Journal of the Institute of Engineering
Volume 17, No. 1
ISSN: 1810-3383
Published: April 2023
© TUTA | IOE | PCU
Printed in Nepal

Multimodal Approach to Analyze Disaster Related Information by using Image & Text Classifications Model on Twitter Data

Bansaj Pradhan [a], Sanjeeb Prasad Panday [b], Aman Shakya [c]

Dept. of Electronics and Computer Engineering, Pulchowk Campus, Institute of Engineering, Tribhuvan University, Nepal

Corresponding Author: [b]sanjeeb@ioe.edu.np, [c]aman.shakya@ioe.edu.np

Received: 2022-03-14 Revised:2023-04-08 Accepted: 2023-04-09

Abstract:
When a natural disaster occurs, we all are eager to know about it, someone willing to help and donate, and someone maybe just curious about it. Multimedia content on social media platforms provides essential information during disasters. Information transmitted includes reports of missing or found people, infrastructure damage, and injured or dead people. Despite the fact that numerous studies have shown how important text and image contents are for disaster response, past research has mostly focused on the text modality with little success with multimodality. The most recent study on the multimodal classification of tweets about disasters makes use of fairly simple models like CNN and VGG16. In order to improve the multimodal categorization of disaster-related tweets, we have gone further in this study work and used cutting-edge text and image classification models. The study focused on two distinct classification tasks: determining if a tweet is informative or not. The various feature extraction techniques from the textual data corpus and the pre-processing of the corresponding image corpus are incorporated into the multimodal analysis process. We then use various classification models to train and predict the output and compare their performances while adjusting the parameters to enhance the outcomes. Models like ResNet and Bi-LSTM for text classification and image classification, respectively, were trained and examined. The Bi-LSTM and ResNet multimodal architecture performs better than models developed utilizing a single modality(ResNet for image or Bi-LSTM for text alone), according to the results. Additionally, it demonstrates that for both classification tasks, our Bi-LSTM and ResNet model outperforms the FastText and VGG-16 baseline model by a respectable margin.

Keywords: Image classification, Text classification, Multimodal fusion, Disasters and analysis, Crisis computing, Social media

1.Introduction

The use of social media during disasters has become increasingly common in recent years, as it provides a platform for people to share critical information and aid response efforts. While text and image content are both essential for effective disaster response, previous research has mainly focused on the text modality. However, recent studies have shown that combining multiple modalities can significantly enhance disaster-related tweet classification. In this thesis, we address the problem of improving the multimodal categorization of disaster-related tweets by utilizing BiLSTM for text classification and ResNet for image classification. Our motivation for this research is to develop a more effective approach to classify tweets during disasters, which can assist emergency response teams in identifying critical information, such as missing persons, infrastructure damage, and injured or deceased individuals. Our objective is to showcase the effectiveness of our proposed multimodal architecture, which leverages BiLSTM and ResNet, in comparison to models that rely on a single modality or other multimodal approaches such as FastText and VGG-16. By doing so, we hope to make a significant contribution towards improving disaster response efforts through social media.

1.1 Background

When a natural disaster occurs, we all are eager to know about it, someone willing to help and donate, someone maybe just curious about it.

In the last 20 years, we lost 1.35 estimated million lives to disaster. Many steps and protocols have been made and adopted during and after post-disaster which helps a lot but still, a real-time analysis system during and post-disaster for damage assessment, the scale of damage, need assessment, etc have yet to be implemented effectively. Humanitarian organizations believe that the data from social media was very useful, especially during the post-crisis period. As we are advancing towards the digital era, the internet plays a big role in communication. People are also using social media to share information in the form of text, images, and videos. Due to the unstructured nature of internet data, duplication is a very prevalent issue. We use information from Nepal's most well-known news sites and Twitter to better focus the data sets.

Users are curious to learn more about natural disasters as soon as they occur. However, for queries concerning such events, search engines offer blue links interface. If users are presented with a structured summary of the most recent events associated with such inquiries, the relevancy of the results for such queries can be greatly increased. Twitter and other microblogging platforms are growing amid emergencies and natural disasters. Many public events, including traffic crashes, recommender systems, disasters, etc.[1] use Twitter as a communication platform. Twitter crisis-related tweet classification is a crucial NLP task. Studies have shown how beneficial the information on Twitter is for many disaster response jobs. Making sense of social media data, however, is a difficult process for a variety of reasons, including the limitations of the tools available for analyzing high-volume and fast-flowing data streams. In this work, textual and audiovisual content from the countless tweets shared on Twitter during the disaster events has undergone a thorough multidimensional analysis. To process the data created during the crisis events, we specifically use a variety of artificial intelligence techniques from the fields of natural language processing and computer vision. These techniques leverage various machine learning algorithms. Our method relies on BiLSTM and ResNet, which exhibits a notable improvement in performance as shown by the results of the validation on seven different disaster-related datasets.

1.2 Problem Statement

On Twitter, a lot of informative and non-informative tweets are posted throughout the crisis. Informative tweets give valuable information on affected people, infrastructure damage, resource availability, and other topics. On the other hand, non-informative tweets don't offer useful information on victims or humanitarian agencies. Finding educational tweets during the disaster is a difficult endeavor. During the crisis, tweets with photographs are frequently posted. For recognizing informative tweets, image elements are just as important as tweet text attributes. However, the majority of approaches in use today either rely on text- or image-based models. Few studies have looked into multimodality approaches to disaster circumstances analysis. A comparative analysis to select a model with better performance can be used to enhance those current multi-model systems. By selecting a model with straightforward model architecture and good accuracy, the existing multi-model architectures can be improved. By taking into account a hybrid model that combines text and visuals, this research suggests a revolutionary strategy. It includes an RNN-based text classification model and a DNN-based picture classification model. To show the efficiency of the suggested strategy, numerous experiments were conducted using Twitter datasets of various crises, including earthquakes in Nepal and Mexico, among others.

1.3 Objectives

The basic objective of the research is to provide a novel approach by considering a hybrid model using text and image classification for classifying crisis-related information from Twitter.

