2019 IEEE Region 10 Humanitarian Technology Conference
Depok, Indonesia | November 12-14, 2019

A Comparative Analysis of Machine Learning Techniques for Disaster-Related Tweet Classification

Abhinav Kumar
Department of CSE
National Institute of Technology Patna
Patna, India
abhinavanand05@gmail.com

Jyoti Prakash Singh
Department of CSE
National Institute of Technology Patna
Patna, India
jps@nitp.ac.in

Sunil Saumya
Department of CSE
IIIT Dharwad
Karnataka, India
sunil.saumya007@gmail.com

***Abstract*—Disaster-related tweets on Twitter during emergencies contain various information about injured or dead people, missing or found people, infrastructure and utility damage that can help government agencies and humanitarian organizations to priorities their help and rescue operations. Because of the huge volume of these tweets, it is essential to construct a model that can classify these tweets into different classes to better organize rescue and relief operations and save lives. In this paper, we have compared various conventional machine learning and deep learning algorithms for classifying disaster-related tweets into six different classes. The models are tested with four different disaster events such as hurricane, earthquake, flood, and wildfire to see the efficiency of the models. The range of F1-score varies from 0.61 to 0.88 for deep neural network-based models whereas it varies from 0.16 to 0.80 for the conventional machine learning classifiers. From this result, it is evident that the deep neural network models are performing significantly well in classifying disaster-related tweets even for imbalanced datasets.**

***Index Terms*—Machine learning, Deep learning, Disaster, Twitter, Tweets**

I. Introduction

Responding to the affected people on time is very important for rescue and relief organizations during natural or man-made disasters. However, this job is very difficult for professional humanitarian communities and government agencies owing to several factors, such as the victim's inaccurate location information, handling an enormous amount of rescue related calls and prioritizing rescue activities based on victim's needs [1]–[5]. The lack of useful information in these emergencies delays the response process [6]. It is found that a huge amount of user-generated data are posted through social media platforms such as Twitter and Facebook during an emergency. People often post their status, report damage to the life and infrastructure, inform about the injured and also ask for assistance through these platforms [2], [3], [6]–[8]. These user-generated pieces of information produced by social networking sites are immensely powerful, rapid and accessible, which can be used to coordinate rescue operation and empower people to become more aware of the situation in case of disaster [9]. A lot of news reports have confirmed that social media plays a crucial role in efforts to assist in disaster relief, finding help and potentially saving lives. For example, a woman was rescued in case of Hurricane Harvey when she tweeted for assistance while the emergency contact number 911 was not in reach[1]. In an American Red Cross survey[2], twenty-eight percent people responded that if emergency contact number (911) was not reachable, they posted their tweets on Twitter for assistance in the event of a disaster.

However, along with the vital information of injured or dead people, missing or found people and damage infrastructures and vehicles, these posts also contain several consolation messages and acknowledgement to different organizations helping them. Because of the huge volume of these posts, it is difficult for humanitarian organizations to manually go through each of the tweets and prioritize rescue and relief activities. This creates an immediate need to develop an intelligent system that can classify tweets into different categories of humanitarian aid. The automatic classification of tweets is very challenging because tweets are restricted to only 280 character space and often includes non-standard abbreviations and spelling mistakes. Because of this, understanding them without adequate context is very hard for the machine learning classifiers [1], [10]. Recently, Several works [3], [6], [9], [11]–[15] have been proposed to classify disaster-related social media texts. Deep neural network-based model [6], [9], [11], [12] are gaining popularity for the task over conventional machine learning techniques [3], [13]–[16]. Convolutional Neural Network (CNN) based models are most popular among the deep learning models while Support Vector Machine, Random Forest, Logistic Regression, Naive Bayes, Decision Tree, and K-NN were most widely used tools in conventional machine learning-based models.

