Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7642781

UNSUPERVISED LEARNING FOR SOCIAL MEDIA TEXT ANALYSIS FORDISASTER MANAGEMENT SYSTEM

Praveen Sharma and Dr. Deepika Pathak
Department of Computer Application, Dr. A. P. J. Abdul Kalam University, Indore (M.P.) - 452016
Corresponding Author Email: praveendadhich.tj@gmail.com

Abstract—Social media is a source of low cost source of news and information. It is also work as information gathering and distribution toolduring the disaster situation.In this paper, we proposed an experimental model for simulating the use of social media data in disaster management. The focus of the model is to design a system which can early detect the natural disaster, identify the help request, getting feedback of response and obtain the situational awareness of the disaster conditions. In order to accomplish such model we implemented a modified Fuzzy C Means (FCM) clustering algorithms. Additionally by enhancing the centroids calculation for different situations the learning of the model is enhanced with the increasing amount of sample data.The experimental model has been implemented and the experiments are conducted on a publically available dataset. Based on the experimental results the proposed model found effective accuracy and acceptable training time. Therefore this model will help to manage the disaster situation from initiation to the disaster response.

Keywords— Disaster Management, Machine Learning Algorithm, Unsupervised Learning, Experimental Study, Performance Comparison.

Introduction

The social media is become one of the most powerful tool for collecting and distributing the information worldwide. Therefore it is now utilized in a number of real world applications for providing ease in human life. Among various humanitarian applications the disaster management is one of the applications where we can utilize the social media data for reducing the human and economical losses. During disasters people utilizesthe social media platform to share information about the disaster situation.This information will help to manage disaster in different phases.In this paper a new model is introduced using the machine learning and text processing techniques, which will filter and extract information regarding the natural disaster in four different administrative tasks.

Early stage natural disaster detection: The aim is to identify the disaster events using the social media post in early phases that will help to identify and provide information to the target location.

Identification of help related social media request:The aim of this module is to categorize the social media data according to help related request. Using this analysis administrator will send alert to the nearest response team to provide timely help and support to victims.

Getting feedback of response: the aim is to getting the feedback related to the response for the help request.

Getting the situational awareness: the social media data can also be used for recovering the updates and situation about the disaster.

In this presented work we proposed a model for computing the discussed objectives using the unsupervised learning techniques and the text processing methods. This section provides the overview of the proposed work involved in this paper, the next section discuss the clustering technique utilized for performing the learning and obtaining the required outcomes.

related work

In this section the recently developed approach for social media data analysis is discussed. The discussion will help to understand the proposed model of disaster management. The proposed work is aimed to develop a machine learning model which is able to perform the following task:

Identify the valuable tweets indicating disaster situation

Extracting the new keywords for incorporating with the learning system

Performing the sentiment analysis to rank the severity of the disaster

The dataset is obtained from the Kaggle [18], the dataset contains 5 attributes and 7503 instances.The attributes are ID, keyword, location, text and target. During preprocessing we eliminate ID.Additionally, the keyword and location has a lot of missing values therefore we also removed them. Finally we utilize the text and target attributes. In order to utilize the text and target first we process the text.

The text preprocessing reduces the noise such as abbreviations, stop words and special characters. In this analysis the hash-tags are very valuable and utilized as potential keywords. Therefore, during preprocessingwe eliminate the stop words, special characters and abbreviations but preserve the hash-tag keywords. Additionally, we used a feature selection process to transform text for utilizing with the ML algorithms.Here, we utilized the Term Frequency and Inverse Document Frequency (TF-IDF) and defined as:

$$TF = \frac{count\ of\ a\ term\ in\ a\ document}{total\ term\ in\ document} \dots\dots\dots (1)$$

And

$$IDF = \log\left(\frac{N}{df(t)}\right) \dots\dots\dots\dots.. (2)$$

Where df(t) is Document frequency of a term t, and N is Number of documents containing the term t.

Finally for selecting features the weights are calculated using:

$$w = tf * IDF \dots\dots\dots\dots. (3)$$

The weights are used for selecting keywords from the text and transformed the text into a learnablevector.

