

Multimodal Analysis of Disaster Tweets

Akash Kumar Gautam
MIDAS, IIIT-Delhi
akash15011@iiitd.ac.in

Ajit Kumar
Adobe Systems
ajikumar@adobe.com

Shashwat Aggarwal
MIDAS, IIIT-Delhi
shashwat.aggarwal9@gmail.com

Luv Misra
MIDAS, IIIT-Delhi
luvm.co@nsit.net.in

Kush Misra
MIDAS, IIIT-Delhi
kushm.co@nsit.net.in

Rajiv Ratn Shah
MIDAS, IIIT-Delhi
rajivrtn@iiitd.ac.in

Abstract—Social media is inevitably the most abundant source of actionable information in times of natural disasters. Most of the data is either available in the form of text, images or videos. Real-time analysis of such data during the events of calamities poses many challenges to machine learning algorithms that require a large amount of data to perform well. Multimodal Twitter Dataset for Natural Disasters (CrisisMMD) is one such novel dataset that provides annotated textual as well as image data to researchers to aid the development of crisis response mechanism which can leverage social media platforms to extract useful information in times of crisis. In this paper, we analyze multimodal data related to seven different natural calamities like hurricanes, floods, earthquakes, *etc.* and propose a novel decision diffusion technique to classify them into informative and non-informative categories. The proposed methodology outperforms the text baselines by more than 4% accuracy and image baselines by more than 3%

Index Terms—Multimodal, CNN, decision fusion, CrisisMMD dataset, disaster analysis

I. INTRODUCTION

The problem of event analysis using social media has been widely studied in the past few years. Event analysis helps us to identify trending topics, understand public sentiments and monitor behavioral changes since an event is something which happens at a particular location involving some entities and people. Natural disasters fall into one such category of events where a large part of the event-related information is derived from peer to peer networks. With the ever-increasing access to mobile devices and penetration of social media platforms, public sharing of texts and images on websites like Twitter, Facebook and Instagram has become much popular. Although such data is useful for humanitarian aid workers and government agencies, rampant sharing of crisis-related posts has led to several other problems. One of the most fundamental problems related to the same is the mixing of non-informative posts with apropos posts.

According to the work [1], humanitarian aid involves assisting people during events of crisis to save lives, reduce suffering, and rebuild affected communities. It ensures that people who need necessities like food, water, shelter, medical assistance, and damage-free critical infrastructure and utilities such as roads, bridges, power-lines, and communication poles



Fig. 1: Sample multimodal tweet from hurricane Harvey belonging to the CrisisMMD dataset

are addressed effectively. Tweets and images that show one or more of the following: cautions, advice, and warnings, rescue, volunteering, or donation request, *etc.* are known to fall into the category of *informative* class label. More information regarding the tweets falling in these categories is given in Table I. While on the other hand, posts that do not throw any light in this context are labeled as *non-informative*. A tertiary label also exists in the CrisisMMD dataset [1] of *Do not know or cannot judge* type which is essentially reserved for non-English text or unclear images.

Often there is a need for posts to get immediate attention. However, due to the vast amounts of both informative and non-informative posts, it is not manually possible to segregate them. Thus it necessitates a method which can automatically find relevant or informative posts from social media. Works [34–37, 47, 48] suggest that a single modality (either text or image) is not able to provide the best results. Hence it is important to combine modalities and evaluate the results. Similarly, Kumar *et al.* [21] showed that exploiting information from multiple views improves the overall system accuracy.









Event Type	Tweet Text	Tweet Image
Caution	<i>California has a large Mediterranean climate... a setup for wildfires.</i>	
Advice	<i>Here is the 11pm advisory for Hurricane Harvey.</i>	
Rescue	<i>Rescue in full swing for the people of Iraq.</i>	
Volunteering	<i>How to help the victims of Sri-Lanka floods.</i>	
Donation Request	<i>Save Mart donating to #California fire victims</i>	
Injured People	<i>8 killed and 23 injured in Hurricane Imra's wrath.</i>	
Dead People	<i>At least 11 dead and 100 missing as wildfire rages across northern california.</i>	
Infrastructure Damage	<i>Mexico earthquake collapse: Rescue operations going on</i>	

TABLE I: Event type with corresponding tweet text and image example.