In this work, we have done a comparative analysis of the conventional machine and deep learning-based classifiers for

[1] http://time.com/4921961/hurricane-harvey-twitter-facebook-social-media/

[2] http://www.ehstoday.com/fire_emergencyresponse/communications/red-cross-social-media-help-disaster-0232

978-1-7281-0834-6/19/$31.00 ©2019 IEEE

classifying various disaster-related tweets. We have used seven different conventional machine learning classifiers and five different deep learning models to compare their performances while classifying the tweets into six different classes. We have also considered the cases of data imbalance as they are the norms of the data collected through Twitter. The distribution of data, in general, is imbalanced across the different classes, so it is very important to see the performance of both conventional and deep learning models in the case of a class imbalance. We have used different combinations of 1-gram, 2-gram, and 3-gram TF-IDF features for conventional machine learning while GloVe [17] and Crisis [18] word vector embedding for deep neural networks. The models are trained and tested with four different disaster events datasets: Hurricane, Earthquake, Flood, and Wildfire. The contribution of this paper can be summarized as follows:

1) Implementing seven different conventional machine learning and five different deep learning models and training and testing them with four different disaster-related datasets.
2) Investigating the role of the TF-IDF feature and two different word embedding vectors in the classification task.
3) Comparing the performance of conventional machine learning and deep learning in case of imbalanced data distribution.

The remainder of the paper is structured as follows: Section II describes the detailed methodology, Section III lists the experimental results. The discussion of findings and conclusion of the work is listed in Section IV.

II. Methodology

This section discusses the methodology employed for classifying disaster-related tweets in detail. We have used seven different machine learning classifiers and five different deep neural networks based models for our system development. In case of machine learning the following classifies are used: (i) Support Vector Machine (SVM), (ii) Random Forest (RF), (iii) Logistic Regression (LR), (iv) K-Nearest Neighbors (KNN), (v) Naive Bayes (NB), (vi) Gradient Boosting (GB), (vii) Decision Tree (DT). We have used following models for deep learning-based classifiers: (i) Convolution Neural Network (CNN), (ii) Long-Short-Term-Memory (LSTM), (iii) Gated Recurrent Unit (GRU), (iv) Bi-directional Gated Recurrent Unit (Bi-GRU), (v) Gated Recurrent Unit-Convolution Neural Network (GRU-CNN).

A. Data description and pre-processing

The dataset published by [19] is used in this study that includes tweets related to seven different disasters: (i) Maria Hurricane , (ii) Harvey Hurricane, (iii) Irma Hurricane, (iv) Iran-Iraq earthquake, (v) Mexico earthquake, (vi) Sri Lanka flood, and (vii) California wildfire. The tweets belonging to each of the disasters are labelled with eight different class information: (i) Infrastructure and Utility Damage (IUD), (ii) Vehicle Damage (VD), (iii) Rescue, Volunteering, or Donation Effort (RVDE), (iv) Injured or Dead People (IDP), (v) Affected Individuals (AI), (vi) Missing or Found People (MFP), (vii) Other Relevant Information (ORI), (viii) Not Relevant or Can't Judge (NRCJ). The class information with label ORI and NRCJ are removed from the dataset because it does not convey the specific information related to the disaster. As the dataset is collected in such a manner that, if the tweet has more than one image URLs, then those images were saved with the same tweet text, so the published dataset contains several duplicate tweet text. Duplicate tweet text is then removed from the dataset. Finally, we merged event-specific datasets into one, which means all hurricane datasets (Hurricane Maria, Hurricane Harvey, and Hurricane Irma) are merged into one. Similarly, Mexico and Iran-Iraq earthquake datasets are merged into one. Finally, we come up with four different groups of datasets. The detail description of the dataset after grouping them is listed in Table I. All the punctuation marks such as @, &, #, !, % are removed from the tweets and then the tweets are converted into the lower case. In all cases, 75 percent of the total data sample was used for training and the remaining 25 percent data sample was used to test model performance.

TABLE I
The description of the datasets used in this study

Class	Hurricane	Earthquake	Flood	Wildfire
Affected Individuals (AI)	329	55	14	75
Infrastructure and Utility Damage (IUD)	906	114	35	154
Injured or Dead People (IDP)	159	204	18	105
Missing or Found People (MFP)	15	11	6	8
Rescue Volunteering or Donation Effort (RVDE)	2628	361	124	184
Vehicle Damage (VD)	50	2	0	2
Total	4087	747	197	528

B. Conventional machine learning algorithms

In the case of conventional machine learning algorithms Term Frequency-Inverse Document Frequency (TF-IDF) vector is used as the input to the classifiers. All the possible combinations of 1-gram, 2-gram, and 3-gram TF-IDF features were extracted to experiment with all classifiers.

