Analysis of Social Media Data using Multimodal Deep Learning for Disaster Response

Ferda Ofli
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
fofli@hbku.edu.qa

Firoj Alam
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
fialam@hbku.edu.qa

Muhammad Imran
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
mimran@hbku.edu.qa

ABSTRACT

Multimedia content in social media platforms provides significant information during disaster events. The types of information shared include reports of injured or deceased people, infrastructure damage, and missing or found people, among others. Although many studies have shown the usefulness of both text and image content for disaster response purposes, the research has been mostly focused on analyzing only the text modality in the past. In this paper, we propose to use both text and image modalities of social media data to learn a joint representation using state-of-the-art deep learning techniques. Specifically, we utilize convolutional neural networks to define a multimodal deep learning architecture with a modality-agnostic shared representation. Extensive experiments on real-world disaster datasets show that the proposed multimodal architecture yields better performance than models trained using a single modality (e.g., either text or image).

Keywords

Multimodal deep learning, Multimedia content, Natural disasters, Crisis Computing, Social media

INTRODUCTION

Information from different modalities often brings complementary signals about a concept, an object, an event, and the like. Learning from these different modalities leads to more robust inference compared to learning from a single modality. Multimodal learning is a well-researched area and has been applied in many fields including audio-visual analysis (e.g., videos) (Poria et al. 2016; Pereira et al. 2016), cross-modal study (Nagrani et al. 2018), and speech processing (e.g., audio and transcriptions) (Chowdhury et al. 2019). Despite the successes of multimodal learning in other areas, limited focus has been given to multimodal social media data analysis until recently (Gautam et al. 2019). In particular, using social media data for social good requires time-critical analysis of the multimedia content (e.g., textual messages, images, videos) posted during a disaster situation to help humanitarian organizations in preparedness, mitigation, response, and recovery efforts.

Figure 1 shows a few tweets with associated images collected from three recent major disasters. We observe that relying on a single modality may often miss important insights. For instance, although the tweet text in Figure 1(f) reports about a 6.1 magnitude earthquake in Southern Mexico, the scale of the damage incurred by this earthquake cannot be inferred from the text. However, if we analyze the image attached to this tweet, we can easily understand the immense destruction caused by the earthquake.

(a) Hurricane Maria turns Dominica into 'giant debris field' https://t.co/rAISiAhMUy by #AJEnglish via <USER>

(b) A friend's text message saved Sarasota man from deadly California wildfire https://t.co/0TNMFgL885

(c) Earthquake leaves hundreds dead, crews combing through rubble in #Mexico https://t.co/XPbAEIBcKw

(d) Corporate donations for Hurricane Maria relief top $24 million https://t.co/w34ZZziu88

(e) California Wildfires Threaten Significant Losses for P/C Insurers, Moodya Says https://t.co/ELUaTkYbzZ

(f) Southern Mexico rocked by 6.1-magnitude earthquake CLICK BELOW FOR FULL STORY... https://t.co/Vkz6fNVe5s...

Figure 1. Tweet text and image pairs from different disaster events with complementary information.

Most of the previous studies that rely on social media for disaster response have mainly focused on textual content analysis, and little focus has been given to images shared on social media (Imran et al. 2015; Castillo et al. 2016). Many past research works have demonstrated that images shared on social media during a disaster event can help humanitarian organizations in a number of ways. For example, Nguyen, Ofli, et al. 2017 use images shared on Twitter to assess the severity of infrastructure damage. Mouzannar et al. 2018 also focus on identifying damages in infrastructure and environmental elements. Taking a step further, Gautam et al. 2019 have recently presented a work on multimodal analysis of crisis-related social media data for identifying informative tweet text and image pairs.

In this study, we also aim to use both text and image modalities of Twitter data to learn (i) whether a tweet is informative for humanitarian aid or not, and (ii) whether it contains some useful information such as a report of injured or deceased people, infrastructure damage, etc. We tackle this problem in two separate classification tasks and solve them using multimodal deep learning techniques.

The typical approach to deal with multimodality includes feature- or decision-level fusion, which is also termed as early and late fusion (Kuncheva 2004; Alam and Riccardi 2014). In deep learning architectures, multimodality is combined at the hidden layers with a different variant of network architecture such as static, dynamic, and N-way classification as can be seen in (Ngiam et al. 2011; Nagrani et al. 2018; Chowdhury et al. 2019). Specifically, Ngiam et al. 2011 explore different architectures for audio-visual data. Their study includes unimodal as well as cross-modal learning (i.e., learning one modality while giving multiple modalities during feature learning), multimodal fusion, and shared representation learning. Nagrani et al. 2018 also study audio-visual data for a biometric matching task while they investigate different deep neural network architectures for developing a multimodal network whereas Chowdhury et al. 2019 analyze audio and transcriptions while concatenating both modalities in a hidden layer.