The authors aim to develop a robust classification system that utilizes both textual as well as visual cues to arrive at a prediction whether the given tweet and the image associated with it is informative, non-informative or cannot judge. For this task, a decision fusion technique has been devised that combines the probability scores of individual text and image classification predictions to arrive at a composite multimodal classification. The approach on which our results have been analyzed are closely related to probability-based late fusion strategies. According to the work [37], one of the advantages of any decision based late fusion strategy is being robust which makes it suitable to use along with any media representations. Hence, the same late decision-based strategy can also be extended to incorporate audio representations as well in the future.

The main contributions of the paper can be summarized as below:

- Analyze the accuracy of pre-trained CNN architectures on CrisisMMD for classification of Images.
- Evaluate the accuracy of standard text features and deep learning methods for classification of tweet text in CrisisMMD.
- Combination of both text and image-based modality for identification of disaster informative tweets in the

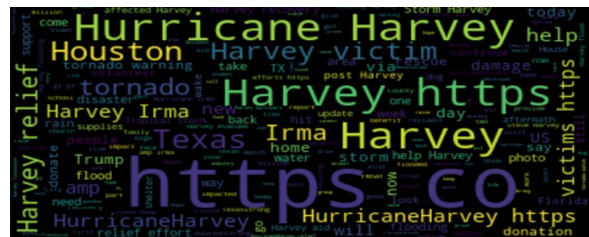


Fig. 2: Word cloud pertaining to the tweets from Hurricane Harvey in the dataset.

CrisisMMD dataset.

The rest of our paper is organized as follows. Section II discusses the literature review and Section III presents the dataset for the task at hand. Section IV contains information on experiments and broken down into the following three main subsections. IV-A discusses when only image modality is used, IV-B discusses the uses of text-only modality, and IV-C presents the decision policy incorporated to account for bi-modal data. Section V lists down insight concluded from the data as well as metrics incorporated to evaluate the results.

II. RELATED WORK

The problem of tweet classification has been explored in the recent literature. Mathur *et al.* [24, 25] used SVM classifier for

tweet classification. Deep-learning architectures have shown to improve the performance as compared to machine learning models as investigated by [26] and [32]. In recent past, Atrey *et al.* [2] has tried to summarize several multimodal fusion strategies such as correlation and independence, confidence level, contextual information, synchronization between different modalities, strategies applied to multimedia applications. A conventional multimodal system is designed to receive multiple representational inputs and should use some form of learning to characterize information. Rasiwasia *et al.* [31] investigated a novel joint modeling the text and image components of multimedia documents using latent Dirichlet allocation bags of visual (SIFT) features which were fused with canonical correlation analysis. It also showed that cross-modal correlations and semantic abstraction both improve multimedia retrieval accuracy. Recently, Yin *et al.* [46] used double fusion technique for scene classification. Thus, the multimodal fusion methodology is very effective.

Chang *et al.* [4] proposed a content-based soft annotation (CBSA) procedure for providing images with semantic labels in order to improve the multimodal image retrieval process by incorporating Support Vector Machines and Bayes Point Machines in tandem. Another interesting work done by [40] introduced the concept of applying Deep Boltzmann Machine for learning a generative model of data consisting of multiple and diverse input modalities which are highly useful for classification and information retrieval tasks. The model works by learning a probability density over the space of multimodal inputs, outperforming the former techniques in the domain. There are also more complex fusion strategies like Restricted Boltzmann Machines (RBMs) [28] and autoencoders [19] that have been employed for congruent tasks. However, these were restricted to video and audio representations only.

Slamet *et al.* [39] proposed a conceptual framework to implement a secure place locator (SPL), which aims to aid disaster management utilizing social media features. Imran *et al.* [15] described automatic methods for identification of information nuggets, brief self information items relevant to disaster response. Their work however focused solely on the textual modality of the tweets. Previously, Albuquerque *et al.* [7] proposed an approach to enhance the identification of relevant messages from social media which relies on volunteered geographic information, geographic features of social media and geography-based features of the disaster derived from authoritative data (sensor data, hydrological data, and digital evaluation models). Various strategies have been utilized in the past [30] for identification of sub-events related to an emergency. All these works, however, have utilized textual modality only, or are working on meta-data extracted from text-based elements of social media. Our work aims to utilize both text and image modalities for identification of informative tweets about a disaster-related scenario.