C. Deep Neural Networks

In the case of the deep neural network, two different pre-trained word vectors GloVe [17] and Crisis [18] are used. In the case of GloVe, 100-dimensional word vector embedding whereas in case of Crisis embedding 300-dimensional word vector embedding is used. The advantage of using a pre-trained word vector is that it reduces the computational overhead of the model. The categorical cross-entropy and Adam is used as the loss function and optimizer respectively for each of the neural network models. For all the hidden layer, we have used ReLU activation function whereas, at the output layer, we have used the softmax activation function in each of the neural networks. The categorical cross-entropy loss, softmax and ReLU activation functions can be defined by Equation 1, Equation 2, Equation 3 respectively.

$$\text{Categorical cross entropy} = -\sum_{i=1}^{N} p_i log(\hat{p}_i) \tag{1}$$

$$\text{Softmax function} = \frac{e^{x_j}}{\sum_{k=1}^{C} e^{x_k}} \tag{2}$$

$$\text{ReLU} = max(0, x_j) \tag{3}$$

where, N is the number of training samples in a batch, $\hat{p}_i$ represents the predicted class probability for the i^{th} training sample and p_i represents the classes in the form of one-hot vector. The role of softmax is to calculate the probabilities of each target class across all possible classes. In the Equation 2, C represents the number of classes and x_j represents a real value calculated at the layer. The ReLU function defined in Equation 3 means for a negative value of x_j it return zero whereas for a positive value of x_j it return x_j itself. The detailed model configuration and hyper-parameter settings can be seen from Table II.

III. Result

The models are implemented at Google Collaboratory[3] with 12 GB NVIDIA Tesla K80 GPU. The conventional machine learning and deep learning models are implemented using scikit-learn[4] and keras[5] python libraries with Tensorflow as back-end respectively.

The performance of the models is measured in terms of Precision (P), Recall (R) and F_1-score (F_1). For the conventional machine learning algorithms, different combinations of 1-gram, 2-gram, and 3-gram TF-IDF features were used. The extensive experiments were performed with each of the disaster-related datasets: (i) Hurricane, (ii) Earthquake, (iii) Flood, and (iv) Wildfire. The results of each of the classifiers with a different combination of the N-gram TF-IDF feature with the mentioned disasters are listed in Table III. Among seven different classifiers: SVM, RF, LR, KNN, NB, GB, and DT, Gradient Boosting (GB) classifier performed best in the case of all the disaster-related events whereas SVM performed worst. The Gradient Boosting classifier achieved an F_1-score of 0.79, 0.80, 0.70, and 0.67 for Hurricane, Earthquake, Flood, and wildfire events respectively.

Next, the experiments were performed with deep neural networks where two different word embedding vectors GloVe and Crisis are used. In the case of Earthquake, Crisis embedding performed best, in case of wildfire GloVe embedding performed best. In the case of the hurricane, both GloVe and Crisis gave comparable results whereas in case of flood both the embedding techniques gave a mixed performance as can be seen from Table IV. The comparison of all the conventional machine learning and deep learning classifiers for Hurricane, Earthquake, Flood, and Wildfire are plotted in Fig. 1, Fig. 2, Fig. 3, and Fig. 4 respectively. As can be seen from the figures, throughout all the disaster events deep neural network-based models outperformed the conventional machine learning techniques. In the case of Hurricane, Bi-directional Gated Recurrent Unit (Bi-GRU) performed best. In the case of Earthquake, the combination of the Gated Recurrent Unit and Convolution Neural Network (GRU-CNN) performed best. In the case of Flood, Long-Short-Term-Memory (LSTM) and Gated Recurrent Unit (GRU) both equally performed best, whereas, in the case of Wildfire, Gated Recurrent Unit with Convolution Neural Network (GRU-CNN) performed best.

