Using the Image-Text Relationship to Improve Multimodal Disaster Tweet Classification

Tiberiu Sosea*
University of Illinois at Chicago
tsosea2@uic.edu

Iustin Sirbu
Politehnica University of Bucharest
iustin.sirbu@stud.acs.upb.ro

Cornelia Caragea
University of Illinois at Chicago
cornelia@uic.edu

Doina Caragea
Kansas State University
dcaragea@ksu.edu

Traian Rebedea
Politehnica University of Bucharest
trebedea@gmail.com

ABSTRACT

In this paper, we show that the text-image relationship of disaster tweets can be used to improve the classification of tweets from emergency situations. To this end, we introduce DisRel, a dataset which contains 4, 600 multimodal tweets, collected during the disasters that hit the USA in 2017, and manually annotated with coherence image-text relationships, such as *Similar* and *Complementary*. We explore multiple models to detect these relationships and perform a comprehensive analysis into the robustness of these methods. Based on these models, we build a simple feature augmentation approach that can leverage the text-image relationship. We test our methods on 2 tasks in CrisisMMD: Humanitarian Categories and Damage Assessment, and observe an increase in the performance of the relationship-aware methods. We make our data and code available at https://github.com/tsosea2/DisRel.

Keywords

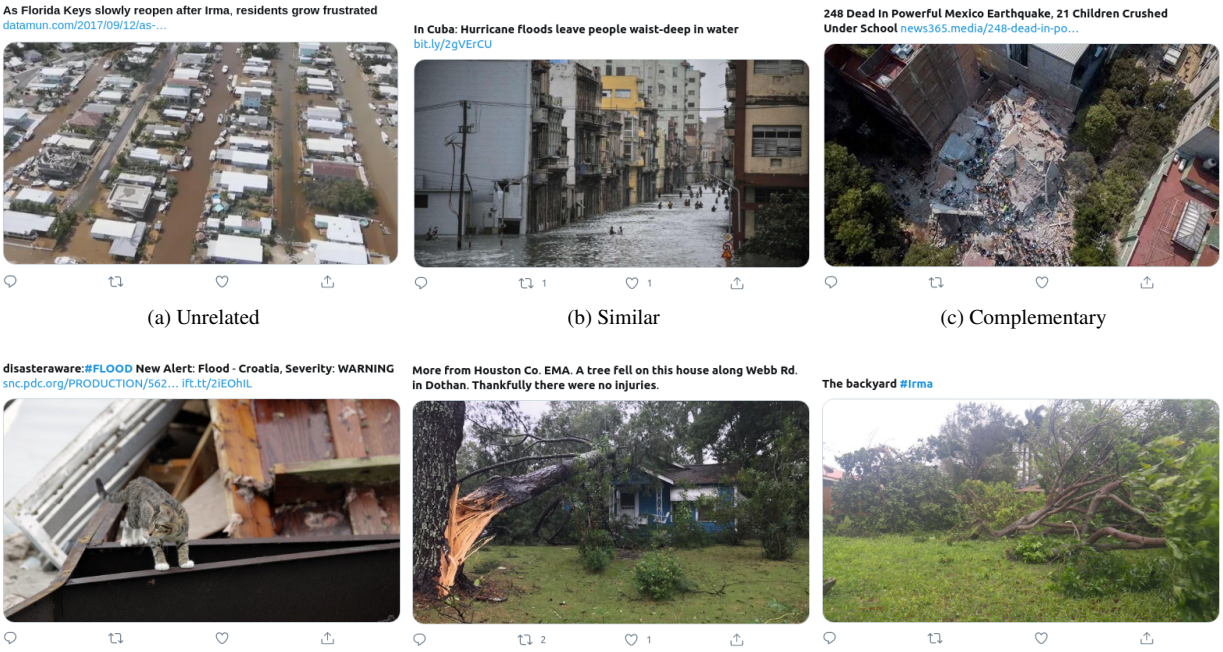
Multimodal disaster tweet classification; image-text coherence relationship prediction; ViLBERT.

INTRODUCTION

According to NOAA National Centers for Environmental Information (NCEI), in 2019 the United States was impacted by 14 natural disasters, while in 2020 there were 22 disasters that produced significant losses and damage. During these times, fast assessment of damaged areas can improve resilience by informing responders and aid agencies about critical areas, guiding the allocation of resources, and improving the real-time response overall. In a recent guide on preliminary damage assessment released in May 2020 (FEMA 2020), the Federal Emergency Management Agency (FEMA) lists several methods for damage assessment, including: self-reporting, door-to-door assessments, windshield surveys, geospatial analysis, remote sensing, and predictive modeling. While very precise, self-reporting and door-to-door assessments are generally performed through surveys and interviews, and lag behind the triggering event by days or even weeks, resulting in a slow recovery.

Individuals affected by disasters often turn to social media, such as Twitter, to report information that can be useful for the relief authorities (e.g., road damage, fallen trees, or downed power lines), as shown in Figure 1. Recent studies focus solely on collecting such data, annotating it for a specific task that can potentially speed up the real-time response of the authorities (e.g., damage assessment, informativeness detection), followed by training

*corresponding author



(a) Unrelated (b) Similar (c) Complementary

(d) Unrelated (e) Similar (f) Complementary

Figure 1. Different types of relationships between the image and the tweet.

machine learning classifiers on these tasks (Alam, Ofli, et al. 2018a; Neppalli et al. 2018; C. Caragea et al. 2016; H. Li, Guevara, et al. 2015; Ashktorab et al. 2014). However, none of these studies take into consideration the relationship between the image and text, or the effects of this relation on downstream tasks.

In this paper, we explore the image-text relationship in multimodal tweets collected during natural disasters, and further investigate if we can leverage this relationship to improve the multimodal disaster tweet classification. To this end, we introduce DisRel, a dataset of 4, 600 disaster tweets, annotated with image-text coherence relations such as *Similar* and *Complementary* using the Amazon Mechanical Turk (AMT) crowdsourcing platform. The dataset is composed of tweets collected during various disasters from the year of 2017 and serves not only as a challenging benchmark for computational models, but also as a valuable resource for analyzing the interrelation between the textual and visual modalities. To provide a better understanding of the task, we discuss a few examples, shown in Figure 1. For example, in Figure 1b the text clearly describes the content of the corresponding image; they are *similar* to each other. However, in Figure 1c, the two modalities *complement* each other; both modalities add information to the other. On one hand, the text adds information by specifying the number of casualties, while on the other hand the image also shows the ongoing relief efforts.