III. DATASET

During any natural or man-made disaster, people often use social media platforms like Twitter to convey their expressions

and relay important information. Many studies have revealed that relevant online information, when accessed in a timely and effective manner, can be worth gold for many humanitarian organizations that carry out relief operations. CrisisMMD dataset [1] is a large corpus of a bi-modal dataset having adequate data samples in both text and image format. The dataset comprises of tweets on seven prominent natural disasters like Hurricane Harvey, Mexico Earthquakes, California Wildfires which took place in the year 2017. Since there are no labels in the dataset for identifying the relevance of combined tweet text and imaged based modality we created a subset of the original dataset [1] containing only those data points which had the same set of ground truth that is they had both text and image-based labeling as *informative*, *non informative* or *cannot judge*. Throughout the paper, we will be referring to this dataset as **CleanCrisisMMD**. The intuition behind having a subset of the original dataset is that it could lead to a less number of false positives and it would also provide a basis for evaluation of modality fusion techniques.

Table II spells out the number of data instances present for each of the natural disasters in the CrisisMMD and CleanCrisisMMD. Alongside, table III portrays the annotation distribution across the text and image modalities for each natural disaster in the original CrisisMMD. There were no data points with label as *cannot Judge* in CleanCrisisMMD. Figure 1 shows a sample tweet pertaining to the Harvey Hurricane in the dataset, and figure 2 shows the word cloud created from the tweets belonging to Hurricane Harvey in the dataset.

IV. METHODOLOGY

Classification of tweets containing multiple modalities can be broken into a series of steps wherein first we try and evaluate the performance of various methods on individual modalities and then combine them using a multimodal decision fusion technique.

A. Image Only

In order to take advantage of features from ImageNet [8] dataset, we use popular pre-trained CNN architectures which have previously produced state of the art results such as VGG-16 [38], VGG-19 [9], ResNet50 [11], InceptionV2 [43], Xception [5] and DenseNet [14] model to extract features, and build a classifier on top of them. We primarily use transfer learning [29] to extend the pre-trained models to work on the CrisisMMD dataset. These models are fine-tuned by adding a Softmax layer which is used to get the final predicted class. In our experiment, we first re-sized the images through ImageDataGenerator [44] function to scale, translate and rotate each image in the dataset for data augmentation. The obtained images are fed into a pre-trained CNN model where the classifier is fine-tuned on the dataset to get the best results. Figure 3 explains this pictorially for VGG-19 pre-trained model.

B. Text Only Modality

As a first step, the tweets were pre-processed using the techniques provided by [25, 33, 45]. The following are the

Disaster Name	CrisisMMD			CleanCrisisMMD
	Text Data Points	Image Data points	Text and Image Data points	Combined Text and Image Data points
Sri Lanka Floods	832	1025	817	861
Hurricane Irma	4041	4525	2918	2799
Hurricane Harvey	4000	4443	3340	3168
Hurricane Maria	4000	4562	3260	3108
California Wildfires	1486	1589	1302	1205
Iran Iraq Earthquake	499	600	483	500
Mexico Earthquake	1239	1382	1192	1121
Total	16097	18126	13312	12762

TABLE II: Number of data points of each disaster in CrisisMMD and CleanCrisisMMD

Disaster	Crisis MMD						Clean Crisis MMD	
	Text			Image			Combined	
	Non-Informative	Informative	Can't Judge	Non-Informative	Non-Informative	Can't Judge	Informative	Non-Informative
Sri Lanka Floods	657	367	1	771	252	2	229	632
Hurricane Irma	957	2564	4	2303	2222	0	2032	767
Hurricane Maria	1718	2815	0	2326	2232	4	1813	1295
California Wildfires	344	1245	0	604	985	0	923	282
Iran Iraq Earthquake	105	493	2	199	400	1	398	102
Mexico Earthquake	1030	352	0	539	841	2	806	315
Hurricane Harvey	1109	2334	0	1978	2461	4	2262	906
Total	5920	10170	7	8720	9393	13	8463	4299

TABLE III: Annotation distributions across text and image modalities for each disaster in CrisisMMD and CleanCrisisMMD.