In this work, we propose to learn a joint representation from two parallel deep learning architectures where one architecture represents the text modality and the other architecture represents the image modality. For the image modality, we use the well-known VGG16 network architecture and extract high-level features of an input image using the penultimate fully-connected (i.e., fc2) layer of the network. For the text modality, we define a Convolutional Neural Network (CNN) with five hidden layers and different filters. Two feature vectors obtained from both modalities are then fed into a shared representation followed by a dense layer before performing a prediction using softmax. In the literature, this type of joint representation is also known as early fusion. The proposed multimodal architecture is trained using three different settings as follows: (i) train a network using input data from both modalities, (ii) train a network using only the text modality, and (iii) train a network using only the image modality.

We perform extensive experiments on a real-world disaster-related dataset collected from Twitter, i.e,. CrisisMMD (Alam, Ofli, et al. 2018a), using the aforementioned three training settings for two different classification tasks: informativeness and humanitarian classification. The test data for the evaluation of all three settings are fixed. The experimental results show that the proposed approach (i.e., multimodal learning) outperforms our baseline models trained on a single modality (i.e., either text or image). For the informativeness classification task, our best model obtained an F1-score=84.2 and for the humanitarian classification task, our best model achieved an F1-score=78.3. Despite the fact that this model outperforms its counter-part unimodal baseline models (i.e., trained on a single modality), we remark that there is a big room for improvement, which we leave as future work. To the best of our knowledge, this is the first study that presents baseline results on CrisisMMD using state-of-the-art deep learning-based unimodal and multimodal approaches for both informativeness and humanitarian tasks, all in one place. We hope that experimental analyses presented in this study will provide guidance for future research using the CrisisMMD dataset.

The rest of the paper is organized as follows. In the *Related Work* section, we provide a review of the literature. Then, in the *Dataset* section, we present details of the dataset used in this study. Next, in the *Experiments* section, we describe the methodology and discuss experimental results. We then present possible applications and future directions in the *Discussion* section. Finally, we conclude the paper with the *Conclusions* section.

RELATED WORK

Many past studies have analyzed social media data, especially textual content, and demonstrated its useful for humanitarian aid purposes (Imran et al. 2015; Castillo et al. 2016). With recent successes of deep learning, research works have started to use social media images for humanitarian aid. For instance, the importance of imagery content on social media has been reported in many studies (Peters and Joao 2015; Daly and Thom 2016; Chen et al. 2013; Nguyen, Alam, et al. 2017; Nguyen, Ofli, et al. 2017; Alam, Imran, et al. 2017; Alam, Ofli, et al. 2018b). Peters and Joao 2015 analyzed the data collected from Flickr and Instagram for the flood event in Saxony, 2013. Their findings suggested that the existence of images within on-topic textual content were more relevant to the disaster event, and the imagery content also provided important information, which was related to the event. Similarly, Daly and Thom 2016 analyzed images extracted from social media data, which is focused on a fire event. They analyzed spatio-temporal meta-data associated with the images and suggested that geotagged information are useful to locate the fire affected areas. The analysis of imagery content shared on social media has been explored using deep learning techniques in several studies (Nguyen, Alam, et al. 2017; Nguyen, Ofli, et al. 2017; Alam, Imran, et al. 2017). Furthermore, Alam, Ofli, et al. 2018b presented an image processing pipeline to extract meaningful information from social media images during a crisis situation, which has been developed using deep learning-based techniques. Their image processing pipeline includes collecting images, removing duplicates, filtering irrelevant images, and finally classifying them with damage severity.

Combining textual and visual content can provide highly relevant information as discussed by Bica et al. 2017 where they explored social media images posted during two major earthquakes in Nepal during April-May 2015. Their study focused on identifying geo-tagged images and their associated damage. Chen et al. 2013 studied the association between tweets and images, and their use in classifying visually relevant and irrelevant tweets. They designed classifiers by combining features from the text, images and socially relevant contextual features (e.g., posting time, follower ratio, the number of comments, re-tweets), and reported an F1-score of 70.5% in a binary classification task, which is 5.7% higher than the text-only classification. Recently, Mouzannar et al. 2018 explored damage detection by focusing on human and environmental damages. Their study explores unimodal as well as different multimodal modeling setups based on a collection of multimodal social media posts labeled with six categories such as infrastructural damage (e.g., damaged buildings, wrecked cars, and destroyed bridges), damage to natural landscape (e.g., landslides, avalanches, and falling trees), fires (e.g., wildfires and building fires), floods (e.g., city, urban and rural), human injuries and deaths, and no damage. Similarly, Gautam et al. 2019 presented a comparison of unimodal and multimodal methods on crisis-related social media data using an approach based on decision fusion for classifying tweet text and image pairs into informative and non-informative categories.

For the tweet classification task, deep learning-based techniques such as Convolutional Neural Networks (CNN) (Nguyen, Al-Mannai, et al. 2017), and Long-Short-Term-Memory Networks (LSTM) (Rosenthal et al. 2017) have been widely used. For the image classification task, state-of-the-art works also utilize different techniques of deep neural networks such as Convolutional Neural Networks (CNN) with deep architectures. Among different CNN architectures, the most popular are VGG (Simonyan and Zisserman 2014), AlexNet (Krizhevsky et al. 2012), and GoogLeNet (Szegedy et al. 2015). The VGG is designed using an architecture with very small (3×3) convolution filters and with a depth of 16 and 19 layers. The 16-layer network is referred to as VGG16 network, which we used in this study.

