

Improving Disaster-related Tweet Classification with a Multimodal Approach

Xukun Li

Department of Computer Science
Kansas State University
xukun@ksu.edu

Doina Caragea*

Department of Computer Science
Kansas State University
dcaragea@ksu.edu

ABSTRACT

Social media data analysis is important for disaster management. Lots of prior studies have focused on classifying a tweet based on its text or based on its images, independently, even if the tweet contains both text and images. Under the assumptions that text and images may contain complementary information, it is of interest to construct classifiers that make use of both modalities of the tweet. Towards this goal, we propose a multimodal classification model which aggregates text and image information. Our study aims to provide insights into the benefits obtained by combining text and images, and to understand what type of modality is more informative with respect to disaster tweet classification. Experimental results show that both text and image classification can be improved by the multimodal approach.

Keywords

Multimodal Model, Tweet Classification, Deep Learning.

INTRODUCTION



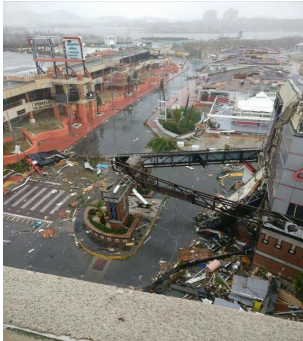
The importance of social media analysis in disaster management has been widely recognized (Velev and Zlateva 2012; Houston et al. 2015; Xiao et al. 2015; Alexander 2014). During a disaster, there is a large volume of communication requests, as people contact families and friends to inform them about disaster developments. Standard communication means may be lost in the beginning of a disaster due to a very heavy workload. This can make the disaster response task very challenging, and presents a need for more efficient communication and data collection tools. Social media has received lots of attention in disaster response. It allows people to share information and to ask for help. Furthermore, it is relatively easy to collect data from social media platforms (e.g., Twitter). If data from social media can be analyzed effectively, the rescue and recovery process can be accelerated because people can be connected with the resources that they need in a timely manner.

There are several studies focused on data collection during a disaster (Alam, Imran, et al. 2017; Alam, Ofli, et al. 2018). Such studies describe detailed steps for data collection and labeling, and make available labeled datasets which can be used for image and/or text classification tasks.

Most of the prior works using such datasets have focused either on text classification (Verma et al. 2011; Ashktorab et al. 2014; Imran et al. 2015; Neppalli et al. 2018; H. Li et al. 2018; Reuter et al. 2018) or image classification (Nguyen et al. 2017; X. Li et al. 2019). However, the text and the image of a tweet presumably hold complementary pieces of information, which can be used together to improve the overall classification of disaster tweets. Because of this relationship between tweet text and images, it is of interest to utilize both text and images together in a model to potentially improve the performance of the models learned from text and/or images separately. Some existing studies proposed to use multimodal data for disaster tweet classification (Agarwal et al. 2020; Mouzannar et al. 2018), and trained models on tweets for which the text and image labels agree. While it is expected that the two

*corresponding author

Table 1. Tweet text and image examples

Text Label	Informative	Not Informative	Informative
Text	puerto rico could be without power for <number> to <number> months after hurricane irma <url> <url>	do n't care if it 's hurricane irma , hypothermia , or even rae sremmurd . i 'm just here to throw picks	hurricane irma destroys ' <number> ' of french part of caribbean island st martin : official <url>
Image Label	Not Informative	Not Informative	Informative
Image			

modalities will enhance each other if their labels agree, it is also of interest to study if they can help each other when the labels don't agree.

Thus, in this paper, we propose a multimodal approach for tweet classification, which makes use of both text and image information in a tweet, when both are available. We evaluate the multimodal approach under two different scenarios. In the first scenario, we only use tweets with matching text and image labels to understand if combining the two modalities results in better classification performance. In the second scenario, we use not only tweets with matching text and image labels, but also tweets with different text and image labels. Here, the goal is to see if one modality helps the classification with respect to the other modality despite potential label mismatches between the two modalities. In addition to understanding the benefits achieved when combining two modalities, we also aim to understand if one modality is more predictive than the other.

