INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Check for updates

OPEN ACCESS

Received: 14-08-2023

Accepted: 11-12-2023

Published: 30-12-2024

Citation: Valliyammai C, Manikandan D, Hariprasanth R, Tamilvanan B, Aravind C (2024) Emergency Information Diffusion of COVID 19 using Machine Learning Techniques. Indian Journal of Science and Technology 17(48): 5091-5098. https://doi.org/10.17485/IJST/v17i48.2067

[*] **Corresponding author**.

cva@mitindia.edu

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Emergency Information Diffusion of COVID 19 using Machine Learning Techniques

C Valliyammai[1]*, D Manikandan[2], R Hariprasanth[3], B Tamilvanan[3], C Aravind[3]

1 Professor, Department of Computer Science and Engineering, College of Engineering, Guindy, Anna University, Chennai, Tamil Nadu, India
2 Research Scholar, Department of Computer Science and Engineering, College of Engineering, Guindy, Anna University, Chennai, Tamil Nadu, India
3 UG Scholar, Department of Computer Technology, Madras Institute of Technology, Anna University, Chennai, Tamil Nadu, India

Abstract

Objectives: To find the real time influential user in Twitter to diffuse emergency information globally during COVID 19 using social media. **Methods:** Twitter API is used to extract the real time COVID related tweets with trending hashtags where 80,000 live tweets are collected to process the data. Machine learning models such as Random Forest, Logistic Regression, Decision Tree, Linear Support Vector Machine, Extreme Gradient Boosting, K-Nearest Neighbor, and Ridge are trained to classify the COVID-related data. K-means clustering groups the communities with respect to zone-wise where the top influencer of the zone is identified using Page Rank influencer score. **Findings:** Random Forest model performs well with 98% accuracy when compared to other models. A total of 16 zones are identified based on K-mean clustering where the page rank algorithm identifies the top influencer user as 38427065 with 0.5741 influential score and it points to zone 14 Perungudi. The emergency information is disseminated through influencers using the proposed influential node identification technique. **Novelty :** Our system exhibits the geo graphical representation of the COVID-affected area with respect to user tweets. The emergency information about COVID is disseminated through top zone influencers.

Keywords: Social Media; COVID 19; Machine Learning; Influencer; Community

1 Introduction

In times of disaster, social media diffuses alert information to various communities in the network, and the technologies help to minimize communication breakdown by relying on a media platform to enhance information dissemination of warnings. In the event of any emergency or disaster, creating and distributing social media posts and polls are used to convey information throughout the network and elicit user replies.

Social media is mostly used as a management tool at the time of disaster which includes a medium to share messages and warnings, replies from the different community members, monitoring of user activities and sharing situational awareness posts. Internet usage and mobile phones provide great opportunities for users and volunteers to share, forward, and aid people in times of disaster. Social media might be useful where volunteers and emergency services can benefit from better collaborations. Detecting images or videos of a bursty or emergency event is unfolding in real time content sharing platforms which can aid with situational awareness. A micro blogging technology, Twitter is used to exchange information in real-time and send out suggestions and cautions.

