A Neural-Based Approach for Detecting the Situational Information From Twitter During Disaster

Article *in* IEEE Transactions on Computational Social Systems · April 2021

DOI: 10.1109/TCSS.2021.3064299

2 authors:

Sreenivasulu Madichetty
National Institute of Technology Tiruchirappalli

Sridevi Muthukumarasamy
National Institute of Technology

A Neural-Based Approach for Detecting the Situational Information From Twitter During Disaster

Sreenivasulu Madichetty and Sridevi M

***Abstract*— Twitter is widely considered an essential social media used during a disaster. For the past five years, there has been a great surge in the use of Twitter during the disaster. A large amount of useful information is posted on Twitter during an emergency, along with the users' sympathies and opinions. Therefore, a reliable methodology is needed for extracting useful information from the posted tweets during the disaster. This article is focused on identifying the situational tweets during a disaster. A neural-based approach is developed based on the combination of the RoBERTa model and feature-based method for identifying the situational tweets during a disaster. Extensive experiments are performed on various disaster data sets such as Typhoon Hagupit, Hyderabad bomb blast, Sandy Hook shooting, Nepal earthquake, and HarDerail accident disaster data set. The proposed method is compared with various deep learning models such as convolutional neural network (CNN), long short term memory (LSTM), bidirectional long short term memory (BLSTM), and bidirectional long short term memory with attention (BLSTM attention). Experimental results demonstrate that the proposed method outperforms the existing methods on different disaster data sets.**

***Index Terms*— Disasters and Twitter, RoBERTa model, situational tweets.**

I. Introduction

NUMEROUS studies [1]–[7] show that people use social media [8] to get the situational awareness (a broad overview of what happens in an event) during disasters. Various people are posting distinct categories of disaster-related posts on Social media. It may be described as sympathy for victims, prayers for people, resource requests [9], [10], useful information [11], spam tweets [12], [13], emotion information [14]–[16], etc. Situational awareness information can be defined as information that allows high-level concerned authorities to understand the situation better during a disaster. It is essential for humanitarian organizations to prepare the relief efforts accordingly, as time is critical during a disaster, especially at the initial stage of the disaster. It gives a broad picture of the disaster, where the resources and medical facilities are needed, and the severity of the disaster. Among them, only a specific category of tweets is essential to humanitarian organizations/governmental organizations contributing to situational awareness during a crisis. Therefore, Verma *et al.* [17] identified the situational tweets by differentiating with nonsituational tweets. The situational tweets include victims, dead people, and availability/requirement of resources. Also, tweets that provide realistic, actionable information [18], [19] lead to situational awareness. These tweets include content that shows knowledge of the severity of the crisis and specific details of the situation. The authors used a feature-based approach to detect situational tweets during a crisis. Later, Rudra *et al.* [20] developed a feature-based approach with the use of an SVM classifier for identifying the situational tweets in both English and Hindi language tweets. However, all existing works have used feature-based methods to detect situational tweets during a disaster.

In a realistic situation, both situational and nonsituational information may exist in a single situational tweet instead of only situational information. These categories of tweets are termed to be raw tweets for this work. Examples of a raw tweet are included in Table I. It contains both situational and nonsituational tweets. The first two situational tweets provide information on the number of casualties or injuries (contribute to the awareness of the situation, which indicates that help is required from the medical organizations) during a disaster. Nonsituational tweets do not contribute to any situational information. It includes tweets about user views and sympathies.

The authors in [20] split the raw tweets into multiple fragment tweets based on exclamatory symbol (**!**), full stop (**.**), and question mark (**?**). They showed that fragments are useful for better summarization of tweets. Examples of fragment tweets are included in Table II. They designed a feature-based approach for identifying both situational raw tweets and fragment tweets. However, their method does not give much performance for identifying the situational raw tweets compared to the situational fragment tweets during a disaster.

Therefore, this article proposes a neural-based approach based on fine-tuned RoBERTa model and a feature-based approach for identifying situational raw tweets and fragments. Furthermore, the proposed method is implemented in both English and Hindi language tweets.

The contributions of this work are summarized as follows:

1) Proposed a neural-based approach, which is the combination of fine-tuned RoBERTa model and feature-based approach for detecting the situational tweets during a disaster.

Manuscript received July 3, 2020; revised February 11, 2021; accepted February 20, 2021. *(Corresponding author: Sreenivasulu Madichetty.)*

The authors are with the National Institute of Technology, Tiruchirappalli 620015, India (e-mail: sreea568@gmail.com).

Digital Object Identifier 10.1109/TCSS.2021.3064299

TABLE I
EXAMPLES OF THE RAW TWEET

Tweet No	Situational tweets
1.	Ewwww. Let's hope not. RT @eyewitnessnyc #BreakingNews Possible school shooting reported in Connecticut
2.	RT @ProvFireVideos: Powerful picture RT @HeidiVoight Kids crying, evacuating Sandy Hook Elementary in NEWTOWN http://t.co/WJzN0o7N via @ ...
	Non-Situational tweets
3.	RT @AltCricket: Which Instagram filter do we use for charred bodies? RT @ndtv Send us your pictures from Hyderabad blasts at @ndtv #hbad.
4.	RT @Naseer25: Dilsuknager is an area where Andhra's people has domination, Its a busiest area of city out skirt away from Charminar.

TABLE II
EXAMPLES OF THE FRAGMENT TWEET

Tweet No	Situational tweets
1.	RT @abpnewstv: BREAKING: 7 feared dead in Hyderabad blast - Reports.
2.	Blasts in hyderabad kill at least 11 , injure dozens.
	Non-Situational tweets
3.	God i pray for my country's safety.
4.	All aircraft to stay away from the philippines.

2) Proposed method outperforms the existing methodologies both in-domain and cross-domain.
3) Proposed a neural-based approach, which is the combination of CNN and feature-based approach for identifying the Hindi language situational tweets during a crisis.
4) Extensive error analysis is performed on various deep learning and feature-based approaches.

This article is organized as follows: Section II explains the related works. Section III explains the neural-based approach for detecting situational tweets during a disaster. Experimental results, discussion, and error analysis of the various models are described in Section IV. Section V concludes this article.

II. Related Work

This section can be divided into two subsections, such as: 1) situational information—discusses the work that strictly focuses on classifying situational and nonsituational information during a disaster and 2) crisis informatics in social media—provide a broader overview of related work in the field of disaster management.

