An Early Synthesis of Deep Neural Networks to Identify Multimodal Informative Disaster Tweets

Fareen Tasneem*
University of Chittagong
fareen.tasneem@gmail.com

Shima Chakraborty
University of Chittagong
shimacse@cu.ac.bd

Abu Nowshed Chy
University of Chittagong
nowshed@cu.ac.bd

ABSTRACT

Twitter is always worthwhile in facilitating communication during disasters. It helps in raising situational awareness and undertaking disaster control actions as quickly as possible to alleviate the miseries. But the noisy essence of Twitter causes difficulty in distinguishing relevant information from the heterogeneous contents. Therefore, extracting informative tweets is a substantial task to help in crisis intervention. Analyzing only the text or image content of the tweet often misses necessary insights which might be helpful during disasters. In this paper, we propose a multimodal framework to address the challenges of identifying informative crisis-related tweets containing both texts and images. Our presented approach incorporates an early fusion strategy of BERT-LSTM and ResNet50 networks which effectively learns from the joint representation of texts and images. The experiments and evaluation on the benchmark CrisisMMD dataset show that our fusion method surpasses the baseline by 7% and substantiates its potency over the unimodal systems.

Keywords

Early fusion, crisis tweets, BERT-LSTM, ResNet50, multimodal framework.

INTRODUCTION

Natural or non-natural crisis events are exceptionally unpredictable and simply inconceivable to foresight. Catastrophic crises such as earthquakes, cyclones, hurricanes, volcanic eruptions, and many more disasters affected the socioeconomic system drastically for years. Since the uncertain nature of the calamities endangers the life of humankind and leads to severe damage, it is necessary to take quick actions to minimize the vulnerabilities as soon as possible.

Communication becomes vital as people attempt to reach the victims in the afflicted areas. However, regular communication infrastructure often becomes compromised following a disaster, making it challenging to take prompt emergency response actions. Over the years, the number of active users on Twitter is increasing at a colossal rate and turning it into a medium for sharing views, news, and different critical information. In particular, Twitter comes in handy by providing real-time updates in various circumstances (Mendoza et al. 2010), (Miyabe et al. 2012), (Murthy and Longwell 2013) and also in the disaster events (Velev and Zlateva 2012). Amid the disaster events, information such as the number of wounded, dead, or missing people, damaged infrastructures, the current state of the disaster, and relief needs. circulates instantly as people start posting tweets. Moreover, Twitter advances information to spread rapidly by applying its real-time features such as retweets, hashtags, and mentions. (Sakaki et al. 2012). Such attributes help people to be aware and enable prompt emergency relief efforts to assist the afflicted. But, it is quite challenging to draw out the relevant information essential for crisis management from

*corresponding author

Twitter which is flooded with miscellaneous data. While there are informative tweets, superfluous and redundant tweets are frequent as well. Two instances of tweets are shown in Figure 1a and Figure 1b. Here, the first tweet is an informative tweet and the other one is related to a crisis event but is not substantially informative to aid in crisis response and management.

Informative

(a) *Tweet text*: *Cyclone Mora* : 4 *killed in Manipur,* 140 *houses destroyed in Mizoram...*

Not-informative

(b) *Tweet text*: *Cool*! *Hero dog watches over goats during California fire...*

Figure 1. Examples of informative and not-informative tweets.

Moreover, with the addition of images to the tweets, the inference of critical information gets more complex. In such cases, it is necessary to address both the text and image contents for extracting significant information. So far, most of the researchers have contributed to this domain either by mining text or images (F. Alam, Joty, et al. 2018; ALRashdi and O'Keefe 2019; F. Alam, T. Alam, et al. 2022). But there are a few pioneering works (Gautam et al. 2019; Ofli et al. 2020; Agarwal et al. 2020) that explored the notion of utilizing the association of text and visual contents of social media during disasters.

In this study, we aim to propose a sound multimodal architecture to identify whether a tweet text and image is informative or not. To combat this challenge we adopt a multimodal dataset, CrisisMMD (F. Alam, Ofli, et al. 2018a) comprising text and images from Twitter during various disaster episodes in 2017. Our unified approach incorporates an early fusion method of BERT-LSTM (Gallo et al. 2020) and ResNet50 (He et al. 2016) to represent texts and images, respectively. We implement this fusion approach since it learns from the shared representations of texts and images and also harnesses the latent correlation between the two modalities. The extensive investigation and outcomes show that our presented framework robustly proves its competence over the state-of-the-art systems in identifying informative crisis tweets.

The organization of this paper is followed by a synopsis of previous studies in *Research Background* section. In the *Proposed Method* section, we elucidate our proposed framework in detail. Next, we brief the results and also evaluate our method against cutting-edge methodologies in the *Experiments and Evaluation* section. Finally, we conclude this paper in *Conclusion and Future Direction* section with some subsequent plans for the future.