Next, task is the identification of the tweets that are potentially belongs to the natural disaster. But there are limited labeled data available for supervised learning therefore unsupervised learning technique is appropriate for learning. Thus, we utilized FCM algorithm, which utilizedan initial centroid and a modified process to update the centroid. Let, the C is the initial centroid:

$$C = \{null\} \dots\dots\dots\dots\dots. (4)$$

In tweeter for trending event users are utilizing specific hash tags. Thus, from the training samples we select keywords with hash tags H and is given by:

$$H = \{k_1, k_2, \dots, k_m\} \dots\dots\dots\dots\dots (5)$$

Where, k_i is the ith keyword selected through hash tags.
But vector H has different complexities like Duplicate words and Word with similar meaning or incomplete spelling. The duplicate words are reduced by:

$$UD = \{H: k_i \notin UD\} \dots\dots\dots (6)$$

Additionally, for words which have the similar spelling we utilized the levenshtein distance L. If the distance among two keywords in the set UD has the similarity greater then threshold T=0.75. Therefore,

$$R = \begin{cases} if\,(L(k_i, k_{i+1}) > 0.75)\;\; then\; add\; to\; R \\ \qquad else\; remove\; from\; list \end{cases} \dots\dots\dots..(7)$$

After refining the keywords from the selected hash tags we get a vector R as:

$$R = \{k_1, k_2, \dots, k_n\} \dots\dots\dots\dots\dots (8)$$

Here, similar size of centroid is used as the length of feature. Let the length of feature is p, then we partitioned the centroid into length of n. Therefore centroid C,

$$C = \begin{cases} if\; p \geq n \;\;\; then\; C = pad(R) \\ if\; P < n\; then\; C = part(R) \end{cases} \dots\dots\dots (9)$$

Where, pad(R) is the sequence padding if the length of centroid is less than the feature, and part(R) is partition of centroid into two or more parts based on length p.
Now we utilize the centroid C and feature $X = \{x_1, x_2, \dots, x_p\}$ obtained by TF-IDF technique. Next, we apply the FCM then we need to minimize the objective function:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \parallel x_i - c_j \parallel \cdots \dots\dots..(10)$$

Where, m is assumed m=2, u_ijis the membership of x_i in the clusterj, x_i is the ithelement ofx,c_j is the centroid, and ||*|| is the similarity.
FCM is optimizing an objective function given in equation (7) and for computing membership u_ij the equation (8) will be used:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left[\frac{\parallel x_i - c_j \parallel}{\parallel x_i - c_k \parallel}\right]^{\frac{2}{m-1}}} \dots\dots\dots\dots.(11)$$

And the centroids are updated usingequation (12):

$$c_j = \frac{\sum_{i=1}^{N} u_{ij} \,.\, x_i}{\sum_{i=1}^{N} u_{ij}^m} \dots\dots\dots\dots\dots (12)$$

After applying the clustering algorithms the tweets related to disaster events are ranked according to the neediness of people thus sentiment analysis is performed for ranking. A sentiment score calculation library namely Valence Aware Dictionary and sEntimentReasoner (VADER) [19] is used for this task. The sentiment score is measured using function "polarity_scores". Itresults in four scores for a sentence or paragraph:Negative (0 – 1), Positive (0 – 1), Neutral (0 – 1), and Compound (-1 - +1). We utilize the compound sentiment score for sentiment analysis. Thus to map the sentiment score into the needs the mapping functionM is used equation (13).

$$M = \begin{cases} if\ S \geq 0\ \ R\ received \\ if\ S < 0\ and > -0.5\ R\ required \ldots\ldots (13) \\ if\ S < -0.5\ then\ IR\ required \end{cases}$$

Where, S is score, R is response required, IR is immediate response needed.

Then we performed experimental evaluation thus a set of training and validation is prepared i.e. 70-30, 75-25 and 80-20,additionally, compared with the baseline FCM algorithm. In this context accuracy is considered as primary matrix, the accuracy is a ratio of correctly predicted class on the total samples and can be calculated using equation (14):

$$accuracy = \frac{correctly\ predicted}{total\ samples} X100 \ldots\ldots (14)$$

Next, for efficiency the training time is calculated. It is the amount of time taken to perform training,and calculated using equation (15):

$$training\ time = end\ time - start\ time \ldots\ldots\ldots\ldots (15)$$

The aim is to obtain the accurate classification of disaster events and identify the urgent request. Thus, a FCM algorithm is modified for categorizing the tweets into disaster and non disaster events. Additionally, a sentiment scores is utilized for identifying the intensity of request. Figure 1(A) and table 1 shows the accuracy of theclustering algorithms.In this figure X axis shows the training and validation sample ratio and Y axis shows thepercentage (%) accuracy. According to the results, the increasing training sample will improve the classification algorithms.The proposed method also enhances the performance as compared to traditional FCM. The reason is that the selection of potential keywords are better to represent information on the other hand the traditional FCM only usage the TF-IDF based features to learn, which can have noisy keywords.

