Classification of Natural Disaster on Online News Data Using Machine Learning

Mauldy Laya
Department of Computer and Informatics Engineering
Politeknik Negeri Jakarta
Depok, Indonesia
mauldy.laya@tik.pnj.ac.id

Mera Kartika Delimayanti
Department of Computer and Informatics Engineering
Politeknik Negeri Jakarta
Depok, Indonesia
mera.kartika@tik.pnj.ac.id

Anggi Mardiyono
Department of Computer and Informatics Engineering
Politeknik Negeri Jakarta
Depok, Indonesia
anggi.mardiyono@tik.pnj.ac.id

Fina Setianingrum
Department of Computer and Informatics Engineering
Politeknik Negeri Jakarta
Depok, Indonesia
fina.setianingurum.tik17@mhsw.pnj.ac.id

Aida Mahmudah
Department of Computer and Informatics Engineering
Politeknik Negeri Jakarta
Depok, Indonesia
aida.mahmudah.tik17@mhsw.pnj.ac.id

Diana Anggraini
Department of Computer and Informatics Engineering
Politeknik Negeri Jakarta
Depok, Indonesia
diana.anggraini.tik17@mhsw.pnj.ac.id

2021 5th International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM) | 978-1-6654-1200-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ELTICOM53303.2021.9590125

***Abstract*— We know neither what the future holds nor what will occur in the future. It is impossible to anticipate what the next instant may bring, and it could be awful. A natural disaster is an unforeseen event, which can have a tremendous impact on human life and the environment. The internet provides many sources that generate vast amounts of news articles daily. As the amount of news stories online increases, it's becoming more difficult for people to access disaster-relevant news, which makes it necessary to extract and classify news to be easily accessed. This paper presents an automated system that scraps news from various online sources and identifies disaster-relevant news. This paper also states the performance evaluation for classifying the natural disaster types based on machine learning in Indonesia's online news. Our results show that relevant Indonesian's online news about natural disasters can be the accuracy around 96% using the Support Vector Machine for three classes of natural disasters. In the Indonesian news data set, the machine learning algorithm that gets the highest value out of all the parameters is the best.**

Keywords—natural disaster, machine learning, online news, classification,

I. INTRODUCTION

Disaster management has dramatically benefited from the rapid advancement of information technology, especially when it comes to hazard reduction measures. The web has expanded quickly in recent years in hazard detection and mitigation. A wealth of information can be found on the Internet, which can be used to the benefit. These data can be used to identify hazardous areas, monitor them regularly, predict disasters early, and prepare for their aftermath. During disasters, the widespread use of social media platforms offers many opportunities for humanitarian organizations, for example, to improve their response. Identifying bystanders and eyewitnesses is one of them[1]. While the technology has advanced dramatically since the Internet began, it requires further study to enhance the provision of catastrophe-related statistics can help future study efforts in scientific research facilities that use enormous Internet data. These fields have been moving increasingly toward exploring enormous datasets and natural language processing, such as from social media and Twitter in recent years. [2], [3]. By monitoring hazard conditions in real-time, technologies like wireless sensor networks in disaster areas provide large amounts of data. The example system is capable of intelligent knowledge base management information and document information, as well as analyzing and refining existing professional knowledge data. [4].

News organizations play an important role in public assistance by disseminating information regarding hazard alerts, emergency planning, critical areas, and charitable organizations. Crawling such sites on the internet to collect and organize hazardous data will aid in making smart decisions in an emergency. This study takes a shrill approach to retrieving news from various news sites about relevant keywords in an emergency relief scenario. News websites broadcast trustworthy and authentic information in comparison with social media. The purpose of the proposed work is to delete content from news items that is not relevant for a disaster scenario and to allow the analysis modules to use only the relevant information. Based on this information, the areas affected can be identified. In some cases, the impact of the disaster could be efforts to recover after the disaster. Another essential advantage of this work is that disaster events can be created in a database and used for various disaster-related studies.

On the other hand, the web is a massive graph with various nodes indicating webpages and edges representing hyperlinks. Web crawling is an ineffective technique of acquiring information because every page contains a vast amount of data in the form of articles, photos, videos, and adverts. The most difficult part of the scraping process is finding meaningful data. This study develops a disaster-relevant crawler. By supporting disaster response that, in turn, aid in early warnings including after relief, news information validates the necessity for "right knowledge at the right time." As a result, crisis management is a snap thanks to the ease with which news data can be gathered and organized.