1.4 Scope of the Study

The scope of the study is as follows:

- Datasets for the research are used from CrisisMMD dataset [28] dataset
- Deep learning system is based on BiLSTM architecture
- Image Classification using various known standard models
- Multimodal Fusion using fully connected

layers+ Softmax

- Using Python as a programming language with Keras, pandas, matplotlib, numpy, etc

2. Literature Review

Many studies employing AI have been done in the realm of managing natural disasters. A large amount of information available online, in the opinion of humanitarian workers, is crucial for efficient disaster management. Making sense of material on social media in time-sensitive situations is a difficult undertaking, even though it could be valuable for response agencies [2]. Due to the vast number and high frequency of social media data streams, it is, for example, unfeasible to manually examine thousands of social media messages [14]. Analyzing social media data to draw out information that can save lives and aid humanitarian groups in emergency preparedness, response, and recovery. Numerous studies have used social site data to examine how people behave in public and aspects of their lives, such as their racial and ethnic backgrounds[8]. Additionally, social media platforms have served as active sensors during emergency situations, including disasters [7] [9]. (e.g., flood, hurricane, & earthquake). To reliably extract high-level information, tweet-level extraction is essential. It may be a good indication that this is the case to see, for example, that many tweets from nearby regions describe the same infrastructure as being destroyed [5].

As online news becomes more popular and more places have their own news websites, the news also carries important information about the tragedy. The majority of previous studies concentrated on cutting-edge techniques for extracting catastrophe information from a certain kind of web material. A solution based on machine learning techniques was described by Valero et al. to enhance the gathering of disaster information from internet news reports [3]. Duplicate information is inevitable when news is gathered from various sources. There are many ways to find duplicate reports, and fuzzy duplication detection is one of them[12]. Due to the diversity of the corpora used to construct the word embeddings, Bidirectional LSTM or CNN may be used for Tweet Classification for Disaster Response in order to leverage different word embeddings. [11]

Handcrafted characteristics were employed in traditional approaches to NLP classification tasks[15, 16, 17, 19]. [19] describes how the authors categorized the situational data during a disaster using bags-of-words models. The vocabulary of the tweets affects the features like n-grams, Parts-Of-Speech tags (POS), etc. Similar to this, the authors in [15] created a platform called Artificial Intelligence Disaster Response (AIDR), which classifies data linked to crises during a disaster using unigram and bi-gram characteristics. In [17], the authors developed features for identifying resource tweets after a disaster by mining the informative terms from the tweets. For classifying the 964 vocabulary terms, the authors [16] created low-level, vocabulary-unrelated syntactic and lexical factors. The use of the wh-word, the use of numerals, the number of non-situational terms, etc. are examples of low-level lexical features, whereas personal pronoun usage and intensifier usage are examples of low-level syntactic features. Utilizing content terms, the SVM classifier summarizes tweets and categorizes situational information. Later, tweets in other languages except English were included. In [5], the authors created a classification ranking method for finding verifiable microblogs in emergency situations. Their approach uses SVM classifiers with linear kernels and hand-crafted features for classification. These techniques, however, only consider tweets from a particular class, and feature engineering is required for classification. Word embeddings attracted a lot of interest as a result of the automatic feature extraction from the tweets. When compared to traditional classifiers, the authors in [20, 21] showed amazing performance in crisis-related tasks when using CNN with word embeddings. A combination of manually created BOW features with ANN and SVM classifiers is used. For classifying the crisis-related data, crisis embeddings are used with MLP-CNN and CNN. Word2vec's skip-gram model [14] was applied to an extensive corpus of almost 57,908 tweets relating to crises for the crisis embedding process. A word embedding-based deep learning model was created by the authors of [22] for identifying informative tweets during a crisis. For the purpose of identifying the informative tweets during the catastrophe, they identified crisis-relevant data during a disaster using crisis-specific word embeddings. In [25], scientists took a more sophisticated technique by classifying the social media data using hybrid algorithms that incorporated both text and image-based models where the author utilized FastText to classify the text, and VGG16 to classify the photos.

Few studies have looked into multimodality approaches to disaster circumstances analysis. However, choosing a classification model with a straightforward architecture and good accuracy can improve those existing multi-model designs. The goal of the study is to improve disaster response through social media by analyzing social media data to extract information that can aid humanitarian groups in emergency preparedness, response, and recovery. The use of multiple modalities, such as text and image data, can provide a more comprehensive understanding of the situation and help to overcome the limitations of using a single modality. Therefore, the proposed approach of using a BiLSTM and ResNet multimodal architecture aims to demonstrate its superiority compared to models developed using a single modality and also FastText and VGG-16 multimodality. By using an advanced and effective model, the study aims to contribute to the ongoing efforts to improve disaster response through social media. ResNet and BiLSTM are chosen for this proposed multimodal architecture due to their ability to capture and learn complex features from different modalities. ResNet, or residual neural network, is a type of deep neural network that is specifically designed to overcome the problem of vanishing gradients, which can occur when training very deep neural networks. ResNet achieves this by using skip connections, which allow the network to bypass some layers and directly propagate information from one layer to another. This architecture has been shown to perform well in a variety of computer vision tasks, including image classification, object detection, and segmentation. In the context of multimodal learning, ResNet can be used to extract features from visual modalities such as images and videos. BiLSTM, or bidirectional long short-term memory, is a type of recurrent neural network that is able to capture both forward and backward dependencies in sequential data. This is achieved by using two separate LSTMs, one that processes the input sequence from beginning to end and another that processes it from end to beginning. BiLSTM has been shown to be effective in a variety of natural language processing tasks, including sentiment analysis, machine translation, and named entity recognition. In the context of multimodal learning, BiLSTM can be used to extract features from textual modalities such as social media posts and news articles.

The combination of ResNet and BiLSTM in a multimodal architecture allows for the learning of complex features from multiple modalities, which can improve the performance of disaster response systems. By extracting features from both visual and textual modalities, the proposed architecture can better capture the diverse and complex nature of disaster-related information on social media sources.

3. Methodology

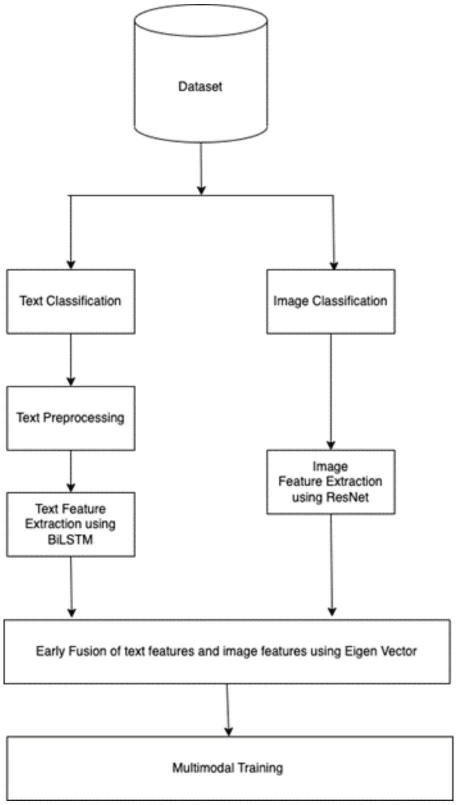

Dataset
Text Classification
Image Classification
Text Preprocessing
Image Feature Extraction using ResNet
Text Feature Extraction using BiLSTM
Early Fusion of text features and image features using Eigen Vector
Multimodal Training

Figure 1: Methodology Block Diagram

3.1 Dataset Preparation

AIDR consists of various disaster-related labeled datasets. A few of the datasets are mentioned below

- CrisisMMD dataset [28]

3.1.1 Dataset Cleaning

- Remove URL - Many tweets consisted of URLs using URL shortener which didn’t carry significance for our related tasks.