The rest of this paper is organized as follows: We describe the proposed model in the "APPROACH" section. We discuss the related work in "RELATED WORK". Experimental results are presented and discussed in the "EXPERIMENTAL RESULTS" section. Finally, we conclude the paper in the "CONCLUSIONS" section.

RELATED WORK

Lots of studies focused on supervised learning on tweet texts and images. Chowdhury et al. (2013) and Stowe et al. (2016) proposed classification models for disaster related tweet classification with traditional machine learning algorithms. Neppalli et al. (2018) and H. Li et al. (2018) proposed the use of deep learning models to conduct text classification. Neppalli et al. (2018) also compared the performance between traditional Naive Bayes models and deep learning models, while H. Li et al. (2018) compared different word embeddings. There are also some studies focused on disaster image analysis. Yang et al. (2011) proposed a hierarchical image classification approach to enhance situational awareness. Barnes et al. (2007) and Vetrivel et al. (2016) analyzed satellite images to identify blocked routes and potential rescue targets. They also analyzed satellite disaster images to assess disaster damage. An image classification model has been proposed by Nguyen et al. (2017), which applied the deep Convolutional Neural Network model on disaster images.

Hu and Flaxman (2018) proposed an multimodal approach to predict emotion word tags for posts made by Tumblr users. Our model is similar with their model but applied on disaster-related tweets. A multimodal approach has been published by Mouzannar et al. (2018) . They trained a CNN model for text and another CNN model for images, and then took the average of the output to generate the final classification decision. At a high level, our proposed approach is similar to the approach in Mouzannar et al. (2018). However, we use different base models for text and images, respectively, and an improved way of combining the predictions, in addition to the average of the predictions. Nalluru et al. (2019) proposed a multimodal model which extracts embeddings for text and images and

