Disaster tweet classification: A majority voting approach using machine learning algorithms

Dasari Siva Krishna (sivakrishna.d@gmrit.edu.in)
GMR Institute of Technoogy
Srinivas Gorla
GITAM University
Prasad Reddy P V G D
Andhra University

Research Article

Keywords:

Posted Date: December 8th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2346186/v1

Abstract

Nowadays, People share their opinions through social media. This information may be informative or non-informative. To filtering the informative information from the social media plays a challenging issue. Nevertheless, in social media especially when a disaster been occurs the peoples will interact more on that particular disaster event. They share their opinion through some textual information such as tweets or posts. In this work, we are proposing a generalized approach for categorizing the informative and non-informative on twitter media. We collected the seven natural disaster events from the crisisNLP. These datasets are different disaster events which contains the people's opinions on that specific event. We preprocess the information which converts the tweet information into machine understandable vectors. These vectors been processed by the different machine learning algorithms. We consider the individual performance of each ML algorithm on different disaster datasets upon chosen the best five algorithms for voting techniques. We tested the performance with matrices such as accuracy, precision, recall and f1-score. We compared our results with existing models in which our proposed model performed better than other existing state of art models.

1. Introduction

In General, Twitter is a place where you can share your information to an individual or group of people via tweets. Twitter can use as a source of data in current and historical for social science but it can also be used traditional data repository in social media networks. The present statistics, there are 500 million tweets per day from the 316 million active users in monthly. This tweet information may be informative or non-informative that depends on the tweet text which posted by the individual people opinion. The tweet information during some natural disaster informative tweet text, which helps under many circumstances.

In most of the related works, the collected information either in text or images formats. Most of the related works in present era is existed on textual information related to ethical, legal, retrieving datasets, cost, spam, unknown, representivity[1,2] etc. In Twitter, The tweet information is collected from the tweet APL[6] which provides twitters API, API key, and Bear token. By make use of these credentials we can collect the tweet information by using relevant hashtags. In disasters, many people wants share their opinions or personal information or past experience information related to current crisis or some emergency information. This information category is challenge because tweet text contains some other information like usernames, hashtags, URLs, special characters; even some tweets are in some other languages based on disaster location. In present Covid pandemic, few literatures works on covid tweets to identify the situational tweets using new approaches[9,10] such as BERT transformers and deep learning algorithms with majority voting and stacking approaches.

In Social Media, identification of damage assessment [9, 21] is one of the important aspects. The tweets posted during any disaster the information play a key role in identification of satiation in for place. The information will around in the world so that some impact on daily routine. Moreover, these information process to categorizing informative or not is also important during natural disasters. The table [1] shown detailed tweet information of different disaster events[12,25] in 2017 year. In past, some literatures focus on various damage assessments [20-22] with manual annotators [27, 28] which categorizes tweet information. They also tested with non-disaster datasets used cricket hashtags #indvspak[2]. Nevertheless, they followed a step approach in which the step is manual classification with annotators[22], after that the collected informative information from the tweet text. In some other related work, the authors [3, 8] proposed new framework to extract the situational information by decision making process by the annotators[29]. Moreover, they considered both english and non-english(Hindi language) tweets for various disaster events such as HDBlast, UFlood, SHShoot, Hagupit, NEquake, HDerail.

Table 1: Sample data about crisis disaster event

Tweet information
Informative: • RT @Live5News: New track puts Hurricane Irma near SC coast by Monday as Cat. 2 or 3 https://t.co/111B28ikoF #chsnews https://t.co/WlI46t16Gs • Tropical Storm Harvey is barely moving, causing MAJOR flooding issues for areas like Houston. https://t.co/1WLMmNBcqZ
Non-Informative: • Harrowing footage shows plane flying through Hurricane Irma https://t.co/zwn45Y0YpT https://t.co/pi4FcWK4Ld • RT @_SusanCarroll: Photo from Jonell Soto in South Houston #Harvey https://t.co/v2YB4DeKAL

The detailed summary of the proposed work as follows:

- We collected datasets from CrisisMMD(Multimodal Crisis Dataset), this CrisisMMD consists of seven disaster event datasets namely hurricane irma, hurricane Harvey, Hurricane Maria, California wildfires, ,mexico earthquake , iraq iran earthquake and sri lanka floods.
- Each disaster dataset consists of tweet information such as informative and non information. The damage assessment had been categories into informative based on human damage and infrastructure damage. In human damage assessment which includes the features such as affected individuals, injured or dead people, missing or found people where as infrastructure damage assessment which includes the features infrastructure, utility damage. Some information which was irrelevant such information had been categorized into non-informative information.   In dataset description described in table 3.
- In this work, we pre-process the tweet information by using pre-processing techniques with a pre-trained BERT tokenizer to convert text information to number of words. After preprocesses the tweet text, each tweet information been converted into some fixed length of words and each word size is greater than 2.

- In Next Stage, we applied vectorization techniques which converted words into vector representation using such as countvectorizer, tf-idf, word2ve and glove.
- In final stage, we tested different ML models on each vectorization technique and taken into consider of best top five ML models for voting and also we compared our results to other state of art models.

2. Related Work

Many of the researchers working with disaster events on twitter data. They used different techniques in exploring the data in natural disaster in some of the crisis. In this session, we discuss the numerous conventional methods that have been employed to evaluate the harm caused by natural disasters.

2.1 Machine learning related work

In the event of a natural disaster, the authors Muhammad Imran et al.[1] suggested extracting meaningful information from social media content. To categorize the text, they collected a variety of Twitter datasets. Additionally, they experimented using cricket hashtags #indvspak with non-disaster datasets. However, they used a step-by-step methodology, with the first step being manual classification with annotators, followed by the collection of instructive data from the tweet content. The information was divided into three categories by the annotators: personal, instructive, and other information. The authors Imran M et al.[3] suggested an automatic method for information extraction during disasters. The authors concentrated on informational gems from the microblog post in this work.