In online social media, user behavior plays a major role in detecting the event at an earlier stage by using COVID 19 symptom dataset. A deep belief network was used to identify the topics of the event and tree convolutional neural networks were used to analyze the behavior and emotions of the user [(1)]. A deep learning model, Bi-LSTM-based sentiment analysis system, was presented to classify the sentiments of the COVID 19 tweets. The system evaluates the public's views in order to make the best decisions and avoid the propagation of misinformation. The system's major flaw is the incorrect prediction of positive and very positive classes [(2)]. Twitter is a predominant source of news, where the information feeds are transmitted freely by the connected users and it reaches others, often ahead of traditional news sources. The users in the network are connected to specific domains such as business, marketing, politics, education, healthcare, and public sector communities are also employed to disseminate information during disasters and emergencies. Tweet analysis system was proposed to detect the newsworthy and social events that happened or not with geographical location, and it identifies the descriptive tweet specified with a particular event and location. The unsupervised UK region tweets were used where the features are extracted into vectors using term frequency-inverse document frequency. DBSCAN and HBDSCAN clustering was performed to cluster the related tweets [(3)]. The COVID 19 severity prediction using machine learning technique was proposed to identify the people who are in need of intensive care and proper mechanical ventilation who suffer from respiration problems. The clinical data, blood panel profile data and social demographic data of the patients are extracted from COVID-19 REDCap repository where Saturated oxygen - SpO_2, C-Reactive Protein, Procalcitonin, d-Dimer, respiratory rate are extracted as the influential features. The random forest algorithm performs well with an accuracy of 80% for patients in need of ICU and 82% accuracy for patients who require mechanical ventilation [(4)]. A hybrid deep and machine learning model was proposed to identify the severity of COVID 19 and other pneumonia with the help of CT images. The dataset was collected from the COVID affected patients where the dataset is divided into three sets to process further and classified into severity level of the patient. The system also helps the patients who are affected with influenza and other pneumonia, etc. [(5)] An open analysis proves that machine learning was most useful in predicting emergency department and electronic health records of patients [(6)]. A systematic survey was performed to overcome the COVID 19 crisis in the field of diagnosis, epidemiology and disease dissemination with the major role of machine learning techniques. The preferred reporting items for systematic reviews and meta-analysis was performed to identify the relevant study on the topics. The PubMed, Web of Science, and CINAHL databases are considered further, in which the publications are divided into 3 themes based on artificial intelligence applications such as early detection and diagnosis of COVID, computational epidemiology of coronavirus and disease progression globally [(7)].

An event detection system determines the location of the users, extracted tweets and topics by utilizing natural language processing. An automatic labeling of urudu tweet using topic modeling was analyzed with multiple approach such as latent semantic analysis, probabilistic latent semantic analysis, latent dirichlet allocation and non-negative matrix factorization. Non-negative matrix factorization with 0.8 million tweets performs well with TF-IDF vectorizer method where linear dirichlet allocation performs well to convert long pseudo documents by merging small texts based on different topics [(8)]. Fast text based sentence -LDA model was used to replace the word embedding approach by overcomes the word co-occurrence information in text. A latent feature of topics is collected and processed with word embedding model [(9)]. The deep neural network based models perform well when compared to other machine learning classifiers even in the case of data imbalance. The deep learning models perform well with geographical based datasets where it belongs to different regions used in the similar type of event [(10)]. Hurricane Harvey in 2017 case study suggests that a kind of emergency plea for assistance posted on Twitter during catastrophes [(11)]. Despite the data was extremely uneven, the successfully created classifiers were used to identify these urgent tweets and the top models attain F1scores. The supervised learning techniques such as Naive Bayes with the meta-heuristic stochastic local search are combined to select the attributes and narrow space of attributes [(12)]. On the Nepal earthquake dataset, a hybrid algorithm was used to classify and rank the tweets. Several strategies have been used to increase the likelihood of extracting and differentiating fact and non-fact-checkable from a group of tweets. The fast text, a text classifier was used in addition to textual features to classify tweets related to flu infection posts [(13)]. The framework was trained using a pre-trained dataset and features like sentiment analysis and predefined keyword occurrences based on the data availability from social networking sites.

The identification of important nodes on the social network was effectively performed using semantic-influence measurement based algorithm which investigates the influence of influential nodes by directing the information to other connected nodes[(14)]. An influencer node is maximized in social network using community detection[(15)]. The degree and clustering coefficient was used to discover influential nodes by taking neighbors and degree into account. The clustering coefficient and weights of degree were computed using entropy technology. The Gaussian kernel approach was used as the co-authorship network to analyze the social network and to identify the prominent nodes. Local similarity measures and a multi-level diffusion algorithm were used for assigning the label to detect the communities in the network based on local information captured[(16)].

The majority of existing papers concentrate primarily on identifying COVID data rather than on disseminating the information globally. The existing sentimental classification of COVID 19 tweets using Bi-LSTM model causes uncertainty in predicting positive and very positive tweets. The existing event detection system using behavioral changes fails and it is impossible to forecast events three days in advance based on their user behavior. To overcome this issue, topic modeling is done in our proposed approach where COVID 19 topics are captured using word embedding and the extracted topics are clustered into sub-topics. The degree centrality is frequently employed in many existing systems to identify prominent nodes, but it fails in the dynamic social media due to the continuous interaction. To overcome this issue, we have used page rank algorithm to identify the top influential node in the dynamic social network.