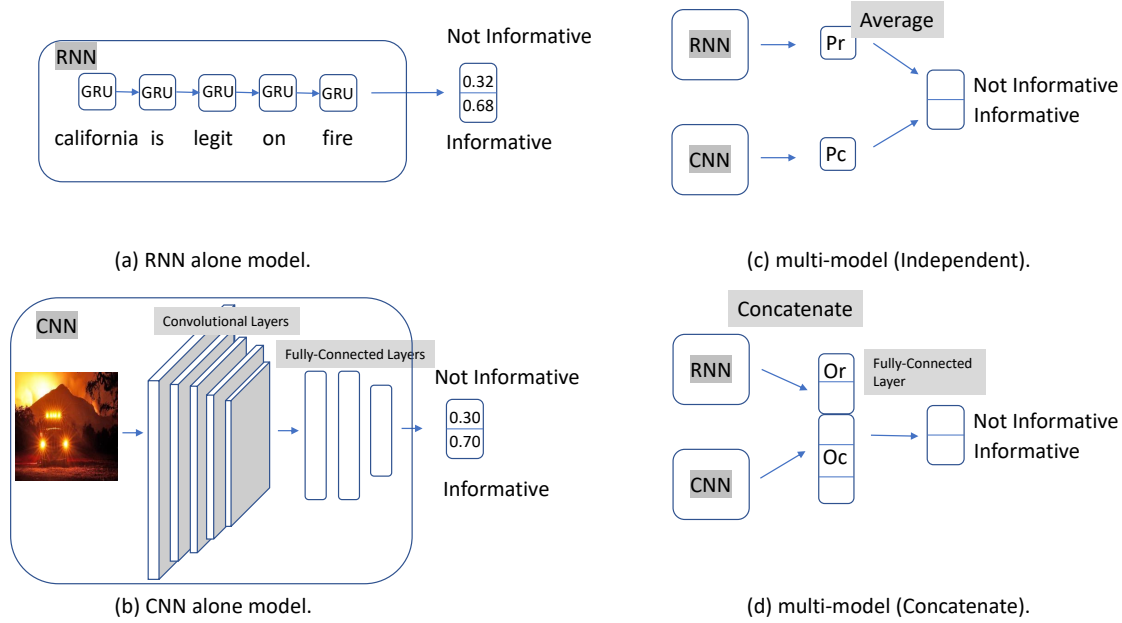


Figure 1. Architecture of proposed model

then applies the lightGBM model, which is an implementation of fast gradient boosting on decision tree (Ke et al. 2017). Compared to this work, we use different embedding extraction methods and apply a fully-connected layer for classification (instead of the lightGBM model), as this approach is widely used in deep learning. Agarwal et al. (2020) proposed a novel end-to-end multimodal framework. Their framework addresses three classification tasks: tweet informativeness, infrastructure damage and damage severity. They combine cues extracted from text and image modalities, and merge them using the attention mechanism. All of the abovementioned works are used for tweets with matching text and image labels, but do not consider the scenario where the labels of the text and images may disagree. Our proposed approach can handle both scenarios and can be used on real Twitter data, where one does not know if the text and image labels match prior to classification but can perform classification of the tweet text and images using multimodal models.

APPROACH

In this section, we introduce some notations and elaborate the details of our approach. Suppose a tweet i consists of both text and image. We denote the tweet text by T_i , while the corresponding label of the text is denoted by y_{t_i} . Similarly, we denote the tweet image by I_i , while the label of the image is denoted by y_{i_i} . In our study, the labels of both text and images can take values *informative* or *non-informative*. For a tweet i which contains both text and an image, the text label y_{t_i} and the image label y_{i_i} can be the same or different, as can be seen in the examples in Table 1. Therefore, the label of a tweet containing both text and an image can be assigned based on the original text labels, or based on the original image labels, unless the text and image have identical labels.

Regardless of the way tweet labels are assigned, a multi-modal labeled tweet instance is denoted by $\{T_i, I_i\}, y_i$ for $i = 1, \dots, n$. Given a set of such multi-modal labeled tweet instances, the goal is to learn a model that uses both the text T_i and image I_i to predict the label y_i . The architecture of our model is shown in Figure 1. As can be seen, the model contains two distinct modules: a recurrent neural network (RNN) for processing the text and a convolutional neural network (CNN) for processing the image. Next, we describe the text model RNN, which predicts label y_i using T_i , and the image model CNN, which predict label y_i using I_i . Then, we describe how we combine the two models into a multi-modal model which predicts label y_i using both modalities of the tweet, $\{T_i, I_i\}$.

RNN

Recurrent Neural Networks (RNNs) have been used extensively in text classification (Liu et al. 2016). An RNN learns sequential patterns by learning information from previous word together with information from the current word. There are many variants of RNN networks. In this study, we use Gated Recurrent Unit (GRU) networks (Cho et al. 2014). The advantage of a GRU is that it can help capture long-term dependencies, while maintaining a

relatively simple architecture. The text T_i consists of a sequence of words $\{w_1, \dots, w_j\}$, where j is the length of the text. As shown in Figure 1 (a), the information of each word (current GRU unit/state) will feed-forward to the next GRU unit. Formally, $g_i = GRU(w_i, g_{i-1})$. The output y of the network will be generated by the last word and the last GRU unit output. $\hat{y} = GRU(w_j, g_j)$. The model output is the prediction class: informative or non-informative.

CNN

Convolutional Neural Networks (CNNs) have been widely used in image classification. They learn spatial patterns in an image by extracting information from the surrounding area of each point through the means of a filter. As shown in Figure 1 (b), a CNN contains convolutional layers and fully-connected layers. The convolutional layer uses a sliding window on the image matrix to “walk” a filter through each image point. The filter identifies predictive information that may be repeated at different points in the image. A convolutional layer is usually followed by a pooling layer, which selects important information by taking the max or average in a window. After several convolutional layers, several fully-connected layers are used to produce final classification decisions. There are several CNN model architectures published. We used AlexNet (Krizhevsky et al. 2012) because it is relatively simple and efficient architecture. However, the CNN component of the model can potentially be improved by utilizing a deeper and more effective CNN architecture.

Multi-Modal Model

A multi-modal model is obtained by combining models corresponding to different modalities into one global model. In this paper, we combined the RNN and CNN models into one global model. As described above, the RNN component consumes text and the CNN component consumes images. The combined model will output one label. There are two ways in which we combine the RNN text and the CNN image models.