For the annotation of the dataset as *Similar* vs. *Complementary*, to ensure high quality annotations, we provide clear definitions and examples for the annotators to carefully understand the task, as discussed in Section Dataset.

Inspired from recent work on cross-modal coherence modeling (Alikhani et al. 2020), which shows impressive improvements in generating captions that take the informational need of a user into account, we present a comprehensive investigation into understanding the relationship between the text and the image using our DisRel dataset. Moreover, we draw ideas from the recent self-supervised multimodal architectures, such as ViLBERT (Lu et al. 2019) and VisualBERT (L. H. Li et al. 2019), which use large amounts of unlabeled data to learn universal, task-agnostic visual-linguistic representations and propose a simple self-supervised relationship-aware architecture based on the ViLBERT model for our disaster context. We show the feasibility of our approach compared to coherence-agnostic methods on CrisisMMD (Alam, Ofli, et al. 2018a), a Twitter multimodal dataset composed of more than 18, 000 image and text tweets collected during natural disasters.

Even though the recent self-supervised multimodal architectures show improved performance compared with existing methods, it still remains unclear if they are robust to adversarial attacks. Moreover, model robustness is particularly important when dealing with the invaluable context of disasters. Therefore, we also perform a comprehensive analysis into the robustness of these self-supervised architectures, and investigate which modality is more vulnerable to adversarial attacks. Interestingly, our experiments show that: **1)** our coherence-aware ViLBERT model obtains the lowest attack success rate compared to other baselines, and **2)** the textual adversarial attacks are

consistently more successful than image attacks. To our knowledge, we are the first to explore model robustness to potential adversarial attacks that may occur in a disaster context.

We summarize our contributions as follows:

- We create and annotate DisRel, a dataset for identifying the relationship (*Similar* or *Complementary*) between the text and image of a disaster tweet.
- We propose a multimodal model based on ViLBERT to identify the text-image relationship using the created dataset.
- We show that the relationship information can be used to improve the performance of multimodal models for disaster tweet classification.
- We investigate common model errors and perform a comprehensive analysis into the robustness of multimodal methods in the disaster domain.

RELATED WORK

Disaster Tweet Classification

Recent research studies have used machine learning and deep learning to show the utility of social media information for disaster management and response teams. For example, the analysis of text data (e.g., tweets) has received significant attention in various recent works (Yin et al. 2012; Guan and C. Chen 2014; Imran, Castillo, et al. 2015; Kryvasheyeu et al. 2016; C. Caragea et al. 2016; H. Li, D. Caragea, et al. 2018; Yuan and Liu 2018; Enenkel et al. 2018). In particular, Kryvasheyeu et al. 2016 used Hurricane Sandy-related tweets, and Enenkel et al. 2018 used Hurricanes Harvey and Irma filtered, geo-located tweets to show that rapid early damage assessment can be facilitated by social media. Similarly, several studies have focused on the analysis of social media images in emergency situations (Lagerstrom et al. 2016; Bica et al. 2017; Alam, Imran, et al. 2017; Nguyen et al. 2017; X. Li, D. Caragea, H. Zhang, et al. 2019; X. Li, D. Caragea, C. Caragea, et al. 2019; Alam, Ofli, et al. 2018b; Chaudhuri and Bose 2020; Weber et al. 2020). Unlike macro-level images, such as satellite images or images taken using drones, social media images provide detailed on-site information from the perspective of the eyewitnesses of the disaster (Bica et al. 2017), and can serve as an ancillary, yet rich source of visual information in disaster damage assessment.

While most of the prior works have focused on either text classification or image classification independently, many tweets posted during disasters contain both text and images, as shown in Figure 1. Generally, the information contained in one modality (e.g., text) can enhance or complement the information contained in the other modality (e.g., images), and together the two modalities (i.e., text and images) can produce more effective models for filtering useful disaster-related information (Imran, Ofli, et al. 2020). This has recently led to multimodal disaster tweet datasets, consisting of both tweet text and images (Alam, Ofli, et al. 2018a; Mouzannar et al. 2018), as well as a surge of studies focused on multimodal models in the disaster space (Mouzannar et al. 2018; Rizk et al. 2019; Gautam et al. 2019; Nalluru et al. 2019; Hao and Wang 2020; Ofli et al. 2020; Agarwal et al. 2020; Abavisani et al. 2020). For example, Mouzannar et al. 2018 developed a multimodal deep learning approach based on text Convolutional Neural Networks (CNN) (Kim 2014) and Inception-v4 networks (Szegedy, Ioffe, et al. 2016) to identify damage related information in social disaster posts. Agarwal et al. 2020 proposed a gated multimodal model based on Recurrent Convolutional Neural Networks (RCNN) for text (Lai et al. 2015) and an Inception-v3 model for images (Szegedy, Vanhoucke, et al. 2016), fused with an attention mechanism (Hua and H.-J. Zhang 2004), and used it on three damage-related tasks introduced by Alam, Ofli, et al. 2018a. Abavisani et al. 2020 proposed a multimodal deep learning framework, which leverages DenseNet (Huang et al. 2017) and BERT (Vaswani et al. 2017) for representing images and text, respectively, equipped with a cross-attention module to filter out irrelevant information from the text and image modalities, and reported state-of-the-art results on the tasks introduced by Alam, Ofli, et al. 2018a. These tasks range from assessing the damage of a disaster to identifying if tweets posted during disasters convey informative messages or not. In this paper, we show that the image-text relationship can improve the performance on these tasks.

Pre-trained Multimodal Models

The Multimodal Bitransformer (MMBT) (Kiela et al. 2019) is a supervised BERT-based model that fuses information from two separate pre-trained text and image encoders. MMBT has obtained very good results on multiple multimodal tasks, hence we consider it as one of our strong baselines. However, the self-supervision

step is carried out separately for each modality. On the another hand, many recent self-supervised multimodal architectures, such as ViLBERT (Lu et al. 2019), VisualBERT (L. H. Li et al. 2019), LXMERT (Tan and Bansal 2019), VL-BERT (Su et al. 2019), Unicoder-VL (G. Li et al. 2020), UNITER (Y.-C. Chen et al. 2019), use large amounts of unlabeled data to learn universal, task-agnostic visual-linguistic representations.

ViLBERT (Lu et al. 2019) is greatly inspired by BERT, and is composed of two parallel BERT-like models that correspond to the text stream and the image stream. The information flows separately through the two models and is fused through a transformer co-attention layer. The co-attention layer is straightforward; after the two transformer blocks compute the query, key and value matrices, the keys and values are exchanged between the two streams. The output consists of two embeddings: a text representation conditioned on the image representation, and an image representation conditioned on the text representation. The pre-training objective of the textual BERT block remains the Masked Language Modeling (Devlin et al. 2018). The image self-supervised training is carried out by masking certain regions extracted by a pre-trained Mask RCNN (He, Gkioxari, et al. 2017) segmentation model. Instead of predicting pixel values, the model learns to predict the class of the masked image region.