RESEARCH BACKGROUND

Research work in the crisis domain is not a new phenomenon. A substantial amount of work has been done in this area. Most of the contributions encompass the textual contents. Chy et al. 2020 exploited transfer learning features with a rich ensemble of hand-crafted features which are then fed to a multi-layer perception network. Also, they explored convolution utilizing multiple kernels on top of nested LSTMs to extract the higher-level features. In WNUT-2020 Task 2, several transformer-based models (Kumar and Singh 2020; Hettiarachchi and Ranasinghe 2020; Jagadeesh and Alphonse 2020; Tasneem et al. 2020) were employed in tackling the challenge of detecting informative COVID-19 English tweets. Li et al. 2021 presented a unified framework of self-training with CNN and BERT transformer models which proved to be improving the performance of tweet classifiers for a target disaster.

While there is a majority of research work concerning the textual information related to crisis events, a handful amount of scholarly work has been introduced gradually. Kyrkou and Theocharides 2020 established a computationally proficient CNN architecture, EmergencyNet. The framework joints multi-resolution convolutions which achieved near cutting-edge performance. Subsequently, F. Alam, T. Alam, et al. 2022 presented a remarkable contribution

towards the disaster image analysis and also in the multi-task learning domain by introducing a large disaster image dataset with whopping 71k images. To deal with the issues of remote sensing images, Yuan et al. 2022 proposed a lightweight SDS-Network algorithm, based on ResNet, for disaster classification.

However, analyzing only texts or images fails to grab meaningful data from social media content during crisis events. Therefore, an ample amount of studies have emerged highlighting the use of joint distribution of both modalities. Gautam et al. 2019 worked on a decision diffusion technique to distinguish multimodal data as informative and not-informative labels harnessing the CrisisMMD dataset. Ofli et al. 2020 delineated a baseline result on this dataset by implementing a joint representation of the CNN framework for text and the VGG16 model for image. Later, another noteworthy work was introduced by Agarwal et al. 2020. Crisis-DIAS is a gated multimodal deep learning composition, developed for the 3 hierarchical tasks in the CrisisMMD dataset. The authors applied a recurrent convolutional neural network (RCNN) and a pre-trained Inception V3 model for the text and image modalities, respectively. Afterward, more works such as Zou et al. 2021 and Krawczuk et al. 2021 also brought out the exploitation of combining features extracted from text and image frameworks.

In contrast to the previous contributions, we employ BERT, a benchmark transfer learning architecture with a hint of LSTM for effective sequential learning of tweet texts. For image modality, we leverage the deep CNN architecture of ResNet50. Finally, we apply an early fusion method to associate the local features extracted from both modalities. The fusion in the early stages facilitates capturing the potential correlation between text and image modalities.

PROPOSED METHOD

In this section, we elucidate the framework designed for the task of multimodal informative disaster tweet classification. We address this task by implementing a multimodal fusion of two architectures for text and visual representations. For text modality, we employ BERT (Devlin et al. 2018) with a layer of LSTM (Hochreiter and Schmidhuber 1997) and a ResNet50 (He et al. 2016) model is employed for image modality. Finally, an early fusion approach merges the input features of text and images and constructively learns the interdependence between the two modalities. The overall methodology is illustrated in Figure 2. The input is a text-image pair that is passed to the corresponding frameworks for both modalities. Then the extracted representations are concatenated in an early fusion following a dense layer and we receive a predicted label (informative or not-informative) from the final softmax layer.