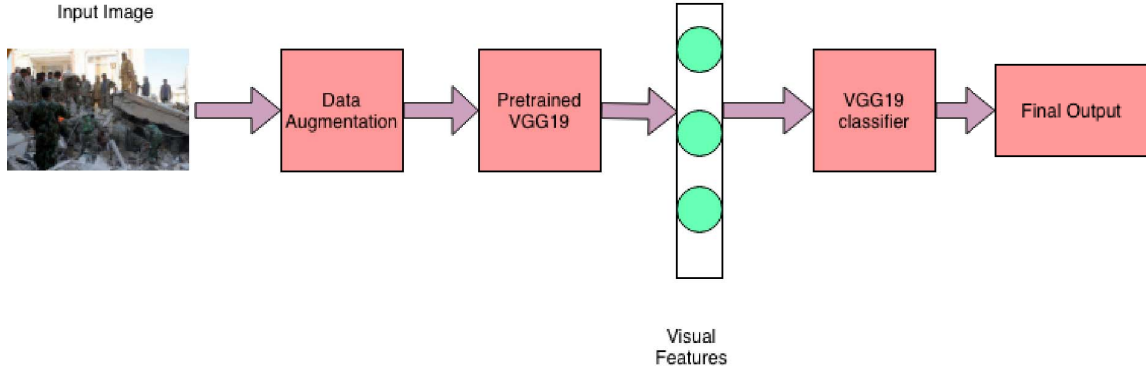


Fig. 3: Transfer learning methodology of image only classification using VGG-19

series of steps through which raw tweet text was passed in the order.

- 1) Removal the tweet if the length of the tweet is less than three words.
- 2) Removal of all the user mentions and URL.
- 3) According to [6] removal of hashtags from tweets which are of length greater than eight because the large size of hashtags will increase the size of vocabulary and may hamper the classification.
- 4) Removal of stop words and punctuation from the text.

The overall methodology for classifying tweets based on the just text has been broken down into two different approaches. The first approach uses N-gram feature representation which is fed to standard machine learning algorithms. The second approach uses sophisticated deep learning models for classification [16, 18, 23, 25, 26]. Figure 4 explains this pictorially.

1) N-gram based methods

After passing raw tweet text through a series of filters unigram and bigram based representations were extracted from the preprocessed text. The features mentioned are used to train Logistic Regression [13], Multinomial Naive Bayes [27] and Random Forest classifier [22]. Further, the problem of identifying tweets that can be informative is broken down into a supervised ternary classification problem.

2) Deep Learning Based Methods

Following deep learning-based methods are used for text classification.

- **LSTM**: These are special kinds of RNN's [42] and possess abilities to handle long term dependencies.
- **Bidirectional LSTM**: Bidirectional LSTM [10] have the ability to process input in two ways from past to future states and from future to past states. Thus two hidden states are able to preserve the required information in a more enhanced manner.

- **CNN+Glove:** This model makes use of pre-training and neural network for text classification. In this architecture [3], we use the pre-trained word embedding model trained on one billion tokens which are then fed to a Convolutional Neural Network for training and testing.

C. Combined Text and Visual Modalities

To handle dual modalities of the dataset, we need to combine the representations from different input sources using some form of merging technique. Zahavy *et al.* [49] summarizes that most techniques for merging the multimodal learning results are based on **features** and **decisions**. In this paper, we have focused on decision level fusion based techniques only since results by [37] suggest that late fusion yields better results than early fusion.

The results by [49] mention that best decision level fusion is the one that learned a decision rule using the class probabilities as input. Once we have the class prediction based probabilities from both text and image representations, we can concatenate them and use different policy systems to evaluate which systems are able to best filter the required information.

Based on the work [20], the architecture of the pre-trained network was chosen to be ResNet50 [12]. In order to establish text representation of the tweets of the disasters, bigram representations have been used. Considering the strategy presented in [17], bigram representations have produced excellent results. The text preprocessing and filtering strategy is the same as mentioned in the previous section. The policy system which was used to perform the late fusion strategies by combining the text and visual representations are mentioned as follows:

After obtaining the image and text representations using the mentioned feature approaches, we obtained the class prediction probabilities on the training data which will be used as an input for our policy system to output the class predictions probabilities of the test dataset. Figure 5 explains this pictorially.

- **Mean Probability Concatenation:** Class prediction probabilities obtained by both textual and visual modalities have been combined, averaged and then thresholded. This technique can serve as a competitive baseline and can provide results which are better than the performance of single modality only.
- **Custom Decision Policy:** After combining the decision probabilities of the training dataset from the text and visual representations, it was passed through a policy system. The policy decision system uses two fully connected layers. The first fully connected layer used Relu and second fully connected layer used Softmax as an activation function to output the class probabilities. The optimizer and loss functions which have been used by our custom policy are Rmsprop and Sparse Categorical Cross-entropy respectively.