LXMERT (Tan and Bansal 2019) uses a two-stream architecture, but employs a multi-component design for the cross-modality model and makes use of additional pre-training tasks.

While ViLBERT employs two different streams for vision and language modalities that can only attend to each other, VisualBERT (L. H. Li et al. 2019) uses a single stream for both modalities. This is achieved by passing both visual and textual embeddings to the multi-layer Transformer, followed by learning alignments between the input text and image regions through a self-attention mechanism.

VL-BERT (Su et al. 2019) makes use of a single-stream unified architecture. However, it differentiates itself from other works by introducing additional pre-training on text-only datasets to improve generalization for long sentences. Moreover, the model does not freeze the weights of the upstream image segmentation network, and further trains it during the self-supervision process.

In this paper, we use as our base models the MMBT model due to its simplicity and the ViLBERT model to explore self-supervised visio-linguistic approaches in our disaster domain.

Text-Image Relationship Datasets

The growth of the social networks focused on text and image sharing (e.g., Twitter, Instagram) has spurred a lot of research into capturing or making use of the coherence relations between images and text. The relationship between these two modalities may be useful in many situations. For example, a classifier that can predict if an image-text pair are closely related is useful for building datasets for various computer vision tasks (e.g. image captioning). Several studies have looked into these relationships from different perspectives. For example, Vempala and Preoţiuc-Pietro 2019 define a way of categorizing the relationship between the images and text of Twitter posts. They investigate whether the text is represented in the image or not and whether an image adds new information to the meaning of the text or not. However, their proposed approach involves using different models for computing image and text representations that are later concatenated. A better fusion technique could be ensured by using models that use joint representations of image and natural language content as proposed in our paper.

Alikhani et al. 2020 also create a coherence-relation annotation protocol for image-caption pairs which they call CLUE. They introduce multiple relations between image-text pairs: Visible, Subjective, Action, Story, and Meta. By allowing overlapping labels, they ensure a very fine-grained classification of text-image pairs. However, this approach was applied for images and their captions automatically extracted from the Web, which already involves a tight connection between the text and the image. On the other hand, social media posts can contain significantly weaker connections between the modalities. For example, when people post images on social media, the text of these images does not necessarily describe the image; the text can be a quote, or can simply give background knowledge for the photo.

Kruk et al. 2019 use Instagram posts to compute the author intent from multimodal data. To this end, they introduce three taxonomies: *Intent taxonomy*, which divides the image-text pairs into eight classes by the speaker's intent: advocative, promotive, exhibitionist, expressive, informative, enterntainment, provocative/discrimination, and provocative/controversial; *Contextual taxonomy*, which categorizes the relationship between the meaning of the image and text into Minimal, Close or Transcendent; *Semiotic taxonomy* captures the relationship between what is signified by the two modalities, which can be Divergent, Additive or Parallel.

Despite all the progress made in studying text-image relationships, the existing taxonomies are not useful in the particular scenario of natural disasters. Our experiments show that these general purpose datasets add no improvement to our domain, reinforcing the importance of DisRel, which enables the exploration of relationship-aware methods in the domain of natural disasters.

(a) View from satellite (b) Conference (c) Chart

Figure 2. Several representative examples of irrelevant tweets.

(a) Similar vs. Complementary (b) Complementary vs. Unrelated

Figure 3. Some representative examples of annotations that did not reach agreement.

DATASET

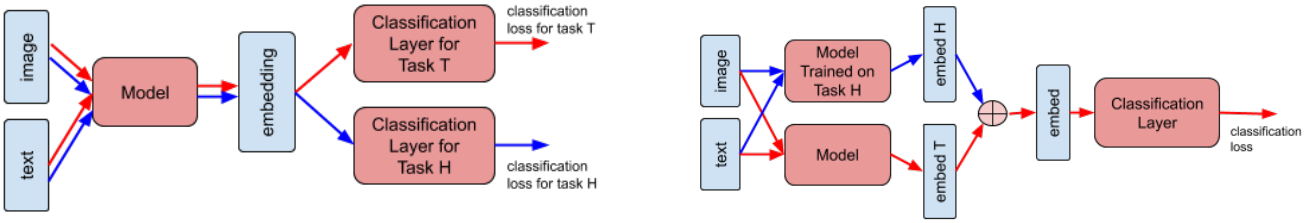
We collect data from the disasters that hit the USA in 2017: Hurricanes Maria, Irma, Harvey, the Mexico Earthquake, and the California Wildfires. We crawl tweets through the Twitter streaming API using keywords such as *#hurricane*, *#Harvey*, etc., and use several methods to clean and filter out duplicates. We filter out retweets (tweets with the "RT" token), normalize the texts, and remove the duplicates from our collection, as well as tweets with more than one image. This process produced 122K tweets with one image and text. A manual inspection of the data revealed that these data are noisy and most of the time irrelevant to disaster response. For example, a satellite photo of a hurricane does not offer information about its impact on the ground, nor does a chart or a meme. We show examples of different types of irrelevant images in Figure 2. In this paper, our main goal is to use the image-text coherence relationships to improve the multimodal disaster tweet classification for a better resource allocation at the time of the disaster. Therefore, we employ another filtering process to remove the noisy and irrelevant image-tweet pairs using classifiers trained on the informative/not-informative classes from CrisisMMD (Alam, Ofli, et al. 2018a). Specifically, we employ two separate classifiers based on the textual and image modalities. For text, we use the BERT (Devlin et al. 2019) base model, while for the image modality we use a ResNet-50 (He, X. Zhang, et al. 2016) model. Both these methods achieve 90% F-1 score on CrisisMMD, and use them as follows: if either the text or the image is classfied in the informative class by one of the classifiers, then we consider it in our next steps. After filtering, we obtain 16,000 examples, which we use to create DISREL.

Text-Image Relationship Types

By manually analyzing the image-text pairs, we derived a taxonomy that captures three broad types of relationships between the modalities relevant for disaster situations. We enumerate these relationships below, and provide some examples in Figure 1:

Unrelated The image and the text do not share any information. Moreover, one modality does not influence the understanding of the other modality.

For instance, in Figure 1a, the image and the text convey unrelated information. On one hand, the text refers to frustrated residents and reopening after Irma, whereas the image depicts a flooded area, with no sign of residents.