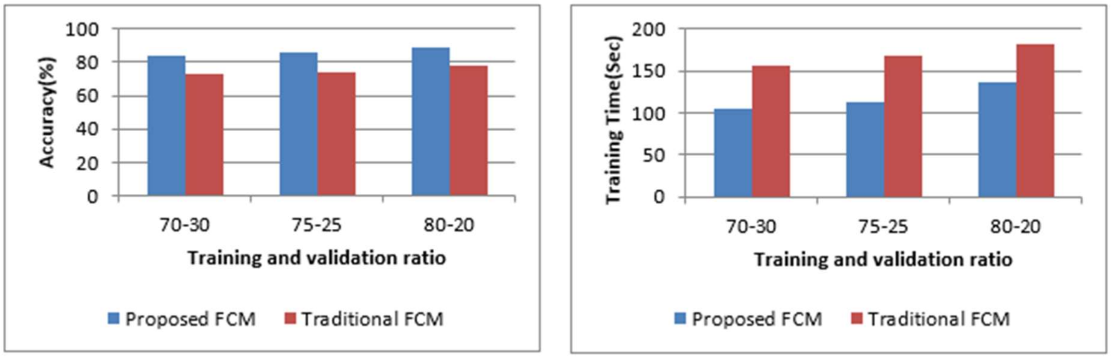
Next,figure 1(B) and table 1 demonstrates the training time of the clustering algorithms.The training time is measured in terms of seconds (Sec). The X axis shows the sample size, and Y axis shows the training time of the models. According to the results, we found the increasing training sample will increase the training time. The proposed technique requires less training time as compared to traditional FCM because the initially created single centroid is updated with the new data. The proposed algorithm optimizing a single cluster but the traditional FCM needed to optimize both the clusters. Thus,we found that the proposed FCM is suitable for more accurate and efficient than the traditional FCM.

Table 1Performance of Clustering Algorithms

		Accuracy (%)		Training time (Sec)	
S. No.	Training and validation ratio	Proposed FCM	Traditional FCM	Proposed FCM	Traditional FCM
1	70-30	83.8	72.7	105	157
2	75-25	86.2	74.4	112	169

3	80-20	89.1	78.2	136	181

The proposed social media content analysis algorithm modifies the traditional FCM to minimize the training time and enhancing the accuracy. The modification involves the centroid selection and optimization by learning the potential keywords.



(A) (B)

Figure 1 shows the performance in terms of (A) Accuracy and (B) Training Time

Proposed work

The aim of the proposed work is enhance the discussed model in previous section. The existing model can identify the tweets indicating disaster, extractionof new keywords for incorporating in the learning, and performing the sentiment analysis to rank the severity of the disaster. But the current model is not able to describe the type of natural disaster event. Additionally introduce framework for providing a plan for application development for disaster management. An overview of the proposed model is demonstrated in figure 2. The proposed model is utilized for disaster management for the administrative point of view. Using the given model the administrator can perform the following key tasks:

Identification of new natural disasters
Obtaining and updating the information about the disaster
Getting requests from the affected area
Managing feedback of request response

Identification of New disaters and it's type

The social media data is essential source of information for collecting and scrutinized according to the needs. In this presented work. We proposed to utilize the social media for designing an effective system for managing the disaster as administrator.

Among first task is to identify the disaster event start happening. In this context, we need to identify the social media post which belongs to the natural disaster events. Therefore, we create a list of keywords which indicate the natural disaster events. The essential keywords are: disaster, earthquake, flood, flash, hurricane, tsunami, thunder storm and cyclone. These keywords are utilized as a centroid for the clustering algorithm. Then, the update process is used to enhance the centroid.