- Convert to lowercase - All the texts were converted to lowercase

- Remove punctuation - Removed the punctuations
- !”#$%&’()*+,-./:;<=>?@[\]ˆ ‘{}˜
- Remove remaining tokens that are not alphabetic
- Filter out stop words - Following stop words were filtered out

  {”couldn’t”, ’him’, ’being’, ’under’, ”she’s”,’me’, ’off’, ”that’ll”, ”aren’t”, ’re’, ’or’, ’we’,’against’, ’nor’, ’to’, ”won’t”, ’own’, ’shan’,’and’, ’by’, ’for’, ”mightn’t”, ’ourselves’,’yourselves’, ’aren’, ’were’, ’between’, ’didn’, ’can’, ’haven’, ’my’, ”didn’t”, ’has’, ’does’, ’did’, ’than’, ’o’, ’yours’, ’won’, ’that’, ’am’, ’all’, ”hasn’t”, ”wasn’t”, ’myself’, ’both’, ’our’, ’mightn’, ’from’, ’ain’, ”weren’t”, ’below’, ’down’, ’such’, ’is’, ”wouldn’t”, ”you’ll”, ’these’, ’isn’, ’her’, ’be’, ’will’, ’d’, ’then’, ’itself’, ’theirs’, ’your’, ”you’d”, ’himself’, ’doing’, ’but’, ’over’, ’needn’, ’them’, ’the’, ’only’, ’through’, ”it’s”, ”you’ve”, ’wouldn’, ’about’, ’up’, ’out’, ’now’, ”hadn’t”, ’hadn’, ’so’, ’it’, ’other’, ’hers’, ’how’, ’he’, ’are’, ’wasn’, ’ve’, ’whom’, ’same’, ’with’, ’what’, ’more’, ’too’, ”shouldn’t”, ’i’, ’yourself’, ’his’, ’shouldn’, ’had’, ”isn’t”, ’where’, ’s’, ’those’, ’while’, ’no’, ’ll’, ”should’ve”, ’y’, ’after’,”you’re”, ”doesn’t”, ”needn’t”, ’above’, ’themselves’, ’having’, ’if’, ’there’, ’should’, ’not’, ’have’, ’before’, ’during’, ’a’, ’which’, ’couldn’, ’an’, ’t’, ’herself’, ’of’, ’some’, ’why’, ’few’, ’they’, ’once’, ’do’, ’weren’, ’she’, ’here’, ”haven’t”, ’this’, ’mustn’, ’very’, ”shan’t”, ’again’, ’been’, ’until’, ”don’t”, ’just’, ’into’, ’ours’, ’who’, ’on’, ’most’, ’its’, ’hasn’,”mustn’t”, ’you’, ’at’, ’don’, ’their’, ’as’, ’m’, ’further’, ’when’, ’any’, ’ma’, ’because’, ’each’, ’was’, ’doesn’, ‘in’}

3.1.2 Word Embedding

Word Embedding is a widely used technique to represent text numerically. It maps each word to a high-dimensional vector, where each dimension represents a particular feature of the word. This technique preserves the semantic relationships between words and captures their meanings. Various models such as Word2Vec, Glove, etc., are available for creating word embeddings. In our study, we used the pre-trained word embeddings provided by crisisNLP, which is a publicly available dataset of crisis-related social media messages. These pre-trained embeddings were trained on a large corpus of crisis-related tweets, which is similar to our dataset. Using these pre-trained embeddings allowed us to leverage the semantic information in the text and improve the performance of our models.

3.2 Models

3.2.1 Text Classification Model

We used BiLSTM for the text classification task. Multiple feature engineering methods are being used, including Count Vectors & TF-IDF Vectors as features. To train our RNN model, pre-trained word-embedding vectors are used.

3.2.1.1 Bidirectional Long Short-Term Memory (BiLSTM)

A Bidirectional Long Short-Term Memory (BiLSTM) is a type of neural network architecture used for sequential data processing tasks like text classification, speech recognition, and time series analysis. BiLSTM is an extension of the standard LSTM network and adds another layer of LSTM cells to process the input sequence in reverse order.

The BiLSTM network consists of two LSTM layers, one that processes the input sequence in a forward direction and another that processes the input sequence in a backward direction. The output of both the forward and backward layers is then concatenated, resulting in a sequence of vectors that captures information from both past and future contexts of the input sequence.

The forward layer processes the input sequence from the beginning to the end, while the backward layer processes the input sequence from the end to the beginning. Each LSTM cell in the BiLSTM network has three gates: the input gate, output gate, and forget gate. These gates control the flow of information through the cell and allow it to selectively remember or forget information from previous time steps.

During training, the BiLSTM network learns to update its internal state based on the input sequence and generate output predictions. The network is optimized by minimizing a loss function that measures the difference between the predicted and actual outputs.

In summary, BiLSTM is a powerful architecture for sequential data processing that can capture both past and future contexts of the input sequence. It has been widely used for text classification tasks, achieving state-of-the-art results on many benchmark datasets.

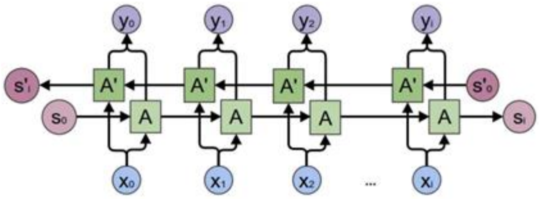


Figure 2: BiLSTM

3.2.2 Image Classification Model

For image classifications we'll be using multiple CNN models. VGG16 and Resnet.

3.2.2.1 VGG16

VGG16 is a deep CNN architecture proposed by the Visual Geometry Group at Oxford in 2014, widely used for image classification. It has achieved state-of-the-art performance on benchmark datasets, consisting of 16 layers: 13 convolutional, 5 max-pooling, and 3 fully connected layers. The convolutional layers extract image features, while the max-pooling layers reduce spatial dimensions. VGG16's key feature is the use of small 3x3 convolutional filters to capture fine-grained features and reduce parameters, allowing for deeper networks to be trained with fewer resources. VGG16 is pre-trained on ImageNet with over 1 million labeled images in 1000 classes. Its pre-trained weights can be fine-tuned on other image classification tasks or retrained on new datasets.