(a) Multitasking Architecture: the red and the blue paths are used alternatively during training. The model used for embedding is either ViLBERT or MMBT.

(b) Feature Augmentation Architecture: only the red path is used during backpropagation. The models used to produce the H and T embeddings are both either ViLBERT or MMBT.

Figure 4. Model Architectures for (a) image-text relationship detection using multitask learning; and (b) disaster tweet classification using feature augmentation based on text-image relationship.

Similar In the early phases of this study, we defined the *Similar* relationship as *the text and image provide the exact same information* (i.e., the text is a caption of the image). For example, both the tweet text and image from Figure 1b show floods that *leave people waist-deep in water*. However, most of the times the tweet text does not precisely describe the image. Therefore, we extended the definition of the *Similar* category to also cover the cases when one modality is partially covered by the other, as long as they have the same focus and the shared information is precise. For instance, in Figure 1e, the text accurately describes a tree that fell on a house, just as in the image, even though it offers the extra information of it causing no injuries. Our annotation process only considers the relaxed version of the *similar* class.

Complementary. Another possible relationship between an image and its corresponding tweet is having two modalities that do not share the same focus, but their information is still related (i.e., one modality helps better understand the other or complement each other). For example, in Figure 1c the two modalities are related as both present information about a school downed by the earthquake. However, the text focuses on the number of victims (not covered in the image), whereas the image illustrates the relief efforts being made (not covered in the text). In this scenario we say the two modalities *complement* each other. There is also the case when one modality offers little information, although it is related to the other as shown in Figure 1f. We consider this situation as *Complementary* as well, rather than *Similar*, as the information shared by both modalities is minimal.

Annotation Process

In order to create a labeled dataset for the aforementioned classes, we uniformly sampled 5,000 image-text pairs from our informative image-tweet pairs, filtered as explained above, and labeled each of them using four different annotators on the Amazon Mechanical Turk crowd-sourcing platform (AMT). We design the AMT form to contain one tweet per page, and the annotators are asked to choose between one of the three labels presented earlier. We take several steps to ensure quality control measures and exclude spurious annotations. First, we compute a trust score for each annotator as the mean number of annotators agreed with. As there are four annotators in total, one annotator can agree with at most three other annotators for a tweet. Therefore, the annotator trust scores lie between 0 and 3. Next, we rejected annotators with trust scores smaller than 1.5, as well as annotators with less than 10 annotated examples. The filtering process ended up removing 55% of the annotators and 24.5% of the annotations, hence each tweet has on average 3 annotations. To assess the agreement between our annotators, we compute the Krippendorff Alpha metric, and obtain an $\alpha = 0.58$. The final label is computed by considering the majority vote of the remaining annotators. The resulted label distribution is unbalanced. We obtained 3135 *similar* examples, 1451 *complementary*, 184 *unrelated*, and for 223 examples an agreement could not be established due to ties between two labels; hence these examples were dropped. To provide additional insights into why agreement was not achieved for some of these tweets, we present some examples in Figure 3. First, we observe in Figure 3a an example where in spite that the image and text have the same focus, one might argue that the text does not offer enough details. Second, Figure 3b shows a tie between the *Complementary* and *Unrelated* classes. Although the image and text might seem unrelated, the text weakly refers to the damage caused by the hurricane, which is reflected in the image.

After computing the final labels, we found that the unrelated class contains less than 4% of the examples, which indicates that informative disaster tweets text and image are rarely unrelated to each other. Due to the very small numbers of unrelated examples, we did not include the *Unrelated* class for our experiments. In consequence, we obtain a dataset of 4,600 image-text pairs annotated with the *Similar* and *Complementary* classes.

APPROACHES

There are two main tasks that we address in this work. First, we aim to identify the text-image relationship for multimodal disaster tweets. Second, we study the usefulness of the relationship with respect to downstream classification tasks on multimodal disaster tweets. We describe the approaches used for each task in what follows.

Text-Image Relationship Detection in Disaster Tweets

To identify the relationship between the textual and visual modalities in DisRel, we explore several models. ConcatBOW concatenates averaged 300-dimensional GloVe (Pennington et al. 2014) word embeddings with the output of a pre-trained ResNet-152 (He, X. Zhang, et al. 2016) network. ConcatBERT concatenates the BERT (Devlin et al. 2018) embeddings with the output of the same ResNet network. For both these models, we project the concatenated embeddings through a linear layer for classification. Next, we explore 2 state-of-the-art multimodal classification methods. First, we use a Supervised Multimodal Bitransformer (Kiela et al. 2019) (MMBT), which projects information from both modalities through a transformer network. Similar to BERT (Devlin et al. 2019), we use the CLS token representation h_{CLS} and pass it through a linear layer for classification. Second, we apply the ViLBERT (Lu et al. 2019) pre-trained visio-linguistic model to our task. ViLBERT produces 2 vector representations of the image and text modalities: h_{IMG} and h_{CLS}. We take an element-wise product between the h_{IMG} and h_{CLS} embeddings produced by the model, then pass this result through a task-specific linear layer for classification. The model is trained end-to-end.

Finally, we investigate a multi-task learning framework using hard parameter sharing (Caruana 1997) to improve the performance on our DisRel task T of identifying coherence relationships, by using information from a helper task H. We show our approach in Figure 4a. We start from the base ViLBERT or MMBT model, on top of which we add 2 task specific classification layers; one for the helper task H, and one for our target task T. At training time, we alternate between batches from task T and task H. A batch from task T updates the parameters of the task-specific layer of T and the base model (ViLBERT or MMBT), while a batch from task H updates the parameters of the classification layer corresponding to H and the base model as well. For the helper task H, we use the original labels of the task. We denote this method by M-H-MT, where M is the name of the model (MMBT or ViLBERT), and H is the name of the helper task. In our context, we experiment with different helper tasks H. First, we set H as the image-text semantic overlap detection task introduced by Vempala and Preoţiuc-Pietro 2019. Their dataset, which we denote by Relationship (or REL for short) in this paper, is composed of $\sim 4,400$ tweets of users from different socio-demographic groups. The data is multimodal, and is annotated with coherence relations such as *all text content is represented in the image*, or *the image adds some information to the text*. The task is very similar to ours; however, the domains differ substantially. Second, we set H as the coherence detection task from the CLUE (Alikhani et al. 2020) dataset, which contains $10,000$ captioned images annotated with coherence relationships inspired by computational models of discourse. These datasets are presented in detail in Section Text-Image Relationship Datasets.