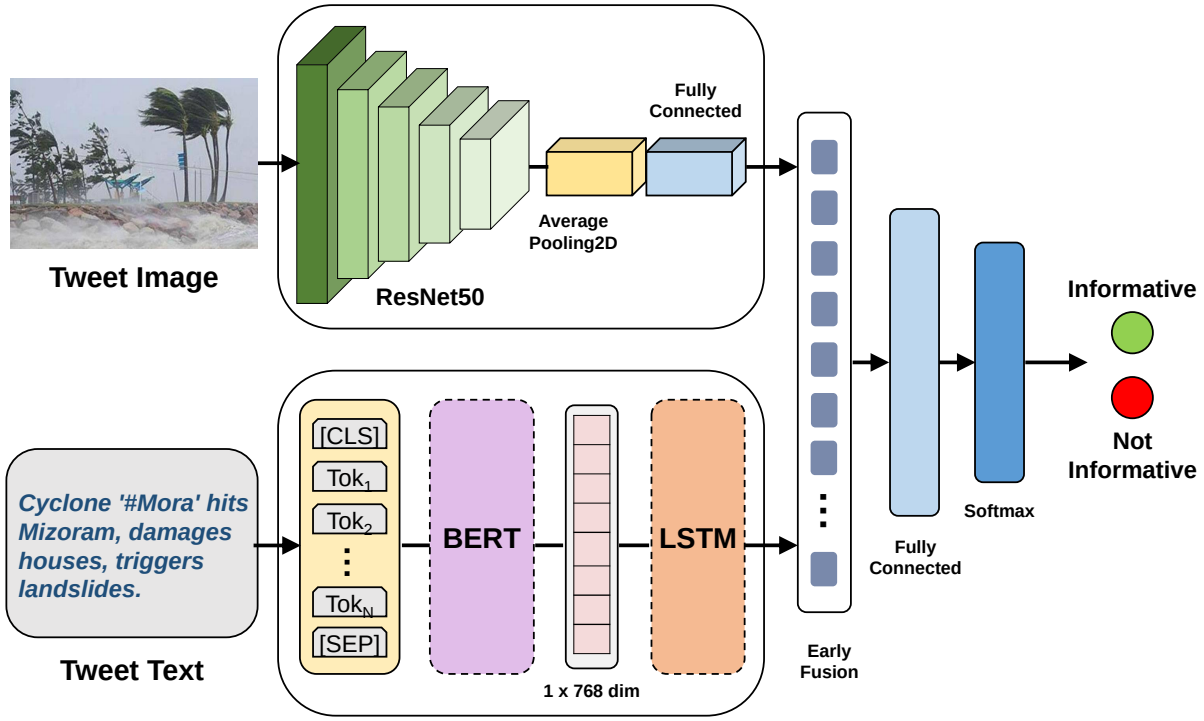


Figure 2. Overview of the proposed method.

BERT-LSTM

Twitter posts are mostly brief texts and sometimes reflect vagueness. During any disaster, some tweets provide crucial information which helps in crisis response and management. Some tweets express condolences and prayers but do not deliver any helpful details. So, it is quite challenging to bring out the gist of the crisis from the short tweets and to distinguish the relevant information. To understand the precise context of disaster-related tweet texts, we exploit a multilayer encoder architecture, BERT (Devlin et al. 2018). It is employed for versatile language-related tasks such as sequence classification, sentence prediction, and question answering. It leverages two revolutionary MLM (Masked Language Modeling) and NSP (Next Sentence Prediction) strategies. Notably, the MLM technique assists in deep bidirectional training of the model that facilitates learning the contextual details of the language proficiently. We utilize the pre-trained 'bert-base-uncased' model with a feature vector of 768 dimensions. The input representation of BERT requires each text to be tokenized. Before tokenization, we employ a few preprocessing steps such as removing punctuations, single alphabetic characters, and unnecessary spaces to facilitate the further processing of input texts. Then we execute this model on the input embeddings and extract a contextual feature vector for each tokenized text. As we leverage the pre-trained weights of the BERT transformer, it may capture plenty of irrelevant features in contrast to the context of a critical situation. To reinforce the meaningful features extracted by the transformer, we add a single-layered and unidirectional LSTM (Hochreiter and Schmidhuber 1997) on top of it. LSTMs are conventional networks that can memorize long-term dependencies for sequentially learning crucial patterns. Hence, the extracted feature vector from BERT is then fed to the additional LSTM layer with 128 hidden units which help in retaining relevant information and eliminating trivial features. Finally, we set the layers as trainable for further learning in the early fusion at the end.

ResNet50

The precedent studies (F. Alam, Ofli, et al. 2018b; Kyrkou and Theocharides 2020) addressed the issue of the twisted nature of the images captured during calamities and mostly proposed using deep CNN networks for extracting critical information. Therefore, for image pipeline, our proposed framework relies on ResNet50 (He et al. 2016) which is a convolutional neural network encasing 50 deep learning layers. The layers include one average-pooling layer, one max-pooling layer, and 48 convolutional layers. This network is composed of accumulating residual blocks and leverages 'skip connection' to resolve the problem of vanishing gradient in deeper networks. Before training, we rescale and reshape the images and discard the softmax layer of the pre-trained model. With the imported weights of ResNet50, we add a 2D average-pooling layer, a dropout layer, and a fully connected layer of 32 units for extracting image features. Finally, we set all the layers as trainable for further learning on the downstream task of disaster image classification.

Early Fusion and Final Predictions

In multimodal fusion, different modalities can be combined either in the early or late phases (Gadzicki et al. 2020). While the early integration emphasizes feature fusion before training the models, the late fusion method utilizes the predictions of the trained models and lacks in utilizing the correlation between the modalities (Pandeya and Lee 2021). Our proposed framework adopts an early fusion technique where the extracted features from BERT-LSTM and ResNet50 architecture are concatenated in a common vector space. The combined feature map is hence passed to a dense layer to generate a shared representation that captures the latent correlation between text and image features. After that, these are passed to a classification layer with a softmax activation function defined as:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad for\ i = 1, 2, \ldots, K \tag{1}$$

Here, z is the input vector and k is the number of classes. The function, σ yields probabilities for the informative or not-informative categories. We train this layer of the network to get the final predictions of the given crisis tweet text and image pairs.