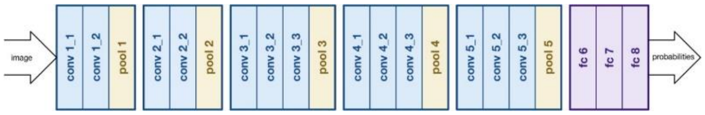


Figure 3: VGG16 Architecture

3.2.2.2 ResNet

ResNet (Residual Network) is a deep neural network architecture that was developed by Microsoft researchers in 2015. The main idea behind ResNet is to address the problem of vanishing gradients that occur when training very deep neural networks. Vanishing gradients occur when the gradient signal becomes smaller and smaller as it backpropagates through the layers of a deep neural network, making it difficult for the network to learn and update its weights. This problem can be particularly severe in very deep networks with dozens or hundreds of layers. ResNet addresses this problem by introducing a ”residual” connection that allows the gradient signal to bypass one or more layers in the network. This means that even if the gradient signal becomes very small, it can still be propagated through the network and used to update the weights. In ResNet, each block of layers in the network contains a ”shortcut” or ”identity” connection that bypasses one or more convolutional layers. This shortcut connection is added to the output of the convolutional layers and then passed through an activation function (usually ReLU) before being added back to the output of the previous block. The addition of these shortcut connections allows the network to learn residual mappings that are easier to optimize than the original mappings, which can result in better accuracy on tasks such as image classification. ResNet has been shown to achieve good performance on a wide range of computer vision tasks, including image classification, object detection, and segmentation.

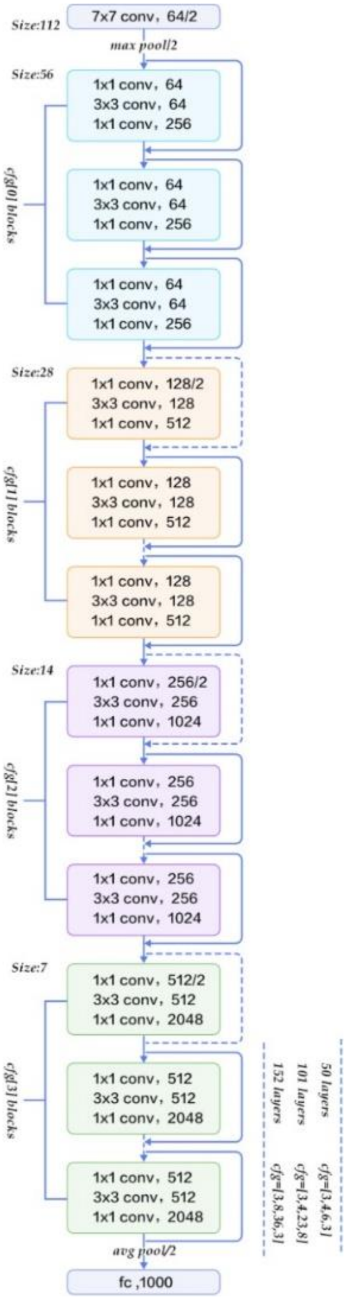


Figure 4: ResNet Architecture

3.2.3 Multimodel Classifier

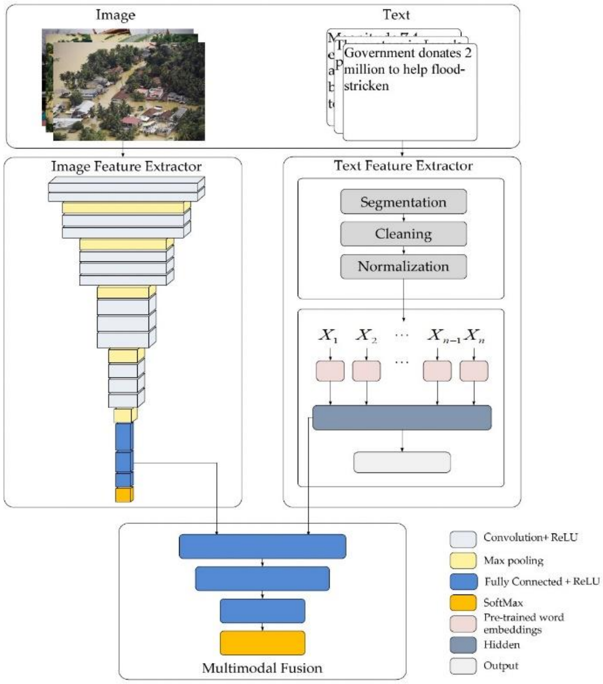


Figure 5: Multimodal Classifier

3.3 Metrics

The metrics used to measure the performance model include: Accuracy, Precision, Recall, F-Measure

3.3.1 Accuracy

It is the percentage of the total percentage of correctly classified Tweets. Accuracy = (TP+TN)/Total Number of Tweets Where TP is True Positive and TN is True Negative.

3.3.2 Precision

It is the proportion of positive cases identified that are real positives. Precision = TP/(TP+FP)

3.3.3 Recall

It is the proportion of actual positive cases that were rightly identified Recall = TP/(TP+FN)

3.3.4 F1 Score

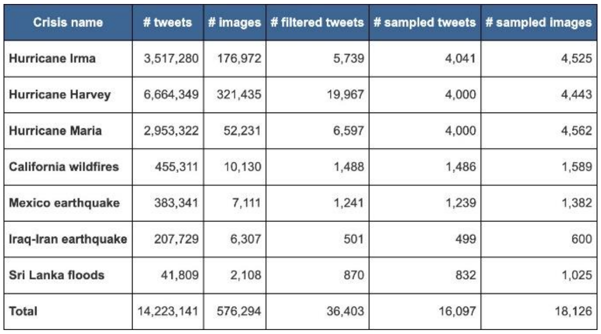
F1 score is used to measure the effectiveness of the

classifier. It is the harmonic mean of precision and recall.
F1 Score = [2*Precision*Recall]/[Precision+Recall]

4. Results and Discussion

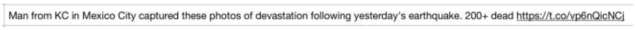
We used RNN model BiLSTM to solve the classification problem.
The datasets contain the following information:

Table.1: Dataset sample distribution



Crisis name	# tweets	# images	# filtered tweets	# sampled tweets	# sampled images
Hurricane Irma	3,517,280	176,972	5,739	4,041	4,525
Hurricane Harvey	6,664,349	321,435	19,967	4,000	4,443
Hurricane Maria	2,953,322	52,231	6,597	4,000	4,562
California wildfires	455,311	10,130	1,488	1,486	1,589
Mexico earthquake	383,341	7,111	1,241	1,239	1,382
Iraq-Iran earthquake	207,729	6,307	501	499	600
Sri Lanka floods	41,809	2,108	870	832	1,025
Total	14,223,141	576,294	36,403	16,097	18,126

Here are some tweets & sample images used in our dataset.