This paper is organized as follows: Section 1 provides an introduction to social media influencers in the network and an analysis of related works, with research gap. Section 2 consists of methodology and proposed work which comprises data preprocessing, word embedding and topic modeling for identifying the emergency event and influential node, classification of emergency events, community based clustering and influencer identification. Section 3 explores the experimental results. Section 4 concludes with future work.

2 Methodology

The proposed emergency alert system helps to detect the COVID 19 tweets and to disseminate the information globally using social media with influential users. The system displays a geographical map of the COVID-affected area in relation to user tweets. The top zone influencers are identified to spread the COVID emergency message. However, in many cases sentiment, emotional and behavioral analysis of the users fails to detect the accurate information of the COVID. Topic modeling is performed with the help of word2vec to extract the features and clustering is performed to model the tweet content based on the COVID-related topics. Page rank algorithm is performed to identify the influencers because the conversation in the social media is dynamic and it continuously updates the rank value to the node and this helps to improve in identifying accurate influencer.

The pre-processing cleans the dataset to remove noises. Tweets contain unwanted lexical terms like URLs, punctuations, unicode, non- english characters, mentions of the users, etc., to be removed. The tweets contain a number of contexts in which they get categorized into different topics using word2vec, a word embedding model. Next, the tweets are classified into emergency or non-emergency events in order to filter out the emergency tweets from the pool of tweets.

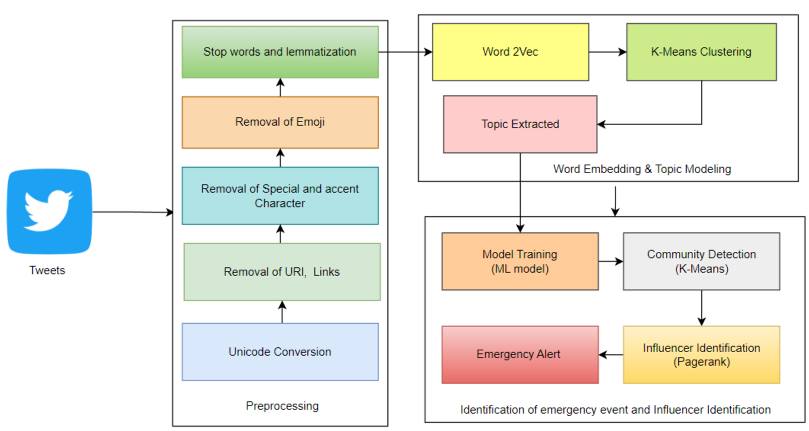


Fig 1. Proposed emergency alert system based on social media with influential users

The influential nodes of Twitter are identified to pass the filtered emergency events in order to create an alert among the public. The proposed emergency alert system based on social media with influential user is shown in Figure 1.

2.1 Data Preprocessing

Textual information contains emotions, accents characters, and text written in different cases. All letters in the text data are converted to lowercase, and words, numbers that incorporate digits are removed because it is difficult for machines to understand texts that combine words and digits. To make it easier to comprehend, substitute an empty string for such an expression. When attempting to evaluate terms from tweets, URLs, punctuations and unicodes are omitted. The stop words are the words that appear repeatedly in the text and they don't provide meaningful information. The stop words are removed using a list of words that is taken from the prebuilt library. The emojis and other special characters in tweets are filtered out of the dataset.

2.2 Word Embedding and Topic Modeling

The Word2Vec word embedding model is used to represent the words as the vectors in which the model generates vectors for the words, distributed numerical representations of word characteristics; these word features indicate the context of particular words in the lexicon. Word embedding assists in forming the relationship of a word with other words which comprises of similar meaning.

