based on keywords from user tweets and divide into two main classes. First, the algorithm will select the keywords of natural disaster (Bencana) which become the root or main class for a tweet that related to the topic of a disaster or not. Second, if the topic is true (yes) then the algorithm will select the next keyword of victims (Korban) which becomes the next level of the previous main class. Third, the algorithm select the location of natural disaster that occur at some region or the area used as a post-disaster secure location. The result of classification process depicted in Figure 2.

Figure 2. Decision tree result for classification.

Figure 2 showed the result of classification for the topic of natural disaster in twitter data. These three level process obtained by decision tree algorithm approach to separate which tweet data contains information based on their keywords that related and unrelated to the natural disaster. Then, we implement this proposed model to predict the twitter data to find the topic that related and unrelated with natural disaster content. The result depicted in Figure 3.
Figure 3. Result of experiment to predict the topic of natural disaster.

The figure 3 showed that the experiment result to predict the twitter data related to natural disaster reach high accuracy. The best result of accuracy from decision tree algorithm in this research is achieve 0.8527. This achieved caused the process of selecting keywords from twitter data more specific than the bag of word technique that used in [7,10].

4. Conclusion
In this research, decision tree was implement to determine related and unrelated topic of natural disaster events based on social media especially Twitter. The best result for the prediction accuracy in classification is achieve 0.8527. This result can be implement for topic selection by three subclass i.e. type of natural disaster level, victim level, and location of natural disaster events or safe location level. This three level can be used as a source of information for the future process in disaster preparedness, real time emergency response, impact of disaster, and recovery process based on social media. For further research, it is recommended to choose a method that can handle the syntax in text processing such as the NLP approach.

References
[1] Amri M R, Yulianti G, Yunus R, Wiguna A, Adi A W, Ichwana A N, Randongkir R E and Septian R T 2016 Risiko Bencana Indonesia
[2] Sakaki T, Okazaki M and Matsuo Y 2010 Earthquake shakes Twitter users Proc. 19th Int. Conf. World wide web - WWW ’10 851
[3] Purwanto T H 2012 Pengembangan Prototipe Infrastruktur Data Spasial dan Sistem Informasi Geografis untuk Manajemen Bencana Alam
[4] Slamet C, Rahman A, Sutedi A, Darmalaksana W, Ramdhani M A and Maylawati D S 2018 Social Media-Based Identifier for Natural Disaster IOP Conference Series: Materials Science and Engineering
[5] Irvan R, Taufiq M, Slamet C, Andrian R, Aulawi H and Ramdhani M A 2018 Early Warning System in Mobile-Based Impacted Areas Int. J. Eng. Technol. 7 118–121
[6] Rodriguez-Espindola O, Albores P and Brewster C 2018 Disaster preparedness in humanitarian logistics: A collaborative approach for resource management in floods Eur. J. Oper. Res. 264 978–993
[7] Basu M, Bandyopadhyay S and Ghosh S 2016 Post disaster situation awareness and decision support through interactive crowdsourcing Procedia Eng. 159 167–173
[8] Anson S, Watson H, Wadhwa K and Metz K 2017 Analysing social media data for disaster preparedness: Understanding the opportunities and barriers faced by humanitarian actors Int. J. Disaster Risk Reduct. 21 131–139

[9] Parilla-ferrer B E, Jr P L F and Iv J T B 2015 Automatic Classification of Disaster-Related Tweets

[10] Palshikar G K, Apte M and Pandita D 2017 Weakly supervised classification of tweets for disaster management CEUR Workshop Proc. 1832 4–13

[11] Yan-yan S O N G and Ying L U 2015 Decision tree methods: applications for classification and prediction Shanghai Arch. Psychiatry 27 130–135

[12] Ross J, Kaufmann Q M and Salzberg S L 1994 Book Review: C4.5: Programs for Machine Learning Mach. Learn. 240 1–6