Man from KC in Mexico City captured these photos of devastation following yesterday's earthquake. 200+ dead https://t.co/vp6nQicNCj

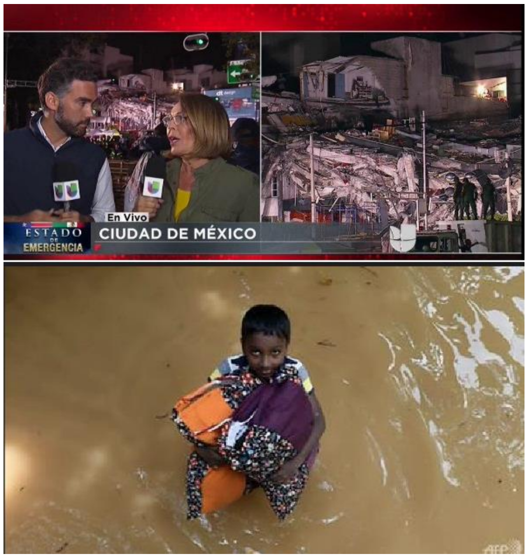


Figure 6: Dataset samples

4.1 Text Classification

AIDR consists of various disaster-related labeled datasets. The following datasets were used:

- CrisisMMD dataset [28]

4.1.1 Dataset Cleaning

The majority of terms in sentences that need cleaning have punctuation at the beginning or end. Due to the trailing punctuation, those words lack embeddings. Words are separated from punctuation by the following symbols: #, @,!,?, +, &, -, $, =, >, ’, (,),[,],*,%,..., ’,.,:, ;
Words that formerly had special characters connected to them are completely eliminated. Contractions have been widened, URLs have been deleted, character entity references have been changed to their actual symbols, typos and slang have been fixed, and informal abbreviations have had their lengthy forms written. Some words have been combined into one, while others have had their acronyms substituted.

4.1.2 Shuffling & Splitting

Pandas DataFrame was used to store the data sets. A 2-dimensional labeled data structure called a ”DataFrame” has columns that could be of many sorts. Before splitting the dataset we shuffled it. Graphs were plotted to show labeled data classification using seaborn.

# of training samples: 9601
# of test samples: 1573
# of valid samples: 1534

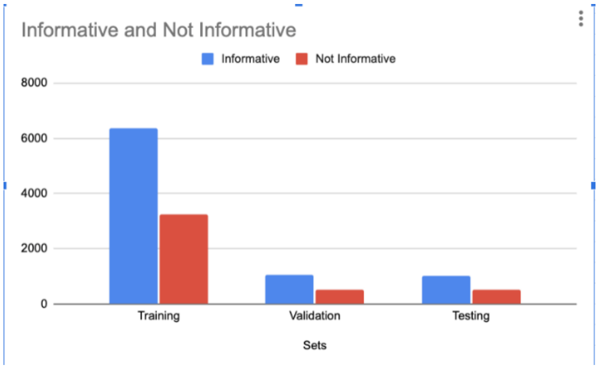


Figure 7: Labeled Data Distribution

4.1.3 Label Encoding

Using Sklearn Library, Label Encoding in Python was **Figure 7:** accomplished. A very effective method for converting the levels of categorical features into numerical values is offered by Sklearn. LabelEncoder encrypts labels with a value between 0 and n-1 classes, where n is the total number of labels. If a label appears more than once, the previous value is assigned when it does.

4.1.4 Data Analytics

Different types of graphs were visualized to analyze the data sets. Some of them are shown in the figure below:

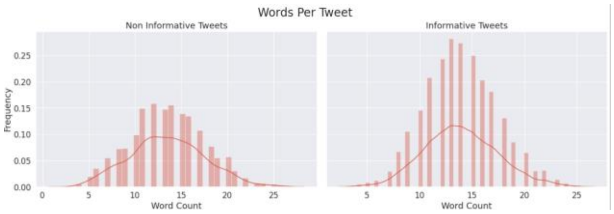


Figure 8: Words Per tweet for informative & non-informative tweets.

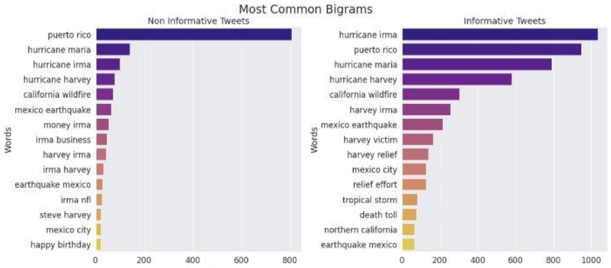


Figure 9: Most common bigrams

4.1.5 Model Building

We have used the pre-trained word-embedding from CrisisNLP vectors as features & used a popular method of RNN called BiLSTM. The padded training sets were used for training x & the label encoded values were used to train with a batch size of 128 and 100 epochs with early stopping criteria. Adam optimizer from Keras was used whose learning rate is 1e-05 by default.

- Hyperparameter Used
  Batch size: 128
  Learning Parameter: 1e-05
  Epochs: 10 (Used Early stopping criteria)
  Loss criteria: Binary Cross entropy
  Optimizer: Adam

4.1.6 Results

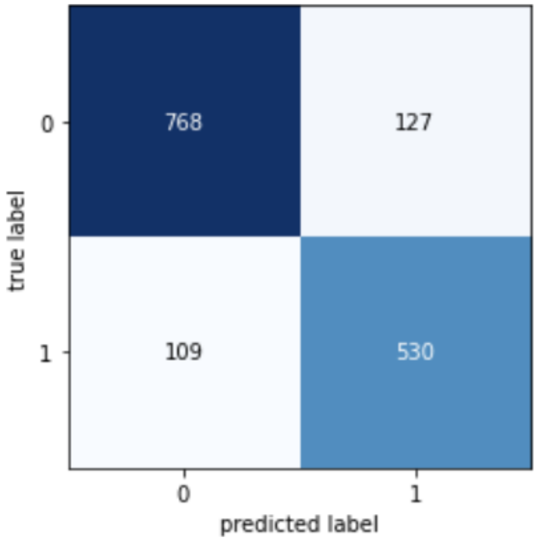


Figure 10 : Confusion Matrix for BiLSTM mode

The confusion matrix is shown in Figure 4.5. Overall, an accuracy of 84.61% was achieved. This is due to the misclassification of some training samples, as evidenced by the confusion matrices