For combining multiple modalities, early and late fusion have been the traditional approaches (Kuncheva 2004). The early-fusion approaches combine features from different modalities while the late-fusion approaches make classification decisions either by majority voting or stacking methods (i.e., combining classifiers' output and making a decision by training another classifier). In deep learning paradigm, typically, hidden layers are combined using approaches such as static or dynamic concatenation as discussed in (Nagrani et al. 2018; Chowdhury et al. 2019). In this study, we follow CNN-based deep learning architectures for both unimodal and multimodal experiments. We extract high-level features from two independent modality-specific networks (i.e., text and image) and concatenate them to form a shared representation for our classification tasks.

Our study differs from previous studies in a number of ways (Mouzannar et al. 2018; Gautam et al. 2019). As such, we experiment with one of the largest, publicly available datasets (i.e., CrisisMMD) to provide baseline results for two popular crisis response tasks, i.e., *informativeness* and *humanitarian categorization*, using multimodal deep learning with a feature-fusion approach. In contrast, Mouzannar et al. 2018 focus only on human and environmental damages using a home-grown dataset, which limits generalization of their findings. As for Gautam et al. 2019, although they also use a subset of the CrisisMMD dataset, they focus only on the informativeness classification task and employ a decision-fusion approach in their experiments. Unfortunately, due to potential differences in data organization (i.e., training, validation, and test splits), our experimental results are not exactly comparable with theirs.

DATASET

We use CrisisMMD[1] dataset, which is a multimodal dataset consisting of tweets and associated images collected during seven different natural disasters that took place in 2017 (Alam, Ofli, et al. 2018a). The dataset is annotated for three tasks: (i) informative *vs.* not-informative, (ii) humanitarian categories (eight classes), and (iii) damage severity (three classes). Because the third task, i.e., damage severity, was applied only on images, we do not consider this task in the current study and focus only on the first two tasks as follows.

Informative vs. Not-informative. The purpose of this task is to determine whether a given tweet text or image, collected during a disaster event, is useful for humanitarian aid purposes[2]. If the given text (image) is useful for humanitarian aid, it is considered as an "informative" tweet (image), otherwise as a "not-informative" tweet (image).

Humanitarian Categories. The purpose of this task is to understand the type of information shared in a tweet text/image, which was collected during a disaster event. Given a tweet text/image, the task is to categorize it into one of the categories listed in Table 1.

An important property of CrisisMMD is that some of the co-occurring tweet text and image pairs have different labels for the same task because text and image modalities were annotated separately and independently. Therefore, in this study, we consider only a subset of the original dataset where text and image pairs have the same label for a given task. As a result of this filtering, some of the categories in the humanitarian task are left with only a few pairs of tweet text and image. This situation skews the overall label distribution and creates a challenging setup for model training. To deal with this issue, we combine those minority categories that are semantically similar or relevant. Specifically, we merge the "injured or dead people" and "missing or found people" categories into the "affected individuals" category. Similarly, we merge "vehicle damage" category into the "infrastructure and utility damage" category. As a result, we are left with five categories for the humanitarian task as shown in Table 1.

Twitter allows attaching up to four images to a tweet, and hence, CrisisMMD contains some tweets that have more than one image, i.e., the same tweet text is associated with multiple images. While splitting data into training, development, and test sets, we need to make sure that no duplicate tweet text exists across these splits. To achieve this, we simply assign all tweets with multiple images only to the training set. This also ensures that there are no repeated data points (i.e., tweet text) in the development and test sets for the unimodal experiments on text modality. While doing so, we maintain a 70%:15%:15% data split ratio for training, development, and test sets, respectively. Table 1 provides the final set of categories, total number of tweet text and images in each category as well as their split into training, development, and test sets.[3] Note that the total number of tweet text and images in the table differ only for the training sets as per the strategy explained above.

EXPERIMENTS

As explained in the previous section, we have two sets of annotations for two separate classification tasks, i.e., informativeness and humanitarian. For each one of these tasks, we perform three classification experiments where we train models using (i) only tweet text, (ii) only tweet image, and (iii) tweet text and image together.

In the following subsections, we first describe the data preprocessing steps and then describe the deep learning architecture used for each modality as well as their training details. To measure the performance of the trained models, we use several well-known metrics such as accuracy, precision, recall, and F1-score.

[1] http://crisisnlp.qcri.org/

[2] A detailed definition of *humanitarian aid* is provided in (Alam, Ofli, et al. 2018a).

[3] The data split used in the experiments can be found online at http://crisisnlp.qcri.org/.

Table 1. List of categories and their data split for different tasks.