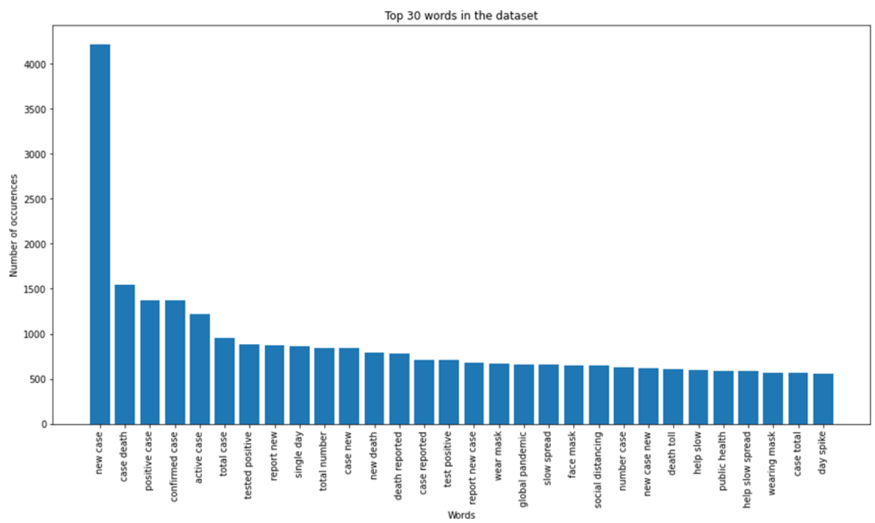


Fig 2. Top 30 bigrams in the Twitter dataset

The Word2Vec effectiveness comes from its ability to combine vectors of other related words together and it tries to predict the word's meaning based on how it appears in the text. The Word2vec generates an embedding vector for each word in the large corpus of text. The word embedding helps in calculating the association of a word with similar meaning words through the created vectors. The words that are frequently repeated that are related to the Covid 19 topics are captured for further processing. The contextual similarities in the tweets are detected based on the topic where the proximity of content is identified using cosine similarity. K-means clustering is performed to cluster the different sub-topics. The top 30 words of topics are captured based on word embedding and topic modeling is shown in Figure 2.

2.3 Identification of Emergency Event and Influential Node

The smart phone influences people more in time of emergency where Twitter is one of the popular communication platforms during emergency events. After the topics are captured, these tweets are clustered. The machine learning algorithms such as Decision Tree, Random forest, Ridge Classifier, Gradient Boosting, Logistic Regression and Linear Support Vector Machine models are trained to classify the tweets into COVID related tweets. The influential persons are identified for disseminating information instantly. The locations are detected using geo-tagged locations which are extracted from Twitter. The community is detected based on the K-mean clustering used in topic modeling where each user is clustered under different subgroups within the dataset. The influencer in the community is identified using centrality measures such as degree, page rank, betweenness, closeness and eigen vector. The Page rank algorithm performs well when compared to the other centrality measures because the influential score increases and decreases for every update.

The Pagerank algorithm uses the link between the pages to connect the relationships in the entire network which is displayed as a web map. Each page sets the same Pagerank value at the beginning of the relationship construction process through the links page of the web map. By counting the iterations, the final rank values for each page are determined in which the pagerank score of the current page is updated after each round of rank calculation.

Algorithm 1. PageRank Algorithm for Influencer Identification
Input. Tweets 'T' collected from the twitter API
Output. Influencer 'I' identification
Procedure. Identifying influential node of community
Initialize. Graph 'G', User nodes 'V', Total number of nodes 'N', Page rank score 'S'
V_j is the collection of user nodes {V_1, V_2, V_3 ____ ____ V_n} in the network.
W_k is represented as incoming nodes { V_{j+1}} of the current node V_j.
Begin
Initialize k =1
for(i=1; i < n-1; i++)
V_i = 1/N
for (j =1; j< n; j++)
if $V_j \leftarrow W_k$ **then**
$W_k \leftarrow W_{k+1}$
$V_j = V_{i+1}$ / N(W_{k+1})
if V_j < threshold value **then**
S = V_j
end if
else
$V_j \leftarrow W_{k+1}$
end if
end for

The updated Pagerank score is then calculated with the new round: the current Pagerank value of each individual page is evenly distributed to the chains contained with the same page so that each link receives the corresponding Pagerank score and each page will sum all chain weights to this page. The updated Pagerank score is calculated, as each page receives a new rank score, each page goes through a round of Pagerank calculations.

$$Rank\_score\,(V_z) = \sum_{j=1}^{n} \frac{(PR\,(V_j)}{N(W_j))} \tag{1}$$

where PR (V_j) represents the page rank value of node, N(W_j) Represents the number of outgoing node.