Firoj Alam et al.[4] authors collected CrisisMMD Datasets related to multimodel twitter datasets from different natural disasters. In this related work, authors proposed 3 types of annotations which address the useful information about crisis response and damage assessment tasks for different humanitarians. Firoj Alam at al.[5] authors performed analysis on collected three datasets on three events Hurricanes Harvey, Irma, and Maria during natural disasters. The random forest classifier categories 5 labels such as “Very Negative”, “Negative”, “Neutral”, “Positive” and “Very Positive”. The classifier performance 80.7 % accuracy consists of fine-grained sentimental labels for 215,154 phrases in the parse trees of 11,855 sentences.

The CrisisMMD Datasets linked to multimodal Twitter datasets from various natural disasters were gathered by Firoj Alam et al. [4] writers. In a companion article, the authors suggested three alternative sorts of annotations that would provide various humanitarians with relevant details on crisis response and damage assessment activities. Three datasets on the three events Hurricanes Harvey, Irma, and Maria during natural disasters were analyzed by Firoj Alam et al. [5] writers. Five labels, including "Very Negative," "Negative," "Neutral," "Positive," and "Very Positive," are categorized by the random forest classifier. Fine-grained emotive labels for 215,154 phrases in the parse trees of 11,855 sentences make up the classifier's performance of 80.7% accuracy.

2.2 BERT based related works

Bert transformers are the most often used and perform better than DL and ML techniques, according to recent study. There are numerous models, including CT-Bert, Bert, Roberta, and distilbet. Malla, S., and Alphonse et al.'s[10,11] proposal was to use an ensemble approach with voting majority and various bert algorithms , including Bert, Roberta, and CT-Bert, to discover covid information. For the classification of tweeters, Madichetty, S. et al.[12] suggested a neural-based solution utilizing DL methods and a fine-tuned Roberts pre-trained model. Authors compared the outcomes with DL models like CNN, LSTM, BLSTM, and BLSTM with attention.

In our work, we made a new attempt that tokenization with different word embedding techniques such as count vectorizer and TF-IDF which improves the other state of art models. In evolution, we compared the proposed models with different ML classifiers such as LogisticRegression(LR), SupportVectorMachines(SVM), DecisionTree(DT), KNeighborsClassifier(KN), RandomForest(RF), Gaussian Naive Bayes(GNB) and XGBoost (XGB).

3. Dataset Description

We collected disaster event datasets from CrisisNLP. The NLP data is multi-model twitter data (CrisisMMD) which contains seven natural disaster events such as wildfires, floods, earthquakes etc. Each dataset contains the information which is related to particular natural disaster tweet information. The detailed information about each dataset is shown the below table [3]

Table 2
CrisisMMD Datasets

S NO	Disaster name	# tweets	# filtered tweets	#informative tweets	Year
1	Hurricane Irma(HI)	3,517,280	4521	584	2017
2	Hurricane Harvey(HH)	6,664,349	4443	616	2017
3	Hurricane Maria(HM)	2,953,322	4562	422	2017
4	California wildfires(CW)	455,311	1590	365	2017
5	Mexico Earthquake(ME)	383,341	1382	274	2017
6	Iraq-Iran Earthquake(II)	207,729	598	169	2017
7	Sri Lanka Floods(SL)	41,809	1025	90	2017

4. Proposed Work

In this section, we are describing the detailed methodology of the proposed methods to detect the informative over tweet information during the natural disaster. The proposed method implementation as shown in step sequences

4.1 Pre-Processing

The following pre-processing techniques are used on the tweet information before the feature vector generation

- The tweet information was converted into lower case letters
- In text cleaning, we have removed the unwanted text information such as removal of hashtags, URLs, digits, punctuations and stop words
- We also removed the works which having length 2
- Tweet converted to tokens using lemmatization and tokenization process
- Converted the tweet text into fixed length of words by applying post padding technique

4.2 Feature extraction

The main problem with the text is that it is not a fixed length text and unstructured So that to convert variable length to fixed length vectors many new approached been introduced. In feature extraction there are many approached that converts the tokens into the numeric constants. In this work, we applied vector based approaches which provides semantics to the text. In this section, the detailed feature extraction techniques are discussed in below

Bag of words: Machine learning or deep learning models works with numeric values rather than textual data. By using bag-of-words technique we can convert a text into an equivalent vector. The vector is generated based on the dictionary size if word not available then for those words it assigns a fixed value.

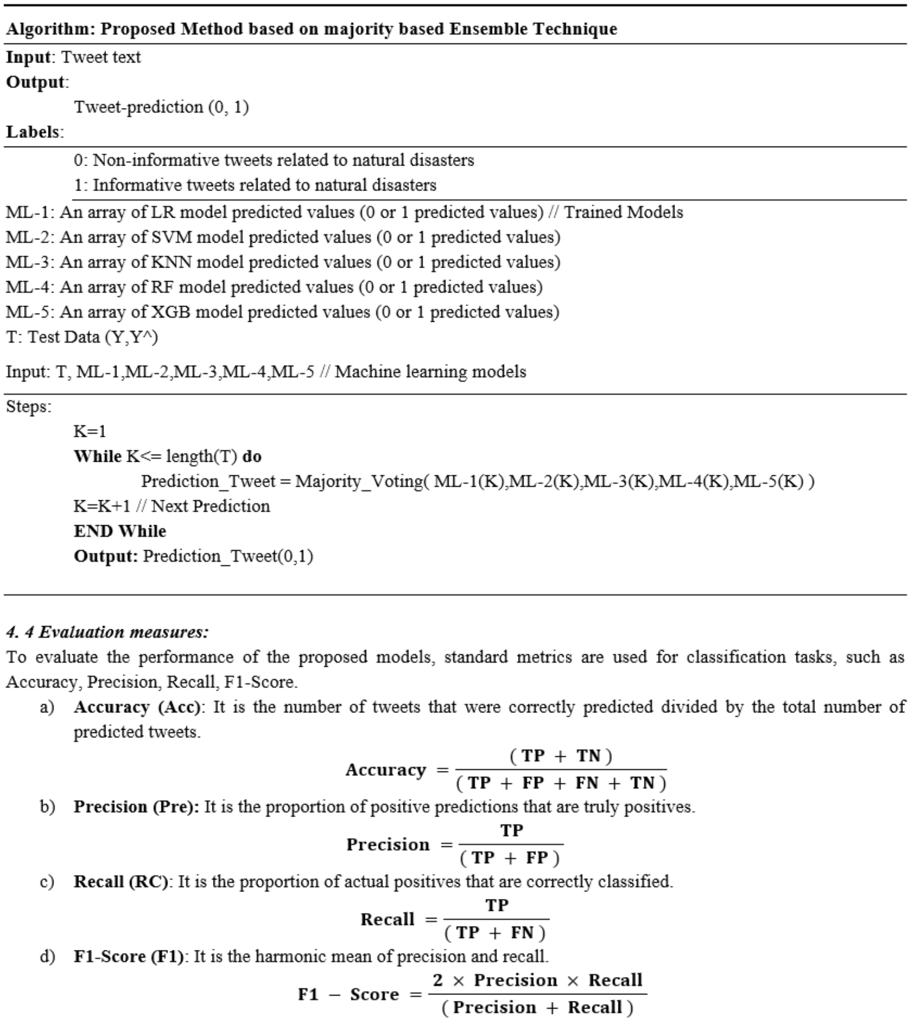
Term frequency and inverse term frequency: In this approach, the weight is given to both word as well as documents. The term frequency refers to number of times word appears in the document divide by the number of words.