	Train (70%)		Dev (15%)		Test (15%)		Total	
	Text	Image	Text	Image	Text	Image	Text	Image
Informative Task								
Informative	5,546	6,345	1,056	1,056	1,030	1,030	7,632	8,431
Not-informative	2,747	3,256	517	517	504	504	3,768	4,277
Total	8,293	9,601	1,573	1,573	1,534	1,534	11,400	12,708
Humanitarian Task								
Affected individuals	70	71	9	9	9	9	88	89
Rescue volunteering or donation effort	762	912	149	149	126	126	1,037	1,187
Infrastructure and utility damage	496	612	80	80	81	81	657	773
Other relevant information	1,192	1,279	239	239	235	235	1,666	1,753
Not-humanitarian	2,743	3,252	521	521	504	504	3,768	4,277
Total	5,263	6,126	998	998	955	955	7,216	8,079

Data Preprocessing

The textual content of tweets is often noisy, usually consisting of many symbols, emoticons, and invisible characters. Therefore, we preprocess them by removing stop words, non-ASCII characters, numbers, URLs, and hashtag signs. We also replace all punctuation marks with white-spaces.

On the image side, we follow the typical preprocessing steps of scaling the pixels of an image between 0 and 1 and then normalizing each channel with respect to the ImageNet dataset (Deng et al. 2009).

CNN: Text Modality

For the text modality, we use Convolutional Neural Network (CNN) due to its better performance in crisis-related tweet classification tasks reported in (Nguyen, Al-Mannai, et al. 2017). Specifically, we create a CNN consisting of 5 hidden layers. To input the network, we zero-padded the tweets for an equal length and then converted them into a word-level matrix where each row represents a word in the tweet extracted using a pre-trained word2vec model discussed in (Alam, Joty, et al. 2018). This word2vec model is trained using the Continuous Bag-of-Words (CBOW) approach of Mikolov et al. 2013 on a large disaster-related dataset of size 364 million tweets with vector dimensions of 300, a context window size of 5 and $k = 5$ negative samples.

The input then goes through a series of sequential layers including the convolutional layer, followed by the max-pooling layer, to obtain a higher-level fixed-size feature representation for each tweet. These fixed-size feature vectors are then passed through one or more fully connected hidden layers, followed by an output layer. In the convolutional and fully-connected layers, we use rectified linear units (ReLU) (Krizhevsky et al. 2012) as the activation function, and in the output layer, we use the softmax activation function.

We train the CNN models using the Adam optimizer (Zeiler 2012). The learning rate is set to 0.01 when optimizing for the classification loss on the development set. The maximum number of epochs is set to 50, and dropout (Srivastava et al. 2014) rate of 0.02 is used to avoid overfitting. We set *early-stopping* criterion based on the accuracy on the development set with the patience of 10. We use 100, 150, and 200 filters with the corresponding window size of 2, 3, and 4, respectively. We use the same pooling length as the filter window size. We also apply batch normalization due to its success reported in the literature (Ioffe and Szegedy 2015).

VGG16: Image Modality

For the image modality, we employ a transfer learning approach, which is an effective approach for visual recognition tasks (Yosinski et al. 2014; Ozbulak et al. 2016). The idea of the transfer learning approach is to use existing weights of a pre-trained model. We use the weights of a VGG16 model pre-trained on ImageNet to initialize our model. We adapt the last layer (i.e., softmax layer) of the network according to the particular classification task at hand instead of the original 1,000-way classification. The transfer learning approach allows us to transfer the features and the parameters of the network from the broad domain (i.e., large-scale image classification) to the specific one, in our case informativeness and humanitarian classification tasks.