Train Text and Image Models Independently, and Average Predictions

The first way of combining the text and image models is to train the RNN and CNN models independently and then take the average of their prediction probabilities to get the final output of the network, as shown in Figure 1 (c).

$$Pr = RNN(T_i)$$

$$Pc = CNN(I_i)$$

$$\hat{y}_i = average(Pr, Pc)$$

where Pr/Pc represent the output probabilities from RNN/CNN. This idea is very straightforward and several works have already evaluated it in disaster management Agarwal et al. 2020. One disadvantage of this model is that it is hard to find tweets where the text and image labels match. Given that the two models are trained independently, their predictions cannot be combined for tweets where the text and image have different labels. Another disadvantage is that the AUC for this model will generally be lower than the best result because of the averaging operator.

Concatenate the Text and Image Representations and Co-train the Corresponding Models

Another way to combine the RNN and CNN models is to concatenate the outputs of their last layers, as shown in Figure 1 (d), and feed the combined representation to a fully connected layer, which makes the final prediction.

$$Or = RNN(T_i); Oc = CNN(I_i)$$

$$\hat{y} = fc(concat(Or, Oc)) = ReLu(W * concat(Or, Oc) + b)$$

where Or/Oc represent the last layers of the RNN/CNN models, respectively, and fc is the fully-connected layer. W and b are the parameters in the fully-connected layer. The activation function is the ReLu function, where $f(x) = \max(0, x)$. Fully-connected layers are widely used in neural networks, especially to connect the final representation/embedding layer to the classification layer. In this model, we extract representations from tweet text and image in the $concat(Or, Oc)$, and predict the label \hat{y} using the fully-connected layer. In this case, the two models will be co-trained together. The advantage of this way of combining the RNN and CNN is that the model will learn useful linking information automatically. If the image is not helpful to the text classification, this will also be presumably captured by the model. Thus, it is expected that the performance with respect to the base models will not be degraded.

Table 2. Statistics of the the dataset CrisisMMD

Before filtering. Original text labels.					After filtering. Matched labels.				
	Dataset	Not Informative	Informative	Total		Dataset	Not Informative	Informative	Total
D0	California Fire	345	1245	1590	D0	California Fire	282	923	1205
D1	Hurricane Harvey	1105	3334	4439	D1	Hurricane Harvey	906	2262	3168
D2	Hurricane Irma	957	3564	4521	D2	Hurricane Irma	767	2032	2799
D3	Hurricane Maria	1714	2844	4558	D3	Hurricane Maria	1295	1813	3108
D4	Iraq Iran Earthquake	104	493	597	D4	Iraq Iran Earthquake	102	398	500
D5	Mexico Earthquake	350	1030	1380	D5	Mexico Earthquake	315	806	1121
D6	Srilanka Floods	655	367	1022	D6	Srilanka Floods	632	229	861

Table 3. Hurricane Irma text/image distribution

txt/image	informative	non-informative
informative	2032	1532
non-informative	190	767

EXPERIMENTAL RESULTS

Dataset

The dataset we used in this study is the CrisisMMD dataset that was published by Alam, Ofli, et al. (2018). The dataset contains tweets crawled during seven disaster events. For each tweet, the text and image of the tweet were labeled as informative/non-informative. The labels for text and image might be different. We show the text label distribution in Table 2 left panel, and the label distribution for text labels matched with image labels in Table 2 right panel. As can be seen, the disaster related tweet datasets are quite unbalanced in terms of informative and not informative instances.

We also show the text/image label distribution in Table 3. We use the Hurricane Irma dataset as an example. The other datasets also show similar patterns. As can be seen in Table 3, if the image is informative, 90% (2032 out of 2222) of text is also informative; if the image is not informative, about 33% of text is not informative. This suggests that if the image is informative, the text is usually informative; if the image is not informative, the text might be informative or not.

Experimental Setup

Our multi-modal model aims to improve performance by learning from text and image together, so that the model will only output one decision for a text/image pair. We evaluate two scenarios in this study. The scenarios differ in how we filter the dataset and select labels for a text-image pair.