It is common for news articles to be organized to put the essential information at the start, some basic details about the event, and some background knowledge at the end. This study aims to calculate the classification of helpful information while filtering out the irrelevant about natural disasters,i.e., flood, earthquake, and forest fires. Because the process needs the extraction of relevant information from an article to work, a classification of retrieved data is necessary. Among the different ways of text classification, machine learning theories are a good fit for this paper. For example, classifying short texts (illustration: tweets, Status on Facebook updates, news headlines), significant texts (example: media articles, blogs),

sentiment analysis, topic labeling, and so on, using machine learning to classify text is a quick and cost-effective way to categorize, organize, and structure text. This project investigates the use of machine learning to discover and organize huge online news material gathered from the web then use a crawling web method and keywords connected to a natural disaster.

II. Related work

A number of studies on crawler design have been conducted. Ideally it is important to use efficiently the resources of a crawler, like the processor, bandwidth, memory, and storage. Websites that publish news items also provide dynamic information for readers, which means that crawlers must be strong enough to collect more than static data. Online articles may include text and pictures, videos and other kinds of data that convey disaster information. The most information from the source should also be extracted from an article to be analyzed. Crawlers must therefore be extendable to allow the management of any data structure[5]. Most of the available literature shows that extracting relevant data from the web is not easy. Significant information must be able to be identified and stored by Web Crawler[6], [7].

This type of disaster classification has been examined by several researchers using Twitter datasets or a specific web site. The writers Delimayanti et al. classify flood natural catastrophe tweets in the Indonesian language Bahasa. The authors compared how these methodologies and algorithms can be used to classify flood disasters into three separate groups[8]. Gopal et al. presented a data scraping technique for acquiring risk-relevant news items from the web, which involved the creation of crawler software and the use of machine learning techniques for filtering out useful material. The crawler software was created and deployed to scour news reporting web pages for hazardous stories.[9]. Using Machine Learning and Natural Language Processing, Domala et al. created an Automated Identification of Disaster News for Crisis Management. Using Natural Language Processing / NLP and machine learning ideas, this system will scrape content from English news websites and detect disaster-related news, that will then be dynamically presented on crisis management websites. [10].

This research team built an Internet distributed and incremental crawling system in China, which crawls knowledge literature, disaster news, and professional data, such as the current whereabouts of home and away from storms, and uses an advanced knowledge management concept to present the information in the form of a tree of knowledge that integrates the authoritative, worldwide typhoon live maps and orthophotos.[11]; Scraped news articles are then fed into a machine-learning algorithm to classify them as disaster or non-disaster. According to the authors of Fernandes et al., Geoparsing is also used to identify the location of interest in the news articles. Named Entity Recognition (NER) is used to create the geoparsing model[7].

III. Materials and Methods

The Internet offers a great many sources of news items every day. The crawler penetrates every website and goes deeper into the article. When an article is found, the content is temporarily parsed, and only the relevant items are eventually downloaded for content analysis. Our work focuses on online disaster data news that has been gathered from current Indonesian online news. Three disasters had a substantial impact on Indonesia in terms of casualties and property loss. We sampled data from three types of disasters to create a broader data set because we want to create a general classification framework for floods, forest fires, and earthquakes. The first step to implement a web crawler is to identify the requirements. It would not help to capture the relevant ones simply by analyzing the news article's source, URL, or title. For a better understanding, the contents of news articles must also be analyzed[9].

The approach suggests that the online data be preprocessed using machine learning techniques. By using a crawler tool, the collected internet news was acquired from Indonesian online news. Implementing a machine learning technique for categorizing online catastrophe news as a meaningful disaster includes several processes. Each stage in the text classification model is critical and determines the outcome. Many researchers have explored open-source online data, in particular, news[7], [10], [11]. The data that had been collected, then be continued for the data preprocessing and the following process as in the Fig.1.