A. Situational Information

This section discusses the existing methods that strictly focused on situational information during a disaster. Several authors [17], [20], [21] have been attempted to classify situational and nonsituational tweets during a crisis, but currently, there is no robust model available. Verma *et al.* [17] used a feature-based approach based on the BOW model to identify the situational tweets during a crisis. Later, Imran *et al.* [22] used the n-grams feature-based method for identifying the user-defined categories of tweets during an emergency. However, it relies on the vocabulary that is present in the training tweets and provides bad results for the new vocabulary that is only present in the testing tweets. It fits nicely for the same disaster case where training and testing are conducted on the same disaster data set only. Later, Rudra *et al.* [21] developed a model that is independent of the vocabulary present in the training tweets to detect the situational tweets during a crisis. Subsequently, the same authors [20] developed a model by appending the two more features, such as the presence of religious words and the presence of slang words, to further improve the performance of the model. The extra added features are mostly present in the nonsituational tweets. And also showed that their model works even for detecting the Hindi language tweets. It is also clarified that the same post includes situational and nonsituational information and demonstrates that the performance is enhanced after the splitting of tweets. However, they noted that their approaches do not enhance the performance of a tweet containing situational and nonsituational information. Even so, it is crucial to monitor situational information for aid groups and survivors during the disaster. Besides that, existing methods cannot distinguish between situational and nonsituational tweets specifically due to the inclusion of nonsituational information in both situations. However, all the existing approaches used feature-based approaches to detect situational tweets during a crisis.

B. Crisis Informatics in Social Media

This section describes a brief overview of different crisis-related information as well as gaps between proposed work and existing works, which adopts deep learning methods for the classification of social media messages during a disaster.

In the literature, there are many works [9], [18], [19], [23]–[26] focused on different types of crisis information using both deep learning and nondeep learning methods. Among them, the authors in [18] and [19] concentrated on detecting the actionable information (it is a part of situational information) during a disaster. The authors in [18] demonstrated that different responders need a different piece of information at a specific time. Also, they also explained that actionable information might differ from one responder to another responder for reacting immediately at the right time during a disaster. For example, Medical organizations required medical information where medical resources such as ambulance, blood, medicines, etc., are needed. Likewise, the World Food Program provided food where food requirement is necessary. Medical information in these examples is the actionable information for medical organizations, and food information is the actionable information for the World Food Program.

Some works [9], [26], [27] mainly focused on detecting the service requests (it is a part of situational information) that are posted on Twitter during a disaster. Identifying the service request tweets based on the priority helps the specific emergency organizations receive multiple service requests in later stages of the disaster. Initially, they built a serviceability model to identify and rank the affected people and users' emergency service requests on behalf of the affected people. For example, they are requesting a resource or information regarding the service. They used generic features, text features, social features, serviceability, and the SVM rank method to rank the service requests. Later, the author [9] extended their work by re-ranking the service requests and reduce redundancy among the service requests by grouping similar tweets using the semantic approach. It provides flexibility to the responders for browsing the service requests. However, they are not using any deep learning models for identifying service requests during a disaster. The authors in [27] classified the disaster-related tweets into donation and rescue request classes that are particularly helpful to the volunteer teams for identifying the victims who need food, medicines, rescue missions, etc. They developed a model based on the term-frequency inverse document frequency (TF-IDF) words as a feature and used classifiers such as naive Bayes, logistic regression, etc. And Chennai and Kerala flood disaster data sets are used for experimentation.

The common limitations of the works mentioned above related to the deep learning methods are shown as follows:

1) They did not focus and discussed using deep learning methods (CNN, LSTM, BLSTM, and BLSTM with Attention mechanism) to detect the situational tweets during a crisis. Similarly, for detecting Hindi language situational tweets also.
2) Besides, various disaster data sets have not been extended to classify situational tweets.

This article proposes a neural-based approach to overcome the drawbacks of existing methods for identifying the situational tweets in English and Hindi language tweets during a disaster.

III. Proposed Method

In this section, we describe our proposed method to detect situational tweets during a disaster. The proposed method comprises three different components: 1) feature-based approach [20]; 2) fine-tuned RoBERTa model; and 3) multiplicative fusion technique. The block of the proposed method is depicted in Fig. 1.

A. Feature-Based Approach

In this approach, initially, tweets are preprocessed by applying the following techniques:

1) Case-folding.
2) Lemmatization.
3) Substitution of modal verbs with variations.
4) Twitter POS tagger is used to identify the POS tags in the tweets.
5) Removal of unnecessary words from the tweets like emotions and URLs.

After getting the preprocessed tweet, we extracted low-level lexical and syntactic features from the tweet such as count of subjective words, presence of personal pronouns, count of numerals, presence of exclamations, presence of question marks, presence of modal verbs, presence of wh-words, presence of intensifiers, presence of nonsituational words, presence of religious words and presence of slangs. A feature vector is constructed, and it is given as an input to the SVM classifier with RBF kernel for training and testing the model. The main reason for selecting specific features and classifiers is that the authors in [20] proved that low-level lexical and syntactic features with the combination SVM classifier give better performance detecting the situational tweets during a disaster. The main drawback of this method is it does not work well if the features are not present in the tweets. However, it gives better results than other feature-based approaches.

B. Fine-Tuned RoBERTa Model

To overcome the drawback of the feature-based approaches, we adopted the pretrained RoBERTa model in the proposed method. We have used a pretrained Robustly optimized BERT approach (RoBERTa) for identifying the situational tweets during a disaster. The last layer of RoBERTa model is replaced with a softmax layer and fine-tuned with our disaster data sets. The RoBERTa approach is a pretrained language model that captures the information not only left to right or right to left directions of a tweet but also deep bidirectional pretraining structure provides the model to learn more about context information of a tweet. The RoBERTa model gives better performance than the previous models such as BERT-base and BERT large due to the following modifications:

1) pretrained the languages with eight times larger batches and ten times more data;
2) used byte-pair encoding (BPE) vocabulary instead of the character-level vocabulary;
3) removed the next sentence prediction (NSP);
4) applied the dynamically changing the masking pattern.

We fine-tuned the RoBERTa model to our disaster data sets for detecting the situational tweets.

C. Multiplicative Fusion

Fine-tuned RoBERTa model and feature-based method give the probability vector of a tweet. Later, the multiplicative fusion technique fuses both vectors into a single vector by performing element-wise multiplication. The resultant vector is used for predicting the tweet label.

IV. Experimental Results

This section describes the data sets used for experimentation, baselines used for comparison, implementation details, and performance measures used to evaluate the performance of the model.

A. Baselines

The following are the baselines used in the comparison of the proposed method:

1) Feature-Based Method: Rudra *et al.* [20] used low-level lexical and syntactic features for identifying the situational tweets during disaster. We experimented with this method for comparison with the proposed method.

2) Deep Learning Models: We experimented with convolutional neural network (CNN), long short term memory (LSTM), bidirectional long short term memory (BLSTM), bidirectional long short term memory with attention mechanism (BLSTMA) with pretrained crisis word embeddings for comparing with the proposed model. We tried different word embeddings such as Glove word embeddings, word embeddings, BERT embeddings and crisis word embeddings. Among them, crisis word embeddings give better performance than the other word embeddings. Therefore, we used crisis word embeddings for experimenting with the deep learning models.