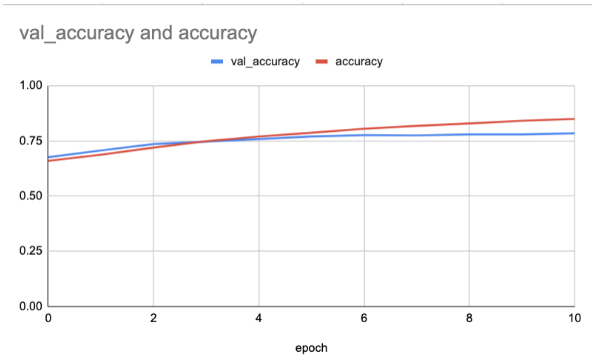


Figure 11: Val Accuracy vs Accuracy for BiLSTM

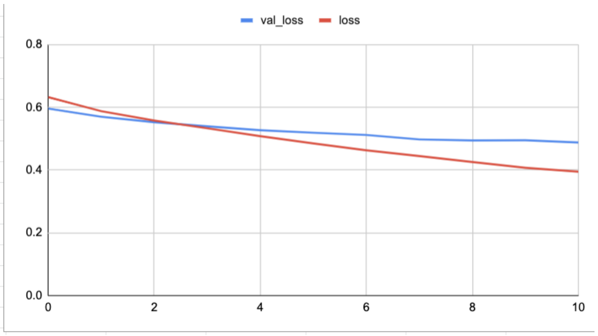
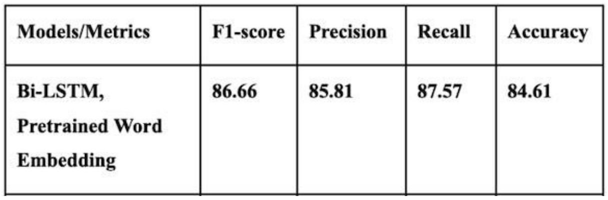


Figure 12: Val Loss vs Loss for BiLSTM

Table 2: Scores for training models for classification of crisis-related data using BiLSTM.



Models/Metrics	F1-score	Precision	Recall	Accuracy
Bi-LSTM, Pretrained Word Embedding	86.66	85.81	87.57	84.61

4.2 Image Classification

4.2.1 Data Preparation

The datasets were downloaded from the crisis NLP datasets. Few images from the dataset are given below:

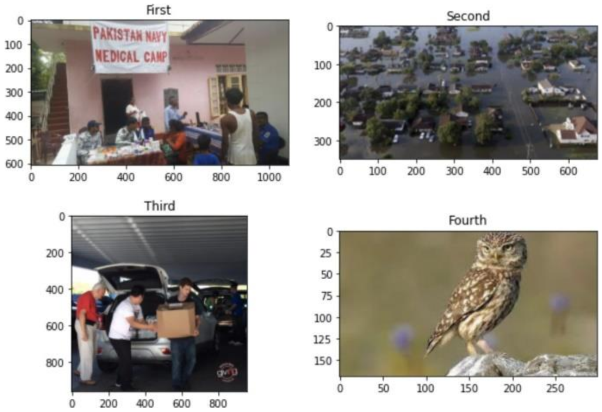


Figure 13: Image dataset samples

4.2.2 Model Building

We have implemented VGG16 & ResNet152 to train our model against the same datasets we used to train our text classification model with the following splits:

# of training samples: 9601
# of test samples: 1573
# of valid samples: 1534

- Hyperparameters Used
  Batch size: 16
  Learning Rate: 1e-6
  Epochs: 100
  Loss criteria: Categorical Cross Entropy
  Optimizer: Adam

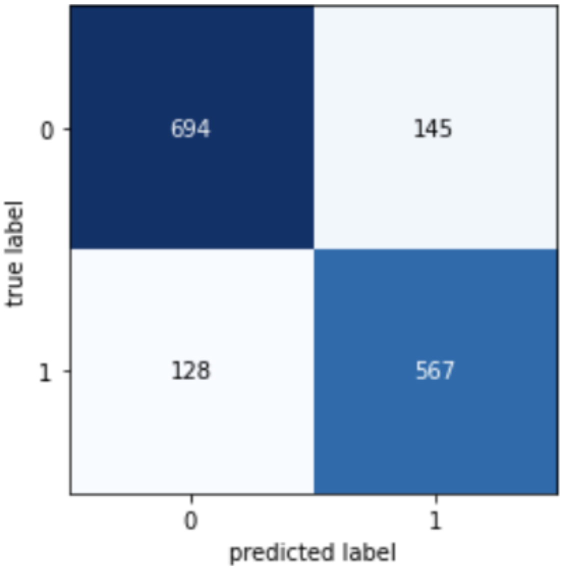


Figure 14: Confusion Matrix for ResNet

The confusion is shown in Figure 4.9. Overall, an accuracy of 82.34% was achieved. This is due to the misclassification of some training samples, as evidenced by the confusion matrices.

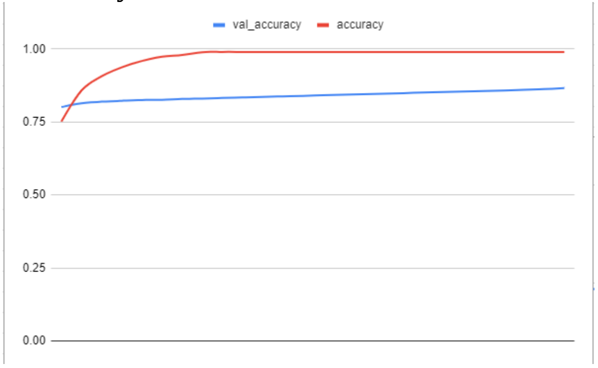


Figure 15: Val Accuracy vs Accuracy for ResNet

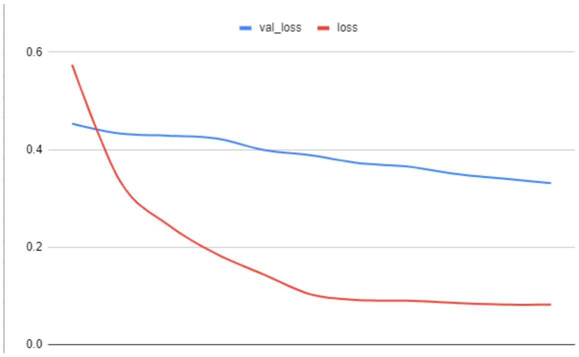
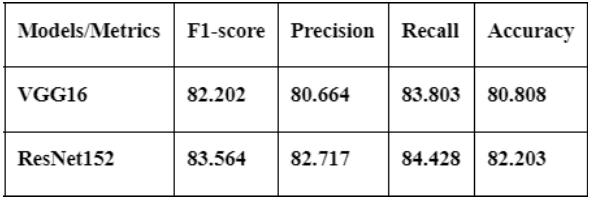


Figure 16: Val Loss vs Loss for ResNet

Table 3: Comparative scores for different image training models



Models/Metrics	F1-score	Precision	Recall	Accuracy
VGG16	82.202	80.664	83.803	80.808
ResNet152	83.564	82.717	84.428	82.203

4.3 Early Fusion

Early fusion is carried out at the feature level. In this instance, various feature vectors from various sources are combined into one substantial feature vector that was later applied to classification. This vector has a lot of features, thus training and classification took longer.