- In the first scenario, we filter out the tweets for which the corresponding text and image labels do not match. The data will be $\{T_i, I_i\}, y_i$ where $i = 1, \dots, n'$ and $n' < n$. The multi-model will predict the matched label y_i from $\{T_i, I_i\}$. We compare three results for this scenario: (1) Text only model with data T_i, y_i ; (2) Image only model with data I_i, y_i ; (3) multi-modal model with data $\{T_i, I_i\}, y_i$.
- In the second scenario, we predict text label using both text and image information, and image label using both text and image information. The text label data will be $\{T_i, I_i\}, y_{t_i}$ where $i = 1, \dots, n$, while the image label data will be $\{T_i, I_i\}, y_{i_i}$ where $i = 1, \dots, n$. This scenario is more realistic, as in practice we don't know the text/image label in advance (i.e., we can't tell if they match or not), and thus we cannot filter out unmatched label tweets. We compare two results for this scenario for text labels: (1) Text only model; (2) multi-modal models. Similarly, we compare two results for image labels: (1) Image only model; (2) multi-modal models. While predicting text or image labels, the multi-model will learn some useful information from the other modality to complement the modality whose labels are predicted.

We evaluate the performance for the proposed model on the seven datasets from CrisisMMD. Each dataset is randomly split into training (70%) and testing (30%) set. To avoid splitting bias, we run each experiment three times on each dataset and average the results over the three runs. We use Hurricane Maria to select hyper-parameters, given that this dataset is larger than others. Specifically, we split Hurricane Maria into training (50%) and development (20%) subsets and select hyper-parameters based on the development subset. Specifically, we use Adam Optimizer with learning rate as 0.001. To avoid over-fitting, we train each RNN model for at 10 epochs and each CNN model for 30 epochs.

Table 4. Experimental results for the first scenario (image/text matched label results). D0: california wildfires D1: hurricane harvey D2: hurricane irma D3: hurricane maria D4: iraq iran earthquake D5: mexico earthquake D6: srilanka floods

		D0	D1	D2	D3	D4	D5	D6	Avg
Text only (RNN)	Acc	75.15	79.86	78.70	76.20	74.25	72.73	90.67	78.22
	AUC	71.51	83.47	81.38	84.33	66.65	72.90	92.21	78.92
	F1	60.45	73.96	70.92	74.81	55.63	64.57	87.68	69.71
Image only (CNN)	Acc	76.68	76.60	78.65	66.75	74.80	74.19	80.48	75.45
	AUC	68.72	77.93	62.41	69.46	70.53	69.03	85.55	71.95
	F1	75.03	75.40	71.54	66.41	73.18	71.57	80.27	73.34
Multimodal (RNN+ CNN) (independent+average)	Acc	78.50	79.32	74.98	74.23	76.15	76.27	84.60	77.72
	AUC	71.93	84.22	75.53	82.37	70.25	79.38	86.04	78.53
	F1	76.04	78.17	72.26	73.89	71.89	73.27	83.90	75.63
Multimodal (RNN+CNN) (concat+co-train)	Acc	75.45	79.78	79.91	76.67	75.06	73.62	91.59	78.87
	AUC	71.49	84.29	81.36	83.64	71.64	76.71	96.50	80.80
	F1	73.20	79.20	79.07	76.59	74.13	72.92	91.85	78.13

Evaluation Metric

We use several standard evaluation metrics as defined below.

Accuracy

The accuracy represents the number of correctly predicted instances divided by the total number of samples:

$$Accuracy = \sum_i I_{\hat{y}_i=y_i} / n$$

where I is the indicator function, $I_{\hat{y}_i=y_i} = 1$ if $\hat{y}_i = y_i$; $I_{\hat{y}_i=y_i} = 0$ otherwise, and n is the sample size. The disadvantage of this metric is that it cannot capture well the performance on unbalanced data. If the model classifies all samples into one class, the accuracy is still high but the model actually does not have good prediction ability.