B. Implementation Details and Performance Measure

We used ktrain package for implementing the RoBERTa model. K-fold cross-validation ($k = 10$) is used for evaluating the model. Keras [28] and Scikit packages are used for implementing the deep learning and machine learning models. We used "roberta-base" model. The following are the hyperparameters used in the proposed model:

1) maximum sequence length = 150;
2) number of epochs = 10;
3) optimizer = "Adam";
4) binary cross entropy is used;
5) batch size = 8.

The following performance measures are used to evaluate the performance of the proposed models, such as accuracy, precision, recall, and F1-score that are mentioned in 1–4, respectively.

$$\text{Accuracy} = \frac{\text{TP}_s + \text{TN}_s}{\text{TP}_s + \text{TN}_s + \text{FP}_s + \text{FN}_s} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}_s}{\text{TP}_s + \text{FP}_s} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{3}$$

$$\text{F1-score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

where TP_s = Total No. of situational tweets detected correctly as situational.

TN_s = Total no. of nonsituational tweets detected correctly as nonsituational.

FP_s = Total no. of nonsituational tweets wrongly detected as situational.

FN_s = Total no. of situational tweets wrongly detected as nonsituational.

s = No. of classes.

C. Data Sets

We adopted the disaster data sets from the work [20]. The data sets are created from various disasters such as the Hyderabad bomb blast, Sandy Hook Shooting, Typhoon Hagupit, Nepal earthquake, and Harderail. Among them, three of the data sets (Hyderabad bomb blast, Sandy Hook Shooting, Typhoon Hagupit) contain English language tweets and the rest (Nepal earthquake and HarDerail accident) are Hindi language tweets. English language tweets include both raw tweets and fragment tweets for three disaster data sets. The details of five disaster data sets are shown in Table III.

D. Raw Tweets

The results show that the bidirectional LSTM with attention mechanism using crisis embeddings performs better than the other deep learning models in case of larger disaster data sets that are included in Table IV across various imbalanced data sets. Hence, bidirectional LSTM with attention mechanism using crisis embeddings is assumed as a typical deep learning model for identifying the situational tweets during disastrous situations. Furthermore, the proposed method gives the best performance compared with the deep learning and feature-based approaches across all parameters that are shown in Table IV.

Table IV shows the comparison of proposed method with the various deep learning models and existing method on unbalanced raw tweets.

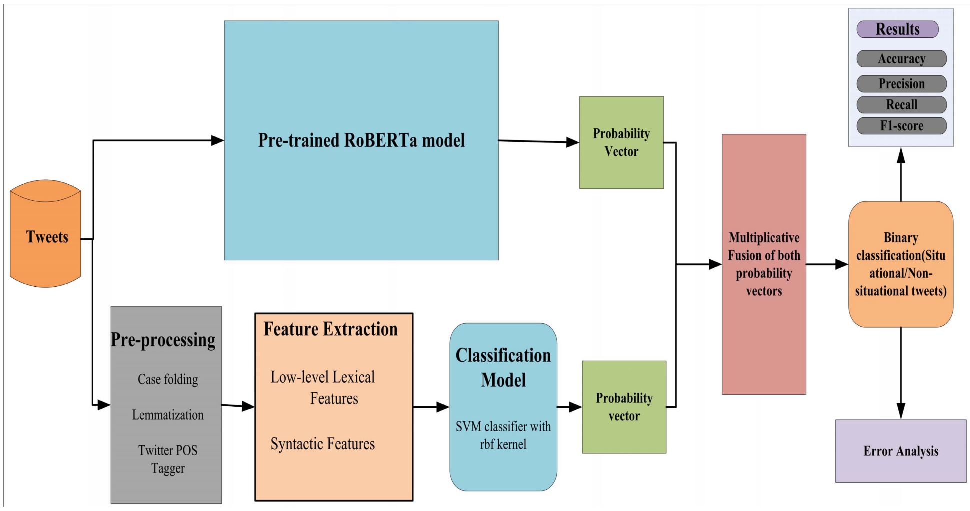

Fig. 1. Framework of the proposed method.

TABLE III
DIFFERENT TYPE OF TWEETS IN EACH DISASTER DATA SET

Types of Tweets	HydBlast	SanHShoot	TypHagupit	NepEquake	HarDerail
Unbalanced Raw Tweets	4,930	4,998	4,996	-	-
Fragment Tweets	832	864	906	562	240

TABLE IV
COMPARISON OF PROPOSED METHOD WITH VARIOUS DEEP LEARNING MODELS AND EXISTING METHOD [20] ON UNBALANCED RAW TWEETS

		Hagupit		
Model	**Precision**	**Recall**	**F1-score**	**Accuracy**
Existing method [20]	64.78%	96.28%	76.85%	62.84%
BLSTM with Attention	99.20%	99.40%	99.44%	97.89%
BLSTM	98.20%	98.80%	98.5%	97.97%
LSTM	98.50%	98.60%	98.60%	98.00%
CNN	98.38%	99.21%	98.79%	98.33%
Proposed method	**100%**	**99.56%**	**100%**	**99.40%**
		Hyderabad		
Existing method [20]	76.66%	95.45%	84.97%	76.19%
BLSTM with Attention	98.60%	99.04%	98.81%	97.70%
BLSTM	97.90%	98.78%	98.37%	97.76%
LSTM	96.87%	99.13%	97.98%	97.22%
CNN	97.90%	98.84%	98.39%	**97.80%**
Proposed method	**98.80%**	**99.46%**	**98.95%**	**98.00%**
		Sandy_hook_frag		
Existing method [20]	75.73%	96.47%	84.77%	75.79%
BLSTM with Attention	**99.07%**	**98.07%**	**98.56%**	97.31%
BLSTM	98.15%	97.63%	97.89%	97.06%
LSTM	97.68%	97.83%	97.75%	96.86%
CNN	98.23%	97.96%	98.09%	**97.35%**
Proposed method	**99.20%**	**99.34%**	**99.18%**	**98.00%**

E. Fragment Tweets

Rudra *et al.* [20] provided the labeled data sets of fragment tweets. We performed various experiments for selecting the appropriate word embeddings to the fragment tweets. After experimentation, we observed that crisis word embeddings are useful to the disaster fragment tweets compared to the other word embeddings such as Glove word embeddings, BERT, and word2vec embeddings.

The comparison results of the proposed method with deep learning and feature-based approaches on Fragment Tweets across various disaster data sets are presented in Table V.