Word2vec: In this model, a word is represented [14] as a feature vector in vector space. The algorithm is combination of CBOW and skip-Gram. Initially, for each for a random number is assigned with a large corpus. Each word will iterate form the document and collect the nearest vector to that particular word on either side and concatenate all those vectors then forwarded through linear+ softmax function. It computes the error between actual and estimated values and back propagate the error then modify the weights of the linear layer and also the vectors or neighbor embedding words. Finally, we will extract the hidden layer weights means of words in the vocabulary. We also tested with different parameter out of which the following parameter outperforms with related parameters shown in table [3-9].

Glove: Global Vectors is based on matrix factorization techniques [15] on word-context matrix developed by Stanford University in year 2014. Word-context matrix is a co-occurrence matrix that derives the semantic relationship between words i.e, conditional probability of words coming together in a corpus. We used glove word embedding from kaggle which contains large corpus with 4 different embedding representations in which 6 billion tokens with 50,100,150 and 200 features out of which 100 dimensions shown out performed shown in table [3-9].

4.3 Ensemble approach:

In ensemble approach, it combines the group of set of diverse models in to a single model. This machine learning models gives a better performance than the individual models. There are different ensemble techniques such as bagging [16] and boosting [18, 19]. In proposed model, we have used boosting technique [17] such XGBoost algorithm that combines the multiple decision trees. In this approach, multiple subsets are created from the original dataset. Each subset of data will create a model and model run in parallel which independent to each other. The final result of prediction will be obtained by combining the results of all predicted models.



Algorithm: Proposed Method based on majority based Ensemble Technique

```
Input: Tweet text
Output:
        Tweet-prediction (0, 1)
Labels:
        0: Non-informative tweets related to natural disasters
        1: Informative tweets related to natural disasters
ML-1: An array of LR model predicted values (0 or 1 predicted values) // Trained Models
ML-2: An array of SVM model predicted values (0 or 1 predicted values)
ML-3: An array of KNN model predicted values (0 or 1 predicted values)
ML-4: An array of RF model predicted values (0 or 1 predicted values)
ML-5: An array of XGB model predicted values (0 or 1 predicted values)
T: Test Data (Y,Y^)
Input: T, ML-1,ML-2,ML-3,ML-4,ML-5 // Machine learning models
Steps:
        K=1
        While K<= length(T) do
                Prediction_Tweet = Majority_Voting( ML-1(K),ML-2(K),ML-3(K),ML-4(K),ML-5(K) )
        K=K+1 // Next Prediction
        END While
        Output: Prediction_Tweet(0,1)
```

4. 4 Evaluation measures:

To evaluate the performance of the proposed models, standard metrics are used for classification tasks, such as Accuracy, Precision, Recall, F1-Score.

a) **Accuracy (Acc)**: It is the number of tweets that were correctly predicted divided by the total number of predicted tweets.

$$\mathbf{Accuracy} = \frac{(\mathbf{TP} + \mathbf{TN})}{(\mathbf{TP} + \mathbf{FP} + \mathbf{FN} + \mathbf{TN})}$$

b) **Precision (Pre):** It is the proportion of positive predictions that are truly positives.

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

c) **Recall (RC)**: It is the proportion of actual positives that are correctly classified.

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

d) **F1-Score (F1)**: It is the harmonic mean of precision and recall.

$$\mathbf{F1} - \mathbf{Score} = \frac{\mathbf{2} \times \mathbf{Precision} \times \mathbf{Recall}}{(\mathbf{Precision} + \mathbf{Recall})}$$

4.5 Experimental results and analysis:

4.5.1. Base-line machine learning approaches:

Table 3: Performance of different machine learning algorithms on Dataset 1

Datasets	CountVectorizer				TF-IDF				Word2Vec				Glove			
	Acc	Pre	RC	F1	Acc	Pre	RC	F1	Acc	Pre	RC	F1	Acc	Pre	RC	F1
HI-LR	82.25	80.76	82.43	81.59	80.34	81.32	80.24	80.78	**81.54**	**78.67**	**81.82**	**80.21**	80.11	77.45	80.13	78.77
HI-SVM	**82.43**	**83.35**	**82.86**	**83.10**	**82.5**	**83.56**	**82.67**	**83.11**	80.63	79.36	80.36	79.86	79.12	82.32	79.13	80.69
HI-DT	74.13	78.36	74.23	76.24	79.42	78.65	79.12	78.88	74.46	73.1	74.62	73.85	79.1	78.9	79.3	79.10
HI-KNN	68.35	75.76	68.76	72.09	79.63	76.28	79.98	78.09	81.27	79.48	81.61	80.53	81.73	80.27	81.17	80.72
HI-RF	80.8	79.1	80.1	79.60	82.1	83.25	82.37	82.81	82.36	81.1	82.6	81.84	**83.24**	**83.37**	**83.15**	**83.26**
HI-GNB	59.47	74.79	59.26	66.13	69.15	74.48	69.51	71.91	72.16	74.38	72.48	73.42	72.3	75.84	72.14	73.94
HI-XGB	81.2	79.73	81.59	80.65	**82.5**	**80.84**	**82.16**	**81.49**	80.15	78.58	80.95	79.75	82.95	81.48	82.49	81.98

Evaluation 1: The performance of the dataset-1 Hurricane Irma(HI) with various ML models with different vectorization and word embedding techniques which shown in table [3]. It shown that accuracy was improves such as CountVectorizer with SVM accuracy 82.43%, TF-IDF with SVM & XGB is accuracy 82.50%, word2vec with LR accuracy 81.54%, Glove with RF accuracy 83.24% respectively. In consideration of other matrices such as precision, recall and f1-score performed differently for different algorithms.