In order to update the centroid, we utilize the pre-labeled disaster event data. And with only natural disaster labeled data is used to prepare the centroid. In this context, we select a random instance of the TF-IDF based vectored data additionally the previously assumed keywords are also merged with it. Thus if the initial keyword list $k_l = (k_1, k_2, \dots, k_l)$ and the selected random instance $I_f = (i_1, i_2, \dots, i_f)$. Then new centroid is created by mearging both the list by replacing the lesss wighted keywords. Thus the new list can be defined as

$$IK_l = \{ik_1, ik_2, \dots, ik_l\} \dots \dots \dots (16)$$

Additionally by comparing the IK_l with the new post feaures $\mathrm{F_l} = (\mathrm{f_1}, \mathrm{f_2}, \dots, \mathrm{f_l})$ we calculate the similar post for identifying the new event if happen. Further, disaster related post is categorized into three main parts:

New identified: when the legitimate disaster categorized post with new noun appeared, categorized post as new disaster detected. Additionally, it is necessary to detect early within 1 hour. Using this information the administrator can issue alert for the effective location.

Happening:after an hour of disaster event the detected post are may be relevant request or a help. Therefore, the administrator is utilizing this information for sending alters to the nearest response team.

Response: after alert by tracking the post the feedback of the response can be collected. In order to detect social media post in early stages by using the centroid IK_l. We consider some newly added keywords by administration related to information obtained from the governmental sources.

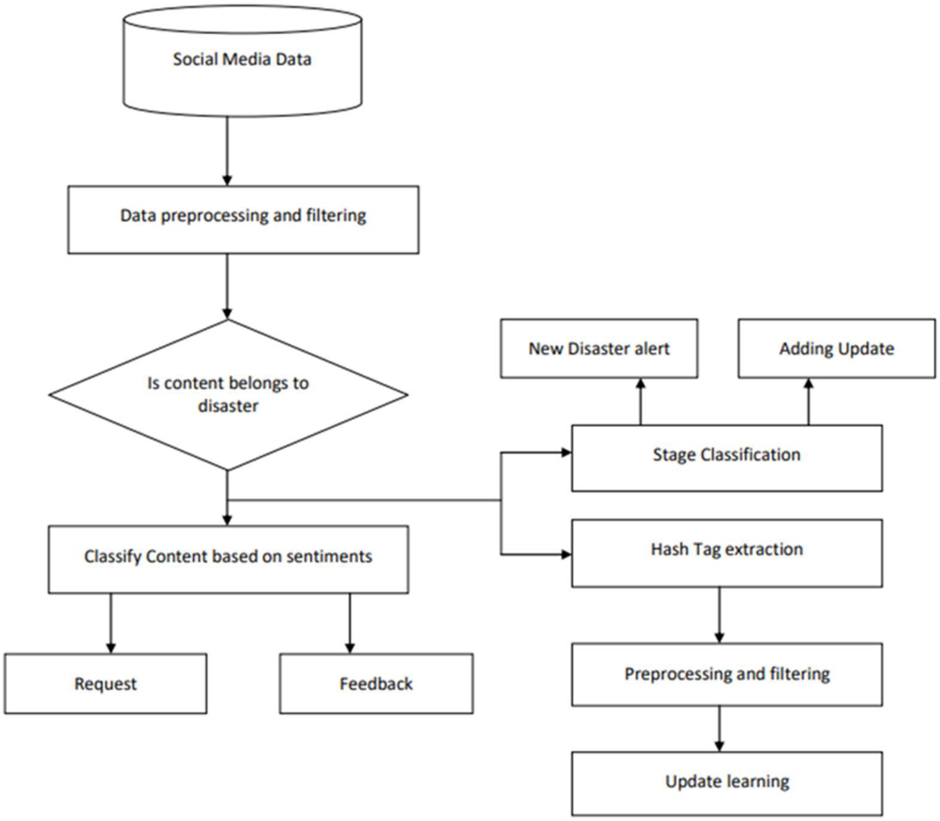

Social Media Data
Data preprocessing and filtering
Is content belongs to disaster
New Disaster alert
Adding Update
Stage Classification
Classify Content based on sentiments
Hash Tag extraction
Request
Feedback
Preprocessing and filtering
Update learning

Figure 2 proposed disaster management system

Managing new updates

In order to manage the updates about the identified disaster the recently identified posts are utilized for identifying the recent disaster related post. Therefore a new centroid is constructed:

$$H_l = \{h_1, h_2, \dots, h_l\} \dots \dots \dots (17)$$

In this context, the social media post after detection of first post is continuously utilized to cluster and update centroids.

