Classification of Natural Disaster on Online News Data Using Machine Learning

by Mera Kartika Delimayanti

Submission date: 14-Jan-2022 10:25PM (UTC+0700)
Submission ID: 1741696203
File name: PAPER-CONFERENCE-ELTICOM-MAULDY-ET-AL-FINAL.pdf (305.61K)
Word count: 3652
Character count: 20089

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.setianingrum.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

***Abstract*— We know neither what the future holds nor what will occur in the future. Nobody knows what the next moment will bring, and it may be disastrous as well. 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. With the increase in the number of online news articles, it has become difficult for users 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 role of the web in hazard detection and mitigation has grown rapidly in recent years. 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 and facilitate research in scientific research institutions using massive Internet data in the future. 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 can analyze and refine the existing professional knowledge data, and it is capable of intelligent knowledge base management information and document information[4].

News organizations are critical in assisting the public by broadcasting information about hazard warnings, emergency preparedness, affected areas, and relief organizations. Crawling such websites to collect and structure hazardous data will help to make swift decisions in emergencies. This research follows a shrill approach to extracting news from various news sites relating to specific keywords in a disaster management scenario. News websites broadcast trustworthy and authentic information in comparison with social media. The proposed work aims at removing from news items non-relevant content specific to a disaster scenario so that only the relevant information can be used for the analysis modules. 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.

However, the web is a vast graph with multiple nodes being pages and edges being hyperlinks. Each page accommodates numerous data in articles, images, videos, and advertisements, making web crawling an inefficient method of obtaining information. Finding relevant data is the biggest challenge in the scraping process as a whole. A disaster-relevant crawler is developed in this paper. News data justifies the need for "right information at the right time" by providing emergency responses that ultimately aid in early warnings and after-effects relief. As a result, gathering and structuring news data make disaster management a breeze.

It is common for news articles to be organized to put the essential information at the beginning, followed by some basic facts about the event and some background information at the end. A crawler that offers only valuable information and filters out the irrelevant is the goal of this paper. A classification of extracted data is required because the system requires the extraction of relevant material from an article to function. Machine learning theories are a good fit for the goal of this paper among the many methods of text classification.

Using machine learning to classify text is a fast and cost-effective way to categorize, organize, and structure text, for example, tagging short texts (example: tweets, Facebook status updates, news headlines), significant texts (example:

media articles, blogs), sentiment analysis, topic labeling and so on. This paper explores the use of machine learning to classify and structure massive online news data extracted from the web using a crawling web approach with keywords specific 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].

Several researchers, using Twitter datasets or a particular web site, have studied this kind of disaster classification. In the Indonesian language of Bahasa, the authors Delimayanti et al. classify tweets of flood natural disasters. These methods and algorithms can be used to classify flood disasters in three different categories, and the authors have compared them[8]. Gopal et al. proposed a data scraping approach for gathering risk-relevant news stories from the web, which included the development of crawler software and the incorporation of machine learning approaches for filtering out insightful information. The crawler software was developed and was used to scour the news reporting web pages to search for hazard stories.[9]. Domala et al. developed an Automated Identification of Disaster News for Crisis Management Using Machine Learning and Natural Language Processing. This system will scrape news from English news websites and identify disaster-related news using natural language processing techniques and machine learning concepts, which will then be dynamically displayed 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 system proposes that the collected online data be processed and analyzed using appropriate machine training techniques. The collected online news was obtained from the online news in Indonesia by crawler tool. Several steps are involved in developing a machine learning model for classifying online disaster news as a relevant disaster. Each step of the text classifier model is crucial and decides the future. 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.