Relationship-Augmented Disaster Tweet Classification

To investigate if the information about the relationship between the textual and visual modalities (considered to be a helper task H in this context) can improve the performance of classification models in disaster-centric domains (representing the target task T here), we explore several approaches. As our main helper task H for image-text relationship detection, we first experiment with the relationship detection in disaster tweets from DisRel. Next, we use the coherence detection task in the CLUE (Alikhani et al. 2020) dataset as well as the Relationship dataset (Vempala and Preoţiuc-Pietro 2019) as the helper task H to understand how a relationship-aware model designed and trained outside of the disaster domain affects the performance of the models. For the target task T, we model two tasks proposed in the CrisisMMD dataset, Humanitarian Categories and Damage Assessment (Alam, Ofli, et al. 2018a).

Introduced by Alam, Ofli, et al. 2018a, CrisisMMD is a multimodal Twitter Dataset from Natural Disasters. It contains ~$18,000$ tweets with image and text, extracted during the following crisis events: Hurricanes Irma, Harvey and Maria, the Mexico and Iraq-Iran earthquakes, California wildfires, and the Sri Lanka floods. The authors removed duplicates and the dataset contains only English tweets with at least two words. CrisisMMD was manually labeled for three classification tasks.

The first task we consider is to determine what kind of humanitarian information is conveyed in the informative messages and images. To this end, the authors define seven categories: infrastructure and utility damage, vehicle damage, rescue, volunteering or donation effort, injured or dead people, affected individuals, missing or found people, and other relevant information. The second task involves assessing the severity of the physical destruction produced

Table 1. Results on *Similar* and *Complementary* tweet classification. In order from top to bottom: (1) Weak Baselines (Top) - single modality and various combinations of these single modality models (2) Strong Baselines (Middle Block) - the vanilla MMBT and ViLBERT models (3) Multitasking Methods (Lower Block) - the ViLBERT and MMBT learned in a multi-tasking scenario using data from CLUE and the Relationship dataset. We assert significance* if $p < 0.05$ under a paired-t test with the counterpart base model (e.g., ViLBERT-MT vs. ViLBERT).

	P SIM	R SIM	F-1 SIM	P COM	R COM	F-1 COM	MICRO F-1	MACRO F-1	ACC
CONCATBOW	0.74	0.85	0.78	0.50	0.36	0.52	0.70	0.65	0.69
CONCATBERT	0.73	0.83	0.77	0.46	0.33	0.39	0.65	0.58	0.67
MMBT	0.80	0.82	0.81	0.61	0.60	0.60	0.74	0.71	0.75
VILBERT	0.82	0.81	0.82	0.61	0.62	0.61	0.75	0.72	0.75
MMBT-CLUE-MT	0.77	0.75	0.76	0.60	0.61	0.60	0.72	0.69	0.72
MMBT-REL-MT	0.81	0.82	0.81	0.62	0.61	0.61	0.75	0.71	0.75
VILBERT-CLUE-MT	0.78	0.76	0.77	0.60	0.61	0.61	0.73	0.70	0.73
VILBERT-REL-MT	0.82	0.82	0.82	0.63	0.62	0.62	0.76*	0.72	0.76*

Table 2. Results of the best models on CRISISMMD. We use †to denote the micro precision, recall and F1, and ‡to denote the macro precision, recall and F1. We assert significance* if $p < 0.05$ under a paired-t test with the counterpart base model (e.g., ViLBERT-AG vs. ViLBERT).

	HUMANITARIAN						DAMAGE ASSESSMENT					
	P†	P‡	R†	R‡	F1†	F1‡	P†	P‡	R†	R‡	F1†	F1‡
MMBT	0.94	0.92	0.91	0.90	0.92	0.90	0.85	0.83	0.83	0.79	0.84	0.80
MMBT-DR-AG	0.95*	0.93*	0.92	0.90	0.94*	0.92*	0.81	0.79	0.86*	0.83*	0.86*	0.83*
MMBT-CLUE-AG	0.92	0.90	0.88	0.88	0.90	0.89	0.84	0.82	0.78	0.79	0.82	0.81
MMBT-REL-AG	0.90	0.89	0.88	0.88	0.89	0.88	0.83	0.82	0.79	0.78	0.81	0.81
VILBERT	0.95	0.93	0.91	0.90	0.93	0.91	0.86	0.85	0.83	0.81	0.84	0.82
VILBERT-DR-AG	0.95	0.94	0.93*	0.94*	0.94	0.94*	0.88*	0.87*	0.86*	0.82	0.87*	0.84*
VILBERT-CLUE-AG	0.90	0.88	0.89	0.86	0.89	0.87	0.83	0.82	0.78	0.78	0.81	0.80
VILBERT-REL-AG	0.92	0.90	0.89	0.88	0.90	0.88	0.83	0.81	0.79	0.79	0.81	0.80

by a disaster. The objective is to classify the type of damage into *severe*, *mild*, or *little to no damage*. However, only the images from the tweets are annotated with this type of damage. We postulate that a relationship-aware multimodal model can leverage the text as well and may improve the performance on the damage assessment task. To this end, we perform a comprehensive set of experiments on this task, and consider the labels of the images to apply to the text-image pairs as well.

In the original version of the CrisisMMD dataset, the two modalities are labeled independently. However, in order to enable multimodal exploration on the humanitarian task, we only consider the image-text pairs where the two modalities share the same label. For humanitarian task, we obtain 3,000 informative tweets split into three categories: infrastructure and utility damage (20%), rescue, volunteering or donation effort (30%) and other relevant information (50%). The other categories were disregarded as they had too few representatives. For the second task, we adopt a binary classification scenario by grouping the *mild* and *severe* categories in a *damage* class, and all the *little to no damage* examples in a *no damage* class. We obtain 6,000 tweets, 70% labeled as *damage* and 30% as *no damage*.

To have a point of reference for the classification performance of computational models on the two CrisisMMD tasks, we first apply our vanilla MMBT and ViLBERT models on each of the two tasks. Next, we explore a feature augmentation approach, which uses information produced by our image-text relationship detection task, assumed to be a helper task H, to improve the performance of the two CrisisMMD target tasks T. First, we train our vanilla MMBT and ViLBERT models on task H. We use these trained classifiers to produce embeddings for the image-text pairs from task T. Following the same notation as in the multi-tasking section, for MMBT, the resulted embedding is the h^H_{CLS} embedding produced by the model, while for ViLBERT, we use the textual embedding h^H_{CLS} and the visual embedding h^H_{IMG}. Finally, to produce our relationship aware representation, we employ an element-wise addition with the embeddings produced on task T:

$$E_{MMBT} = h^H_{CLS} \oplus h^T_{CLS} \quad (1)$$

$$E_{ViLB} = (h^H_{CLS} \otimes h^H_{IMG}) \oplus (h^T_{CLS} \otimes h^T_{IMG}) \quad (2)$$

where $\otimes$ is the element-wise multiplication, and $\oplus$ is the element-wise addition. These embeddings are then projected through a linear layer for classification. We denote these models by M-H-AG, where M is the name of the model (ViLBERT or MMBT), and H is the name of the helper task (CLUE, REL or DisRel). We show our approach in Figure 4b.