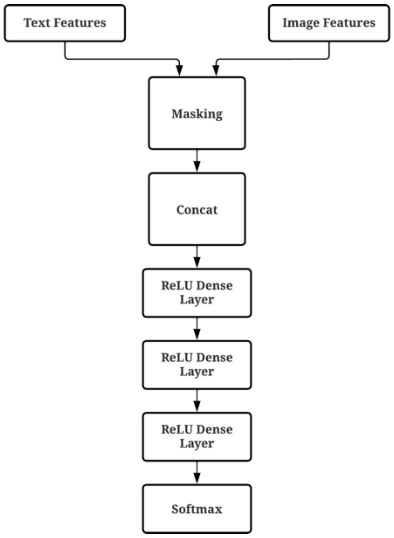
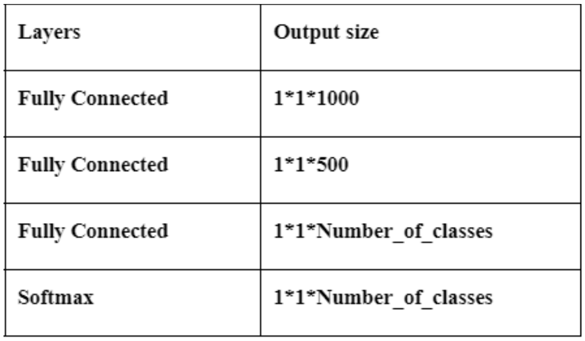


Figure 17: Block diagram of Early fusion process

The BiLSTM for text feature extraction and ResNet152 for image feature extraction multimodal approach for text and image classification consists of an early fusion strategy that combines text and image features. The text features are fed into the bi-directional LSTM network, which learns the sequential relationships between words in the text data. The image data is passed through the ResNet network, which learns hierarchical representations of the image features. The image features extracted is then exposed to dimension reduction technique by using Eigen vectors. The outputs from the bi-directional LSTM and ResNet network after dimension reduction are then combined through a softmax layer to make the final prediction. This approach has shown promising results in accurately classifying text and image data. Three completely interconnected layers make up our multimodal fusion, together with a SoftMax layer. We choose to combine two 500-dimensional eigenvectors into one 1000-dimensional eigenvector using a straightforward concatenation in series rather than intricate eigenvector alignment techniques. Finally, it was possible to achieve the final prediction results using the four network layers mentioned above.

Table 4: Multimodal fusion model architecture



Layers	Output size
Fully Connected	1*1*1000
Fully Connected	1*1*500
Fully Connected	1*1*Number_of_classes
Softmax	1*1*Number_of_classes

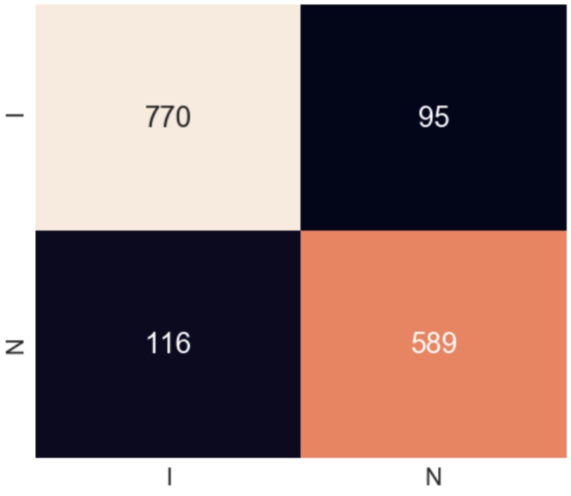
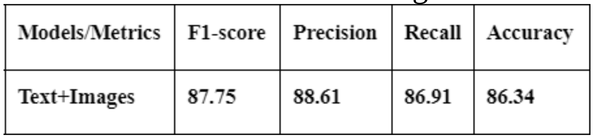


Figure 18: Confusion Matrix for fused multimodal

The confusion is shown in Figure 4.13. Overall, an accuracy of 86.34% was achieved. This is due to the misclassification of some training samples, as evidenced by the confusion matrices.

The overall metrics calculated in this training are given in the table below:

Table 5: Scores for text & image fusion model

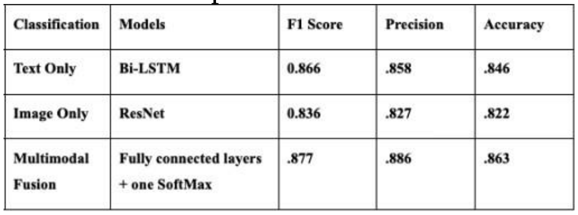


Models/Metrics	F1-score	Precision	Recall	Accuracy
Text+Images	87.75	88.61	86.91	86.34

4.4 Comparison between single modality and multimodality

Using BiLSTM for test classification and the Resnet152 for image classification, the multimodal is achieved by early fusion means between them. We computed the same metrics on the validation and test sets in order to compare the text classification using BiLSTM, image classification using ResNet to the multimodal classification. The comparison between the models is displayed in the table below.

Table 6: Comparison with baseline model



Classification	Models	F1 Score	Precision	Accuracy
Text Only	Bi-LSTM	0.866	.858	.846
Image Only	ResNet	0.836	.827	.822
Multimodal Fusion	Fully connected layers + one SoftMax	.877	.886	.863

5. Conclusion

Social media is flooded with posts from victims of disasters detailing what they have gone through and seen. Sorting through the vast social media data to find pertinent information can aid relief personnel in determining the severity of a disaster. A mechanism is required to extract information that is pertinent to disasters because the material on social media is frequently huge and disorganized. However, the majority of earlier research on this subject has concentrated on either text analysis or picture analysis, and just a small number of studies have used multimodal techniques. Their classification accuracy, even when employing multimodal techniques, is not particularly good. In this study, we proposed a multimodal technique for categorizing disaster images. In order to perform the classification tasks, the text features were simultaneously combined with the image features that were extracted using the deep learning method. The efficiency of the suggested strategy is demonstrated by experimental findings using datasets from actual disasters. The proposed multimodal strategy outperforms the unimodal one in terms of overall performance and accuracy compared to the existing multimodal technique. The model's architecture also makes training it easier.