- **Logistic Regression Decision Policy:** Suk *et al.* [41] summarized in their work that simple concatenation of different representations of text and visual modalities can be used efficiently for combining the information conveyed by individual text and visual modalities.

V. RESULTS AND OBSERVATIONS

We divided the dataset into 70, 10, and 20 ratios for training, validation, and testing set, respectively. To measure the performance of algorithms, accuracy was used as a metric. There have been two sets of experiments performed: the first one reports the accuracy on the original dataset and the second one analyzes the accuracy on CleanCrisisMMD. The results are reported in table IV and V.

Baseline Analysis

1) Image Only Modality

Table IV and V summarize the baseline results of various deep learning based techniques for different disasters on CrisisMMD and CleanCrisisMMD respectively. Using the pre-trained deep learning models, it can be concluded that the best accuracy for image based modality is achieved by the InceptionV3 model.

2) Text Only Modality

Table VI list down baseline accuracy of standard text based features on original dataset. Table VII do the same for CleanCrisisMMD.

It can be observed that while considering bigram representations of text, Random Forest-based classification approach outperforms other standard machine learning-based approaches in several cases, however, there is no clear classifier which outperforms others in unigram representation since unigram representations are very naive and are not able to capture text relationships efficiently. On the other hand, the bigram LSTM model has the best results which can be attributed to the ability of the model to capture long-term dependencies. Throughout the experimentation stage, the outperformance of the CrisisClean of considering analogous text and image labels is consistent with our hypothesis.

3) Combined Text and Visual Modality

Table VIII shows that Logistic Regression based decision policy outperforms the rest. It was able to obtain the best results in 5 out of 7 disaster instances when compared alongside mean probability and custom decision policy. It can be concluded that logistic regression in conjunction with bigram representations for text modality and ResNet50 architecture for image modality is able to provide the best set of results. Also, comparison of logistic regression-based decision policy with Inception V3, which is giving best results in case of image only modality, leads to a firm conclusion that decision-based policy outperforms the latter by at least 3%. Based on the improved accuracy score it can be concluded that logistic regression-based decision policy for incorporating text and image-based

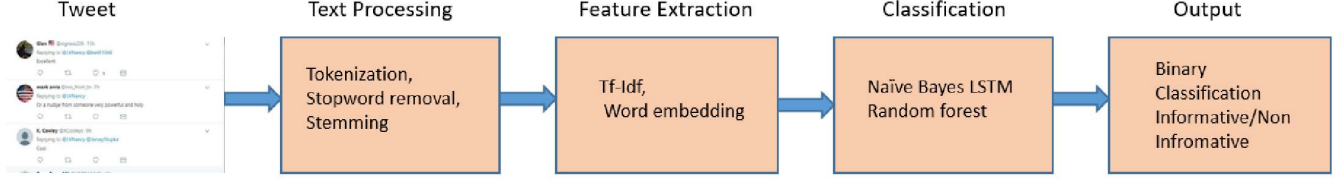


Fig. 4: Text only classification pipeline

Disaster	VGG16	ResNet50	VGG19	Inception v3	DenseNet121	Xception
Sri Lanka Floods	71	71	71	71	69.4	71.5
Hurricane Irma	48	48	49.3	77	61	67.4
Hurricane Harvey	51	51	51	76	76	68
Hurricane Maria	50	50	50	68.5	63	53
California Wildfires	64	64	64	69.3	68	68
Iran Iraq Earthquake	61	61	63.2	71	68.6	70.1
Mexico Earthquake	61	61	64	74.3	69	68.6

TABLE IV: Accuracy for image based models on CrisisMMD

Disaster	VGG16	ResNet50	VGG19	Inception v3	DenseNet121	Xception
Sri Lanka Floods	71	71	69	71	69	69.4
Hurricane Irma	68	68	71	68	71	68
Hurricane Harvey	72	72	73	72	73	71
Hurricane Maria	57	57	60	61	61	55.3
California Wildfires	73	73	71	73	71	70.1
Iran Iraq Earthquake	79	79	77	79	77	79
Mexico Earthquake	74	74	73	70	66	69