EXPERIMENTS AND RESULTS

In this section, we present our experimental setting, then investigate the performance on two target tasks: First, we analyze the models trained to predict the text-image relationship in DisRel. Then, we explore if using the relationship between the text and image improves the downstream classification performance.

Experimental Setting

We use pre-trained 300 dimensional GloVe embeddings for the BOW approaches, and use the BERT-base model (Devlin et al. 2018) trained on Wikipedia and Bookcorpus. For both tasks, we follow the best reported hyper-parameters for MMBT (Kiela et al. 2019), freeze 3 transformer layers for the image modality, and use a batch size of 16. For ViLBERT (Lu et al. 2019), we follow the initial values used by the authors for the *Caption-Based Image Retrieval* task, then vary the learning rate in steps of $1e^{-5}$. Due to memory restrictions, we use a batch size of 8. In order to obtain statistically significant results, we repeat the experiments 10 times and report the average results. We perform all the experiments on an Nvidia V100 GPU.

To spur further research on the exploration of coherence relations in the disaster domain, we create a 60/20/20 train, development, and test split for DisRel, which we will make available alongside our data.

Predicting the Relationship between the Text and Image of a Disaster Tweet

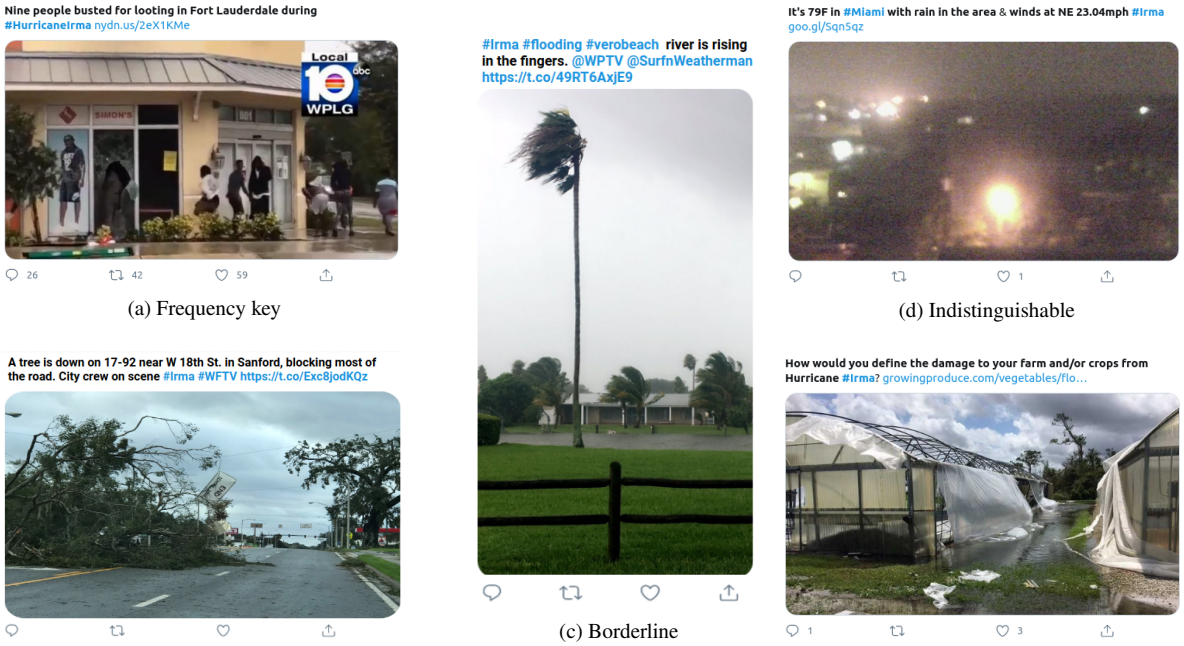
We evaluate the performance of the presented methods on DisRel, and show the results in Table 1. First, the ConcatBERT and ConcatBOW models, which use concatenated embeddings from two separate networks, perform significantly worse than the other models, which jointly learn image and text representation during the training time. Surprisingly, the ConcatBOW model greatly outperforms the ConcatBERT model by as much as 5% in micro F1, even though it is much simpler. ViLBERT obtains a slight improvement over the MMBT model, which shows the benefit of training visio-linguistic representations during the self-training step. The ViLBERT-REL-MT model however, manages to outperform all the other methods, and obtains a 1% improvement in F1 and accuracy compared to the ViLBERT model. Interestingly, even though the domain of the multi-tasking helper task H for ViLBERT-REL-MT differs substantially from our task, this approach still manages to improve the performance of the task. On the other hand, the ViLBERT-CLUE-MT model sees a decrease in performance compared to the base ViLBERT model. We assume this is due to the fact that the Relationship task is more similar to ours. The Relationship dataset is annotated with image-text relations such as *the image adds some information to the text* which can be directly mapped to our *Similar* and *Complementary* classes. On the other hand, the CLUE task can have labels which conflict with our task. For example, the *Meta* class, which employs that the text allows the reader to draw inferences for both the presented scene, as well as the production of the image, can belong to either the *Similar* or *Complementary* classes.

Using the Text-Image Relationship to Improve Disaster Tweet Classification

We show the results of our experiments on the two CrisisMMD tasks in Table 2 and observe a few patterns. **Models augmented with features from the disaster domain consistently outperform methods that use features from other domains**. Moreover, these models manage to leverage the relationship between the text and image to improve the downstream performance. The ViLBERT-DR-AG, which uses features learned on the DisRel dataset, outperforms the task-agnostic VilBERT model by 3% in macro F1 on the humanitarian task, and 3% in micro F1 on the damage assessment task. The MMBT-DR-AG also sees an improvement on both tasks. On the other hand, the ViLBERT-CLUE-AG and ViLBERT-REL-AG models, which leverage features from disaster unrelated domains see a 5% drop in micro F1 on the Humanitarian task, and a dramatic 6% drop in macro F1 on the Damage Assessment task. These results show the importance of our dataset, DisRel, which provides an opportunity for the exploration of coherence-aware methods in disaster-centric domains.