AUC

We also use AUC (Area Under The Curve) to measure the model's classification performance. As the name suggests, AUC is the area under a curve, specifically the curve which plots TPR vs. FPR at all possible thresholds, where True Positive Rate (TPR) is

$$TPR = \frac{TP}{TP + FN}$$

and False Positive Rate (FPR) is

$$FPR = \frac{FP}{FP + TN}.$$

This is a measure which balances the True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN), and better captures the performance for unbalanced data.

F1 Score

F1 Score is another measure which balances the precision and recall. In disaster management, the dataset is generally unbalanced. This might be a better measure than Accuracy to make the comparison.

Experimental Results

The experimental results for the first scenario are shown in Table 4. We compare the results for text alone, image alone and two types of multi-modal models. These results are performed on the matched label tweets for which the statistics are shown in Table 2, right panel. Because the datasets are unbalanced, we focus our discussion on the AUC and F1 scores. As can be seen, the RNN alone network has better AUC than the CNN alone network, while the CNN alone network has better F1-score. However, the combination of the two networks has better performance overall. As can be seen, the average performance of the multimodal model over all experiments/datasets has better results than both the CNN model and RNN model. This shows that the combination of text and image information

Table 5. Text label results. D0: california wildfires D1: hurricane harvey D2: hurricane irma D3: hurricane maria D4: iraq iran earthquake D5: mexico earthquake D6: srilanka floods

		D0	D1	D2	D3	D4	D5	D6	Avg
Text only (RNN)	Acc	75.90	79.17	78.81	70.41	76.23	74.90	89.84	77.89
	AUC	66.08	77.60	73.60	75.74	70.47	71.17	96.07	75.82
	F1	57.82	68.33	62.19	66.83	55.64	65.23	90.19	66.60
Multimodal (RNN+ CNN) (independent+average)	Acc	77.62	76.34	76.15	67.11	81.11	74.91	78.26	75.93
	AUC	67.35	71.65	63.49	69.47	67.98	73.11	83.47	70.93
	F1	74.29	73.48	72.61	65.05	77.04	71.47	78.44	73.20
Multimodal (RNN+ CNN) (concatenate+co-train)	Acc	73.65	77.86	79.69	71.56	80.89	75.35	89.86	78.41
	AUC	68.59	77.79	73.90	76.81	72.28	72.95	95.99	76.90
	F1	71.45	76.42	77.18	70.72	79.57	74.77	89.97	77.15

Table 6. Image label results. D0: california wildfires D1: hurricane harvey D2: hurricane irma D3: hurricane maria D4: iraq iran earthquake D5: mexico earthquake D6: srilanka floods

		D0	D1	D2	D3	D4	D5	D6	Avg
Image only (CNN)	Acc	76.68	76.83	78.65	64.33	83.33	74.19	81.20	76.46
	AUC	68.72	70.79	62.41	65.11	76.30	69.03	85.47	71.72
	F1	75.03	73.14	71.54	60.18	80.43	71.57	80.93	73.26
Multimodal (RNN+ CNN) (independent+average)	Acc	77.13	77.08	77.74	65.58	81.33	75.81	78.40	76.15
	AUC	67.41	71.96	65.30	66.40	70.13	72.31	84.76	71.18
	F1	73.92	73.15	73.75	62.28	75.25	72.00	78.27	72.66
Multimodal (RNN+ CNN) (concatenate+co-train)	Acc	75.11	78.42	80.38	71.42	80.00	73.39	92.00	78.67
	AUC	70.11	77.92	72.24	77.35	72.28	75.32	96.09	77.33
	F1	74.09	77.68	78.38	69.63	78.01	73.34	92.06	77.60




leads to better classification labels. The concatenation+co-training multimodal model gives the best results overall. By looking at specific datasets, we can see that the multimodal model has better results for hurricane disasters (D1, D2, D3), as compared to the results for the fire disaster (D0). The reason for better performance might be the fact that the hurricane datasets are larger and enable better learning of the models. Based on the results in Table 4, we can conclude that the proposed multimodal model can improve the classification performance for the matched label scenario. However, this scenario is not very useful from a practical point of view, as we don't know if the text and image labels will match when we try to use this model for an ongoing disaster.