EXPERIMENTS AND EVALUATION

Dataset

In consideration of the challenges related to this study, we utilize CrisisMMD (F. Alam, Ofli, et al. 2018a), a novel multimodal dataset with manually labeled tweets containing text and image pairs. It was released in two versions. In one version, tweet text and image pairs have the same annotations and in the other, the pairs may have different or similar annotations. In our study, we only make use of tweets with the same annotations of text and image. However,

it encompasses tweets from seven crisis events such as Hurricane Irma, Hurricane Harvey, Hurricane Maria, the Mexico earthquake, the California wildfires, the Iraq-Iran earthquake, and the Sri Lanka floods that occurred in the year 2017. Along with the dataset, three multimodal tasks were introduced: Task 1: Informative vs Not-informative, Task 2: Humanitarian categories, and Task 3: Damage severity assessment. We emphasize our focus on task 1 which is a binary classification task to identify the informative and not-informative crisis tweets. The overall summary of the task 1 dataset is demonstrated in Table 1. We can observe that the dataset contains 11,400 texts and 12,708 images in total for task 1. The difference in the number of texts and images in the training set occurred because of omitting the duplicate tweet texts. It is also evident that the dataset is relatively imbalanced since 67% of the samples are of the informative category. As per the baseline (Ofli et al. 2020), we consider weighted F1-score as our evaluation measure.

Table 1. Summary of CrisisMMD (F. Alam, Ofli, et al. 2018a) dataset.

	Train		**Trial**		**Test**		**Total**	
	Text	Image	Text	Image	Text	Image	Text	Image
Informative	5546	6345	1056	1056	1030	1030	7632	8431
Not-informative	2747	3256	517	517	504	504	3768	4277
Total	8293	9601	1573	1573	1534	1534	11400	12708

Experimental Settings

In the table 2, we report the hyperparameter configurations of our proposed model. The proposed model was trained on GPU (Owens et al. 2008) to optimize the simultaneous processing of vast textual and visual data for training. In the early fusion method, we apply a hidden layer of unit size 256 with an activation function of 'relu'. We also set the initial learning rate since we employed the ReduceLROnPlateau class of the Keras. It reduces the learning rate by a factor of 0.1 till it reaches the minimum learning rate of $1e-5$. Before training, the tweet texts and images are assembled in 32 samples per batch. Finally, we train the batches for 20 epochs.

Table 2. Experimental settings of the proposed model.

BERT-LSTM	ResNet50	Early Fusion
1. Max length: 40 2. LSTM units: 128 3. BERT model: 'bert-base-uncased'	1. Image width and height: 299 2. Model: ResNet50 3. Pool size: (8, 8) 4. Dropout rate: 0.4 5. Dense layer unit: 32	1. Dense layer unit: 256 2. Activation: 'relu' 3. Initial learning rate: $1e-3$ 4. Optimizer: SGD 5. Minimum learning rate: $1e-5$ 6. Batch size: 32 7. Epoch: 20

Evaluation

The evaluation of our proposed framework for the task of informative crisis tweet detection is stated in Table 3. The performance is analyzed against the baseline reported in Ofli et al. 2020. It shows that the unimodal text framework, BERT-LSTM moderately outperforms the baseline model, CNN by 4%. Whereas, the ResNet50 architecture for image modality marginally improves the performance compared with the baseline VGG16 method. Remarkably, the proposed early fusion-based approach stands out from the baseline fusion method of CNN and VGG16 by an impressive margin of 7% approx. It certainly implies the potency of our presented early fusion technique of BERT-LSTM and ResNet50. The unification of the features extracted from both modalities robustly identifies the multimodal informative disaster tweets and validates it's efficacy.

In Table 4, we elucidate a comparison with the early fusion of different methods for both modalities to validate the significance of our proposed fusion method of BERT-LSTM and ResNet50. From this table, we can observe that

Table 3. Evaluation of our proposed methodology against the baseline models.

Methods	Recall	Precision	F1-score	Accuracy
Unimodal text				
CNN (Baseline)	0.81	0.81	0.81	0.81
BERT-LSTM	0.92	0.86	**0.85**	0.91
Unimodal image				
VGG16 (Baseline)	0.83	0.83	0.83	0.83
ResNet50	0.88	0.88	**0.84**	0.84
Multimodal fusion				
Early fusion (Baseline)	0.84	0.84	0.84	0.84
Early fusion (Proposed method)	0.94	0.93	**0.91**	0.91

our proposed approach surpassed all the other attempted fusion methods. In the first three cases, we demonstrated the impact of the proposed BERT-LSTM model for text modality in the fusion by tuning it with different settings. In the first strategy, we discarded the LSTM layer from the BERT-LSTM module of the fusion. But we incorporated a two-layered CNN instead of a single-layer LSTM with BERT in the next case. Even though the differences in performance of these strategies with our proposed fusion model are minimal, it implies how the addition of a single-layered LSTM with BERT can be useful. Subsequently, to highlight the efficacy of BERT in the fusion architecture, we replaced it with an ELECTRA (Clark et al. 2020) transformer which in turn shrinks the performance drastically. In the next two cases, we analyzed the performance of the ResNet50 model in the early fusion by superseding it with two CNN architectures e.g. Xception (Chollet 2017) and EfficientNetB0 (Tan and Le 2019) respectively. But, both of these two configurations for image modality resulted in remote drops in the performance of the fusion method.