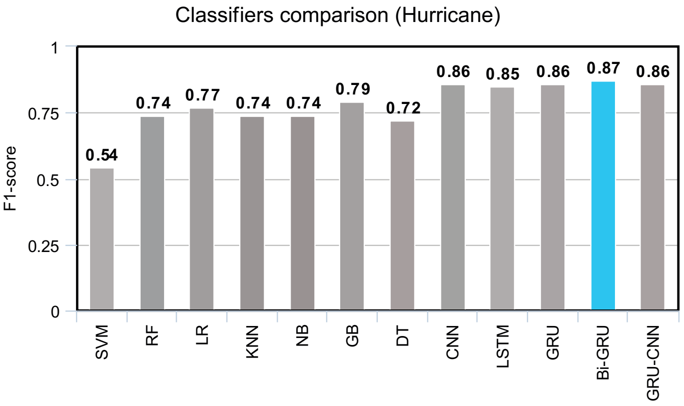

Fig. 1. Classifiers comparison in case of Hurricane dataset

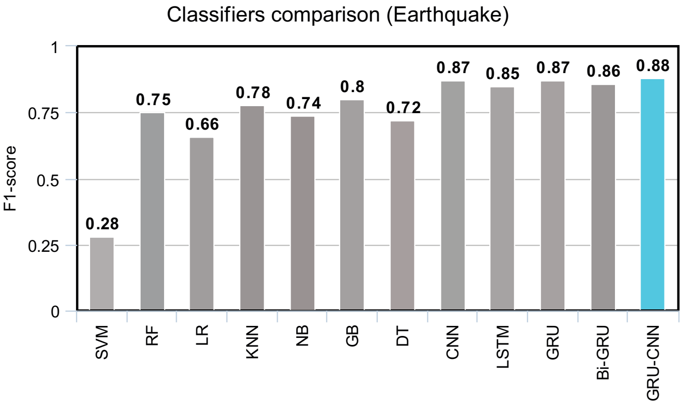

Fig. 2. Classifiers comparison in case of Earthquake dataset

The data distribution between the different classes is imbalance in nature as can be seen from Table I. This research also compares the effectiveness of conventional machine learning with deep learning networks to see how classifiers perform across each class. For each of the disaster events Hurricane, Earthquake, Flood, and Wildfire, the result of best performing conventional machine and deep learning are shown in Table V, Table VI, Table VII, and Table VIII respectively. In the case of the hurricane (see Table I), the minority classes AI, IDP, MFP, and VD have 329, 159, 15, and 50 samples respectively. As can be seen from Table V, even in the case of fewer data samples, the deep learning-based model has learned

[3] https://colab.research.google.com

[4] https://scikit-learn.org

[5] https://keras.io

TABLE II
THE DETAILED MODEL CONFIGURATION AND HYPER-PARAMETER SETTINGS FOR EACH NEURAL NETWORK ARCHITECTURE

Models	#Layers	#Units	#Filters	Filter size	Pooling window	#Dense layer	#Neuron (Dense)	Activation	Learning rate	Optimizer	Loss	Batch size	Epochs
CNN	2	-	128	2,3,4	5 (Max Pooling)	2	128,6	ReLU, Softmax	0.001	Adam	Categorical crossentropy	4	200
LSTM	2	128	-	-	-	2	32,6	ReLU, Softmax	0.001	Adam	Categorical crossentropy	4	200
GRU	2	128	-	-	-	2	64,6	ReLU, Softmax	0.001	Adam	Categorical crossentropy	4	200
Bi-GRU	2	128	-	-	-	2	64,6	ReLU, Softmax	0.001	Adam	Categorical crossentropy	4	200
Bi-GRU-CNN	1,1	128	128	3	Global Avg. Pooling	2	64,6	ReLU, Softmax	0.001	Adam	Categorical crossentropy	4	200