Table 3. Attack success rates on perturbations applied on images and text.

	Image Perturbation				Text Perturbation		
	S & P	Gaussian	Speckle	Poisson	Random	List	Gs-Gr
concatbert	5%	28%	29%	10%	13%	15%	94%
mmbt	1%	2%	2%	1%	9%	12%	90%
vilbert	1%	1%	1%	1%	8%	9%	85%



(a) Frequency key (b) Named entity (c) Borderline (d) Indistinguishable (e) Syntactic

Figure 5. Some errors of our models.

Robustness

We explore the robustness of our models under textual and image perturbations. We ask the following questions: Is our ViLBERT model more robust compared to other baselines? Are text perturbations more successful than image perturbations? We investigate various methods to generate adversarial examples (which we call *attacks)*. The purpose of an attack is to make a correct model prediction on the **test set** to be incorrect by adding a small perturbation to the input. We experiment with our two best models MMBT and ViLBERT, as well as a weak baseline which uses both modalities: ConcatBERT.

Image Perturbation. Our image perturbations are straightforward: we add different types of noise to the images. We experiment with Gaussian, Salt and Pepper, Speckle and Poisson noise.

Textual Perturbation. We define a textual attack as changing only **one** word from the input sentence: **1) Random** replaces one word from the input sentence with a random word from the vocabulary. The results reported are the average of $1,000$ different runs. **2) List** (Alzantot et al. 2018) iteratively replaces each word in a sentence with another word extracted from a **list** of semantically similar words. A successful attack is reported if one of the replaced synonyms leads to an incorrect prediction. **3) GS-GR** (greedy select + greedy replace) (Yang et al. 2018) first finds the weak spot in an input sentence: it iteratively zeroes out each word in the input sentence and analyzes how the model output probabilities change. Next, it picks the word that produces the highest change in probabilities as the weak spot, and replaces it with a random word from the vocabulary. This process is repeated $1,000$ times, and a successful attack is reported if one of the words manages to change the output of the model.

We denote by *success rate* the percentage of correct predictions on the test set that become incorrect under the perturbation, and report our results in Table 3. Models that sustain lower attack success rates are more robust in the face of adversarial attacks.

First, we observe from the table that ViLBERT is more robust under all attacks, consistently outperforming the other models. ConcatBERT is the most vulnerable model, especially on image perturbations such as Gaussian or

Speckle noise. Interestingly, the textual perturbations are significantly more successful than image attacks on the best performing MMBT and ViLBERT models.

ERROR ANALYSIS

Common Errors To get an insight into the drawbacks of our methods, we perform a comprehensive error analysis of common errors on the *Similar* and *Complementary* detection task. Understanding these errors is paramount and can help create better coherence-aware models on different tasks. Therefore, we sample 100 test errors of our best performing ViLBERT-CMMD-MT model on DisRel, and manually group them into different error categories: low frequency key textual terms (39%), textual named entity presence (28%), borderline (9%), hard-to-distinguish image (9%), lexical and syntactic cues (8%), other reasons (7%).

Low Frequency Key Textual Terms We observe a large number of errors when the textual modality contains a rare keyword. For instance, in Figure 5a, the word *looting* appears only two times in the training set, and is a key information that determines the label of the example.

Textual Named Entity Presence We found that numerous mislabeled examples contain named entities. For instance, in Figure 5b, even though the tweet text clearly describes the image, we believe that the named entity *17-92 near W 18th St. in Sanford* manages to mislead the model.

Borderline There are a few examples where the difference between the *Similar* and *Complementary* classes is minor. For instance, in Figure 5c, even though the text describes the image well, one can argue that the focus of the image is not the rising river, but the strong winds.

Hard-to-distinguish Images There are quite a few indistinguishable images that manage to mislead the models. For instance, in Figure 5d, even though the example is clearly in the *Complementary* class, the model mislabels it as *Similar*.

Lexical and Syntactic Cues As ViLBERT is a BERT-based model, the processing of the textual modality still relies on surface-level lexical features. We found that errors sometime arise from misspelled words, or punctuation that changes the textual meaning. We show an example of this phenomenon in Figure 5e. Although the text information can be found in the image, the question mark completely changes the label of this example to the *Complementary* class. This could be seen as a type of attack at character level, i.e., with minor changes to the data, the model is faked.

We observe the same pattern as in Subsection Robustness; a significant amount of the errors produced by our model originate from anomalies in the textual modality. This error analysis opens up directions for further investigation to improve models' robustness to address adversarial types of attacks that could occur in disasters contexts.

CONCLUSION

We introduced DisRel, a disaster-centric dataset for detecting coherence relations such as the *Similar* and *Complementary* classes. Composed of 4,600 image-tweet pairs from the disasters that hit the USA in 2017, we show that this dataset can be used to improve the performance on other disaster related tasks, and propose a simple approach to learn relationship-aware multimodal models on CrisisMMD (Alam, Ofli, et al. 2018a). Therefore, we believe our data provides an invaluable context and can help relief authorities better assess the damage produced by disasters. In the future, we plan to leverage the large amount of data that circulates the social media around the time of these disasters. Specifically, even though we collected 122K tweets, due to cost restrictions, we only annotated about 4% of these. Hence, we will investigate semi-supervised learning techniques to use the large amount of unlabeled data to improve our classification task.

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., 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. (2018a). "CrisisMMD: Multimodal Twitter Datasets from Natural Disasters". In: *Proc. of the International AAAI Conference on Web and Social Media*. ICWSM. Stanford, California, USA.

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.

Alikhani, M., Sharma, P., Li, S., Soricut, R., and Stone, M. (2020). "Clue: Cross-modal Coherence Modeling for Caption Generation". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6525–6535.

Alzantot, M., Sharma, Y., Elgohary, A., Ho, B.-J., Srivastava, M., and Chang, K.-W. (Oct. 2018). "Generating Natural Language Adversarial Examples". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 2890–2896.

Ashktorab, Z., Brown, C., Nandi, M., and Culotta, A. (2014). "Tweedr: Mining twitter to inform disaster response". In: *Proc. of ISCRAM*.

Bica, M., Palen, L., and Bopp, C. (2017). "Visual Representations of Disaster". In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '17. Portland, Oregon, USA, pp. 1262–1276.