TABLE V: Accuracy for image based models on CleanCrisisMMD

Text Features	Unigram Features			Bigram Features			Deep Learning methods		
	Logistic Regression	Multinomial NB	Random Forest	Logistic Regression	Multinomial NB	Random Forest	LSTM	Bidirectional LSTM	CNN+Glove
Sri Lanka Floods	63	62.5	60.5	61.5	60	56.3	65	65	54
Hurricane Irma	66.3	62	57.7	61	65.7	66	77	77	72
Hurricane Harvey	65.1	60	66.8	61.1	63	59	76	78	72
Hurricane Maria	50.4	52	55.2	51	52.4	53	63	63	72
California Wildfires	58	61	63	65.3	66	69.5	75	81	74
Iran Iraq Earthquake	71.3	66.5	69	67	69	70	79	82	75
Mexico Earthquake	59.2	61	62.5	55	51	59	69	69	67

TABLE VI: Accuracy of text models on CrisisMMD

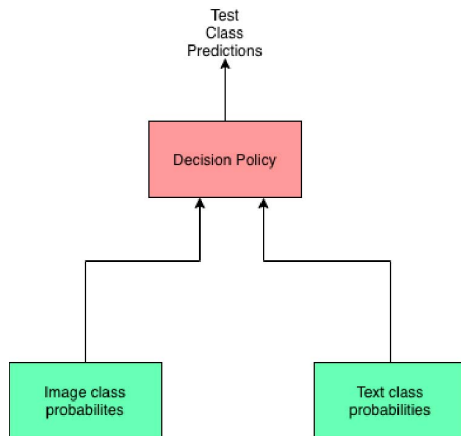


Fig. 5: Decision level fusion technique

modality is better than both only image-based and only text-based modality techniques. Figures 6, 7 and 8 shows the tweet image, text, actual label and predicted labels on our logistic regression-based multimodal fusion technique when evaluated on CleanCrisisMMD dataset. Based on the features learned from both text and image of the tweet, predicted label is found to be same as that of the actual label. However, in figures 9 and 10 the actual label in the dataset was not found to be same as the predicted label.

One of the major reasons for the wrong classification could be the presence of heterogeneous noise in the tweet image. A general inference is the ability of multi-modal fusion techniques to better identify tweets associated with a disaster.

VI. CONCLUSION AND FUTURE WORK

This paper proposes a multimodal methodology for analyzing tweets by using a set of features, both text, and image-based for identification of informative tweets using deep learning. The major contribution of this work is the improved

Text Features	Unigram Features			Bigram Features			Deep Learning methods		
	Logistic Regression	Multinomial NB	Random Forest	Logistic Regression	Multinomial NB	Random Forest	LSTM	Bidirectional LSTM	CNN+Glove
Sri Lanka Floods	60.3	58	61	58.0	59.4	60.1	69	69	67
Hurricane Irma	61.5	67.4	59	65	62	63	71	71	68
Hurricane Harvey	65.8	60	59.4	63	65.3	66.7	73	73	71
Hurricane Maria	52	50.5	56.3	56.2	53	54	61.8	63.2	74.7
California Wildfires	63	63.5	64	66.2	66.9	65	71	77	71
Iran Iraq Earthquake	62	64	68	70	68	54	77	77	74
Mexico Earthquake	61.3	62.7	57.6	64	65	66	73	74	70

TABLE VII: Accuracy of text models on CleanCrisisMMD

Disaster	Mean Probability	Custom Decision Policy	Logistic Regression Decision Policy
Sri Lanka Floods	72.6	70.8	74.14
Hurricane Irma	73.39	71.4	80.2
Hurricane Harvey	76.0	73.5	79.2
Hurricane Maria	74.79	57.7	79.4
California Wildfires	79.2	75.0	75.3
Iran Iraq Earthquake	73.5	80.2	75.2
Mexico Earthquake	77.3	74.6	77.9

TABLE VIII: Accuracy of various multimodal decision level fusion techniques on CleanCrisisMMD



Fig. 6: Disaster: California Wildfires
Text: Wildfires raging through Northern California are terrifying.
Actual Label: Informative
Predicted Label: Informative



Fig. 8: Disaster: Sri-Lanka Floods
Text: Authorities and military aid required in areas of Rakhine State.
Actual Label: Informative
Predicted Label: Informative



Fig. 7: Disaster: California Wildfires
Text: I just had to evacuate my home in California due to the wildfire.
Actual Label: Non-Informative
Predicted Label: Non-Informative



Fig. 9: Disaster: California Wildfires
Text: Firefighters are fighting desperately to stop the fire.
Actual Label: Informative
Predicted Label: Non-Informative

performance of logistic regression-based decision policy for incorporating both texts and image-based modality of tweets. In the future, there is scope for utilizing feature-based fusion techniques for combining modalities and using more advanced models.