TABLE III
RESULTS FOR VARIOUS MACHINE LEARNING CLASSIFIER FOR DIFFERENT DISASTER

Model	N-Gram	Hurricane			Earthquake			Flood			California wildfire		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
SVM	1	0.45	0.67	0.54	0.21	0.45	0.28	0.44	0.66	0.52	0.11	0.33	0.16
	2	0.45	0.67	0.54	0.21	0.45	0.28	0.44	0.66	0.52	0.11	0.33	0.16
	3	0.45	0.67	0.54	0.21	0.45	0.28	0.44	0.66	0.52	0.11	0.33	0.16
	1,2	0.45	0.67	0.54	0.21	0.45	0.28	0.44	0.66	0.52	0.11	0.33	0.16
	2,3	0.45	0.67	0.54	0.21	0.45	0.28	0.44	0.66	0.52	0.11	0.33	0.16
	1,2,3	0.45	0.67	0.54	0.21	0.45	0.28	0.44	0.66	0.52	0.11	0.33	0.16
Random Forest	1	0.74	0.77	0.74	0.74	0.77	0.75	0.56	0.64	0.59	0.46	0.47	0.45
	2	0.63	0.68	0.64	0.61	0.62	0.54	0.60	0.68	0.57	0.46	0.45	0.41
	3	0.57	0.64	0.58	0.71	0.55	0.44	0.60	0.68	0.57	0.55	0.42	0.38
	1,2	0.73	0.76	0.73	0.66	0.70	0.67	0.50	0.66	0.56	0.55	0.55	0.52
	2,3	0.62	0.68	0.64	0.50	0.58	0.49	0.61	0.70	0.60	0.52	0.46	0.43
	1,2,3	0.68	0.72	0.68	0.74	0.68	0.61	0.60	0.68	0.57	0.53	0.53	0.51
Logistic Regression	1	0.82	0.81	0.77	0.74	0.72	0.66	0.60	0.68	0.57	0.61	0.65	0.60
	2	0.66	0.70	0.61	0.51	0.62	0.53	0.44	0.66	0.52	0.53	0.53	0.48
	3	0.55	0.67	0.54	0.51	0.53	0.42	0.44	0.66	0.52	0.37	0.35	0.25
	1,2	0.73	0.78	0.72	0.72	0.67	0.58	0.44	0.66	0.52	0.60	0.62	0.57
	2,3	0.62	0.69	0.58	0.52	0.60	0.51	0.44	0.66	0.52	0.48	0.45	0.39
	1,2,3	0.73	0.76	0.70	0.52	0.66	0.56	0.44	0.66	0.52	0.59	0.58	0.53
KNN	1	0.75	0.74	0.74	0.79	0.78	0.78	0.64	0.62	0.62	0.65	0.61	0.62
	2	0.73	0.72	0.72	0.69	0.66	0.67	0.59	0.64	0.62	0.61	0.55	0.57
	3	0.65	0.65	0.65	0.64	0.61	0.61	0.53	0.52	0.51	0.46	0.41	0.42
	1,2	0.74	0.74	0.73	0.78	0.75	0.75	0.62	0.62	0.61	0.61	0.57	0.59
	2,3	0.70	0.69	0.69	0.68	0.65	0.65	0.57	0.60	0.59	0.52	0.48	0.50
	1,2,3	0.73	0.72	0.72	0.71	0.67	0.68	0.56	0.62	0.58	0.63	0.60	0.61
Naive Bayes	1	0.68	0.68	0.68	0.70	0.70	0.70	0.61	0.60	0.60	0.60	0.61	0.60
	2	0.71	0.71	0.71	0.72	0.68	0.70	0.62	0.56	0.59	0.61	0.49	0.53
	3	0.70	0.56	0.61	0.72	0.55	0.62	0.61	0.40	0.47	0.58	0.42	0.46
	1,2	0.72	0.76	0.73	0.75	0.76	0.74	0.58	0.60	0.58	0.63	0.64	0.63
	2,3	0.71	0.71	0.71	0.71	0.69	0.69	0.63	0.56	0.59	0.58	0.48	0.51
	1,2,3	0.73	0.76	0.74	0.76	0.76	0.74	0.59	0.60	0.59	0.66	0.66	0.65
Gradient Boosting	1	0.78	0.80	**0.79**	0.81	0.81	**0.80**	0.70	0.72	**0.70**	0.64	0.59	0.61
	2	0.68	0.68	0.65	0.63	0.60	0.56	0.44	0.54	0.48	0.55	0.48	0.47
	3	0.62	0.68	0.59	0.52	0.53	0.46	0.43	0.64	0.52	0.44	0.43	0.37
	1,2	0.80	0.79	**0.79**	0.80	0.79	0.78	0.62	0.62	0.59	0.66	0.67	**0.67**
	2,3	0.67	0.70	0.66	0.61	0.60	0.55	0.43	0.62	0.51	0.58	0.56	0.54
	1,2,3	0.80	0.78	**0.79**	0.80	0.78	0.78	0.62	0.64	0.57	0.66	0.59	0.62
Decision Tree	1	0.73	0.70	0.71	0.70	0.68	0.69	0.61	0.58	0.58	0.49	0.48	0.46
	2	0.66	0.61	0.62	0.64	0.55	0.57	0.50	0.32	0.35	0.45	0.41	0.41
	3	0.57	0.58	0.57	0.50	0.51	0.46	0.60	0.68	0.60	0.41	0.39	0.31
	1,2	0.72	0.70	0.71	0.75	0.67	0.69	0.56	0.60	0.56	0.55	0.52	0.52
	2,3	0.64	0.58	0.60	0.66	0.50	0.54	0.45	0.32	0.35	0.45	0.42	0.41
	1,2,3	0.72	0.72	0.72	0.74	0.71	0.72	0.55	0.54	0.53	0.54	0.44	0.45