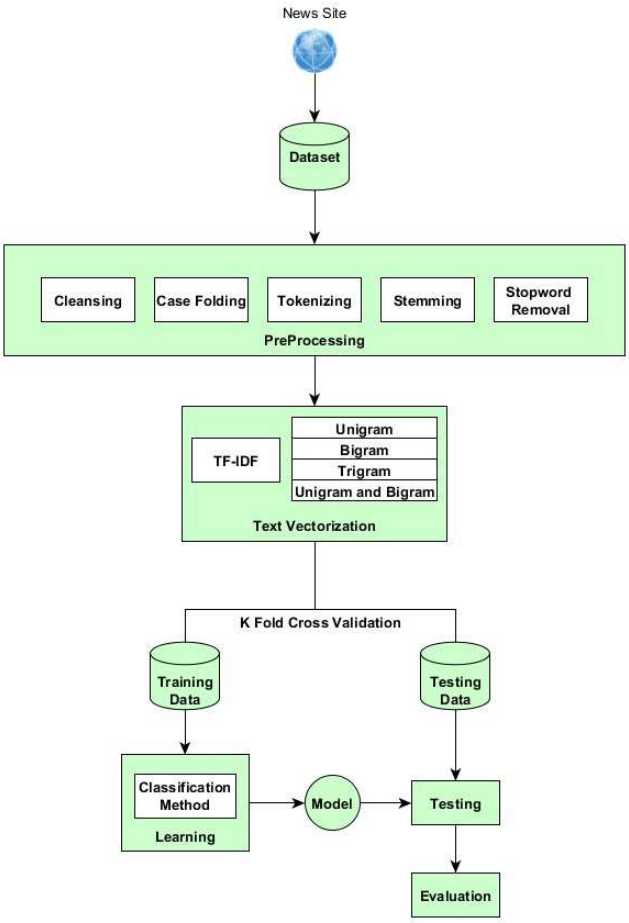

Fig. 1. The Flowchart of The Proposed System

The raw data used is the most recent news captured with Beautiful Soup. *Beautiful Soup* is a Python library that extracts

information from HTML and XML files. Kompas.com, and Liputan6.comis the news websites used. Kompas.com and Liputan6.com are popular news websites among Indonesian internet users. The raw data is then saved in a MySQL database. The first stage involves preprocessing input data in classified news, which includes data cleansing, case folding, tokenizing, stemming, and stopword removal. Various preprocessing techniques have previously been experimented with, and the data is cleaned so that the model does not learn the unnecessary characters in news articles.

The goal of case folding is to create any word pattern look the same. By altering the word to lowercase or a lowercase letter, the case was folded. The tokenizing technique was used to continue identifying the article's content in words. Following that, we do stemming on the text, which normalizes it. A sentence can be written in different of tenses while maintaining its meaning. As a consequence, stemmer aids in the removal of those tenses by aligning the sentences' meanings. In this scenario, talla stemming is applied.[12]. Stopword removal filters of many familiar words or words that are not standard and have no meaning (stopword). Two types of terminal dictionaries are used to remove it. Stopword dictionary id.stopwords.02.01.2016 was downloaded at https://github.com/masdevid/ID-Stopwords.git.

The following process is text vectorization. The text data will be converted into a numerical form in this step. TF is the Term Frequency, calculating the frequency of each term for the whole document, as the name implies. IDF knows Inverse Document Frequency, and the principal concept behind classification can differentiate relevant and irrelevant terminologies in a document. TF-IDF removes common words and removes vital features from the corpus. It makes commonly used words important but compensates for that number by the number of documents in which they appear, thereby causing a low standard of commonly used words. The relevance of words is evaluated in a document[13]. The frequency with which a word appears in a document is defined as TF. The IDF takes advantage of the fact that there are less meaningful and instructive words accessible. In word and character level, the two types are employed with unigram, bigram, trigram, and unigram combinations. As an example, the authors employed n-grams, which are neighboring letters or groups of words that help predict the following item in a series. N-grams represent a language's structure, such as which character or word comes follows the one before it. The N-goal gram's is to construct a word vector depending on the text's context. Figure 2 depicts an example of the n-gram process. The text is being vectorized while the data is being processed. When determining term weights, consider document length normalization. Because of the normalization process, any weight of the vectors document is worth something (0-1). The cosine normalization formula is used to do normalization in the equation.

$$W_{k\,j} \equiv \frac{tfidf(t_k, d_j)}{\sqrt{\sum_{s=1}^{f}(tfidj(t_k, d_j))^2}} \qquad (1)$$

Where $W_{k\,j}$ a normalized wight of many words k in the document j, $tfidf(t_k, d_j)$ is TF-IDF of word k in document *j*, *r* is the number of words in document *j*.