The results highlighted that the proposed method works well than the deep learning models and feature-based approaches for three disaster data sets on fragment tweets. Among the deep learning models, bidirectional LSTM with attention mechanism works well in some cases, and CNN works well

TABLE V

COMPARISON OF PROPOSED METHOD WITH DEEP LEARNING AND FEATURE-BASED APPROACHES ON FRAGMENT TWEETS

		Hagupit		
Model	**Precision**	**Recall**	**F1-score**	**Accuracy**
Existing method [20]	-	94.00%	87.00%	85.86%
BLSTM with Attention	88.49%	**94.13%**	87.98%	87.96%
BLSTM	88.32%	86.76%	87.43%	87.51%
LSTM	30.05%	60.00%	39.96%	50.33%
CNN	88.80%	88.16%	88.71%	88.63%
Proposed method	**95.00%**	90.00%	**93.00%**	**93.00%**
		Hyderabad		
Existing method [20]	-	85.00%	84.00%	84.26%
BLSTM with Attention	84.38%	**90.59%**	87.25%	86.77%
BLSTM	85.80%	87.20%	86.27%	86.18%
LSTM	30.00%	60.00%	39.96%	50.12%
CNN	87.16%	87.37%	85.16%	86.21%
Proposed method	**98.00%**	85.00%	**91.00%**	**90.00%**
		Sandy_hook_frag		
Existing method [20]	-	87.00%	89.00%	90.04%
BLSTM with Attention	92.09%	92.12%	91.95%	92.01%
BLSTM	94.04%	92.11%	92.95%	93.04%
LSTM	33.00%	70.00%	45.58%	47.57%
CNN	93.55%	91.76%	92.58%	92.82%
Proposed method	**95.00%**	**95.00%**	**95.00%**	**95.00%**

in some other cases. The explanation for inconsistent findings is that the size of the disaster data set is tiny. On the other hand, LSTM gives a more unsatisfactory performance on the fragment tweets due to the length of the tweet is small. However, the proposed method provides better results than the existing methods for detecting situational tweets during a crisis.

Furthermore, to know the impact of threshold (confidence) scores on classification performance, plots are drawn between confidence scores and classification parameters for various deep learning models to take advantage of the best classification performance on different data sets shown in Fig. 2. It is observed that the precision value for Typhoon Hagupit and Sandy Hook shooting data set is increased by 8.94% and 4.79%, respectively, at a threshold value of 0.95, while other parameters are not getting significant improvement. In the Hyderabad data set, the accuracy parameter is slightly improved by 0.85% at a threshold value of 0.75. However, the proposed method achieved better performance than the existing methods for identifying the situational tweets during a disaster.

F. Hindi Tweets

Rudra *et al.* [20] explained that detecting the Hindi language tweets is also essential at the early stage of the disaster. Therefore, various deep learning models such as CNN, LSTM, BLSTM, and BLSTM with attention mechanism based on the Hindi word embeddings [29], [30] has experimented with detecting the Hindi language situational tweets during a disaster. Two disaster data sets such as Nepal Earthquake and Harderail accident, are used for experimentation, and the results are shown in Table VI. CNN provides the better performance compared to the other deep learning models on detecting the situational tweets in th Hindi language. It is due to the other deep learning models are complex architectures that need a large amount of data. Hence, we proposed a model based on the combination of CNN and a feature-based approach for detecting the Hindi tweets during the disaster.

In order to know the impact of threshold scores on the classification performance of Hindi tweets, plots are drawn between threshold scores and classification performance using different parameters for the various deep learning models shown in Fig. 3. It is evident that there is an improvement in the precision value of the Nepal Earthquake by 1.7% at 0.75 threshold value, while there is no significant improvement for other parameters. However, the proposed method gives a better performance than the existing works.

G. Cross-Domain

In this work, we consider the Sandy Hook shooting and Hyderabad bomb blast disaster data sets are past disaster data sets for training the proposed model, and the rest are used for testing the model. The comparison of proposed method with the deep learning and feature-based approach on English tweets are shown in Tables VII and VIII. In the case of a cross-domain scenario, our experiments provide superior results to the existing method [20].

Besides, the impact of threshold values is checked on classification performance on English tweets, and plots are shown in Fig. 4. The precision value is increased by 7.75%,