Table 4: Performance of different machine learning algorithms on Dataset 2

Datasets	CountVectorizer				TF-IDF				Word2Vec				Glove			
	Acc	Pre	RC	F1	Acc	Pre	RC	F1	Acc	Pre	RC	F1	Acc	Pre	RC	F1
HH-LR	80.31	80.23	78.24	79.22	81.34	85.35	81.23	83.24	80.54	81.23	78.32	79.75	79.12	78.24	79.42	78.83
HH-SVM	81.24	78.78	76.35	77.55	82.25	85.65	82.77	84.19	82.23	82.38	81.72	82.05	80.32	80.83	80.25	80.54
HH-DT	78.87	80.23	78.67	79.44	83.23	83.64	83.67	83.65	76.23	77.02	76.09	76.55	74.37	73.24	74.89	74.06
HH-KNN	56.65	80.47	56.34	66.28	80.36	79.23	80.64	79.93	80.34	79.23	80.82	80.02	80.14	81.63	80.17	80.89
HH-RF	72.23	74.13	70.57	72.31	82.35	84.75	82.23	83.47	83.34	83.12	83.65	83.38	83.76	83.75	83.23	83.49
HH-GNB	59.23	76.42	59.67	67.01	68.32	75.12	68.72	71.78	74.24	77.12	74.41	75.74	75.13	75.32	75.53	75.42
HH-XGB	**82.23**	**78.42**	**82.65**	**80.48**	**85.35**	**85.25**	**85.78**	**85.51**	**84.76**	**84.36**	**84.78**	**84.57**	**83.89**	**82.12**	**83.85**	**82.98**

Evaluation 2: The performance of the dataset-2 Hurricane Harvey(HH)) with various ML models with different vectorization and word embedding techniques which shown in table [4]. It shown that performance was improved with XGB algorithms compared with other ML models.

Table 5: Performance of different machine learning algorithms on Dataset 3

	CountVectorizer				TF-IDF				Word2Vec				Glove			
Datasets	Acc	Pre	RC	F1	Acc	Pre	RC	F1	Acc	Pre	RC	F1	Acc	Pre	RC	F1
MM-LR	81.45	81.89	79.23	80.54	81.22	82.4	81.24	81.82	76.52	75.13	74.72	74.92	75.56	74.23	72.56	73.39
MM-SVM	**82.23**	**81.24**	**80.65**	**80.94**	**82.34**	**82.42**	**82.22**	**82.32**	79.12	78.14	76.41	77.27	76.12	75.12	73.12	74.11
MM-DT	77.13	76.42	75.75	76.08	80.23	79.12	80.12	79.62	68.12	66.89	66.12	66.50	71.14	69.86	69.2	69.53
MM-KNN	68.41	71.56	71.78	71.67	79.8	79.34	79.12	79.23	76.31	75.21	71.2	73.15	72.86	72.24	67.53	69.81
MM-RF	81.6	80.4	79.56	79.98	80.52	80.36	76.42	78.34	**80.13**	**80.56**	**77.14**	**78.81**	76.45	75.34	73.24	74.28
MM-GNB	60.4	67.46	66.3	66.87	73.45	72.12	68.4	70.21	72.24	71.31	71.64	71.47	68.23	67.12	68.2	67.66
MM-XGB	81.21	80.32	79.23	79.77	65.35	66.25	67.45	66.84	79.54	78.75	76.12	77.41	**77.12**	**76.56**	**74.21**	**75.37**

Evaluation 3: The performance of the dataset-3 Hurricane Maria(HM) with various ML models with different vectorization and word embedding techniques which shown in table [5]. It shown that accuracy was improves such as CountVectorizer with SVM accuracy 82.23%, TF-IDF with SVM is accuracy 82.34%, word2vec with RF accuracy 80.13%, Glove with XGB accuracy 77.12% respectively. In consideration of other matrices such as precision, recall and f1-score performed differently for different algorithms.

Table 6: Performance of different machine learning algorithms on Dataset 4

Datasets	CountVectorizer				TF-IDF				Word2Vec				Glove			
	Acc	Pre	RC	F1	Acc	Pre	RC	F1	Acc	Pre	RC	F1	Acc	Pre	RC	F1
CW-LR	79.25	75.51	79.61	77.51	79.14	62.47	79.62	70.01	79.15	77.52	79.37	78.43	75.15	72.62	75.58	74.07
CW-SVM	79.25	76.48	79.58	78.00	**80.63**	**78.47**	**80.15**	**79.30**	79.25	62.25	79.14	69.69	74.62	55.14	74.37	63.33
CW-DT	68.15	72.37	68.48	70.37	75.41	69.47	75.25	72.24	64.51	70.22	64.52	67.25	65.37	66.26	65.74	66.00
CW-KNN	73.37	73.37	73.35	73.36	78.26	72.25	78.62	75.30	79.26	74.26	79.15	76.63	74.85	68.15	74.41	71.14
CW-RF	**80.63**	**77.23**	**80.47**	**78.82**	80.25	78.36	80.36	79.35	**81.25**	**80.14**	**81.41**	**80.77**	**76.15**	**79.15**	**76.25**	**77.67**
CW-GNB	65.32	70.25	65.47	67.78	65.24	69.47	65.1	67.21	69.26	72.17	69.38	70.75	72.83	70.71	72.17	71.43
CW-XGB	79.37	76.74	79.4	78.05	80.27	76.48	80.36	78.37	80.51	80.37	78.51	79.43	75.48	71.27	75.44	73.30

Evaluation 4: The performance of the dataset-4 California wildfires (CW) with various ML models with different vectorization and word embedding techniques which shown in table [6]. It shown that accuracy was improves such as CountVectorizer with RF accuracy 80.63%, TF-IDF with SVM is accuracy 80.63%, word2vec with RF accuracy 81.25%, Glove with RF accuracy 76.15% respectively. In consideration of other matrices such as precision, recall and f1-score performed differently for different algorithms.