TABLE IV
RESULTS OF THE VARIOUS DEEP NEURAL NETWORK BASED MODEL

Model	Embedding	Hurricane			Earthquake			Flood			California wildfire		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
CNN	Crisis	0.84	0.85	0.84	0.87	0.87	0.87	0.63	0.74	0.68	0.70	0.65	0.66
	Glove	0.86	0.86	0.86	0.77	0.83	0.80	0.74	0.78	0.76	0.76	0.78	0.77
LSTM	Crisis	0.85	0.85	0.84	0.89	0.83	0.85	0.68	0.78	0.72	0.60	0.66	0.61
	Glove	0.85	0.85	0.85	0.83	0.82	0.83	0.77	0.84	**0.80**	0.74	0.75	0.74
GRU	Crisis	0.84	0.86	0.84	0.88	0.87	0.87	0.83	0.80	**0.80**	0.74	0.76	0.74
	Glove	0.86	0.86	0.86	0.87	0.83	0.85	0.61	0.72	0.66	0.79	0.80	0.79
Bi-directional GRU	Crisis	0.86	0.88	**0.87**	0.86	0.86	0.86	0.76	0.78	0.75	0.71	0.75	0.72
	Glove	0.87	0.87	**0.87**	0.83	0.83	0.83	0.69	0.74	0.70	0.78	0.78	0.78
GRU-CNN	Crisis	0.86	0.87	0.86	0.88	0.88	**0.88**	0.65	0.76	0.70	0.78	0.78	0.78
	Glove	0.86	0.87	0.86	0.82	0.83	0.82	0.70	0.76	0.73	0.82	0.80	**0.81**

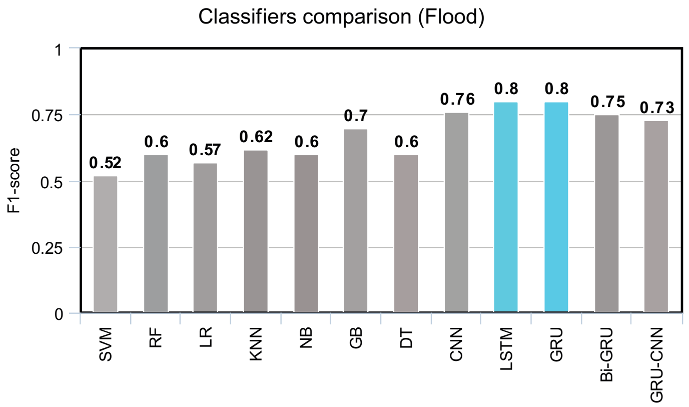

Fig. 3. Classifiers comparison in case of Flood dataset

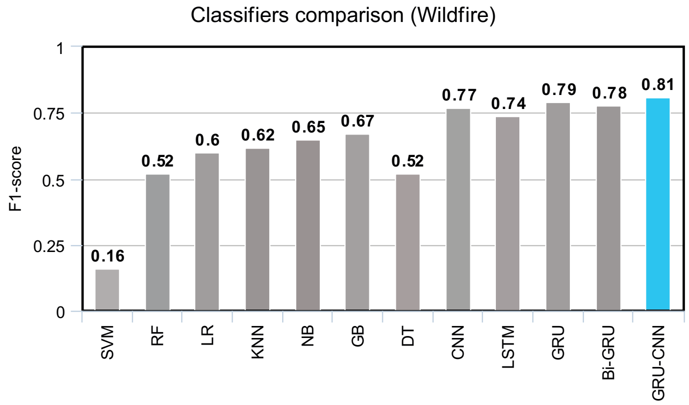

Fig. 4. Classifiers comparison in case of wildfire dataset

better than conventional machine learning models. A similar observation can be seen in Table VI, Table VII, and Table VIII for earthquake, flood, and wildfire events respectively. These findings are evidence that the deep neural network learned better than conventional machine learning classifiers in the situation of imbalanced data distribution.