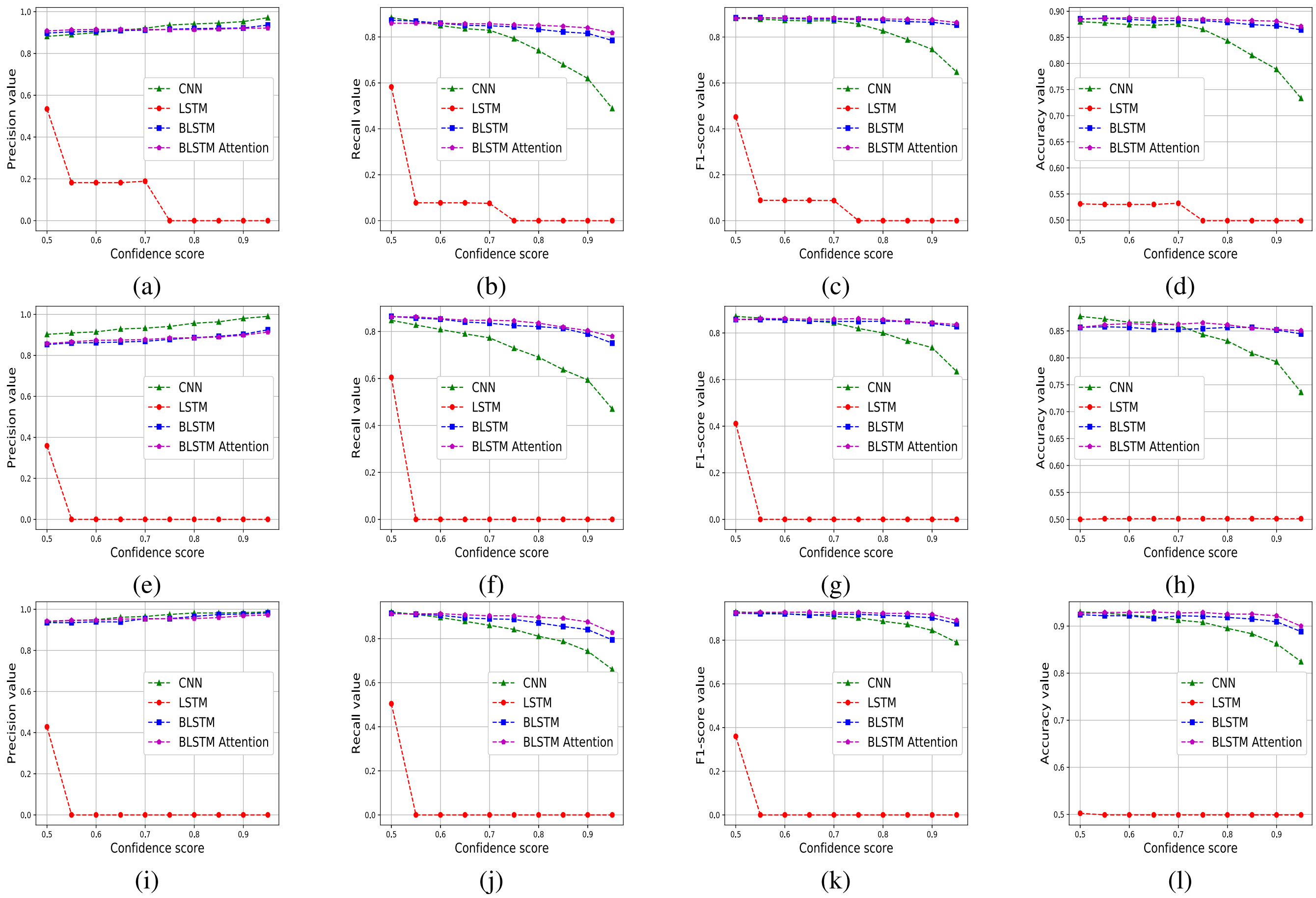

Fig. 2. Classification results versus confidence score of different deep learning model on English fragment data sets. (a), (b), (c), and (d) represent precision, recall, F1-score, and accuracy on Hagupit data set, respectively. (e), (f), (g), and (h) represent precision, recall, F1-score, and accuracy on Hyderabad bomb blast data set, respectively. (i), (j), (k), and (l) represent precision, recall, F1-score, and accuracy on Sandy Hook shooting data set, respectively.

TABLE VI

COMPARISON OF PROPOSED METHOD WITH THE DEEP LEARNING MODELS AND FEATURE-BASED APPROACH [20] ON HINDI TWEETS FOR PRECISION AND ACCURACY PARAMETERS

	Precision		Accuracy	
Model	**Nepal Earthquake**		**Hderail**	
Existing work [20]	94.67%	89.40%	81.30%	74.22%
BLSTM with Attention	68.41%	71.00%	68.27%	45.00%
BLSTM	58.39%	53.00%	72.76%	58.55%
LSTM	57.77%	69.66%	59.02%	54.49%
CNN	94.67%	89.40%	68.65%	66.71%
Proposed method	**96.54%**	**92.32%**	**85.43%**	**75.17%**

TABLE VII

COMPARISON OF PROPOSED METHOD WITH FEATURE-BASED AND DEEP LEARNING APPROACHES (DEEPMODEL) ON TRAINING SANDY HOOK SHOOTING (PAST) DATA SET AND TESTING DIFFERENT DATA SET (FUTURE) TWEETS FOR RAW TWEETS

Testing Dataset	Recall			F1-score			Accuracy		
	Koustav [20]	DeepModel	Proposed Method	Koustav [20]	DeepModel	Proposed Method	Koustav	DeepModel	Proposed Method
Hagupit	29.46%	64.64%	**100%**	43.13%	70.82%	**82.12%**	57.56%	73.36%	**69.76%**
Hyderabad	47.50%	80.16%	**100%**	54.67%	81.14%	**83.44%**	24.20%	81.30%	**78.09%**

4.58%, 11.37%, and 1.93% for cross-domain-1 (training and testing on Sandy Hook shooting and Typhoon Hagupit data set), cross-domain-2 (training and testing on Sandhy shooting and Hyderabad data set), cross-domain-3 (training and testing on Hyderabad and Typhoon Hagupit data set), cross-domain-4 (training and testing on Hyderabad and Sandy Hook shooting data set) at threshold values of 0.96, 0.80, 0.80, and 0.90, respectively. However, the proposed method gives

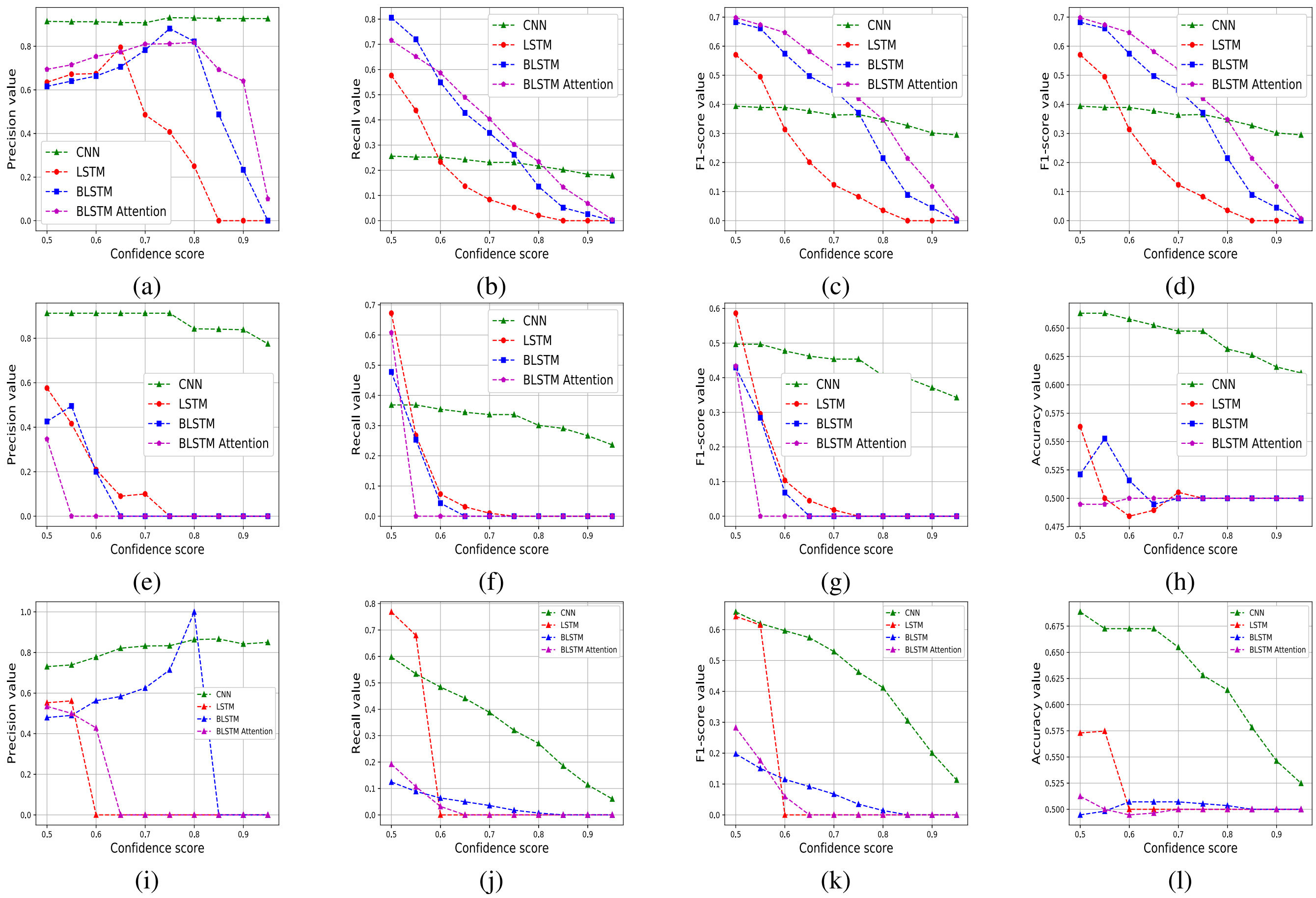

Fig. 3. Classification results versus confidence score of different deep learning model on Hindi data sets. (a), (b), (c), and (d) represent the precision, recall, F1-score, and accuracy on Nepal Earthquake data set. (e), (f), (g), and (h) represent the precision, recall, F1-score, and accuracy on HarDerail accident data set. (i), (j), (k), and (l) represent the precision, recall, F1-score, and accuracy parameters, when training and testing on HarDerail accident and Nepal Earthquake data set.