Table 7: Performance of different machine learning algorithms on Dataset 5

Datasets	CountVectorizer				TF-IDF				Word2Vec				Glove			
	Acc	Pre	RC	F1	Acc	Pre	RC	F1	Acc	Pre	RC	F1	Acc	Pre	RC	F1
ME-LR	76.6	75.36	76.36	75.86	76.14	80.65	73.36	76.83	73.14	71.64	73.31	72.47	76.57	73.51	76.74	75.09
ME-SVM	**77.47**	**77.53**	**77.83**	**77.68**	78.25	78.58	73.11	75.75	75.14	77.74	75.31	76.51	77.36	83.14	77.57	80.26
ME-DT	70.36	68.83	70.37	69.59	71.27	69.48	71.73	70.59	69.16	68.27	69.48	68.87	70.15	70.15	70.15	70.15
ME-KNN	70.72	66.26	70.27	68.21	74.87	74.38	72.59	73.47	74.58	73.69	74.44	74.06	77.54	76.25	77.58	76.91
ME-RF	74.43	71.37	74.59	72.94	74.48	79.26	74.73	76.93	74.42	73.48	74.38	73.93	**80.16**	**83.2**	**80.15**	**81.65**
ME-GNB	69.15	70.48	69.48	69.98	68.16	69.37	68.58	68.97	66.38	67.73	66.47	67.09	71.63	71.47	71.36	71.41
ME-XGB	75.58	73.52	75.58	74.54	**79.3**	**80.5**	**78.05**	**79.26**	**76.25**	**75.58**	**76.84**	**76.20**	76.14	75.37	77.62	76.48

Evaluation 5: The performance of the dataset-5 Mexico Earthquake(ME) with various ML models with different vectorization and word embedding techniques which shown in table [7]. It shown that accuracy was improves such as CountVectorizer with SVM accuracy 77.47%, TF-IDF with XGB is accuracy 79.30%, word2vec with XGB accuracy 76.25%, Glove with RF accuracy 80.16% respectively. In consideration of other matrices such as precision, recall and f1-score performed differently for different algorithms.

Table 8: Performance of different machine learning algorithms on Dataset 6

Datasets	CountVectorizer				TF-IDF				Word2Vec				Glove			
	Acc	Pre	RC	F1	Acc	Pre	RC	F1	Acc	Pre	RC	F1	Acc	Pre	RC	F1
II-LR	87.12	86.34	88.43	87.37	87.87	88.23	87.32	87.77	86.31	81.63	86.86	84.16	84.23	75.52	84.67	79.83
II-SVM	**88.25**	**89.17**	**88.37**	**88.77**	88.56	90.44	88.35	89.38	87.13	88.56	87.67	88.11	87.75	75.1	87.3	80.74
II-DT	85.24	82.22	85.24	83.70	88.23	87.24	88.78	88.00	82.32	83.35	82.12	82.73	78.12	79.35	78.13	78.74
II-KNN	77.42	80.23	77.22	78.70	88.12	90.42	88.35	89.37	75.21	78.12	75.13	76.60	86.34	81.13	86.78	83.86
II-RF	89.12	90.12	89.14	89.63	88.92	86.23	88.87	87.53	88.56	90.53	88.27	89.39	**88.15**	**87.26**	**88.13**	**87.69**
II-GNB	83.46	83.63	83.76	83.69	85.14	80.75	85.47	83.04	88.14	86.12	88.32	87.21	78.21	84.14	78.12	81.02
II-XGB	85.12	80.24	85.25	82.67	**85.46**	**82.12**	**85.35**	**83.70**	**88.15**	**86.16**	**88.74**	**87.43**	88.13	87.14	88.67	87.90

Evaluation 6: The performance of the dataset-6 Iraq-Iran Earthquake (II) with various ML models with different vectorization and word embedding techniques which shown in table [8]. It shown that accuracy was improves such as CountVectorizer, TF-IDF, word2vec and Glove with RF accuracy 89.12%, 88.92%, 88.56%, 88.15% respectively. In consideration of other matrices such as precision, recall and f1-score performed differently for different algorithms.

Table 9: Performance of different machine learning algorithms on Dataset 7

Datasets	CountVectorizer				TF-IDF				Word2Vec				Glove			
	Acc	Pre	RC	F1	Acc	Pre	RC	F1	Acc	Pre	RC	F1	Acc	Pre	RC	F1
SL-LR	92.26	92.36	92.73	92.54	92.36	92.25	92.26	92.25	92.61	92.22	92.12	92.17	91.25	91.37	91.71	91.54
SL-SVM	93.73	93.58	93.95	93.76	**93.85**	**93.84**	**93.58**	**93.71**	**93.48**	**93.59**	**93.95**	**93.77**	92.27	92.38	92.26	92.32
SL-DT	88.56	86.73	88.58	87.65	91.58	91.52	91.48	91.50	88.51	88.74	88.15	88.44	86.48	86.26	86.47	86.36
SL-KNN	84.47	87.85	84.73	86.26	93.27	93.59	93.28	93.43	82.22	82.47	82.8	82.63	89.17	89.72	89.23	89.47
SL-RF	**95.37**	**95.84**	**95.85**	**95.84**	93.76	93.69	93.37	93.53	92.58	92.95	92.26	92.60	**93.69**	**93.47**	**93.48**	**93.47**
SL-GNB	86.47	87.7	86.36	87.02	91.28	91.69	91.4	91.54	90.48	90.47	90.69	90.58	89.47	89.62	89.26	89.44
SL-XGB	87.6	86.8	82.35	84.52	87.58	88.68	87.25	87.96	90.2	90.25	90.58	90.41	93.36	93.25	93.5	93.37

Evaluation 7: The performance of the dataset-7 Sri Lanka Floods(SL) with various ML models with different vectorization and word embedding techniques which shown in table [7]. It shown that accuracy was improves such as CountVectorizer with RF accuracy 95.37%, TF-IDF with SVM is accuracy 93.85%, word2vec with SVM accuracy 93.38%, Glove with RF accuracy 93.69% respectively. In consideration of other matrices such as precision, recall and f1-score performed differently for different algorithms.