The information present on Twitter relating to disaster can vary greatly. Often there is also the scope of irrelevant or misleading information being distributed across the platform. Humanitarian organizations also do not want to deal with noisy data which are of personal nature and do not contain any important information. The classification provided by the proposed methodology can be used to gain situation awareness about a particular disaster which is happening in the world.

Also depending on the nature of the roles, the classification system can help extract fine-grained information about specific incidents which accompany the main disaster.

Such information type can be used effectively to provide financial and medical support to the affected areas on time. In near future feature selection techniques similar to those adopted by [32] can be experimented for improving the performance of the present systems on various disasters. Improvement in automated machine learning-based systems can ease the stress on limited human resources and can improve the process of disaster relief. The findings of this

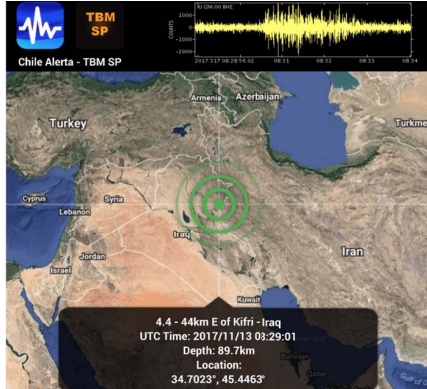


Fig. 10: Disaster: Iran-Iraq Earthquake
Text: New earthquake of magnitude 4.4 looks like a counter strike game.
Actual Label: Non-Informative
Predicted Label: Informative

study will be useful for regulating the process of the tweet based filtering and identification of information related tweets in an emergency related scenario. The systems developed in the paper can also be extended to other disaster-related scenario's like landslides, forest fires with minor modifications and further training of deep-learning models.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of NVIDIA Corporation with the donation of a Titan XP GPU used for this research. Rajiv Ratn Shah is partly supported by the Infosys Center for AI, IIIT Delhi and ECRA Grant by SERB, Government of India.

REFERENCES

- [1] Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisis-mmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM)*, June 2018.
- [2] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El-Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16:345–379, 2010.
- [3] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee, 2017.
- [4] Edward Chang, Kingshy Goh, Gerard Sychay, and Gang Wu. Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(1):26–38, 2003.
- [5] François Chollet. Xception: Deep learning with depth-wise separable convolutions. *arXiv preprint*, pages 1610–02357, 2017.
- [6] Anqi Cui, Min Zhang, Yiqun Liu, Shaoping Ma, and Kuo Zhang. Discover breaking events with popular hashtags in twitter. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1794–1798. ACM, 2012.
- [7] Joao Porto De Albuquerque, Benjamin Herfort, Alexander Brenning, and Alexander Zipf. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International Journal of Geographical Information Science*, 29(4):667–689, 2015.
- [8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.
- [9] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Advances in neural information processing systems*, pages 262–270, 2015.
- [10] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.
- [11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [13] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.
- [14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, volume 1, page 3, 2017.
- [15] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Extracting information nuggets from disaster-related messages in social media. In *Iscram*, 2013.
- [16] Hitkul Jangid, Shivangi Singhal, Rajiv Ratn Shah, and Roger Zimmermann. Aspect-based financial sentiment analysis using deep learning. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 1961–1966. International World Wide Web Conferences Steering Committee, 2018.
- [17] Ioannis Kanaris and Efstathios Stamatatos. Webpage genre identification using variable-length character n-grams. In *Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE International Conference on*, volume 2, pages 3–10. IEEE, 2007.
- [18] Raghav Kapoor, Yaman Kumar, Kshitij Rajput, Rajiv Ratn Shah, Ponnuram Kumaraguru, and Roger Zimmermann. Mind your language: Abuse and offense