Getting help request from the affeced area

This method for extracting help related social media post is given in section II. This method is utilized after identification of first post. That extracts the help request from the disaster victim.

Managing the response

After one hour of alter or help request the social media text is used to identify the action taken by the nearest team and/ or current status of the action taken. Therefore, the sentiment based classification is performed for identifying the response satisfaction of the disaster response. In order to perform this sentiment analysis task the equation (13) is used.

In order to organize these functions into an application the required flow is demonstrated in figure 3. In this given model the social media stream is used as input. The social media data is first preprocessed and vectored based on the TF-IDF technique. After that the centroid as defined in equation (16) is used to perform clustering. The resultant cluster is indicating new event identification. After an hour of first social media post detection the three different consequences can be recovered getting updates about the disaster effect, identification of help request, and sentiment analysis to get feedback of the disaster victims. Therefore, first we utilize the FCM clustering with the centroid H_l as described in equation (17). Next we utilize the FCM clustering using the with the centroid C as described in equation (9), and then sentiment analysis of post have been performed for identifying the sentiments or feedback of the disaster victims.

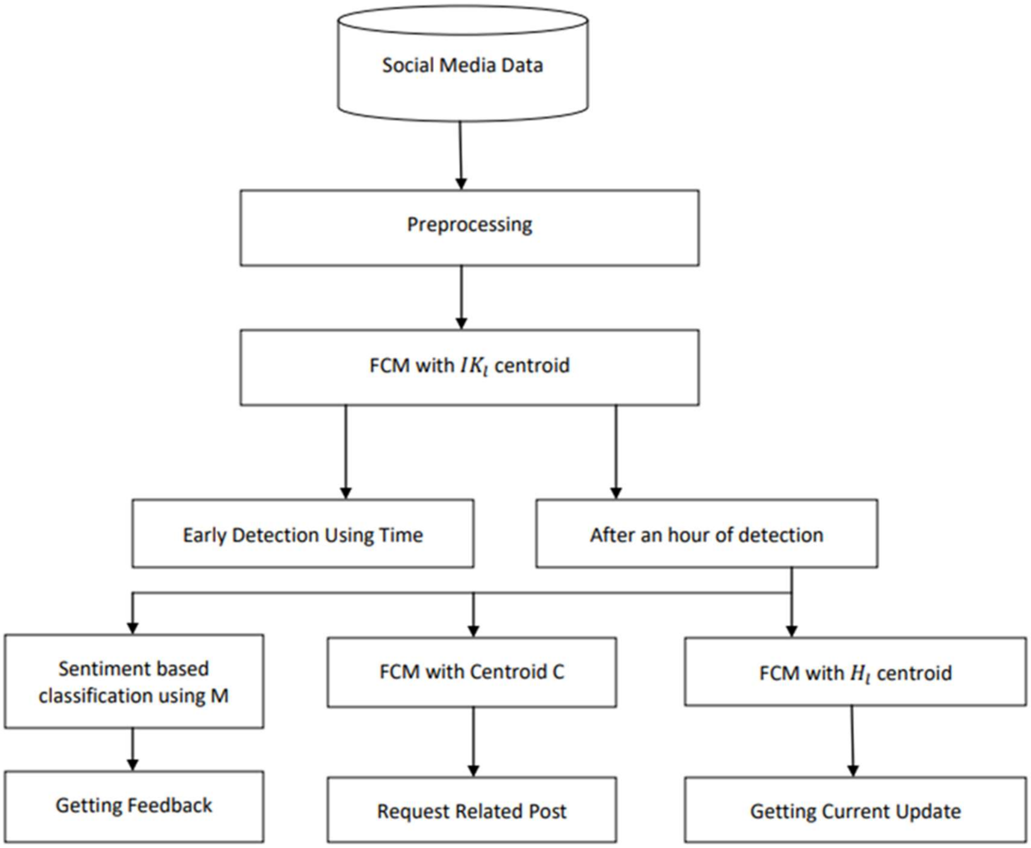


Figure 3 The processes involve step of identifying different administrative insights for disaster management

Results analysis

After implementation of the simulation as described in the above section the performance of the model is calculated. Therefore the FCM algorithm is used with two different scenarios:

For early detection of natural disasters thus simple IK_l is used as centroid

For getting updates and request related post thus FCM clustering with two centroids are performed, where first centroid is H_l and second is C.