Table 10: Performance of ML models with various embedding techniques

S NO	Data Set	Algorithms	Acc	Pre	RC	F1
1	Hurricane Irma(HI)	Glove -XGB	83.24	83.37	83.15	83.26
2	Hurricane Harvey(HH)	TF-IDF-XGB	85.35	85.25	85.78	85.51
3	Hurricane Maria(HM)	TF-IDF-SVM	82.34	82.42	82.22	82.32
4	California wildfires(CW)	Word2Vec-RF	81.25	80.14	81.41	80.77
5	Mexico Earthquake(ME)	Glove-RF	80.16	83.2	80.15	81.65
6	Iraq-Iran Earthquake(II)	CV-RF	89.12	90.12	89.14	89.63
7	Sri Lanka Floods(SL)	CV-RF	95.37	95.84	95.85	95.84

In the evaluation of different datasets with various ML models performed significantly differ from one to other. The above Table [10] shows that dataset-1 with Glove with XGB algorithms out performed compared with the other ML models. Similarly for remaining datasets 2,3,4,5,6 and 7 TF-IDF with XGB, TF-IDF with SVM, word2Vec with RF, Glove with RF, CV with RF and CV with RF accuracy respectively.

4.5.2 Proposed Methodology

In the proposed method, we consider the best performed vectorization technique along with better performed ML model chosen for the majority voting. We also tested different vectorization with different ML model with significantly not improves. In this work, TF-IDF with LR, SVM, KNN, XGB and RF ML models performance shown in table [11].

Table 11: Performance of proposed work

S NO	Data Set	Acc	Pre	RC	F1
1	Hurricane Irma(HI)	82.35	80.63	82.16	81.39
2	Hurricane Harvey(HH)	85.58	83.16	85.27	84.20
3	Hurricane Maria(HM)	82.26	80.71	79.83	80.27
4	California wildfires(CW)	80.72	81.48	79.84	80.65
5	Mexico Earthquake(ME)	79.92	79.8	78.73	79.26
6	Iraq-Iran Earthquake(II)	88.47	87.41	88.58	87.99
7	Sri Lanka Floods(SL)	93.85	91.15	93.58	92.35

4.5.3 Comparison between proposed methodologies with existing state of art models on CrisisMMD

In this work, We observed that, different ML models on different disaster events performs significantly differ. The table [10] shown that a single model is not achieved better performance than the other models. In the experimental investigation, we tested datasets with different vectorization techniques such as Countvectorizer, TF_IDF, word2vec and glove. In this investigation, all vectorization techniques are significantly differ with different datasets. The table [10] shown that performance of the different ML model with different vectorization techniques on different natural disasters are significantly differ.

Table 12: Comaparsion between proposed work with existing works

Dataset	Authors	Acc	Pre	RC	F1	Year
Hurricane Irma(HI)	Imran et al. [26]	68.04	70.37	68.04	67.04	2014
	HI Rudra et al. [25]	48.51	36.73	48.33	40.07	2018
	Alam et al. [24]	–	70.04	63.4	64.8	2019
	Abinav et al. [23]	–	66	63.4	57.8	2019
	Firoj et al. [22]	–	67.1	68.29	67.6	2020
	Sreenivasulu M et al.[21]	77.13	77.27	77.13	77.1	2021
	Proposed Method	**82.35**	**80.63**	**82.16**	**81.39**	**Present**
Hurricane Harvey(HH)	Imran et al. [26]	65.18	70.41	65.18	62.82	2014
	Rudra et al. [25]	45.18	29.03	43.75	33.69	2018
	HH Alam et al. [24]	–	70.04	63.4	64.8	2019
	Abinav et al. [23]	–	66	63.4	57.8	2019
	Firoj et al. [22]	–	67.1	68.29	67.6	2020
	Sreenivasulu M et al.[21]	77.32	76.14	76.72	76.28	2021
	Proposed Method	**85.58**	**83.16**	**85.27**	**84.20**	**Present**
Hurricane Maria(HM)	Imran et al. [26]	64.68	69.34	64.68	62.43	2014
	Rudra et al. [25]	65	62.17	63.33	61.13	2018
	HM Alam et al. [24]	–	70.04	63.4	64.8	2019
	Abinav et al. [23]	–	66	63.4	57.8	2019
	Firoj et al. [22]	–	67.1	68.29	67.6	2020
	Sreenivasulu M et al.[21]	79.14	79.41	79.14	79.1	2021
	Proposed Method	**82.26**	**80.71**	**79.83**	**80.27**	**Present**
California Wildfires(CW)	Imran et al. [26]	51.48	25.74	50	33.98	2014
	Rudra et al. [25]	55.17	52.5	55.83	50.02	2018
	CW Abinav et al. [23]	–	44.4	46.2	44.8	2019
	Firoj et al. [22]	–	67.1	68.29	67.6	2020
	Sreenivasulu M et al. [21]	73.31	73.78	73.45	73.24	2021
	Proposed Method	**80.72**	**81.48**	**79.84**	**80.72**	**Present**
Mexico Earthquake(ME)	Imran et al. [26]	71.16	73.93	71.16	70.11	2014
	Rudra et al. [25]	55.83	41.67	52.5	43.33	2018
	Mexico Alam et al. [24]	–	70.04	63.4	64.8	2019
	Abinav et al. [23]	–	55.6	63.4	57.8	2019
	Firoj et al. [22]	–	67.1	68.29	67.6	2020
	Sreenivasulu M et al. [21]	79.37	79.9	79.37	79.26	2021
	Proposed Method	**79.92**	**79.8**	**78.73**	**79.26**	**Present**
Iraq–Iran Earthquake(II)	Imran et al. [26]	61.7	30.84	50	38.16	2014
	Rudra et al. [25]	56.67	28.33	50	36	2018
	Iraq–Iran Alam et al. [24]	–	70.04	63.4	64.8	2019
	Abinav et al. [23]	–	55.6	63.4	57.8	2019
	Firoj et al. [22]	–	67.1	68.29	67.6	2020
	Sreenivasulu M et al. [21]	79.58	78.73	78.86	78.48	2021
	Proposed Method	**88.47**	**87.41**	**88.58**	**87.99**	**Present**
Sri Lanka Floods(SL)	Imran et al. [26]	90	92.27	90	89.76	2014
	Rudra et al. [25]	50	35	50	40	2018