TABLE V
BEST PERFORMING DEEP AND CONVENTIONAL MACHINE LEARNING MODELS FOR HURRICANE

Hurricane: Deep Learning vs Conventional Machine Learning							
Class	Bi-GRU			Gradient Boosting			# data sample
	P	R	F_1	P	R	F_1	
AI	0.52	0.41	0.46	0.35	0.15	0.22	71
IUD	0.81	0.84	0.82	0.77	0.65	0.70	213
IDP	0.97	0.67	0.79	0.91	0.71	0.80	45
MFP	0.00	0.00	0.00	0.00	0.00	0.00	2
RVDE	0.91	0.95	0.93	0.83	0.93	0.88	683
VD	1.00	0.50	0.67	0.45	0.62	0.53	8
Weighted Avg.	0.87	0.87	0.87	0.78	0.80	0.79	1022

TABLE VI
BEST PERFORMING DEEP AND CONVENTIONAL MACHINE LEARNING MODELS FOR EARTHQUAKE

Earthquake: Deep Learning vs Conventional Machine Learning							
Class	GRU-CNN			Gradient Boosting			# data sample
	P	R	F_1	P	R	F_1	
AI	0.56	0.45	0.50	0.50	0.18	0.27	11
IUD	0.81	0.81	0.81	0.85	0.59	0.70	37
IDP	0.91	0.91	0.91	0.92	0.85	0.88	53
MFP	0.50	1.00	0.67	0.00	0.00	0.00	1
RVDE	0.93	0.94	0.94	0.78	0.98	0.86	85
VD	0.00	0.00	0.00	0.00	0.00	0.00	0
Weighted Avg.	0.88	0.88	0.88	0.81	0.81	0.80	187

TABLE VII
BEST PERFORMING DEEP AND CONVENTIONAL MACHINE LEARNING MODELS FOR FLOOD

Flood: Deep Learning vs Conventional Machine Learning							
Class	GRU			Gradient Boosting			# data sample
	P	R	F_1	P	R	F_1	
AI	0.50	1.00	0.67	0.00	0.00	0.00	2
IUD	0.56	0.62	0.59	0.38	0.38	0.38	8
IDP	1.00	0.60	0.75	1.00	0.60	0.75	5
MFP	1.00	0.50	0.67	1.00	1.00	1.00	2
RVDE	0.88	0.88	0.88	0.76	0.85	0.80	33
Weighted Avg.	0.83	0.80	0.80	0.70	0.72	0.70	50

TABLE VIII
BEST PERFORMING DEEP AND CONVENTIONAL MACHINE LEARNING MODELS FOR WILDFIRE

Wildfire: Deep Learning vs Conventional Machine Learning							
Class	GRU-CNN			Gradient Boosting			# data
	P	R	F_1	P	R	F_1	sample
AI	0.46	0.65	0.54	0.23	0.18	0.20	17
IUD	0.81	0.89	0.85	0.69	0.75	0.72	44
IDP	0.95	0.77	0.85	0.91	0.81	0.86	26
MFP	0.00	0.00	0.00	0.00	0.00	0.00	1
RVDE	0.92	0.84	0.88	0.68	0.74	0.71	43
VD	0.00	0.00	0.00	0.00	0.00	0.00	1
Weighted Avg.	0.82	0.80	0.81	0.66	0.67	0.67	132

IV. Discussion and Conclusion

This work classifies the disaster-related tweets into six different classes that can help in humanitarian tasks. Several conventional machines and deep learning techniques are compared in the paper. In the case of conventional machine learning different combinations of N-gram TF-IDF feature vector whereas in case of deep learning two different GloVe and Crisis word embedding vectors are used. In the case of conventional machine learning, Gradient Boosting (GB) classifier outperformed all other classifiers in all the disaster events as can be seen from Table III. In the case of GB, the 1-gram TF-IDF feature is played a major role in classifying tweets. Gradient Boosting classifier performed significantly well when 1-gram, 1,2-gram, and 1,2,3-gram TF-IDF features are used in the comparison of 2-gram, 3-gram, and 2,3-gram TF-IDF features as can be seen from Table III. That means in classifying tweets 1-gram features are one of the important features in the case of conventional machine learning classifiers.

The use of different embedding played a major role in the classification task, Crisis embedding performed best in case of earthquake and GloVe embedding performed best in case of wildfire. So this study suggests future researchers to choose appropriate word embedding vectors to achieve better performance. The implemented deep neural network-based models outperformed all the conventional machine learning classifiers even in the case of data imbalance. The models were tested on four different categories of the datasets where event-specific datasets are merged into one. All the merged event-specific datasets belong to different geographical regions. Considering that the deep learning models have performed well in the case of datasets belonging to different geographical regions, this research also shows that different geographic event datasets can be used for a similar type of event where very limited labelled data of the current event are available.