Table 4. Performance analysis of the proposed early fusion of BERT-LSTM and ResNet50 against some different settings of text and image modalities in the fusion.

Methods	Recall	Precision	F1-score	Accuracy
BERT-LSTM+ResNet50	0.94	0.93	**0.91**	0.91
BERT+ResNet50	0.92	0.93	0.90	0.91
BERT-CNN+ResNet50	0.95	0.91	0.90	0.90
ELECTRA-LSTM+ResNet50	0.91	0.86	0.84	0.84
BERT-LSTM+Xception	0.93	0.91	0.89	0.89
BERT-LSTM+EfficientNetB0	0.93	0.89	0.88	0.88

Contribution of Early Fusion

The noteworthy performance of the early fusion shown in Table 3 indicates the significance of multimodal learning as opposed to the unimodal approaches. The fusion method progresses the performance of unimodal text and image frameworks by about 6% and 7%, respectively. Since the early fusion of the features leverages unique attributes of both modalities, it succeeds in capturing better contextual details in detecting informative tweets.

To further inspect the contribution of the early fusion method, we demonstrate a comparison with the late fusion approach in Table 5. In the late fusion method, we first train the BERT-LSTM and ResNet50 for text and image classification individually and we extract the output features from the classification layer. Then we merge these features and pass them to a dense layer following a softmax layer to get the final predicted labels. The late fusion of the two modalities yields a mediocre performance against the proposed early fusion. Micro-blogging sites' textual and visual content are often mismatched and do not follow the contexts. Therefore, it is required to utilize the correspondence between the text and images in multimodal classification. In the late fusion, the predictions are combined to make the final decision, instead of leveraging the input features from both modalities. So, it lacks the required correlation between disaster text and images gathered from Twitter. But in early fusion, the raw input features are concatenated which potentially utilizes the correlation and identifies the informative tweets which are strongly related to crisis events.

Table 5. Comparison of early fusion and late fusion method.

Methods	Recall	Precision	F1-score	Accuracy
Early fusion	0.94	0.93	**0.91**	0.91
Late fusion	0.98	0.78	0.78	0.81

Besides, we compare our proposed early fusion with other multi-modal methods attempted in this task. The comparative results are presented in Table 6. The first attempted model (Shah and Chy 2020) is an ensemble of a multi-kernel CNN with LSTM and two MLP (multi-layer perceptron) modules. The input to the MKCNN-LSTM module is a set of word embedding extracted from the tweet texts using word2vec (Mikolov et al. 2013) and the inputs of the MLP modules are feature vectors extracted from BERT and VGG16 (Simonyan and Zisserman 2014) for text and image representations respectively. Likewise, the second method reflects the same configurations but an additional MLP module is utilized here which is fed with a feature matrix extracted from the ELECTRA transformer for textual modality. Both methods obtained quite identical outcomes even though an additional MLP module is exploited in the second case. Whereas, in the third approach, a weighted average-based fusion of spaCy with the RoBERTa (Liu et al. 2019) transformer model for texts and the Inception V3 (Szegedy et al. 2016) model for images, achieved a moderate performance compared to the previous cases. However, the performance of our proposed early fusion method compared with these attempted approaches signifies the robustness in the multimodal task of informative crisis tweet identification.

Table 6. Performance analysis of the proposed early fusion against other multi-modal approaches.

Methods	Recall	Precision	F1-score	Accuracy
Proposed Early Fusion	0.94	0.93	**0.91**	0.91
MKCNN-LSTM+MLP	0.99	0.68	0.81	0.68
MKCNN-LSTM+MLP_2	0.99	0.66	0.80	0.68
spaCy-RoBERTa+Inception	0.91	0.79	0.85	0.78

Comparison with Related Works

To substantiate the performance of our proposed system, we present a comparative analysis with the existing state-of-the-art contributions on task 1 of the CrisisMMD dataset in Table 7. Abavisani et al. 2020 proposed a cross-attention scheme for combining text and image features from BERT and DenseNet and also a stochastic shared embedding as a regularization technique. Though it introduces a novel framework, it is computationally infeasible and marginally lags in performance. Zou et al. 2021 adopts a multimodal fusion of FastText and VGG16 which overlooks the advantage of leveraging deeper networks like ResNet50 and the transformer architecture as we endeavored. On the other hand, Khattar and Quadri 2022 apply a cross-attention fusion mechanism including Bi-LSTM network and VGG16. They performed an average of F1-scores over the seven disaster events which largely impacted the outcomes compared to the state-of-the-art results.