	Sri Lanka Alam et al. [24]	–	70.04	63.4	64.8	2019
	Abinav et al. [23]	–	76.5	68	67	2019
	Firoj et al. [22]	–	67.1	68.29	67.6	2020
	Sreenivasulu M et al. [21]	86.92	87.26	86.92	86.89	2021
	Proposed Method	**93.85**	**91.15**	**93.58**	**92.35**	**Present**

In this investigation, the dataset-1(HI) significantly improves 5% rather than the other existing works. For dataset-2(HH), dateset-3(HM), dataset-4(CW),dataset-5(ME),dataset-6(II) and dataset-7(SL) it also improves 7%, 3%, 7%, 0.5%, 10%, 7% accuracy respectively comparing with other state of art models. We evaluated our present methods with best performed ML models with improved vectorization techniques for chosen the ML model for voting. Moreover, We also carefully investigated the tweets while preprocessing the information. This helps to improve the performance of the ML models in other matrices such as precision, recall and F1-score.

Conclusion

In the present, the social media information plays impact on the regular activities of the human life. This information categorizes is also a challenge in present many scenarios. Moreover, particularly when a natural disaster been occurs social media information plays vital role in the society. In this work, we consider the different natural disaster events for identification informative information. This information helps to the people or an individual in know the impact of a particular natural disaster. In this work, we focused on ML models with different vectorization techniques which significantly improve the performance compared with other state of art models.

Limitation And Future Scope

In this work, we consider CrisisMMD dataset which includes the natural disaster in year of 2017. Moreover, these datasets are imbalanced datasets and having duplicates. We faced the problem in preprocessing stage to convert word into the best fit tokens. In this work, we considered the pre-trained word2vec and glove corpus in mapping word to corpus. In the future, we can prepare a corpus for all related disaster event mapping the works into vectors. There was a chance of significant improvement into results while considered into new research works. We also consider into further to explore the research in other aspects while improving the results in other new research interests.

Declarations

Conflict of interest: The authors declare that they have no confict of interest.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors

References

1. Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013, May). Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd international conference on world wide web* (pp. 1021–1024).
2. Rudra, K., Ghosh, S., Ganguly, N., Goyal, P., & Ghosh, S. (2015, October). Extracting situational information from microblogs during disaster events: a classification-summarization approach. In *Proceedings of the 24th ACM international on conference on information and knowledge management* (pp. 583–592).
3. Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013). Extracting information nuggets from disaster-Related messages in social media. Iscram, *201*(3), 791–801.
4. Alam, F., Ofli, F., & Imran, M. (2018, June). Crisismmd: Multimodal twitter datasets from natural disasters. In *Twelfth international AAAI conference on web and social media*.
5. Alam, F., Ofli, F., Imran, M., & Aupetit, M. (2018). A twitter tale of three hurricanes: Harvey, irma, and maria. *arXiv preprint arXiv:1805.05144*.
6. Imran, M., Mitra, P., & Castillo, C. (2016). Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. *arXiv preprint arXiv:1605.05894*.
7. Nguyen, D. T., Al Mannai, K. A., Joty, S., Sajjad, H., Imran, M., & Mitra, P. (2017, May). Robust classification of crisis-related data on social networks using convolutional neural networks. In *Eleventh international AAAI conference on web and social media*.
8. Alam, F., Joty, S., & Imran, M. (2018). Domain adaptation with adversarial training and graph embeddings. *arXiv preprint arXiv:1805.05151*.
9. Nguyen, D. T., Ofli, F., Imran, M., & Mitra, P. (2017, July). Damage assessment from social media imagery data during disasters. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017* (pp. 569–576).
10. Malla, S., & Alphonse, P. J. A. (2022). Fake or real news about COVID-19? Pretrained transformer model to detect potential misleading news. The European Physical Journal Special Topics, 1–10.
11. Malla, S., & Alphonse, P. J. A. (2021). COVID-19 outbreak: an ensemble pre-trained DL model for detecting informative tweets. Applied Soft Computing, *107*, 107495.