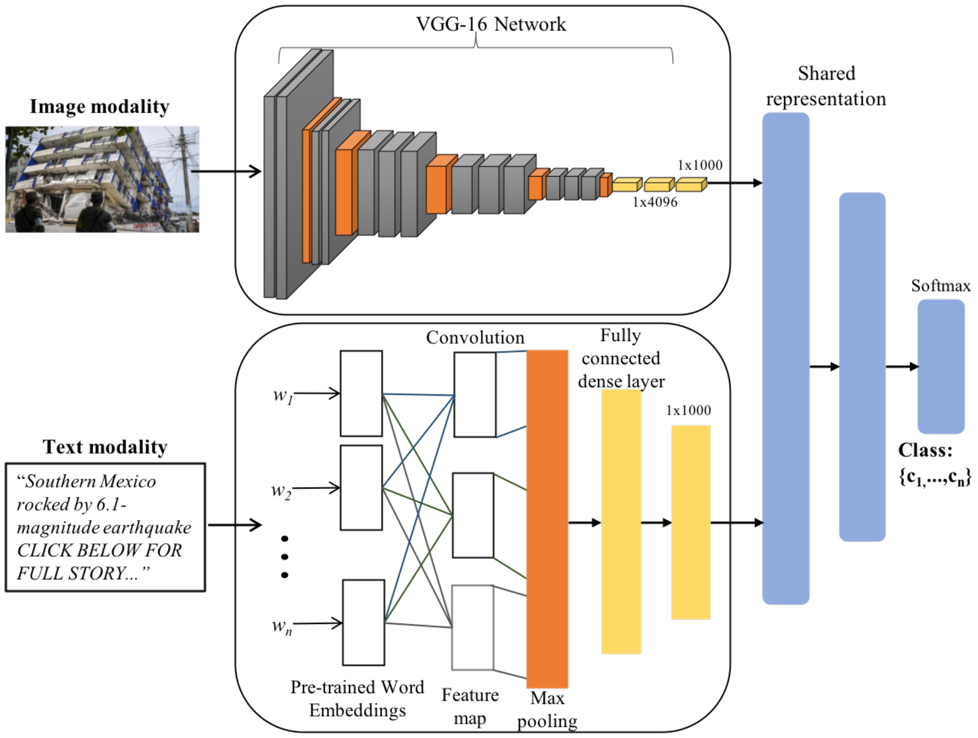

VGG-16 Network
Image modality
Shared representation
1x1000
1x4096
Softmax
Convolution
Fully connected dense layer
1x1000
Text modality
"Southern Mexico rocked by 6.1-magnitude earthquake CLICK BELOW FOR FULL STORY..."
w_1
w_2
w_n
Class: $\{c_1,\ldots,c_n\}$
Pre-trained Word Embeddings
Feature map
Max pooling

Figure 2. The multimodal architecture for the classification task using both text and image as input to the system.

We train the image models using the Adam optimizer (Zeiler 2012) with an initial learning rate of 10^{-6}, which is reduced by a factor of 0.1 when accuracy on the development set stops improving for 100 epochs. We set the maximum number of epochs to 1,000 with an early-stopping criterion.

Multimodal: Text and Image

In Figure 2, we present the architecture of the multimodal deep neural network that we use for the experiment. As can be seen in the figure, for the image modality we use the VGG16 network. For the text modality, we use a CNN based architecture. Before forming the shared representations from both modalities we have another hidden layer of size 1,000 from each side. The reason to choose the same size is to have an equal contribution from both modalities. In the current experimental setting, there is no specific reason for choosing the size of 1,000, which can be optimized empirically. After the concatenation of both modalities, we have one hidden layer before the softmax layer.

We use the Adam optimizer with a minibatch size of 32 for training the model. In order to avoid overfitting, we use early-stopping condition, and as an activation function, we choose ReLU. For this experiment, we do not tune any hyper-parameter (e.g., the size of hidden layers, filter size, dropout rate, etc.). Hence, there is room for further improvement in future studies.

Results

In Tables 2 and 3, we present the performance results achieved for different tasks and modalities. In the unimodal experiments, the image-only models perform better than the text-only models in both informativeness and humanitarian tasks. Specifically, the improvement is more than 2% on average in the informativeness task whereas it is more than 6% on average in the humanitarian task. In the multimodal experiments, we observe additional improvements in performance in both tasks. Specifically, multimodal model performs about 1% better than the image-only model in all measures for the informativeness task and about 2% better than the image-only model in all measures for the humanitarian task. These results confirm that multimodal learning approach provides further performance improvement over the unimodal learning approach.

Overall performance of the humanitarian classification models is lower than the informativeness classification models due to the relatively more complex nature of the former task. However, it is important to note that the results presented in this study are obtained using basic network architectures and should be considered as a baseline study.

Table 2. Results for the informativeness classification task.

Training mode	Modality	Accuracy	Precision	Recall	F1-score
Unimodal	Text	80.8	81.0	81.0	80.9
	Image	83.3	83.1	83.3	83.2
Multimodal	Text + Image	84.4	84.1	84.0	84.2

Table 3. Results for the humanitarian classification task.

Training mode	Modality	Accuracy	Precision	Recall	F1-score
Unimodal	Text	70.4	70.0	70.0	67.7
	Image	76.8	76.4	76.8	76.3
Multimodal	Text + Image	78.4	78.5	78.0	78.3

DISCUSSION

Deep learning models are data hungry. Given our initial condition to consider only the tweet text and image pairs that have the same labels for a given task, we are left with a limited subset of the original CrisisMMD dataset. The proposed multimodal joint learning showed comparable performance over unimodal models for both text and image modalities. Furthermore, while designing a model for multiple modalities an important challenge is to find concatenation strategies that better capture important information from both modalities. Towards this direction, we design the model by concatenating hidden layers into another layer to form a joint shared representation.

We further analyzed the performance of the three models (i.e., text-only, image-only, and text + image) by examining their confusion matrices. Table 4 shows three confusion matrix for the three models for the first task (i.e., informative vs. not-informative). From the emergency managers' point of view, it is important that the machine does not miss any useful and relevant message/tweet. The three confusion matrices (a, b, & c) reveal that our text-only and image-only models missed 155 and 114 instances, respectively, whereas multimodal model missed only 101 instances. These are the instances where machine says "not informative", but the ground-truth labels (i.e., human annotators) say "informative" (a.k.a. false negatives). The image-only model made significant improvements over the text-only model, however, when text and image modalities are combined in the multimodal case, the error rate dropped significantly (i.e., from 155 to 101).