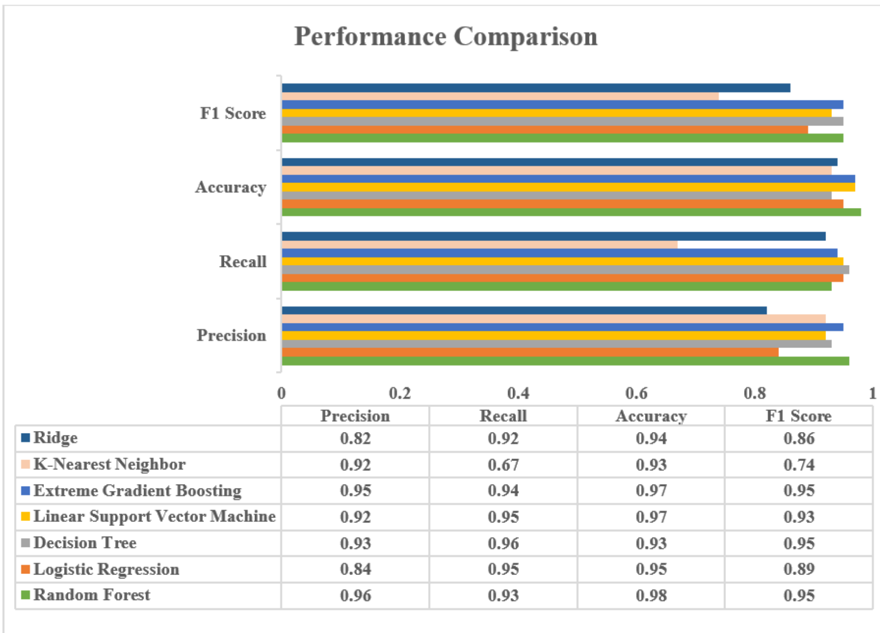
A link to the page counts as one vote for that page and the amount of votes a page receives based on the importance of that page is to all of the pages it is linked to. The importance of all pages linked to a page determines its PageRank using a recursive method. After fixing a method to identify the influential nodes, classified communities are passed to the Pagerank algorithm and then influential nodes are calculated. Finally, the emergency tweet is sent through the identified influential nodes to send an alert to the users. Emergency information is disseminated through twitter.

3 Experimental Results

The machine learning models such as random forest, logistic regression, decision tree, support vector machine, extreme gradient boosting, k-nearest neighbor, ridge classifier models are trained with COVID 19 data to classify whether the tweet is related to emergency or not. The accuracy of the classification algorithms is shown in Figure 3. After splitting the training and testing data as 80:20, Random forest classifier scored a good accuracy of 98%, followed by SVM and gradient boosting classifier having accuracy around 97%.

Precision is an important metric which calculates the quality of a positive prediction made by the model. The recall score is used to compute the model performance in terms of evaluating the count of true positives out of all the actual positive values.

The accuracy is used to identify the overall performance of the system. The random forest classification model outperforms well with other machine learning model such as logistic regression, decision tree, linear support vector machine, gradient boosting algorithm, K-nearest neighbour and ridge classifier. F1-score is a common metric used to evaluate the model quality, and it is calculated as the harmonic mean between precision and recall which gives a better measure of the incorrectly classified cases and ensures the robustness and preciseness of the proposed model. F1 score of any model is represented as F1 score = 2 * Precision * Recall / (Precision + Recall). After clustering of tweet, using the k-means algorithm, all the user-ids are clustered into different clusters based on geo-coordinates. For each community out of sixteen communities, an influential node is identified through the Pagerank algorithm as shown in Table 1.



	Precision	Recall	Accuracy	F1 Score
Ridge	0.82	0.92	0.94	0.86
K-Nearest Neighbor	0.92	0.67	0.93	0.74
Extreme Gradient Boosting	0.95	0.94	0.97	0.95
Linear Support Vector Machine	0.92	0.95	0.97	0.93
Decision Tree	0.93	0.96	0.93	0.95
Logistic Regression	0.84	0.95	0.95	0.89
Random Forest	0.96	0.93	0.98	0.95

Fig 3. Performance Metrics of Different Classification Algorithm

Table 1. Identification of Influential users in Community

User id	Influencer Score	Location
208056970	0.4903	Zone 7 Ambattur
8431652	0.5431	Zone 13 Adyar
38427065	0.5741	Zone 14 Perungudi
185461758	0.5735	Zone 3 Madavaram
151624811	0.4534	Zone 12 Alandur

The Folium is a python based library for creating various types of leaflet maps where the data can be manipulated and visualized in the leaflet map. It helps the users to generate a base map of customized height and width. To create a basic folium map, first initiate an empty map and then loop through the rows in the data frame to add markers to the map based on each influential user's latitude and longitude. Within the marker function, specify the data to be shown in the popup. The visualization of various clusters is shown in Figure 4.