Table 7. Comparison of the proposed method with the state-of-the-art models.

Methods	Recall	Precision	F1-score	Accuracy
Proposed Early fusion	0.94	0.93	**0.91**	0.91
SSE-Cross-BERT-DenseNet (Abavisani et al. 2020)	-	-	0.89	0.89
VGG16+FastText (Zou et al. 2021)	0.88	0.88	0.88	0.88
Cross-Attention Multimodal Fusion (Khattar and Quadri 2022)	0.84	0.86	0.84	0.88

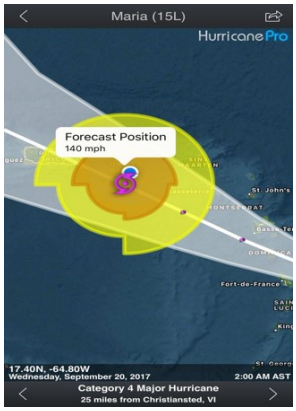
Discussion

To qualitatively analyze the significance of our multimodal approach against the unimodal models, we demonstrate some sample tweets in Figure 3 classified by our system. In the sample in Figure 3a, we can observe that the tweet text is identified as not informative because contextually it doesn't reflect any information about the crisis. But the

image modality correctly classified the image recognizing the pattern of after-disaster destruction. Whereas in the second example in Figure 3b, the ResNet50 model identified the tweet as informative because of the fire region in the image but the BERT-LSTM captured the ambiguous nature of the text. However, when the features from tweet text and image are fused in the early fusion, it yields in true positive. Hence, this analysis implies the influence of concatenating both modalities in the early fusion since it captures the different insights of text and image. On the other hand, in Figure 3c, both the BERT-LSTM and ResNet50 networks incorrectly predicted the tweet as not-informative and consequently resulted in erroneous prediction by the early fusion method. However, in the case of the sample in Figure 3d, the unimodal text and image frameworks correctly predicted the tweet as informative while the early fusion method fails to get an accurate prediction. The plausible reason for such incorrect prediction by the proposed early fusion is the infused noise or the redundant features of the multiple modalities.

(a) *Tweet text*: *A #Drone video of a lonely mailman reveals California's post-wildfire dystopia https://t.co/DhFxAaxX7f https://t.co/7DukKyvwnn*
Gold: *Informative*
Early Fusion: *Informative*
BERT-LSTM: *Not informative*
ResNet50: *Informative*

(b) *Tweet text*: *Healthcare? North Korea? Puerto Rico? The Cactus Curtain? Don't blame the petulant crybaby. #SAD Bloatos. https://t.co/Z675PwAZGS*
Gold: *Not informative*
Early Fusion: *Not informative*
BERT-LSTM: *Not informative*
ResNet50: *Informative*



(c) *Tweet text*: *I like to wake up to good news. This is NOT it. #Irma #StCroix https://t.co/LKnbs168xT*
Gold: *Informative*
Early Fusion: *Not informative*
BERT-LSTM: *Not informative*
ResNet50: *Not informative*

(d) *Tweet text*: *Fire swept through Northern California early Monday morning. https://t.co/crmtxj8IPa https://t.co/FxpdGjKXek*
Gold: *Informative*
Early Fusion: *Not informative*
BERT-LSTM: *Informative*
ResNet50: *Informative*

Figure 3. Performance analysis of our proposed system as opposed to unimodal models.

Further, we provide another piece of analysis in Figure 4, regarding the enumerations of correctly predicted tweets in the seven disaster events by our proposed fusion architecture. The inspection shows that the majority of the informative tweets were posted during hurricane Harvey and our proposed model successfully detected approximately 94.1% informative tweets. Also, the rate of not informative tweets was relatively low during this disaster. Tweets concerning California wildfires also suggest the same pattern. However, it indicates the severity of these calamities and how fervently people responded. On the other hand, during the Sri Lanka floods, the ratio of not informative tweets was higher than the informative ones which caused hindrances in extracting valuable information from Twitter. But a system can be biased toward a few crisis events because of such imbalanced samples of different crises. We can see that the difference between the gold and predicted labels is marginal in case of the all seven crisis incidents. This analysis implies the fact that our proposed architecture efficiently overcomes the limitation of the imbalanced dataset.