The experimental results for the second scenario are shown in Table 5 and Table 6 for text labels and image labels, respectively. As can be seen in Table 5, for text labels, the average multimodal model has better overall F1-score than the RNN alone model. However, the AUC for the average multimodal model is overall worse than that of the RNN alone model. The concatenation+co-train multimodal model has AUC performance better than that of the RNN alone model, and has much better performance in terms of the F1 Score. The concatenation+co-train multimodal model also has better performance than the average multimodal model overall. Considering specific datasets, the multimodal model has better performance on hurricane datasets, earthquake datasets and comparable results on flood datasets. As in the first scenario, this shows that our proposed model benefits from larger datasets. Based on these results, our conclusion is that the multimodal model (based on concatenation+co-training) has better performance than the text only model. Thus, we suggest to use the multimodal model when predicting the labels of tweet texts.

Similar patterns can be seen in Table 6 for image labels. When using text to supplement image information, the results are comparable between the image only model and the average multi-modal model. The concatenation+co-train multimodal model has better performance overall. Similar individual pattern can be seen here. The multimodal model benefits most from the larger hurricane datasets. The results suggests that text can also help the prediction of image. Therefore, the concatenation+co-train multimodal model can be used in image classification.

We also show some text prediction examples in Table 7 to get insights into how the multimodal model helps. In the first two examples, the text is informative but the RNN model misclassifies it as not-informative. The corresponding images show a damaged church, and a hurricane, respectively, and help the multimodal model assign the correct labels. The text in the third example is somewhat confusing cannot be used to predict the tweet as not informative. However, the corresponding image is clearly not informative with respect to a disaster and helps the multimodal

Table 7. Text Classification Examples: the multimodal model helps improve the classification

Text Label	Informative	Informative	Not Informative
Text	Baptist church damaged by #harvey can't use sanctuary	A piece of a #Harvey rain band drifts north through Breton Sound	Hurricane Harvey Relief Supply Drive...and efforts towards IRMA
Image Label	Informative	Informative	Not Informative
Image			
RNN prediction	Not Informative	Not Informative	Informative
Multimodal prediction	Informative	Informative	Not Informative

model correctly classify the tweet. These examples support our conclusion that the information from an image can help the prediction of text labels.

CONCLUSIONS

In this paper, we propose a multi-modal model which combines text and image information through the means of RNN and CNN models. We evaluate different scenarios and identify a scenario where it is useful to use the multi-modal model as opposed to specific modality models. Specifically, we show that our proposed multimodal model can be used to improve the text classification by adding the image information and improve the image classification by adding the text information. The future work for this study will be focused on explore different text/image classification models to improve the multi-modal model architecture. We will evaluate our approach on disaster-specific data (e.g., learn a model for hurricanes, another model for earthquakes, etc.) and also on cross-disaster data (e.g., learn a model on hurricanes and use it on a new type of disaster), so that we can get more insights into the model performance. We will also evaluate our proposed models on other datasets, including multi-class datasets for predicting situational awareness categories.

ACKNOWLEDGEMENTS

We thank the National Science Foundation and Amazon Web Services for support from grant IIS-1741345, which enabled the research and the computation in this study. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either express or implied, of the National Science Foundation or Amazon Web Services.

REFERENCES

- 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 34th American Association for Artificial Intelligence (AAAI 2020)*.
- Alam, F., Imran, M., and Ofli, F. (2017). "Image4act: Online social media image processing for disaster response". In: *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pp. 601–604.