Caragea, C., Silvescu, A., and Tapia, A. H. (2016). "Identifying Informative Messages in Disasters using Convolutional Neural Networks". In: *13th Proceedings of the International Conference on Information Systems for Crisis Response and Management, Rio de Janeiro, Brasil, May 22-25, 2016*. Ed. by A. H. Tapia, P. Antunes, V. A. Bañuls, K. A. Moore, and J. P. de Albuquerque. ISCRAM Association.

Caruana, R. (1997). "Multitask learning". In: *Machine learning* 28.1, pp. 41–75.

Chaudhuri, N. and Bose, I. (2020). "Exploring the role of deep neural networks for post-disaster decision support". In: *Decision Support Systems* 130, p. 113234.

Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. (2019). "Uniter: Learning universal image-text representations". In: *arXiv preprint arXiv:1909.11740*.

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*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.

Enenkel, M., Saenz, S. M., Dookie, D. S., Braman, L., Obradovich, N., and Kryvasheyeu, Y. (2018). "Social Media Data Analysis and Feedback for Advanced Disaster Risk Management". In: *Social Web in Emergency and Disaster Management*.

FEMA (2020). *FEMA Preliminary Damage Assessment Guide*.

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.

Guan, X. and Chen, C. (2014). "Using social media data to understand and assess disasters". In: *Natural hazards* 74.2, pp. 837–850.

Hao, H. and Wang, Y. (2020). "Leveraging Multimodal Social Media Data for Rapid Disaster Damage Assessment". In: *International Journal of Disaster Risk Reduction*, p. 101760.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). "Mask r-cnn". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969.

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.

Hua, X.-S. and Zhang, H.-J. (2004). "An attention-based decision fusion scheme for multimedia information retrieval". In: *Pacific-Rim Conference on Multimedia*. Springer, pp. 1001–1010.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708.

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

Imran, M., Ofli, F., Caragea, D., and Torralba, A. (2020). *Using AI and Social Media Multimodal Content for Disaster Response and Management: Opportunities, Challenges, and Future Directions*.

Kiela, D., Bhooshan, S., Firooz, H., and Testuggine, D. (2019). "Supervised multimodal bitransformers for classifying images and text". In: *arXiv preprint arXiv:1909.02950*.

Kim, Y. (2014). "Convolutional neural networks for sentence classification". In: *arXiv preprint arXiv:1408.5882*.

Kruk, J., Lubin, J., Sikka, K., Lin, X., Jurafsky, D., and Divakaran, A. (2019). "Integrating text and image: determining multimodal document intent in instagram posts". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 4622–4632.

Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., and Cebrian, M. (2016). "Rapid assessment of disaster damage using social media activity". In: *Science advances* 2.3.

Lagerstrom, R., Arzhaeva, Y., Szul, P., Obst, O., Power, R., Robinson, B., and Bednarz, T. (2016). "Image Classification to Support Emergency Situation Awareness". In: *Frontiers in Robotics and AI* 3, p. 54.

Lai, S., Xu, L., Liu, K., and Zhao, J. (2015). "Recurrent convolutional neural networks for text classification". In: *Twenty-ninth AAAI conference on artificial intelligence*.

Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D., and Zhou, M. (2020). "Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training." In: *AAAI*, pp. 11336–11344.

Li, H., Caragea, D., Caragea, C., and Herndon, N. (2018). "Disaster response aided by tweet classification with a domain adaptation approach". In: *Journal of Contingencies and Crisis Management* 26.1, pp. 16–27.

Li, H., Guevara, N., Herndon, N., Caragea, D., Neppalli, K., Caragea, C., Squicciarini, A. C., and Tapia, A. H. (2015). "Twitter Mining for Disaster Response: A Domain Adaptation Approach". In: *12th Proceedings of the International Conference on Information Systems for Crisis Response and Management, Krystiansand, Norway, May 24-27, 2015*. Ed. by L. Palen, M. Büscher, T. Comes, and A. L. Hughes. ISCRAM Association.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. (2019). "Visualbert: A simple and performant baseline for vision and language". In: *arXiv preprint arXiv:1908.03557*.

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 2019)*. Valencia, Spain.

Li, X., Caragea, D., Zhang, H., and Imran, M. (2019). "Localizing and quantifying infrastructure damage using class activation mapping approaches". In: *Social Network Analysis and Mining* 9.1, p. 44.

Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks". In: *Advances in Neural Information Processing Systems*, pp. 13–23.

Mouzannar, H., Rizk, Y., and Awad, M. (2018). "Damage Identification in Social Media Posts using Multimodal Deep Learning". In: *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2018)*. Rochester, NY.

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, Rochester, NY, USA, May 20-23, 2018*. Ed. by K. Boersma and B. M. Tomaszewski. ISCRAM Association.

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.

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*.

Pennington, J., Socher, R., and Manning, C. D. (2014). "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Rizk, Y., Jomaa, H. S., Awad, M., and Castillo, C. (2019). "A computationally efficient multi-modal classification approach of disaster-related Twitter images". In: *Proceedings of the 34th ACM/SIGAPP symposium on applied computing*, pp. 2050–2059.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. (2019). "Vl-bert: Pre-training of generic visual-linguistic representations". In: *arXiv preprint arXiv:1908.08530*.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. (2016). "Inception-v4, inception-resnet and the impact of residual connections on learning". In: *arXiv preprint arXiv:1602.07261*.

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, H. and Bansal, M. (2019). "Lxmert: Learning cross-modality encoder representations from transformers". In: *arXiv preprint arXiv:1908.07490*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). "Attention is all you need". In: *Advances in neural information processing systems*, pp. 5998–6008.

Vempala, A. and Preoţiuc-Pietro, D. (July 2019). "Categorizing and Inferring the Relationship between the Text and Image of Twitter Posts". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 2830–2840.

Weber, E., Marzo, N., Papadopoulos, D. P., Biswas, A., Lapedriza, A., Ofli, F., Imran, M., and Torralba, A. (2020). "Detecting natural disasters, damage, and incidents in the wild". In: *arXiv preprint arXiv:2008.09188*.

Yang, P., Chen, J., Hsieh, C.-J., Wang, J.-L., and Jordan, M. I. (2018). "Greedy Attack and Gumbel Attack: Generating Adversarial Examples for Discrete Data". In: *CoRR* abs/1805.12316. arXiv: 1805.12316.

Yin, J., Lampert, A., Cameron, M., Robinson, B., and Power, R. (2012). "Using social media to enhance emergency situation awareness". In: *IEEE intelligent systems* 6, pp. 52–59.

Yuan, F. and Liu, R. (2018). "Feasibility study of using crowdsourcing to identify critical affected areas for rapid damage assessment: Hurricane Matthew case study". In: *International Journal of Disaster Risk Reduction*.