The limitation of this work is that we have used only English language tweets for the classification task whereas during disaster several tweets are posted by the users in their regional languages. So a multi-lingual system can be made for future work. One other limitation is that for all the deep neural network-based model we performed the experiments with fixed batch size, learning rate, and optimizer. In the future, these hyper-parameters can be tuned further to get better performance from the deep neural models.

References

[1] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, "Processing social media messages in mass emergency: A survey," *ACM Computing Surveys (CSUR)*, vol. 47, no. 4, p. 67, 2015.

[2] A. Kumar and J. P. Singh, "Location reference identification from tweets during emergencies: A deep learning approach," *International Journal of Disaster Risk Reduction*, vol. 33, pp. 365–375, 2019.

[3] J. P. Singh, Y. K. Dwivedi, N. P. Rana, A. Kumar, and K. K. Kapoor, "Event classification and location prediction from tweets during disasters," *Annals of Operations Research*, May 2017. [Online]. Available: https://doi.org/10.1007/s10479-017-2522-3

[4] A. Singh, U. K. Bera, and D. Sarma, "Two stages post-disaster humanitarian logistics," in *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, Dec 2017, pp. 259–262.

[5] D. Sarma, U. K. Bera, A. Singh, and M. Maiti, "A multi-objective post-disaster relief logistic model," in *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, Dec 2017, pp. 205–208.

[6] D. T. Nguyen, S. Joty, M. Imran, H. Sajjad, and P. Mitra, "Applications of online deep learning for crisis response using social media information," *arXiv preprint arXiv:1610.01030*, 2016.

[7] A. Kumar, J. P. Singh, and N. P. Rana, "Authenticity of geo-location and place name in tweets," in *Proceedings of the 23rd americas conference on information systems*, 2017.

[8] A. Kumar and N. C. Rathore, "Relationship strength based access control in online social networks," in *Proceedings of International Conference on Information and Communication Technology for Intelligent Systems: Volume 2*. Cham: Springer International Publishing, 2016, pp. 197–206.

[9] C. Caragea, A. Silvescu, and A. H. Tapia, "Identifying informative messages in disaster events using convolutional neural networks," in *International Conference on Information Systems for Crisis Response and Management*, 2016, pp. 137–147.

[10] D. T. Nguyen, F. Alam, F. Ofli, and M. Imran, "Automatic image filtering on social networks using deep learning and perceptual hashing during crises," *arXiv preprint arXiv:1704.02602*, 2017.

[11] M. Yu, Q. Huang, H. Qin, C. Scheele, and C. Yang, "Deep learning for real-time social media text classification for situation awareness using hurricanes sandy, harvey, and irma as case studies," *International Journal of Digital Earth*, vol. 0, no. 0, pp. 1–18, 2019. [Online]. Available: https://doi.org/10.1080/17538947.2019.1574316

[12] D. Nguyen, K. A. A. Mannai, S. Joty, H. Sajjad, M. Imran, and P. Mitra, "Robust classification of crisis-related data on social networks using convolutional neural networks," 2017. [Online]. Available: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15655

[13] Q. Huang and Y. Xiao, "Geographic situational awareness: mining tweets for disaster preparedness, emergency response, impact, and recovery," *ISPRS International Journal of Geo-Information*, vol. 4, no. 3, pp. 1549–1568, 2015.

[14] H. Li, N. Guevara, N. Herndon, D. Caragea, K. Neppalli, C. Caragea, A. C. Squicciarini, and A. H. Tapia, "Twitter mining for disaster response: A domain adaptation approach." in *Proceedings of the 12th ISCRAM Conference*, 2015.

[15] Z. Ashktorab, C. Brown, M. Nandi, and A. Culotta, "Tweedr: Mining twitter to inform disaster response." in *Proceedings of the 11th ISCRAM Conference*, 2014, pp. 354–358.

[16] C. Caragea, N. Mcneese, A. Jaiswal, G. Traylor, H. woo Kim, P. Mitra, D. Wu, A. H. Tapia, L. Giles, B. J. Jansen, and J. Yen, "Classifying text messages for the haiti earthquake," in *Proceedings of the 8th International Conference on Information Systems for Crisis Response and Management (ISCRAM2011)*, 2011.

[17] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[18] M. Imran, P. Mitra, and C. Castillo, "Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, May 2016.

[19] F. Alam, F. Ofli, and M. Imran, "Crisismmd: Multimodal twitter datasets from natural disasters," in *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM)*, June 2018.