Fig. 2. The Example of the N-Gram Process

After that, the dataset must be split into two sections: training data and testing data. With k = 10, k-cross-validation can be used to train and test sets. For 10-fold cross-validation, the data is separated into ten pieces, with nine serving as training data and one serving as testing data. There are ten 10-fold cross-validation repetitions, each with a distinct set of testing data. The final stage in this classification model is to use the model established in the previous stage to train a classifier. The following classifiers are being used by system:

a. Naïve Bayes Classifier

A Naive Bayes classifier is a probabilistic model for classification issues. It is utilized because the Naive Bayes method is one of the standard Naive Bayes variants used in text classification and has been used for distributed multinomial data.[14].

b. Support Vector Machine Classifier

The SVM algorithm is used to find a hyperplane in N-dimensional space (N — features) that classifying the data points, i.e. the various categories of natural catastrophe news. [15].

c. Random Forest Classifier

Random forest classification is accomplished using multiple decision trees. Since it is composed of multiple decision trees, it is a very robust model. Random forest models produce reasonably good results out of the box[16].

IV. Result and Discussion

The primary data used include reports about floods, earthquakes, and forest fires in Bahasa Indonesia from Kompas and Liputan6. The overall news took 149 news about Beautiful Soup for three months. Python is a lovely Soup library that pulls HTML and XML data. Kompas.com and Liputan6.com were the news sites chosen. The news websites Kompas.com and Liputan6.com are commonly accessed by users in Indonesia. In the MySQL database, the raw data is then stored. All news items were scraped and marked manually.

The data set is cleansed and pre-processed by case folding, tokenizing, stemming, stop word removal before this data is provided to train the model. The model was constructed using three classification algorithms: Multinomial Naive Bayes, Support Vector Machine, and Random Forest. To evaluate the performance of each algorithm, we used 10- cross cross-validation to divide the data into training and testing sets.

It is crucial to evaluate the predictions made to adjust the disaster classification model from the online news. These

models are performed to determine how the output of disaster news reports is good and reliable. This generates confusion for the test dataset predictions and calculates the precision, accuracy, recall, and F1 score to better understand our model's performance.

The accuracy describes what parts of the articles during the classification are correctly predicted. Since the percentage of non-disaster items is usually considerably higher than those related to disasters, accuracy is not sufficient to assess the performance of a model. The accuracy described the proportion of articles that were projected to be disaster related. More accuracy would ensure that a smaller number of non-relevant items in our news feed will exist. The callback provided an accurate classification of the proportion of items related to disasters in the disaster class. Remember that it has proven to be the most crucial metric because news articles related to disasters have fewer articles than other categories. The model must not classify disaster articles incorrectly. F1 mark is the harmonic mean of accuracy and reminder that measures the two-performance metrics balanced.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$FI\ score = \frac{2 + Precision + Recall}{Precision + Recall} \quad (4)$$

TP =True Positive
FP = False Positive
FN = False Negative

From the table 1, we can see that Support Vector Machine (SVM) provides the best results for the classification model. For the future works, the model has ultimately been integrated in the real-time system to filter the disaster news from the raw news scraped.

TABLE 1. Result of The Classification Model

Model	Precision	F1 Score	Recall	Accuracy
Multinomial Naïve Bayes	88.00%	88.00%	88.00%	86.00%
SVM	94.00%	95.00%	97.00%	96.00%
Random Forest	94.00%	86.00%	83.00%	90.00%

V. CONCLUSION

Several online sources are available to supply news and manually classify it as natural disaster news. It is a complicated process, and it takes time. In this paper, we presented a model for classifying news articles in Indonesia that deal with natural disasters. Their natural disasters include floods, forest fires, and earthquakes. This model is an automated scraper that continually scraps news from 3 websites and stores it in the MySQL database. A model was trained to classify the types of natural disasters in three classes using machine learning algorithms. The process of the future, this model, will automatically become a system that can scrap social media and Twitter from the Internet. The system's primary purpose is to support civilians in gathering information and news about a particular disaster-related to their location and minimizing time spent on numerous various online sources. The classification models performed appropriately, and the Support Vector Machine (SVM), which was better compared to the other models, achieved 96% accuracy. This work has contributed several times. Firstly, we proposed a general environment for creating an automated disaster-specific news monitoring and classification system, a key innovation for early disaster detect.