Table 2 Performance Results of the implemented model for both the experimental scenario

Sample Size	Accuracy (%)		Training time	
	Scenario 1	Scenario 2	Scenario 1	Scenario 2
1000	67.8	66.2	25.4	29.6
2000	69.5	70.1	42.7	48.2
5000	72.9	74.5	89.6	101.5
7503	76.4	77.8	133.8	151.3

The experiments with both the scenario have been performed and the performance in terms of accuracy and training time is measured. The accuracy can be estimated using the equation (14) and the training time is calculated using the equation (15).The figure 4 and table 2 demonstrate the obtained results from both the experimental scenarios. The figure 4(A) demonstrate the accuracy measured in terms of percentage (%). According to the obtained results the accuracy of the scenario 2 is higher as compared to scenario 1. Additionally the accuracy of the model is enhancing with the increase in experimental sample. Therefore, inclusion of the more data with the time will improve the performance.

On the other hand the training time of both the models are given in figure 4(B). The training time of the models are measured in terms of seconds. The training time of both the models are increasing with the increasing amount of training samples. But the training time of scenario 2 is higher as compared to scenario 1. The centroid update process is most time consuming task. Therefore in order toupdate the two different centroids the scenario 2 utilize more time.

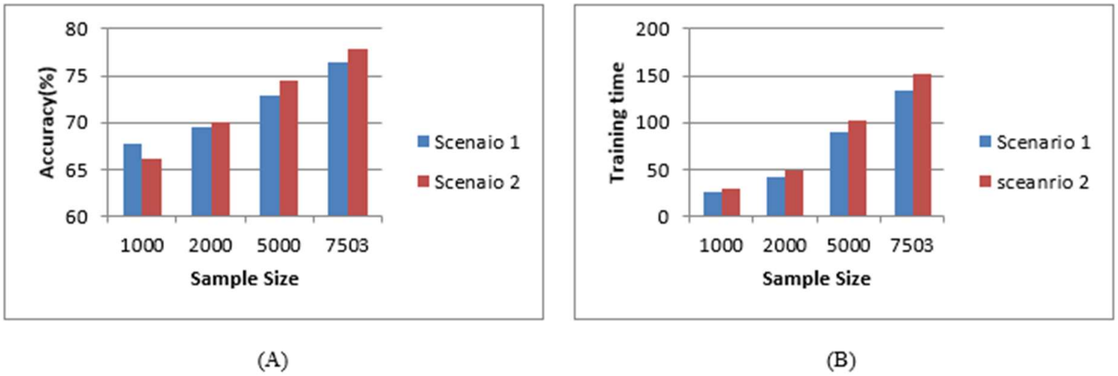


Figure 4 Performance in terms of (A) accuracy and (B) Training time

conclusion

The disaster management is one of the critical applications for saving human and economical loss during natural disasters. In this context, we proposed an unsupervised learning approach based on fuzzy c means clustering for utilizing the social media data and obtain the administrative insights for providing timely help and relief to the disaster victims. The model is not only providing ease in timely relief management, it is also useful for early disaster detection, identification of help request, collecting the response feedback, and updates on the current disaster situation.

The proposed model is implemented on simulation level and the performance of the implemented model in two experimental scenarios has been carried out. The experimental results demonstrate the FCM for early stage detection of disaster is performing less accurate as compared to categorization and identification of help related post and update collection. The reason is that in early stage the new keywords are appeared with the natural disaster which will work as noise for the FCM algorithm. However, the model is at the experimental level but the refined and tuned model will help the disaster management teams to locate and enhance the coordination in disaster management.

References

M. Aqib, R. Mehmood, A. Albeshri, A. Alzahrani, “Disaster Management in Smart Cities by Forecasting Traffic Plan Using Deep Learning and GPUs”, Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018, LNICST 224, pp. 139–154, 2018.