Acknowledgements

This research work has been supported by the University Grants Commission, Nepal under a Collaborative Research Grant (UGC Award No. CRG-74/75-Engg-01).

References

[1] Imran, M., Castillo, C., Diaz, F., Vieweg, S., (2015). Processing social media messages in mass emergency: A survey. ACM Computing Surveys (CSUR) 47, 67.

[2] Hiltz, S. R. and Plotnick, L. (2013). "Dealing with information overload when using social media for emergency management: Emerging solutions." In: Proceedings of the 10th international conference on information systems for crisis response and management (ISCRAM2013). ISCRAM.

[3] Téllez, V.A.; Manuel, M.Y.G.; Villaseñor, P.L. Using Machine Learning for Extracting Information from Natural Disaster News Reports. Comput. Y Sist. 2009, 13, 33–44.

[4] Imran, M.; Castillo, C.; Diaz, F.; and Vieweg, S. 2015. Processing social media messages in mass emergency: A survey. ACM Computing Surveys 47(4):67.

[5] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, Patrick Meierm, 2013, Practical Extraction of Disaster-Relevant Information from Social Media

[6] S. Wakamiya, R. Lee, and K. Sumiya. ”Crowd-based urban characterization: extracting crowd behavioral patterns in urban areas from twitter.” In Proceedings of the 3rd ACM SIGSPATIAL international workshop on location-based social networks, pp. 77-84. ACM, 2011.

[7] A. Kongthon, C. Haruechaiyasak, J. Pailai, and S. Kongyoung. ”The role of Twitter during a natural disaster: Case study of 2011 Thai Flood.” In 2012 Proceedings of PICMET’12: Technology Management for Emerging Technologies, pp. 2227-2232. IEEE, 2012.

[8] L. Mitchell, M. R. Frank, K. D. Harris, P. S. Dodds, and C. M. Danforth. ”The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place.” PloSone 8, no. 5 (2013): e64417.

[9] T. Sakaki, M. Okazaki, and Y. Matsuo. ”Earthquake shakes Twitter users: real-time event detection by social sensors.” In Proceedings of the 19th international conference on World wide web, pp. 851-860. ACM, 2010.

[10] S. E. Vieweg, ”Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications.” 2012.

[11] Reem ALRashdi, Simon O’Keefe.2018.Deep Learning and Word Embeddings for Tweet Classification for Crisis Response

[12] Per Runeson, Magnus Alexandersson and Oskar Nyholm (2007) “Detection of Duplicate Defect Reports Using Natural Language Processing”, Software Engineering Research Group, Lund University, Box 118, SE-221 00 Lund, Sweden

[13] Qatar Computing Research institute, Referenced from https://crisisnlp.qcri.org/, 2019

[14] Hiltz, S. R., Kushma, J., and Plotnick, L. (2014). “Use of Social Media by U.S. Public Sector Emergency Managers: Barriers and Wish Lists”. In: University Park, Pennsylvania, USA.

[15] Imran, M., Castillo, C., Lucas, J., Meier, P., Vieweg, S., (2014). Aidr: Artificial intelligence for disaster response, in: Proceedings of the 23rd International Conference on World Wide Web, ACM. pp. 159–162.

[16] Rudra, K., Ghosh, S., Ganguly, N., Goyal, P., Ghosh, S., (2015). 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, ACM. pp. 583–592.

[17] Sreenivasulu, M., Sridevi, M., (2017). Mining informative words from the tweets for detecting the resources during disaster, in: International Conference on Mining Intelligence and Knowledge Exploration, Springer. pp. 348–358.

[18] Sreenivasulu, M., Sridevi, M., (2018). A survey on event detection methods on various social media, in: Recent Findings in Intelligent Computing Techniques. Springer, pp. 87–93.

[19] Verma, S., Vieweg, S., Corvey, W.J., Palen, L., Martin, J.H., Palmer, M., Schram, A., Anderson, K.M., (2011). Natural language processing to the rescue? extracting” situational awareness” tweets during mass emergency., Citeseer. pp. 385–392.

[20] Caragea, C., Silvescu, A., Tapia, A.H., (2016). Identifying informative messages in disaster events using convolutional neural networks, in: International Conference on Information Systems for Crisis Response and Management.
[8] Chollet, F., et al., (2015). Keras. https://github.com/fchollet/keras.

[21] Nguyen, D.T., Mannai, K.A.A., Joty, S., Sajjad, H., Imran, M., Mitra, P., (2016). Rapid classification of crisis-related data on social networks using convolutional neural networks. arXiv preprint arXiv:1608.03902

[22] Madichetty, S., Sridevi, M., (2019). Detecting informative tweets during disaster using deep neural networks, in: 2019 11th International Conference on Communication Systems & Networks (COMSNETS), IEEE. pp. 709–713

[23] Christopher Olah. Understanding LSTM Networks. Available Online: https://colah.github.io/posts/2015-08-Understanding-LSTMs/. Taken on Dec 1, 2021

[24] Jannik Ro¨ ßler. LSTM with Keras. Available Online: https://mc.ai/lstm-with-keras/. Taken on Dec 1, 2020.

[25] Zhiqiang Zou , Hongyu Gan, Qunying Huang , Tianhui Cai, Kai Cao, (2021). Disaster Image Classification by Fusing Multimodal Social Media Data. International Journal of Geo-Information, ISPRS Int. J. Geo-Inf. 2021, 10, 636.

[26] Tilon, S.; Nex, F.; Kerle, N.; Vosselman, G. Post-Disaster Building Damage Detection from Earth Observation Imagery Using Unsupervised and Transferable Anomaly Detecting Generative Adversarial Networks. Remote Sens. 2020, 12, 4193.

[27] Gautam, A.K.; Misra, L.; Kumar, A.; Misra, K.; Aggarwal, S.; Shah, R.R. Multimodal Analysis of Disaster Tweets. In Proceedings of the 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), Singapore, 11–13 September 2019; pp. 94–103.

[28] Alam, F.; Ofli, F.; Imran, M. CrisisMMD: Multimodal twitter datasets from natural disasters. In Proceedings of the 12th International AAAI Conference on Web and Social Media, ICWSM 2018, Palo Alto, CA, USA, 25–28 June 2018; pp. 465–473.

[29] Siddharth Das Available Online: https://medium.com/analytics-vidhya/cnns-architectures-lenet-alexnet-vgg-googlenet-resnet-and-more-666091488df5 Taken on Dec 30, 2021