Another important aspect is related to information overload on emergency managers during a disaster situation. Specifically, it happens when the machine says a message is informative, but according to ground-truth labels it is not (a.k.a. false positives). The confusion matrices in Table 4 show these mistakes made by the three models as 139 by the text-only, 145 by image-only, and 139 in the multimodal case. We do not observe any improvements from the multimodal approach as observed in the false negative case.

Table 5 shows confusion matrices from the three models for the humanitarian categorization tasks. One prominent and important column to observe is "N", which corresponds to the "not-humanitarian" category, and shows all instances where the model prediction is "not-humanitarian". In particular, if we look at the number of instances where actual label is "infrastructure and utility damage" (denoted as "I") but the model prediction is "not-humanitarian" (i.e., the value of the cell at the intersection of row "I" and column "N"), we see that the text-only model has 41 false negative instances in Table 5(a) whereas the image-only and multimodal models have 13 and 10 instances in Tables 5(b) and 5(c), respectively. This indicates that the image modality helps models better understand the "infrastructure and utility damage" category, and hence, significantly reduce the errors in the predictions. A similar phenomenon can be observed in favor of the text modality for the case where the actual label is "rescue, volunteering

Table 4. Confusion matrices resulted for the informativeness task: Informative, Not-informative.

(a) Text-only

		Predicted	
		Inf	Not-inf
Human	Inf	875	155
	Not-inf	139	365

(b) Image-only

		Predicted	
		Inf	Not-inf
Human	Inf	916	114
	Not-inf	145	359

(c) Text + Image

		Predicted	
		Inf	Not-inf
Human	Inf	929	101
	Not-inf	139	365

Table 5. Confusion matrices resulted for the humanitarian task: Affected individuals, Infrastructure and utility damage, Not-humanitarian, Other relevant information, Rescue, volunteering or donation effort.

(a) Text-only

		Predicted				
		A	I	N	O	R
Human	A	0	0	5	1	3
	I	0	17	41	12	11
	N	0	1	458	20	25
	O	0	6	105	112	12
	R	0	2	37	2	85

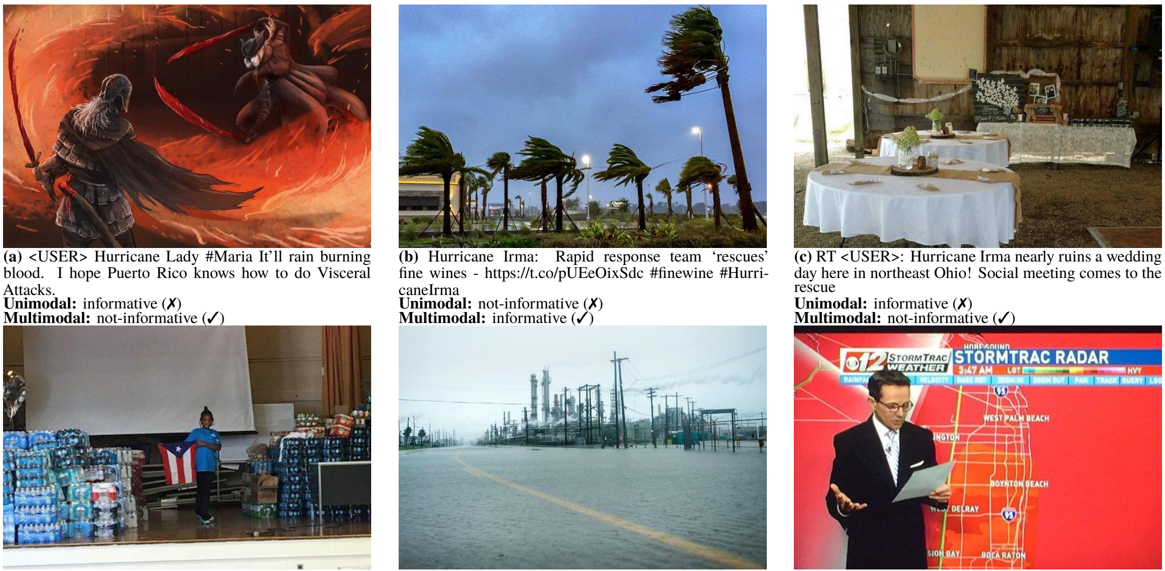
(b) Image-only

		Predicted				
		A	I	N	O	R
Human	A	1	0	4	0	4
	I	1	56	13	6	5
	N	0	13	437	22	32
	O	0	5	50	178	2
	R	0	5	43	5	73

(c) Text + Image

		Predicted				
		A	I	N	O	R
Human	A	1	0	3	0	5
	I	1	61	10	4	5
	N	0	17	426	26	35
	O	0	3	49	180	3
	R	0	9	33	3	81

or donation effort" (denoted as "R") whereas the predicted label is "not-humanitarian" (i.e., the value of the cell at the intersection of row "R" and column "N"). Specifically, the image-only model has 43 false negative instances in Table 5(b) while the text-only and multimodal models have 37 and 33 instances in Tables 5(a) and 5(c), respectively. In general, we see that the number of such errors are minimized by the third model which uses both text and image modalities together. However, there are still some cases where significant improvements can be achieved. For instance, the "other relevant information" category (denoted as "O") seems to create confusion for all the models, which needs to be investigated in a more detailed study.