The Markers is customized by including HTML and codes to the icon parameter. The separate function is used to create an HTML page for each marker in the dataset, for an influential user where it comprises of userid and community to which the user belongs to. Then pass the HTML to Folium map markers and make the icons colors look different for each community. The customized markers for each community are shown in Figure 5.

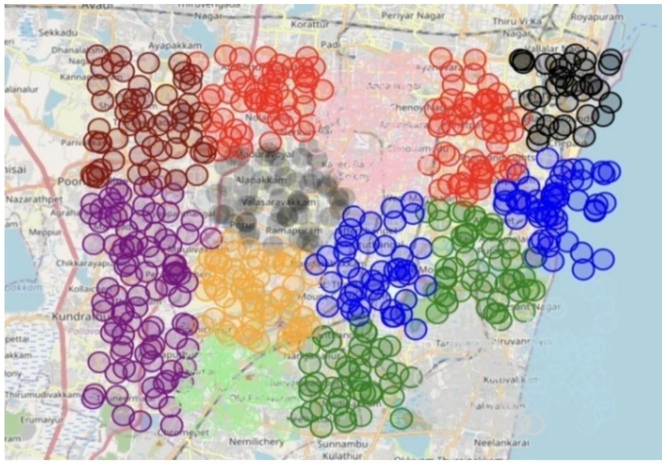


Fig 4. Visualization of Community based Clustering

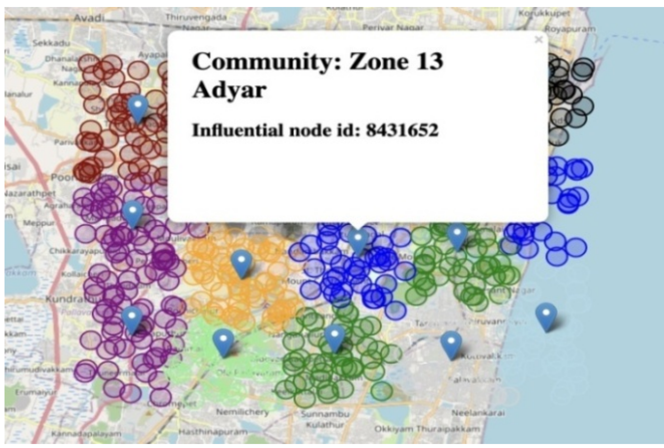


Fig 5. Visualization of Communities with Influential nodes

4 Conclusion and Future work

Covid-19 tweets are categorized using word embedding and topic modeling is done where different subtopics are clustered using K-means algorithm and Random forest model performs better for classifying the emergency tweets which gives the maximum accuracy of 98%. The influential users in each community are identified in order to send an alert about emergency event. The communities are detected using K-mean clustering and it is listed based on the zone wise where influencers are identified using the page rank score and it is visualized using leaflet map. The proposed work can be extended to send an alert to the influencer users even if there is no internet connection in that area. The emergency tracking can be added to the proposed system which will be helpful to the rescue teams.

Acknowledgements

This research work is supported by Science and Engineering Research Board (SERB), Government of India (File No.SPF/2021/000068). One of the authors C.Valliyammai is thankful to SERB for the award of POWER(Promoting Opportunities for Women in Exploratory Research) Fellowship to carry out this research.

References

1) Rosa RL, De Silva MJ, Silva DH, Ayub MS, Carrillo D, Nardelli PH, et al. Event detection system based on user behavior changes in online social networks: Case of the covid-19 pandemic. *IEEE Access*. 2020;8:158806–158825. Available from: https://doi.org/10.1109/ACCESS.2020.3020391.