TABLE VIII

COMPARISON OF PROPOSED METHOD WITH FEATURE-BASED AND DEEP LEARNING APPROACHES (DEEPMODEL) ON TRAINING HYDERABAD BOMB BLAST (PAST) DATA SET AND TESTING DIFFERENT DATA SET (FUTURE) TWEETS FOR RAW TWEETS

Testing Dataset	Recall			F1-score			Accuracy		
	Koustav [20]	DeepModel	Proposed Method	Koustav [20]	DeepModel	Proposed Method	Koustav	DeepModel	Proposed Method
Hagupit	49.10%	**98.72%**	94.08%	58.82%	74.37%	**93.06%**	26.82%	65.98%	**90.37%**
Sandy	87.87%	**95.56%**	85.33%	69.87%	79.53%	**86.35%**	41.13%	75.41%	**81.61%**

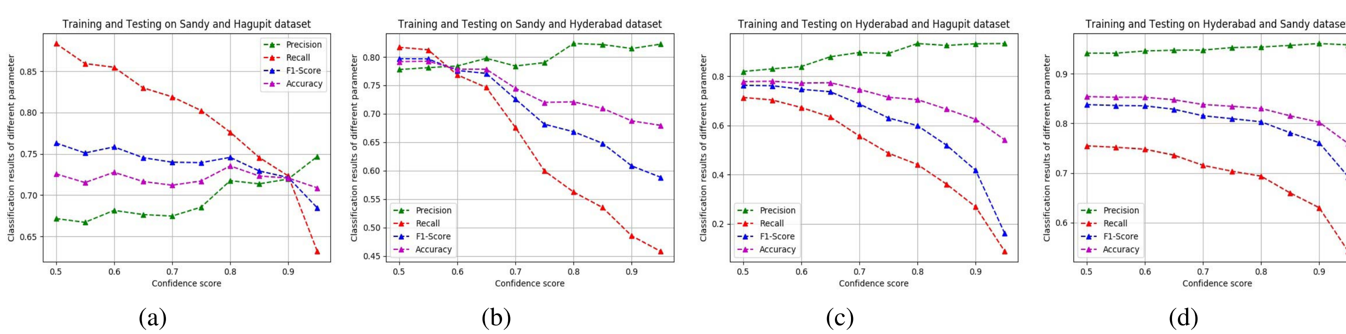

Fig. 4. Classification results versus confidence score of proposed deep learning model for training and testing on past and future event data sets. (a) represents training and testing on Sandy Hook shooting and Typhoon Hagupit data set, (b) represents training and testing on Sandy Hook shooting and Hyderabad bomb blast data set, (c) represents training and testing on Hyderabad bomb blast and Typhoon Hagupit data set, and (d)represents training and testing on Hyderabad bomb blast and Sandy Hook shooting data set.

better results compared to the deep learning and feature-based approaches.

It is tested for training and testing of past and future event data sets on Hindi tweets in the cross-domain, respectively. The precision value of the deep learning model and the existing method is 73.04% and 65.94%, respectively. Furthermore, the impact of threshold values is checked on the classification performance of Hindi tweets in cross-domain (training on HarDerail and testing on Nepal Earthquake data set) that is shown in Fig. 3. The precision value is increased by 13.63% and 28.06% for CNN and BLSTM models at 0.85 and 0.80 threshold values, respectively. Although the BLSTM model improves in most cases, it delivers less performance compared to CNN other than cross-domain (training on HarDerail and testing on Nepal Earthquake data set). CNN gives better performance compared to the other deep learning models with the precision parameter. Therefore, CNN has experimented with the combination of a feature-based approach (proposed approach for Hindi tweets) in cross-domain also, and it improves the performance further 5.4% of the precision value.

H. Discussion and Error Analysis

We have seen in the experiments and the results, deep learning models outperform the traditional machine learning algorithms to identify the situational tweets during a crisis. But in the case of fragments, deep learning models outperform the conventional machine learning algorithms except for LSTM to detect situational information. Among the deep learning techniques, the BLSTM attention mechanism works effectively by automatically extracting the features from the tweets by capturing the information in both directions of the tweets and giving the weights to the most influential words in the tweets. It is also seen that it works with different categories of tweets (raw and fragment tweets). Furthermore, it is also seen that word embeddings also influence the performance of the model that is described in Section IV-B. However, the proposed model provides better results than the existing methods.

In the case of Hindi tweets, we reported that proposed CNN with the combination of a feature-based approach provides a better result than other deep-learning models in detecting situational tweets. It is believed that the proposed model would also help to detect situational tweets in the cross-domain.

Traditional machine learning methods [20], [21] will fail to identify the situational tweets during a crisis if there is an absence of low-level lexical (count of numerals, count of nonsituational words, etc.) and syntactic features (count of modal verbs, count of intensifiers, etc.) within the tweets. Table IX provides the example tweets that are misclassified by the traditional machine learning methods and correctly classified by the deep learning models.

Subsequently, we will address a set of error types with example tweets where the various deep learning models are successful and unsuccessful in identifying the situational tweets during a disaster reported in Table X. The first, second, third, and fourth columns represent the serial number, error type, example tweet of the error type, and the actual label of the tweet. The explanation of different error types for various deep learning models is given below:

1) Tweets that fall under error type-1 are incorrectly predicted by the CNN model. It is because CNN captures local features from the tweets that contain ambiguous information, while other models work well by capturing the word order information of the tweets. The example tweet contains ambiguous information in the form of words such as "Dilsukh," "Nagar," and "Hyderabad." Such words are used in both situational and nonsituational tweets.
2) Tweets that fall under error type-2 are incorrectly predicted by the LSTM model. This is due to the fact that LSTM extracts information from left to right direction in a tweet and the right side information dominates the left side information, whereas other models (CNN, BLSTM, and BLSTM attention) perform well by capturing information from both directions in a tweet. The example tweets contain informative content that is present on the left side and the noninformative content on the right side.
3) Tweets that fall under error type-3 are wrongly predicted by both the LSTM and BLSTM models. This is because the models capture the sequence information of a tweet in which confusing words are present, while other models perform well, either because they capture the most insightful information or because they give weight to the important words. Example tweets reflect much of the confusing words in the center of the tweets.
4) Tweets that come under error type 4 are incorrectly predicted by CNN, LSTM, and BLSTM. It is because the models use the noisy words present in a tweet, while BLSTM Attention works well because it neglects the noisy words by assigning weight to the most important words. Example tweets show that noisy words, such as prayers, changed, anything, and same, will dominate the other important words in a tweet.