(a) <USER> Hurricane Lady #Maria It'll rain burning blood. I hope Puerto Rico knows how to do Visceral Attacks.
Unimodal: informative (✗)
Multimodal: not-informative (✓)

(b) Hurricane Irma: Rapid response team 'rescues' fine wines - https://t.co/pUEeOixSdc #finewine #HurricaneIrma
Unimodal: not-informative (✗)
Multimodal: informative (✓)

(c) RT <USER>: Hurricane Irma nearly ruins a wedding day here in northeast Ohio! Social meeting comes to the rescue
Unimodal: informative (✗)
Multimodal: not-informative (✓)

(d) 6th grade Maryland student collects 3,000 cases of drinking water for Puerto Rico https://t.co/x57AeLHeaC
Text-only: not-humanitarian (✗)
Image-only: not-humanitarian (✗)
Multimodal: rescue, volunteering or donation effort (✓)

(e) Hurricane Harvey's impact on the US oil industry https://t.co/zxVWR3u0fU
Text-only: other relevant information (✗)
Image-only: not-humanitarian (✗)
Multimodal: infrastructure and utility damage (✓)

(f) #Breaking Tornado warning for Lantana Rd south to Boca Raton. #BeSafe <USER>
Text-only: not-humanitarian (✗)
Image-only: not-humanitarian (✗)
Multimodal: other relevant information (✓)

Figure 3. Example tweet text and image pairs where joint modeling of the input modalities yield better predictions. The symbol (✗) indicates an incorrect prediction and the symbol (✓) indicates a correct prediction.

Figure 3 shows example image and text pairs that illustrate how joint modeling of image and text modalities can yield better predictions, and hence, lead to performance improvements over unimodal models. For instance in (a), we reckon that the image-only model thinks the image is informative because it shows *fire-like* patterns whereas the text-only model thinks the text is informative because it mentions *rain burning blood*. However, when these two modalities are evaluated together, they do *not* really provide any consistent evidence for the model to label this image-text pair as informative any more. Similarly, in (d), evaluating image alone or text alone does not result in correct predictions whereas joint evaluation of image and text yields the correct label, i.e., "rescue, volunteering or donation effort". Furthermore, we observe another interesting case in (e): text-only model thinks there is potentially useful information for humanitarian purposes by predicting "other relevant information" whereas the image-only model thinks there is nothing related to humanitarian purposes by predicting "not-humanitarian". However, the multimodal model effectively fuses the information coming from both modalities to make the correct prediction, i.e., "infrastructure and utility damage". We believe these examples provide further insights about the success of the proposed multimodal approach for modeling crisis-related social media data.

Challenges and Future Work

In contrast to other popular multimodal tasks such as image captioning or visual question answering where there is strong alignment or coupling between text and image modalities, social media data are not warranted to have such strong alignment or coupling between co-occuring text and image pairs. In some cases, each modality conveys different type of information, which may even be contradicting the other modality. Therefore, it is important *not* to assume the existence of strong correspondences between social media text and images. To this date, this is a relatively less explored phenomenon that needs more attention from the research community since all of the existing multimodal classification approaches assume that there always exists a common label for data coming from different modalities. As such, a challenging future direction is to design a multimodal learning algorithm that can be also trained on heterogeneous input, i.e., tweet text and image pairs with disagreeing labels, in which case CrisisMMD can be used for model training in its full form.

CONCLUSION

Important informative signals gathered from different data modalities on social media can be highly useful for humanitarian organizations for disaster response. Although images shared on social media contain useful information, past studies have largely focused on text analysis, let alone combining both modalities to get better performance. In this work, we proposed to learn a joint representation using both text and image modalities of social media data. Specifically, we use state-of-the-art deep learning architectures to learn high-level feature representations from text and images to perform two classification tasks. Several experiments performed on real-world disaster-related datasets reveal the usefulness of the proposed approach. In summary, our study has two main contributions: (i) It provides baseline results, all in one place, using unimodal and multimodal approaches for both informativeness and humanitarian tasks on the CrisisMMD dataset, and (ii) it shows that a feature-fusion-based multimodal deep neural network can outperform the unimodal models on the challenging CrisisMMD dataset for both tasks, which underlines the importance of multimodal analysis of the crisis-related social media data. Despite the fact that our multimodal classifiers achieve better performance than the unimodal classifiers, we remark that there is still big room for improvement, which we leave for future work.

REFERENCES

Alam, F. and Riccardi, G. (May 2014). “Fusion of acoustic, linguistic and psycholinguistic features for Speaker Personality Traits recognition”. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 955–959.