2) Arbane M, Benlamri R, Brik Y, Alahmar AD. Social media-based COVID-19 sentiment classification model using Bi-LSTM. *Expert System Applications*. 2023;212. Available from: https://doi.org/10.1016/j.eswa.2022.118710.
3) Thilagavathi G, Priyadharshini G, Akshatha M, Boopika AM, Swetha SV. Detection of Social and Newsworthy events using Tweet Analysis. In: Third International Conference on Inventive Research in Computing Applications (ICIRCA), 02-04 September 2021, Coimbatore, India. IEEE. 2021;p. 769–774. Available from: https://ieeexplore.ieee.org/document/9544726.
4) Patel D, Kher V, Desai B, Lei X, Cen S, Nanda N, et al. Machine learning based predictors for COVID-19 disease severity. *Scientific Reports*. 2021;11. Available from: https://doi.org/10.1038/s41598-021-83967-7.
5) Park D, Jang R, Chung MJ, An HJ, Bak S, Choi E, et al. Development and validation of a hybrid deep learning-machine learning approach for severity assessment of COVID-19 and other pneumonias. *Scientific Reports*. 2023;13(1). Available from: https://www.nature.com/articles/s41598-023-40506-w.
6) Xie F, Zhou J, Lee JW, Tan M, Li S, Rajnthern LSO, et al. Benchmarking emergency department prediction models with machine learning and public electronic health records. *Scientific Data*. 2022;9. Available from: https://doi.org/10.1038/s41597-022-01782-9.
7) Syeda HB, Syed M, Sexton KW, Syed S, Begum S, Syed F, et al. Role of Machine Learning Techniques to Tackle the COVID-19 Crisis: Systematic Review. *JMIR Medical Informatics*. 2021;9(1). Available from: https://medinform.jmir.org/2021/1/e23811/.
8) Zoya, Latif S, Shafait F, Latif R. Analyzing LDA and NMF Topic Models for Urdu Tweets via Automatic Labeling. *IEEE Access*. 2021;9:127531–127547. Available from: https://ieeexplore.ieee.org/document/9536731.
9) Yang Y, Wang X, Chen Y, Hu M, Ruan C. A novel centrality of influential nodes identification in complex networks. *IEEE Access*. 2020;8:58742–58751. Available from: https://ieeexplore.ieee.org/document/9045973.
10) Shannag FB, Hammo BH. Lessons learned from event detection from Arabic tweets: the case of Jordan flash floods near dead sea. In: and others, editor. IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), 09-11 April 2019, Amman, Jordan. IEEE. 2019;p. 806–811. Available from: https://ieeexplore.ieee.org/document/8717425.
11) Kumar A, Singh JP, Saumya S. A comparative analysis of machine learning techniques for disaster-related tweet classification. In: and others, editor. IEEE R10 Humanitarian Technology Conference (R10-HTC)(47129), 12-14 November 2019, Depok, West Java, Indonesia. IEEE. 2019;p. 222–227. Available from: https://ieeexplore.ieee.org/document/9042443.
12) Zhang F, Gao W, Fang Y, Zhang B. Enhancing short text topic modeling with FastText embeddings. In: and others, editor. 2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), 12-14 June 2020, Fuzhou, China. IEEE. 2020;p. 255–259. Available from: https://ieeexplore.ieee.org/document/9196423.
13) Devaraj A, Murthy D, Dontula A. Machine-learning methods for identifying social media-based requests for urgent help during hurricanes. *International Journal of Disaster Risk Reduction*. 2020;51. Available from: https://doi.org/10.1016/j.ijdrr.2020.101757.
14) Roy S, Bhowmik A, Ghosh S, Roy B. Statistical Analysis & Categorization for Tweets during Natural Disaster using Classification and Ranking Approach. In: and others, editor. International Conference for Emerging Technology (INCET), 05-07 June 2020, Belgaum, India. IEEE. . Available from: https://ieeexplore.ieee.org/document/9154122.
15) Kumar S, Mallik A, Panda BS. Influence maximization in social networks using transfer learning via graph-based LSTM. *Expert Systems with Applications*. 2023;212. Available from: https://doi.org/10.1016/j.eswa.2022.118770.
16) Ali M, Hassan M, Kifayat K, Kim JY, Hakak S, Khan MK. Social media content classification and community detection using deep learning and graph analytics. *Technological Forecasting and Social Change*. 2023;188. Available from: https://doi.org/10.1016/j.techfore.2022.122252.