Later, we attempted to identify systematic error patterns of the BLSTM with an attention mechanism to motivate future work. For this analysis, we compare predictions of the BLSTM with the attention mechanism and the actual label on the test data with the weakest absolute performance. The reasons for the incorrect prediction of the BLSTM with attention mechanism are described below, followed by a sample tweet and a description of the following sample tweets:

1) *Mobile Number and Landline Numbers:*
Example Tweet: *RT @ibnlive: State Helplines for Hyderabad Blasts: 040 27854771, 040-27853408 040 27852435-36.*
Explanation: This tweet is labeled as a situational tweet but detected as a nonsituational tweet as it contains most of the landline and mobile numbers, and the proposed model treated all types of numerals equally that misclassified the tweets.
2) *Location and People Names:*
Example Tweet: *RT @iCASumit: Needs AB + ve blood For: Farida Narayana Hrudayalaya, Suraram, Jeedimetla, Call: 9676595836 @JoinAA*
Explanation: This tweet is labeled as a situational tweet but detected as a nonsituational tweet. Tweets contain the words that do not include the pretrained crisis word embeddings. During the disaster, affected people post location and person names of their requirements. The names of the people and location from one disaster to another disaster. Most of the time, nonsituational

TABLE IX
EXAMPLE TWEETS WHERE THE EXISTING METHODS [20], [21] FAILED TO DETECT SITUATIONAL INFORMATION

Tweet No	Situational Information
1.	RT @robinrajsingh: Bomb blast in dilsukhnagar (hyderabad) near venkatadri theatre...many feared dead. http://t.co/CpwIySnSn6
2.	Seven killed in #Hyderabad #blast http://t.co/ZMsb0WUXwh at #worldsnap #breakingnews #india #topnews

TABLE X
ERROR ANALYSIS FOR DIFFERENT DEEP LEARNING MODELS (WHERE "FAIL" INDICATES THAT THE MODEL PREDICT THE TWEETS INCORRECTLY AND "PASS" INDICATES THAT MODEL PREDICTS THE TWEET CORRECTLY)

S No.	Error Type	Example tweet	Actual Label	CNN	LSTM	BLSTM	BLSTM Attention
1.	Ambiguity information in local features	RT @BDUTT: Two blasts in Dilsukh Nagar in Hyderabad, several casualties, 50 reported injured.	**Situational**	Fail	**Pass**	**Pass**	**Pass**
2.	Less informative content at the end of the tweet	RT @ibnlive: Police suspect one of the bombs may have been kept on a motorcycle; the other in a tiffin box	**Situational**	**Pass**	**Fail**	**Pass**	**Pass**
3.	Confusion words in sequence information	RT @amitbhawani: Lot of traffic moving around and 7 confirmed dead. Stay	**Situational**	**Pass**	Fail	Fail	**Pass**
4.	More number of Noisy words	Blasts reported in hyderabad, Dilsukh Nagar. Its the same area as last blasts Has anything changed since then?	**Situational**	Fail	Fail	Fail	**Pass**

tweets also contain both locations and names. And also, it includes the names of the persons and locations such as Farid, narayana, Hrudayalaya, suraram, and jeedimetla that are not present in the crisis word embeddings. And also, it may not present in the training tweets due to experiments are performed in cross-domain (Training and Testing in different disaster event data sets).

3) *Some Tweets Contain Sarcastic Information:*
Example tweet: *RT @timesofindia: Hyderabad bomb blasts rock Parliament, opposition attacks govt for alleged lapses—The Times of India [URL].*
Explanation: This tweet is labeled as a nonsituational tweet but detected as a situational tweet. It is due to most of the tweet content reveals the situational information.

However, the proposed method (neural-based approach based on the RoBERTa model and feature-based approach) works well in all situations where the existing methods do not correctly detect the situational tweets. Also, we have investigated failure cases of the proposed method. We have identified the reasons for the misprediction of tweets by the proposed model. It is beneficial to further improve the model in the future. The reasons for the misclassification of tweets by the proposed model are as follows:

1) *Misleading Hashtags and Keyword Influence:*
Example Tweets: *RT @NASA: We caught* 3 *days of Typhoon #Hagupit's motion over Philippines [URL] #EarthRightNow [URL]*
Explanation: This tweet is labeled as a nonsituational tweet but detected as a situational tweet. The classifier is confused by hashtag words like "Haguput," "EarthRightNow," and keywords like Typhoon, which might be appeared in situational tweets. A lot of nonsituational tweets misclassified as situational tweets when the experiments are performed in in-domain (training and testing on the same disaster event).
2) *Need to Understand the URL's:*
Example Tweets: *RT @eumetsat: New Metop image of #Hagupit/#RubyPH from 01:25 UTC today (Friday) https://t.co/JSRcpJOrba http://t.co/mhv9YXBvBs*
Explanation: This tweet is labeled as a situational tweet but detected as a nonsituational tweet. The classifier needs to understand the URL for determining the tweet is situational or nonsituational.

The above-mentioned incorrect predictions are not only for the BLSTM with attention mechanism model as well as existing approaches, including the other deep learning models such as CNN, LSTM, and BLSTM. In Hindi tweets, the proposed deep learning model (CNN with the combination of feature-based method) misclassified the situational tweets, where most of the tweets contain very few words. The main reason for the lower performance of the deep learning model is due to fewer training tweets.

V. CONCLUSION

This article proposed a neural-based approach based on the RoBERTa model and feature-based method to identify situational tweets during the disaster, especially for English language tweets. Extensive experiments on two different language tweets (English and Hindi) are performed. Besides, it is concluded that the CNN based on crisis embeddings with the combination of the feature-based method produces better output for Hindi language tweets. We also performed error analysis on the various deep learning models and the proposed method where the model failed to identify the situational tweets during a crisis. It is beneficial for further enhancement in the future.

REFERENCES

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

[2] T. H. Nazer, F. Morstatter, H. Dani, and H. Liu, "Finding requests in social media for disaster relief," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2016, pp. 1410–1413.

[3] M. Sreenivasulu and M. Sridevi, "Mining informative words from the tweets for detecting the resources during disaster," in *Proc. Int. Conf. Mining Intell. Knowl. Explor.* Hyderabad, India: Springer, 2017, pp. 348–358.

[4] M. Sreenivasulu and M. Sridevi, "A survey on event detection methods on various social media," in *Recent Findings in Intelligent Computing Techniques*. Singapore: Springer, 2018, pp. 87–93.