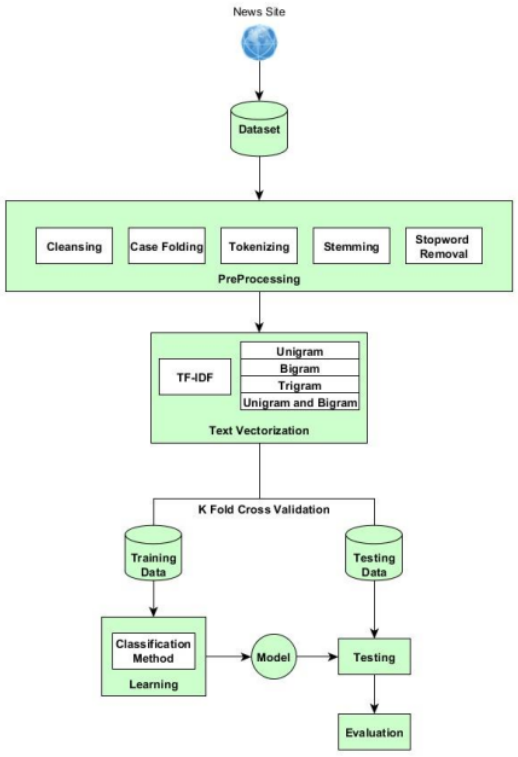

Fig. 1. The Flowchart of The Proposed System

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.

Case folding is intended to make any word shape the same. The case was folded by changing the word to a lower case or a lowercase letter. The following process continued to identify the content of the article in words, namely tokenizing. Furthermore, after that, we perform stemming on the text, which normalizes it. A sentence can be written in a variety of tenses and still have the same meaning. As a result, stemmer assists in removing those tenses by bringing the sentences into the same meaning. Talla stemming is used in this case[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 also 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]. TF is defined as how often the word is written in the document. The IDF relies on the fact that fewer informative and significant words are available. The two types are used with unigram, bigram, trigram, and unigram combinations in word and character level. The authors used n-grams as an argument, which are adjacent letters or words in a text that aid in predicting the next item in a sequence. N-grams capture the structure of a language, such as which letter or word comes after the previous one. The purpose of the N-gram is to generate a word vector based on the context of the text. The example of the n-gram process is as follows in Fig. 2. The data is now being processed, and the text is being vectorized. Take document length normalization into account when determining term weights. The normalization process means that any weight of the vector document is worth (0-1). Normalization is carried out in the equation with the cosine normalization formula as follows.

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

Where $W_{k\,j}$ a normalized wight of word k in 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

The dataset must then be divided into two parts: training data and testing data. The training and testing sets can be k-cross-validation with k = 10. The data is divided into ten parts for 10-fold cross-validation, with nine parts serve as training data and one as testing data. There are ten repetitions of 10-fold cross-validation, with each repetition using different testing data. The last step in this classification framework is to train a classifier using the model created in the previous step. The system employs the following different classifiers:

a. Naïve Bayes Classifier

A classifier of Naive Bayes is a probabilistic model used for classification problems. It is employed because the Naïve Bayes algorithm has been used for the distributed multinomial data and is one of the classic Naive Bayes versions used in text classification[14].

a. Support Vector Machine Classifier

The support vector machine (SVM) algorithm is used to locate a hyperplane in N-dimensional space (N — features) that categorizes the data points, i.e. the different types of natural disaster news[15].

b. 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 raw data used is the latest news in Bahasa Indonesia from Kompas and Liputan6 related to floods, earthquakes, and forest fires. The total news was 149 news for three months taken using Beautiful Soup. Beautiful Soup is a Python library for pulling data from HTML and XML files. The news sites used to consist of Kompas.com, and Liputan6.com. Kompas.com and Liputan6.com are news sites frequently visited by Indonesia's netizens. Then, the raw data is stored in the MySQL database. All the news articles were manually scraped and labeled.

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

Classification of Natural Disaster on Online News Data Using Machine Learning

ORIGINALITY REPORT

14%	7%	10%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

2%

★ Kanish Shah, Henil Patel, Devanshi Sanghvi, Manan Shah. "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification", Augmented Human Research, 2020

Publication

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		