CONCLUSION AND FUTURE DIRECTION

Retrieving valuable information from Twitter can be extremely helpful during crises aiding emergency management and many humanitarian activities. In this study, we demonstrated our contribution to the task of multimodal informative crisis tweet detection. We addressed this challenge utilizing a unimodal textual framework of LSTM on top of BERT architecture and a deep CNN, ResNet50 for image modality. In the end, an early fusion mechanism is applied to fuse the local features of two modalities and to attain a final label for the tweets. The evaluation based on the novel CrisisMMD dataset infers that our approach outperforms the state-of-the-art models and the baseline. Moreover, our comprehensive analysis demonstrates how the multimodal framework excelled in accurate predictions while the individual models lacked. In the future, we anticipate extending the proposed method to classify the humanitarian categories and for damage severity assessment. We also intend to enhance the scope of disaster events to generalize the learning of the multimodal system.

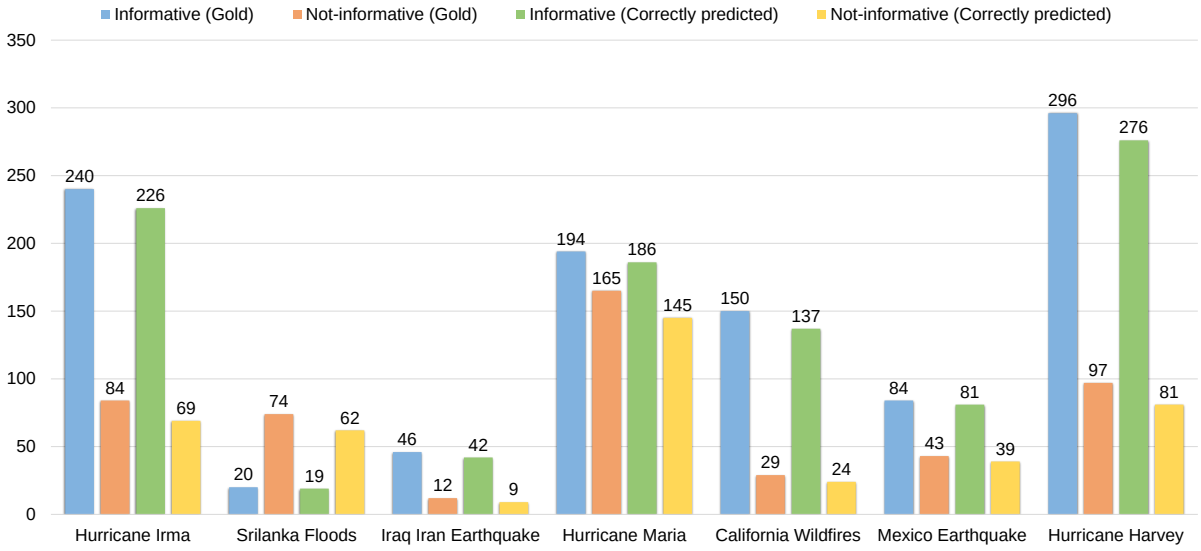


Figure 4. Enumeration of correctly detected tweets by our proposed model over seven disaster events.

REFERENCES

Abavisani, M., Wu, L., Hu, S., Tetreault, J., and Jaimes, A. (2020). "Multimodal categorization of crisis events in social media". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14679–14689.

Agarwal, M., Leekha, M., Sawhney, R., and Shah, R. R. (2020). "Crisis-dias: Towards multimodal damage analysis-deployment, challenges and assessment". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 01, pp. 346–353.

Alam, F., Alam, T., Hasan, M., Hasnat, A., Imran, M., Ofli, F., et al. (2022). "MEDIC: a multi-task learning dataset for disaster image classification". In: *Neural Computing and Applications*, pp. 1–24.

Alam, F., Joty, S., and Imran, M. (2018). "Graph based semi-supervised learning with convolution neural networks to classify crisis related tweets". In: *Twelfth International AAAI conference on web and social media*.

Alam, F., Ofli, F., and Imran, M. (2018a). "Crisismmd: Multimodal twitter datasets from natural disasters". In: *Twelfth international AAAI conference on web and social media*.

Alam, F., Ofli, F., and Imran, M. (2018b). "Processing social media images by combining human and machine computing during crises". In: *International Journal of Human–Computer Interaction* 34.4, pp. 311–327.

ALRashdi, R. and O'Keefe, S. (2019). "Deep learning and word embeddings for tweet classification for crisis response". In: *arXiv preprint arXiv:1903.11024*.

Chollet, F. (2017). "Xception: Deep learning with depthwise separable convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258.

Chy, A. N., Siddiqua, U. A., and Aono, M. (2020). "A Neural Network Model to Identify the Crisis-related Actionable Informative Tweets for Disaster Management". In: *2020 IEEE Region 10 Symposium (TENSYMP)*. IEEE, pp. 390–393.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). "Electra: Pre-training text encoders as discriminators rather than generators". In: *arXiv preprint arXiv:2003.10555*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Gadzicki, K., Khamsehashari, R., and Zetzsche, C. (2020). "Early vs late fusion in multimodal convolutional neural networks". In: *2020 IEEE 23rd international conference on information fusion (FUSION)*. IEEE, pp. 1–6.