[5] I. Varga *et al.*, "Aid is out there: Looking for help from tweets during a large scale disaster," in *Proc. ACL*, 2013, pp. 1619–1629.

[6] R. Dong, L. Li, Q. Zhang, and G. Cai, "Information diffusion on social media during natural disasters," *IEEE Trans. Comput. Social Syst.*, vol. 5, no. 1, pp. 265–276, Mar. 2018.

[7] L. Li *et al.*, "Characterizing the propagation of situational information in social media during COVID-19 epidemic: A case study on Weibo," *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 2, pp. 556–562, Apr. 2020.

[8] M. Hasan, M. A. Orgun, and R. Schwitter, "Real-time event detection from the Twitter data stream using the TwitterNews+Framework," *Inf. Process. Manage.*, vol. 56, no. 3, pp. 1146–1165, May 2019.

[9] H. Purohit, C. Castillo, and R. Pandey, "Ranking and grouping social media requests for emergency services using serviceability model," *Social Netw. Anal. Mining*, vol. 10, no. 1, pp. 1–17, Dec. 2020.

[10] R. Dutt, M. Basu, K. Ghosh, and S. Ghosh, "Utilizing microblogs for assisting post-disaster relief operations via matching resource needs and availabilities," *Inf. Process. Manage.*, vol. 56, no. 5, pp. 1680–1697, Sep. 2019.

[11] S. Roy, S. Mishra, and R. Matam, "Classification and summarization for informative tweets," in *Proc. IEEE Int. Students' Conf. Elect., Electron. Comput. Sci. (SCEECS)*, Feb. 2020, pp. 1–4.

[12] S. Madisetty and M. S. Desarkar, "A neural network-based ensemble approach for spam detection in Twitter," *IEEE Trans. Comput. Social Syst.*, vol. 5, no. 4, pp. 973–984, Dec. 2018.

[13] H. Gupta, M. S. Jamal, S. Madisetty, and M. S. Desarkar, "A framework for real-time spam detection in Twitter," in *Proc. 10th Int. Conf. Commun. Syst. Netw. (COMSNETS)*, Jan. 2018, pp. 380–383.

[14] S. Madisetty and M. S. Desarkar, "An ensemble based method for predicting emotion intensity of tweets," in *Proc. Int. Conf. Mining Intell. Knowl. Explor.* Hyderabad, India: Springer, 2017, pp. 359–370.

[15] S. Madisetty and M. S. Desarkar, "Nsemo at emoint-2017: An ensemble to predict emotion intensity in tweets," in *Proc. 8th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, 2017, pp. 219–224.

[16] S. Madisetty and M. S. Desarkar, "Exploiting meta attributes for identifying event related hashtags," in *Proc. 9th Int. Joint Conf. Knowl. Discovery, Knowl. Eng. Knowl. Manage.*, vol. 1, A. L. N. Fred and J. Filipe, Eds. Madeira, Portugal: SciTePress, Nov. 2017, pp. 238–245, doi: 10.5220/0006502602380245.

[17] S. Verma *et al.*, "Natural language processing to the rescue? extracting 'situational awareness' tweets during mass emergency," in *Proc. 5th Int. AAAI Conf. Weblogs Social Media*, 2011, pp. 385–392.

[18] H. Zade, K. Shah, V. Rangarajan, P. Kshirsagar, M. Imran, and K. Starbird, "From situational awareness to actionability: Towards improving the utility of social media data for crisis response," in *Proc. ACM Hum.-Comput. Interact.*, vol. 2, Dec. 2018, pp. 1–18.

[19] M. Avvenuti, S. Cresci, F. Del Vigna, T. Fagni, and M. Tesconi, "CrisMap: A big data crisis mapping system based on damage detection and geoparsing," *Inf. Syst. Frontiers*, vol. 20, no. 5, pp. 993–1011, Oct. 2018.

[20] K. Rudra, P. Goyal, N. Ganguly, S. Ghosh, and M. Imran, "Extracting and summarizing situational information from the Twitter social media during disasters," *ACM Trans. Web*, vol. 12, no. 3, p. 17, Jul. 2018.

[21] K. Rudra, S. Ghosh, N. Ganguly, P. Goyal, and S. Ghosh, "Extracting situational information from microblogs during disaster events: A classification-summarization approach," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2015, pp. 583–592.

[22] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, "Aidr: Artificial intelligence for disaster response," in *Proc. 23rd Int. Conf. World Wide Web*, 2014, pp. 159–162.

[23] D. T. Nguyen, K. A. Al Mannai, S. Joty, H. Sajjad, M. Imran, and P. Mitra, "Robust classification of crisis-related data on social networks using convolutional neural networks," in *Proc. 11th Int. AAAI Conf. Web Social Media*, 2017, pp. 632–635.

[24] S. Subramani, H. Wang, H. Q. Vu, and G. Li, "Domestic violence crisis identification from facebook posts based on deep learning," *IEEE Access*, vol. 6, pp. 54075–54085, 2018.

[25] F. Alam, S. Joty, and M. Imran, "Graph based semi-supervised learning with convolution neural networks to classify crisis related tweets," in *Proc. 12th Int. AAAI Conf. Web Social Media*, 2018, pp. 556–559.

[26] H. Purohit, C. Castillo, M. Imran, and R. Pandey, "Social-EOC: Serviceability model to rank social media requests for emergency operation centers," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 119–126.

[27] M. Hasan, M. A. Orgun, and R. Schwitter, "Real time Twitter based disaster response system for indian scenarios," in *Proc. 26th Int. Conf. High Perform. Comput., Data Analytics Workshop (HiPCW)*, Dec. 2019, pp. 82–86.

[28] F. Chollet. (2015). *Keras*. [Online]. Available: https://github.com/fchollet/keras

[29] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," 2016, *arXiv:1607.04606*. [Online]. Available: http://arxiv.org/abs/1607.04606

[30] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, 2018, pp. 1–5.

Sreenivasulu Madichetty received the B.Tech. and M.Tech. degrees in computer science and engineering from the Jawaharlal Nehru Technological University, Ananthapur, India, in 2013 and 2015, respectively. He is currently pursuing the Ph.D. degree in computer science and engineering at the National Institute of Technology, Tiruchirappalli, India.

He worked as a Lecturer with CSE Department at IIIT RK Valley and SVNE, Tirupathi, India, from 2015 to 2016. His research interests include information retrieval, social media analysis, natural language processing, and machine learning.

Sridevi M received the M.E. degree in computer science and engineering from Annamalai University, Chidambaram, India, in 2006 and the Ph.D. degree from the National Institute of Technology, Tiruchirappalli, India, in 2015.

She is an Assistant Professor with the Department of Computer Science and Engineering, National Institute of Technology, Tiruchirappalli, India, and has 13 years of teaching and research experience. Her research interests include image processing, parallel algorithms, and soft computing techniques.