Alam, F., Imran, M., and Ofli, F. (Aug. 2017). “Image4Act: Online Social Media Image Processing for Disaster Response.” In: *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 1–4.

Alam, F., Joty, S. R., and Imran, M. (2018). “Graph Based Semi-supervised Learning with Convolution Neural Networks to Classify Crisis Related Tweets”. In: *Proc. of the 12th ICWSM, 2018.* AAAI press.

Alam, F., Ofli, F., and Imran, M. (Jan. 2018a). “CrisisMMD: Multimodal twitter datasets from natural disasters”. English. In: *Proc. of the 12th ICWSM, 2018.* AAAI press, pp. 465–473.

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.

Bica, M., Palen, L., and Bopp, C. (2017). “Visual Representations of Disaster.” In: *Proc. of the CSCW*, pp. 1262–1276.

Castillo, C., Imran, M., Meier, P., Lucas, J. K., Srivastava, J., Leson, H., Ofli, F., and Mitra, P. (2016). “Together We Stand—Supporting Decision in Crisis Response: Artificial Intelligence for Digital Response and MicroMappers”. In: ed. by OCHA and partners. Istanbul: Tudor Rose, World Humanitarian Summit, pp. 93–95.

Chen, T., Lu, D., Kan, M.-Y., and Cui, P. (2013). “Understanding and classifying image tweets”. In: *ACM International Conference on Multimedia*, pp. 781–784.

Chowdhury, S. A., Stepanov, E. A., Danieli, M., and Riccardi, G. (2019). “Automatic classification of speech overlaps: Feature representation and algorithms”. In: *Computer Speech and Language* 55, pp. 145–167.

Daly, S. and Thom, J. (2016). “Mining and Classifying Image Posts on Social Media to Analyse Fires”. In: *International Conference on Information Systems for Crisis Response and Management*, pp. 1–14.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

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.

Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (2015). “Processing social media messages in mass emergency: A survey”. In: *ACM Computing Surveys* 47.4, p. 67.

Ioffe, S. and Szegedy, C. (2015). “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *arXiv preprint arXiv:1502.03167*.

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.

Kuncheva, L. I. (2004). *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons.

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

Mouzannar, H., Rizk, Y., and Awad, M. (2018). “Damage Identification in Social Media Posts using Multimodal Deep Learning”. In: *15th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2018)* May, pp. 529–543.

Nagrani, A., Albanie, S., and Zisserman, A. (2018). “Seeing Voices and Hearing Faces: Cross-Modal Biometric Matching”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 8427–8436.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). “Multimodal deep learning”. In: *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 689–696.

Nguyen, D. T., Alam, F., Ofli, F., and Imran, M. (May 2017). “Automatic Image Filtering on Social Networks Using Deep Learning and Perceptual Hashing During Crises”. In: *International Conference on Information Systems for Crisis Response and Management (ISCRAM)*.

Nguyen, D. T., Al-Mannai, K., Joty, S. R., Sajjad, H., Imran, M., and Mitra, P. (2017). “Robust Classification of Crisis-Related Data on Social Networks Using Convolutional Neural Networks.” In: *ICWSM*, pp. 632–635.

Nguyen, D. T., Ofli, F., Imran, M., and Mitra, P. (Aug. 2017). “Damage Assessment from Social Media Imagery Data During Disasters”. In: *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 1–8.

Ozbulak, G., Aytar, Y., and Ekenel, H. K. (Sept. 2016). “How Transferable Are CNN-Based Features for Age and Gender Classification?” In: *International Conference of the Biometrics Special Interest Group*, pp. 1–6.

Pereira, M. H. R., Pádua, F. L. C., Pereira, A. C. M., Benevenuto, F., and Dalip, D. H. (2016). “Fusing Audio, Textual and Visual Features for Sentiment Analysis of News Videos”. In: *Tenth International AAAI Conference on Web and Social Media (ICWSM)*, pp. 659–662.

Peters, R. and Joao, P. d. A. (2015). “Investigating images as indicators for relevant social media messages in disaster management”. In: *International Conference on Information Systems for Crisis Response and Management*.

Poria, S., Cambria, E., Howard, N., Huang, G. B., and Hussain, A. (2016). “Fusing audio, visual and textual clues for sentiment analysis from multimodal content”. In: *Neurocomputing* 174, pp. 50–59.

Rosenthal, S., Farra, N., and Nakov, P. (2017). “SemEval-2017 task 4: Sentiment analysis in Twitter”. In: *Proc. of the 11th SemEval, 2017)*, pp. 502–518.

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

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). “Dropout: a simple way to prevent neural networks from overfitting.” In: *Journal of MLR* 15.1, pp. 1929–1958.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). “How Transferable Are Features in Deep Neural Networks?” In: *Advances in Neural Information Processing Systems*, pp. 3320–3328.

Zeiler, M. D. (2012). “ADADELTA: an adaptive learning rate method”. In: *arXiv preprint arXiv:1212.5701*.