Gallo, I., Ria, G., Landro, N., and La Grassa, R. (2020). "Image and text fusion for upmc food-101 using bert and cnns". In: *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*. IEEE, pp. 1–6.

Gautam, A. K., Misra, L., Kumar, A., Misra, K., Aggarwal, S., and Shah, R. R. (2019). "Multimodal analysis of disaster tweets". In: *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. IEEE, pp. 94–103.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hettiarachchi, H. and Ranasinghe, T. (2020). "Infominer at wnut-2020 task 2: Transformer-based covid-19 informative tweet extraction". In: *arXiv preprint arXiv:2010.05327*.

Hochreiter, S. and Schmidhuber, J. (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.

Jagadeesh, M. and Alphonse, P. (2020). "NIT_COVID-19 at WNUT-2020 Task 2: Deep Learning Model RoBERTa for Identify Informative COVID-19 English Tweets." In: *W-NUT@ EMNLP*, pp. 450–454.

Khattar, A. and Quadri, S. (2022). "CAMM: Cross-Attention Multimodal Classification of Disaster-Related Tweets". In: *IEEE Access* 10, pp. 92889–92902.

Krawczuk, P., Nagarkar, S., and Deelman, E. (2021). "CrisisFlow: multimodal representation learning workflow for crisis computing". In: *2021 IEEE 17th International Conference on eScience (eScience)*. IEEE, pp. 264–266.

Kumar, P. and Singh, A. (2020). "Nutcracker at wnut-2020 task 2: Robustly identifying informative covid-19 tweets using ensembling and adversarial training". In: *arXiv preprint arXiv:2010.04335*.

Kyrkou, C. and Theocharides, T. (2020). "EmergencyNet: Efficient aerial image classification for drone-based emergency monitoring using atrous convolutional feature fusion". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13, pp. 1687–1699.

Li, H., Caragea, D., and Caragea, C. (2021). "Combining self-training with deep learning for disaster tweet classification". In: *The 18th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2021)*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692*.

Mendoza, M., Poblete, B., and Castillo, C. (2010). "Twitter under crisis: Can we trust what we RT?" In: *Proceedings of the first workshop on social media analytics*, pp. 71–79.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781*.

Miyabe, M., Miura, A., and Aramaki, E. (2012). "Use trend analysis of twitter after the great east japan earthquake". In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*, pp. 175–178.

Murthy, D. and Longwell, S. A. (2013). "Twitter and disasters: The uses of Twitter during the 2010 Pakistan floods". In: *Information, Communication & Society* 16.6, pp. 837–855.

Ofli, F., Alam, F., and Imran, M. (2020). "Analysis of social media data using multimodal deep learning for disaster response". In: *arXiv preprint arXiv:2004.11838*.

Owens, J. D., Houston, M., Luebke, D., Green, S., Stone, J. E., and Phillips, J. C. (2008). "GPU computing". In: *Proceedings of the IEEE* 96.5, pp. 879–899.

Pandeya, Y. R. and Lee, J. (2021). "Deep learning-based late fusion of multimodal information for emotion classification of music video". In: *Multimedia Tools and Applications* 80, pp. 2887–2905.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2012). "Tweet analysis for real-time event detection and earthquake reporting system development". In: *IEEE Transactions on Knowledge and Data Engineering* 25.4, pp. 919–931.

Shah, S. and Chy, A. N. (2020). "Fusion of Hand-crafted Features and Deep Semantic Features in a Unified Neural Model for Irony Detection in Microblogs". In: *2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC)*. IEEE, pp. 1–6.

Simonyan, K. and Zisserman, A. (2014). "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826.

Tan, M. and Le, Q. (2019). "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International conference on machine learning*. PMLR, pp. 6105–6114.

Tasneem, F., Naim, J., Tasnia, R., Hossain, T., and Chy, A. N. (2020). "CSECU-DSG at WNUT-2020 task 2: Exploiting ensemble of transfer learning and hand-crafted features for identification of informative COVID-19 English tweets". In: *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pp. 394–398.

Velev, D. and Zlateva, P. (2012). "Use of social media in natural disaster management". In: *International Proceedings of Economic Development and Research* 39, pp. 41–45.

Yuan, J., Ma, X., Han, G., Li, S., and Gong, W. (2022). "Research on lightweight disaster classification based on high-resolution remote sensing images". In: *Remote Sensing* 14.11, p. 2577.

Zou, Z., Gan, H., Huang, Q., Cai, T., and Cao, K. (2021). "Disaster image classification by fusing multimodal social media data". In: *ISPRS International Journal of Geo-Information* 10.10, p. 636.