V. Linardos, M. Drakaki, P. Tzionas, Y. L. Karnavas, “Machine Learning in Disaster Management: Recent Developments in Methods and Applications”, Mach. Learn. Knowl. Extr. 2022, 4, 446–473

L. Dwarakanath, A. Kamsin, R. A. Rasheed, A. Anandhan, L. Shuib, “Automated Machine Learning Approaches for Emergency Response and Coordination via Social Media in the Aftermath of a Disaster: A Review”, IEEE Access, VOLUME 9, 2021

H. S. Munawar, “Flood Disaster Management: Risks, Technologies, and Future Directions”, Machine Vision Inspection Systems (Vol. 1): Image Processing, Concepts, Methodologies and Applications, (115–146) © 2020 Scrivener Publishing LLC

R. Veloso, J. Cespedes, A. Caunhye, D. Alem, “Brazilian disaster datasets and real-world instances for optimization and machine learning”, Data in Brief 42 (2022) 108012

S. Nanda, C. R. Panigrahi, B. Pati, A. Mishra, “A Novel Approach to Detect Emergency Using Machine Learning”, Progress in Advanced Computing and Intelligent Engineering, Advances in Intelligent Systems and Computing 1199, 2021

B. Resch, F. Usländer, C. Havas, “Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment”, Cartography and Geographic Information Science, 2018, Vol 45, No 4, 362–376

A. Madubedube, S. Coetzee, V. Rautenbach, “A Contributor-Focused Intrinsic Quality Assessment of OpenStreetMap in Mozambique Using Unsupervised Machine Learning”, ISPRS Int. J. Geo-Inf. 2021, 10, 156

H. Li, D. Caragea, C. Caragea, N. Herndon, “Disaster response aided by tweet classification with a domain adaptation approach”, J Contingencies and Crisis Management. 2018;26:16–27

C. Choi, J. Kim, J. Kim, D. Kim, Y. Bae, H. S. Kim, “Development of Heavy Rain Damage Prediction Model Using Machine Learning Based on Big Data”, Hindawi Advances in Meteorology Volume 2018, Article ID 5024930, 11 pages

B. Gokaraju, R. A. A. Nóbrega, D. A. Doss, A. C. Turlapaty, R. C. Tesiero, “Data Fusion of Multi-Source Satellite Data Sets For Cost Effective Disaster Management Studies”, 978-1-4673-9558-8/15/$31.00 ©2015 IEEE

J. Samuel, G. G. Md. N. Ali, M. M. Rahman, E. Esawi, Y. Samuel, “COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification”, Information 2020, 11, 314;

J. R. Ragini, P. M. R. Anand, V. Bhaskar, “Big data analytics for disaster response and recovery through sentiment analysis”, International Journal of Information Management, 42, 13–24, 2018

M. Imran, C. Castillo, F. Diaz, S. Vieweg, “Processing Social Media Messages in Mass Emergency: Survey Summary”, International World Wide Web Conference Committee, WWW ’18 Companion, April 23–27, 2018, Lyon, France

A. Khattar, P. R. Jain, S. M. K. Quadri, “Effects of the Disastrous Pandemic COVID 19 on Learning Styles, Activities and Mental Health of Young Indian Students - A Machine Learning Approach”, Proceedings of the International Conference on Intelligent Computing and Control Systems, IEEE, 2020

M. Rahman, C. Ningsheng, M. M. Islam, A. Dewan, J. Iqbal, R. M. A. Washakh, T. Shufeng, “Flood Susceptibility Assessment in Bangladesh Using Machine Learning and Multi criteria Decision Analysis”, Earth Systems and Environment, 2019

B. Choubin, A. Mosavi, E. H. Alamdarlood, F. S. Hosseinid, S. Shamshirband, K. Dashtekiang, P. Ghamisih, “Earth fissure hazard prediction using machine learning models”, Environmental Research 179 (2019) 108770

https://www.kaggle.com/competitions/nlp-getting-started/data?select=train.csv

S. Elbagir, J. Yang, “Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment”, Proceedings of the International MultiConference of Engineers and Computer Scientists 2019, IMECS 2019, March 13-15, 2019, Hong Kong