- detection for code-switched languages. *arXiv preprint arXiv:1809.08652*, 2018.
- [19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
 - [20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
 - [21] Yaman Kumar, Mayank Aggarwal, Pratham Nawal, Shin’ichi Satoh, Rajiv Ratn Shah, and Roger Zimmermann. Harnessing ai for speech reconstruction using multi-view silent video feed. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 1976–1983. ACM, 2018.
 - [22] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.
 - [23] Debanjan Mahata, Jasper Friedrichs, Rajiv Ratn Shah, and Jing Jiang. Detecting personal intake of medicine from twitter. *IEEE Intelligent Systems*, 33(4):87–95, 2018.
 - [24] Puneet Mathur, Meghna Ayyar, Sahil Chopra, Simra Shahid, Laiba Mehnaz, and Rajiv Shah. Identification of emergency blood donation request on twitter. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop and Shared Task*, pages 27–31, 2018.
 - [25] Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. Did you offend me? classification of offensive tweets in hinglish language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 138–148, 2018.
 - [26] Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. Detecting offensive tweets in hindi-english code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26, 2018.
 - [27] Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.
 - [28] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.
 - [29] Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.
 - [30] Daniela Pohl, Abdelhamid Bouchachia, and Hermann Hellwagner. Automatic sub-event detection in emergency management using social media. In *Proceedings of the 21st international conference on world wide web*, pages 683–686. ACM, 2012.
 - [31] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 251–260. ACM, 2010.
 - [32] Ramit Sawhney, Prachi Manchanda, Raj Singh, and Swati Aggarwal. A computational approach to feature extraction for identification of suicidal ideation in tweets. In *Proceedings of ACL 2018, Student Research Workshop*, pages 91–98, 2018.
 - [33] Ramit Sawhney, Puneet Mathur, and Ravi Shankar. A firefly algorithm based wrapper-penalty feature selection method for cancer diagnosis. In *International Conference on Computational Science and Its Applications*, pages 438–449. Springer, 2018.
 - [34] Rajiv Shah and Roger Zimmermann. *Multimodal analysis of user-generated multimedia content*. Springer, 2017.
 - [35] Rajiv Shah and Roger Zimmermann. *Multimodal analysis of user-generated multimedia content*. Springer, 2017.
 - [36] Rajiv Ratn Shah, Debanjan Mahata, Vishal Choudhary, and Rajiv Bajpai. Multimodal semantics and affective computing from multimedia content. In *Intelligent Multi-dimensional Data and Image Processing*, pages 359–382. IGI Global, 2018.
 - [37] Rajiv Ratn Shah, Yi Yu, and Roger Zimmermann. Advisor: Personalized video soundtrack recommendation by late fusion with heuristic rankings. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 607–616. ACM, 2014.
 - [38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
 - [39] Cepy Slamet, Ali Rahman, Ade Sutedi, Wahyudin Darmalaksana, Muhammad Ali Ramdhani, and Dian Sa’adillah Maylawati. Social media-based identifier for natural disaster. In *IOP Conference Series: Materials Science and Engineering*, volume 288, page 012039. IOP Publishing, 2018.
 - [40] Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012.
 - [41] Heung-Il Suk and Dinggang Shen. Deep learning-based feature representation for ad/mci classification. In Kensaku Mori, Ichiro Sakuma, Yoshinobu Sato, Christian Barillot, and Nassir Navab, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, pages 583–590. Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.
 - [42] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*, 2012.
 - [43] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017.
 - [44] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption

- generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.
- [45] Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1980–1984. ACM, 2012.
 - [46] Yifang Yin, Rajiv Ratn Shah, and Roger Zimmermann. Learning and fusing multimodal deep features for acoustic scene categorization. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 1892–1900. ACM, 2018.
 - [47] Yi Yu, Suhua Tang, Kiyoharu Aizawa, and Akiko Aizawa. Category-based deep cca for fine-grained venue discovery from multimodal data. *IEEE transactions on neural networks and learning systems*, 30(4):1250–1258, 2018.
 - [48] Yi Yu, Suhua Tang, Francisco Raposo, and Lei Chen. Deep cross-modal correlation learning for audio and lyrics in music retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(1):20, 2019.
 - [49] Tom Zahavy, Alessandro Magnani, Abhinandan Krishnan, and Shie Mannor. Is a picture worth a thousand words? A deep multi-modal fusion architecture for product classification in e-commerce. *CoRR*, abs/1611.09534, 2016.