12. Madichetty, S., & Sridevi, M. (2021). A neural-based approach for detecting the situational information from Twitter during disaster. IEEE Transactions on Computational Social Systems, *8*(4), 870–880.
13. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
14. Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.
15. Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).
16. Dasari, S.K., Chintada, K.R., Patruni, M. (2018). Flue-Cured Tobacco Leaves Classification: A Generalized Approach Using Deep Convolutional Neural Networks. In: Cognitive Science and Artificial Intelligence. SpringerBriefs in Applied Sciences and Technology(). Springer, Singapore. https://doi.org/10.1007/978-981-10-6698-6_2
17. Dasari, S.K., Prasad, V. A novel and proposed comprehensive methodology using deep convolutional neural networks for flue cured tobacco leaves classification. Int. j. inf. tecnol. **11**, 107–117 (2019). https://doi.org/10.1007/s41870-018-0174-4
18. Sree Ram Kiran Nag, M., Srinivas, G., Venkata Rao, K., Vakkalanka, S., Nagendram, S. (2022). An Efficient Procedure for Identifying the Similarity Between French and English Languages with Sequence Matcher Technique. Lecture Notes on Data Engineering and Communications Technologies, vol 86. Springer, Singapore. https://doi.org/10.1007/978-981-16-5685-9_4.
19. G AppaRao, G Srinivas, K VenkataRao and P V G D Prasad Reddy," Characteristic mining of Mathematical Formulas from Document - A Comparative Study on Sequence Matcher and Levenshtein Distance procedure", International Journal of Computer Sciences and Engineering, Apr 2018, Volume-6, Issue-4, pp 400–403.
20. Madichetty, S. (2021). A stacked convolutional neural network for detecting the resource tweets during a disaster. Multimedia tools and applications, *80*(3), 3927–3949.
21. Madichetty, S., & Sridevi, M. (2021). A novel method for identifying the damage assessment tweets during disaster. Future Generation Computer Systems, *116*, 440–454.
22. Alam, F., Sajjad, H., Imran, M., & Ofli, F. (2021, April). CrisisBench: Benchmarking Crisis-related Social Media Datasets for Humanitarian Information Processing. In *ICWSM* (pp. 923–932).
23. Kumar, A., Singh, J. P., & Saumya, S. (2019, November). A comparative analysis of machine learning techniques for disaster-related tweet classification. In *2019 IEEE R10 Humanitarian Technology Conference (R10-HTC)(47129)* (pp. 222–227). IEEE.
24. Alam, F., Ofli, F., & Imran, M. (2020). Descriptive and visual summaries of disaster events using artificial intelligence techniques: case studies of Hurricanes Harvey, Irma, and Maria. Behaviour & Information Technology, *39*(3), 288–318.
25. Rudra, K., Ganguly, N., Goyal, P., & Ghosh, S. (2018). Extracting and summarizing situational information from the twitter social media during disasters. ACM Transactions on the Web (TWEB), *12*(3), 1–35.
26. Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, Sarah Vieweg, AIDR: Artificial intelligence for disaster response, in: Proceedings of the23rd International Conference on World Wide Web, ACM, 2014, pp.159–162.
27. K.N.Brahmaji Rao, G.Srinivas, P.V.G.D Prasad Reddy., "An Experimental Study with Tensor Flow Characteristic mining of Mathematical Formulae from a Document", EAI Endorsed Transactions on Scalable Information Systems,03 2019–06 2019 | Volume 6 | Issue 21 | e6.
28. K.N.Brahmaji Rao, G.Srinivas, P.V.G.D Prasad Reddy, T.surendra "A Heuristic ranking of different Characteristic mining based Mathematical Formulae retrieval models", Volume-9 Issue-1, October 2019.

Figures

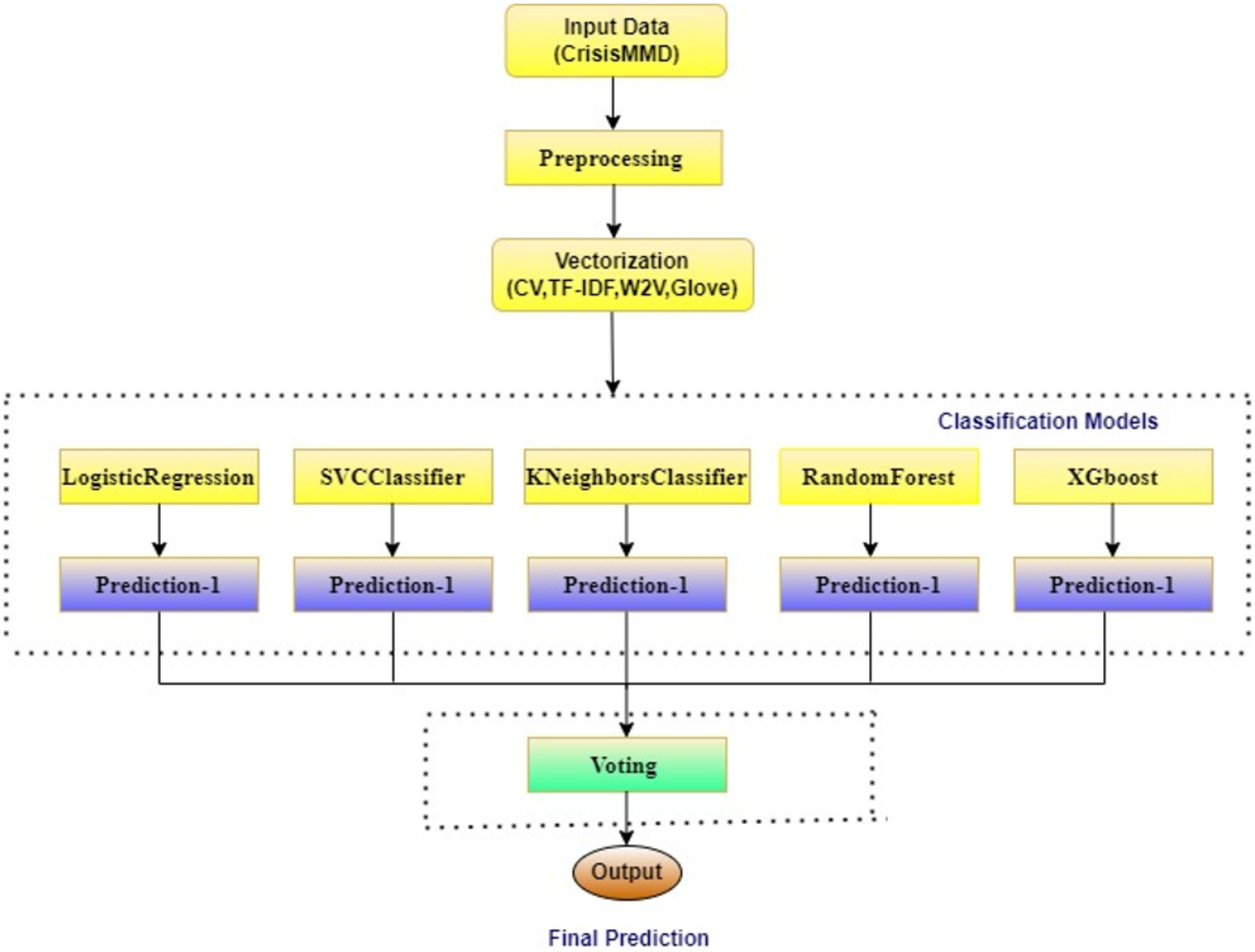


Figure 1

The framework of proposed methods