More analytics could be carried out in the future by extracting more social media sources. The precise location of the natural disaster can be provided with these data. It can be automatically processed using a different machine and natural language models, helping citizens understand the disaster situation better and help disaster management make well-informed decisions.

ACKNOWLEDGMENT

We would like to express our highest gratitude to Politeknik Negeri Jakarta for supporting this research through Penelitian Produk Teknologi Terapan (PPTT) Research Schema 2021.

REFERENCES

[1] K. Zahra, M. Imran, and F. O. Ostermann, "Automatic identification of eyewitness messages on twitter during disasters," *Inf. Process. Manag.*, vol. 57, no. 1, p. 102107, Jan. 2020, doi: 10.1016/j.ipm.2019.102107.

[2] M. K. Delimayanti, Sari, Risna, Laya, Mauldy, Faisal, M. Reza, Pahrul, and Naryanto, R. Fitri, "The Effect of Pre-Processing on the Classification of Twitter's Flood Disaster Messages Using Support Vector Machine Algorithm," presented at the International Conference on Applied Engineering (ICAE), Batam, Indonesia, Oct. 2020.

[3] A. A. Khaleq and I. Ra, "Twitter Analytics for Disaster Relevance and Disaster Phase Discovery," in *Proceedings of the Future Technologies Conference (FTC) 2018*, vol. 880, K. Arai, R. Bhatia, and S. Kapoor, Eds. Cham: Springer International Publishing, 2019, pp. 401–417. doi: 10.1007/978-3-030-02686-8_31.

[4] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification," *Augment. Hum. Res.*, vol. 5, no. 1, p. 12, Dec. 2020, doi: 10.1007/s41133-020-00032-0.

[5] H. Srivastava and R. Sankar, "Information Dissemination From Social Network for Extreme Weather Scenario," *IEEE Trans. Comput. Soc. Syst.*, vol. 7, no. 2, pp. 319–328, Apr. 2020, doi: 10.1109/TCSS.2020.2964253.

[6] B. S. Jayasri and G. R. Raghavendra Rao, *Int. J. Electr. Comput. Eng.*, vol. 8, no. 6, pp. 4637–4645, 2018, doi: 10.11591/ijece.v8i6.pp.4637-4645.

[7] C. Fernandes, J. Fernandes, S. Mathew, S. Raorane, and A. Srinivasaraghavan, "Automated Disaster News

Collection Classification and Geoparsing," *SSRN Electron. J.*, 2021, doi: 10.2139/ssrn.3852688.

[8] 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.

[9] L. S. Gopal, R. Prabha, D. Pullarkatt, and M. V. Ramesh, "Machine Learning based Classification of Online News Data for Disaster Management," in *2020 IEEE Global Humanitarian Technology Conference (GHTC)*, Seattle, WA, USA, Oct. 2020, pp. 1–8. doi: 10.1109/GHTC46280.2020.9342921.

[10] 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.

[11] Q. Chen, Y. He, Q. Su, and T. He, "Building A Natural Disaster Knowledge Base Expert System based on the Distributed and Incremental Crawling Technology," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 435, p. 012024, Feb. 2020, doi: 10.1088/1755-1315/435/1/012024.

[12] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," p. 55.

[13] P. Bafna, D. Pramod, and A. Vaidya, "Document clustering: TF-IDF approach," in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, Chennai, India, Mar. 2016, pp. 61–66. doi: 10.1109/ICEEOT.2016.7754750.

[14] N. Bayes, *Naïve Bayes*. [Online]. Available: https://scikit-learn.org/stable/modules/naive_bayes.html

[15] S. V. Machine, *Support Vector Machine*. [Online]. Available: https://scikit-learn.org/stable/modules/svm.html

[16] R. Forest, *Random Forest*. [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