- Alam, F., Ofli, F., and Imran, M. (2018). "Crisismmd: Multimodal twitter datasets from natural disasters". In: *Twelfth International AAAI Conference on Web and Social Media*.
- Alexander, D. E. (2014). "Social media in disaster risk reduction and crisis management". In: *Science and engineering ethics* 20.3, pp. 717–733.
- Ashktorab, Z., Brown, C., Nandi, M., and Culotta, A. (2014). "Tweedr: Mining twitter to inform disaster response". In: *ISCRAM*.
- Barnes, C. F., Fritz, H., and Yoo, J. (2007). "Hurricane disaster assessments with image-driven data mining in high-resolution satellite imagery". In: *IEEE Transactions on Geoscience and Remote Sensing* 45.6, pp. 1631–1640.
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *arXiv preprint arXiv:1406.1078*.
- Chowdhury, S. R., Imran, M., Asghar, M. R., Amer-Yahia, S., and Castillo, C. (2013). "Tweet4act: Using incident-specific profiles for classifying crisis-related messages." In: *ISCRAM*.
- Houston, J. B., Hawthorne, J., Perreault, M. F., Park, E. H., Goldstein Hode, M., Halliwell, M. R., Turner McGowen, S. E., Davis, R., Vaid, S., McElderry, J. A., et al. (2015). "Social media and disasters: a functional framework for social media use in disaster planning, response, and research". In: *Disasters* 39.1, pp. 1–22.
- Hu, A. and Flaxman, S. (2018). "Multimodal sentiment analysis to explore the structure of emotions". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 350–358.
- Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (2015). "Processing social media messages in mass emergency: A survey". In: *ACM Comp. Surveys (CSUR)*.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in neural information processing systems*, pp. 3146–3154.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*, pp. 1097–1105.
- Li, H., Li, X., Caragea, D., and Caragea, C. (2018). "Comparison of Word Embeddings and Sentence Encodings as Generalized Representations for Crisis Tweet Classification Tasks". In:
- Li, X., Caragea, D., Caragea, C., Imran, M., and Ofli, F. (2019). "Identifying Disaster Damage Images Using a Domain Adaptation Approach". In: *Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management (ISCRAM), Valencia, Spain. Academic Press*.
- Liu, P., Qiu, X., and Huang, X. (2016). "Recurrent neural network for text classification with multi-task learning". In: *arXiv preprint arXiv:1605.05101*.
- Mouzannar, H., Rizk, Y., and Awad, M. (2018). "Damage Identification in Social Media Posts using Multimodal Deep Learning." In: *ISCRAM*.
- Nalluru, G., Pandey, R., and Purohit, H. (2019). "Relevancy Classification of Multimodal Social Media Streams for Emergency Services". In: *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, pp. 121–125.
- Neppalli, V. K., Caragea, C., and Caragea, D. (2018). "Deep Neural Networks versus Naive Bayes Classifiers for Identifying Informative Tweets during Disasters." In: *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2018)*.
- Nguyen, D. T., Ofli, F., Imran, M., and Mitra, P. (2017). "Damage assessment from social media imagery data during disasters". In: *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ACM, pp. 569–576.
- Reuter, C., Hughes, A. L., and Kaufhold, M.-A. (2018). "Social Media in Crisis Management: An Evaluation and Analysis of Crisis Informatics Research". In: *Int. J. of Human–Computer Interaction* 34.4, pp. 280–294.
- Stowe, K., Paul, M., Palmer, M., Palen, L., and Anderson, K. M. (2016). "Identifying and categorizing disaster-related tweets". In: *Proceedings of The fourth international workshop on natural language processing for social media*, pp. 1–6.
- Velev, D. and Zlateva, P. (2012). "Use of social media in natural disaster management". In: *Intl. Proc. of Economic Development and Research* 39, pp. 41–45.

- Verma, S., Vieweg, S., Corvey, W. J., Palen, L., Martin, J. H., Palmer, M., Schram, A., and Anderson, K. M. (2011). "Natural Language Processing to the Rescue? Extracting " Situational Awareness" Tweets During Mass Emergency." In: *ICWSM*.
- Vetrivel, A., Kerle, N., Gerke, M., Nex, F., and Vosselman, G. (2016). "Towards automated satellite image segmentation and classification for assessing disaster damage using data-specific features with incremental learning". In:
- Xiao, Y., Huang, Q., and Wu, K. (2015). "Understanding social media data for disaster management". In: *Natural hazards* 79.3, pp. 1663–1679.
- Yang, Y., Ha, H.-Y., Fleites, F., Chen, S.-C., and Luis, S. (2011). "Hierarchical disaster image classification for situation report enhancement". In: *2011 IEEE International Conference on Information Reuse & Integration*. IEEE, pp. 181–186.