Enhancing Emergency Post Classification through Image Information Amplification via Large Language Models

Pablo Giaccaglia*
Politecnico di Milano, DEIB
pablo.giaccaglia@mail.polimi.it

Carlo A. Bono*
Politecnico di Milano, DEIB
carlo.bono@polimi.it

Barbara Pernici
Politecnico di Milano, DEIB
barbara.pernici@polimi.it

ABSTRACT

Real-time information extracted from social media platforms can be highly valuable during emergencies. For example, reports and direct witnesses can help build situational awareness in the early phases of an emergency, with the potential to save lives. However, suitable techniques for selecting relevant data are needed to gather this information from large-scale social media streams and utilize it effectively. Given the multimedia nature of these streams, selection techniques should simultaneously understand textual and image information, as previous studies highlighted. Leveraging recent advances in language and vision models, we propose and evaluate a method working with a homogeneous, text-only representation for the different modalities of social media posts. Experiments on established and novel datasets, including video data, show that the proposed method achieves state-of-the-art performances while providing a highly general and plug-and-play approach to multimodal data filtering.

Keywords

social media, multimodal data filtering, LLMs

INTRODUCTION

Nowadays, the widespread availability of portable devices and mobile connections has made information production and sharing ubiquitous, making social media a valuable source of information for various purposes, from staying informed to gathering a general overview of a phenomenon. Social media can enhance situational awareness during crisis events by highlighting critical situations, damage to things or people, urgent needs, and other information useful for managing response operations, potentially saving lives. This information often comes in the form of multimedia, such as images and videos. However, the large volume of items posted during crisis events requires proper resources and technologies to rapidly isolate and utilize this information, since their amount is exceptionally high in the first hours following the onset of an emergency.

Several past works investigated the usefulness of social media data for obtaining actionable information during an emergency. The use of social media for collecting up-to-date descriptions during emergencies has been widely advocated over the last decade (Stollberg and De Groeve, 2012), often utilizing a "human as a sensor" paradigm (Avvenuti et al., 2016). Many studies aim to classify social media content according to different categorizations of interest, leveraging various forms of data, including text and multimedia content such as images and videos, to obtain a timely description of ongoing events. Classifying which content is relevant for a given categorization is perceived as a principal barrier to exploiting social media data (Stieglitz et al., 2018).

*corresponding authors

In the present study, we aim to contribute to the task of understanding the informativeness of social media content. Specifically, we look at how augmenting multimodal social media posts with text descriptions derived from Large Language Models (LLMs) might improve their informativeness. We derive textual descriptions from photographic scenes using automatic text generation and combine this additional information with the original text of the posts. Our methodology is tested against two relevant scenarios: a well-known multimodal Twitter dataset (Alam et al., 2018b) and a manually labeled dataset of videos acquired from Reddit. We compare our approach with existing studies and conduct an explainability analysis to understand the nature of the additional information introduced.

The remainder of this work is structured as follows. We summarize literature relevant to our research in the *Related Work* section. Our proposed approach for classifying social media posts is detailed in the *Method* section. In the homonymous sections, we provide the details of the *Datasets* utilized, the *Experiments* performed, and the *Results* obtained by the empirical evaluation of the method. The findings are then examined in the *Discussion* section. Finally, *Conclusions* are drawn in the last section, together with future directions.

RELATED WORK

The prevalent approaches for classifying disaster-related social media posts are grounded in supervised learning. One of the classic works in literature (Sakaki et al., 2013) proposes an algorithm for monitoring tweets to detect earthquake events, focusing on the timeliness of the detection. Over time, approaches based on deep neural networks gained attention (Wiegmann et al., 2020), and production-level platforms for the automatic extraction of relevant posts have been proposed (Imran et al., 2014; Havas et al., 2017). The ability of supervised methods to generalize, especially among different types of events, is one of the critical obstacles, highlighting the need for general-purpose approaches, such as few-shot models (Wiegmann et al., 2020; Kruspe et al., 2020). This challenge also holds for works related to classifying image data or image and text data together (Imran et al., 2020). Recent technological advances have widely upgraded the discerning of both textual and image data. In the context of emergency management, Convolutional Neural Networks (CNNs) have been employed for content classification and damage assessment. The relevance of images during crisis events has been highlighted and exploited by many studies (Peters and De Albuquerque, 2015; Alam et al., 2018a; Bono et al., 2022; Nguyen et al., 2017).

Multimodal fusion techniques aim to enhance classification performance by combining information from both visual and textual modalities (Ramachandram and Taylor, 2017). In recent years, these approaches have been broadly evaluated in emergency scenarios (Abavisani et al., 2020). The integration of modalities can be categorized into two primary strategies. The first and most prevalent involves using distinct backbone networks for each modality, with multimodal fusion occurring either at the classification score level (Wang et al., 2021) or through the combination of high-level features from each network (Kiela et al., 2018; Mozannar et al., 2018), using techniques like addition, outer product, cross-gating, and tensor fusion. The primary limitation of this strategy is the minimal interaction between the different modalities. The second strategy focuses on merging the mid-level features of each modality for a more detailed inter-modal interaction (Huang et al., 2020). Notably, recent multimodal BERT models incorporate transformer layers over these mid-level features, enhancing a fine-grained interplay between text and image features through the attention mechanism (Li et al., 2019). Finally, Liang et al., 2022 propose a method for adapting pre-trained unimodal models like DenseNet and BERT to process image-text pairs for multimodal recognition by integrating cross-modal features without altering the original network structures.

Our study takes a distinct approach from previous ones. Rather than customizing an architecture for combining multimodal data, we convert image information into text with a recent large multimodal model. We then use text-only representations to train a transformer-based classifier. Furthermore, our method is tested with both image and video content, applying it to practical scenarios and including a type of media not frequently considered in similar studies.

METHOD

In the present work, we investigate whether automatically derived textual content, such as captions and descriptions obtained from LLMs, can support multimedia classification and improve the ability to distinguish informative content published on social media. We utilize Large Language and Vision Assistant (LLaVA) (H. Liu et al., 2023), a state-of-the-art large multimodal model, for generating image-related descriptions, and Robustly Optimized BERT Approach (RoBERTa) (Y. Liu et al., 2019) for training the classifier on the resulting representation. We focus on a binary classification setup, aiming at differentiating "*informative*" and "*not informative*" content, according to a context-dependent definition of informativeness. The generated text is concatenated to the original text of the post, thus obtaining a homogeneous, augmented textual representation. A RoBERTa classifier is then trained over this representation. The architecture of the proposed method is illustrated in Figure 1.

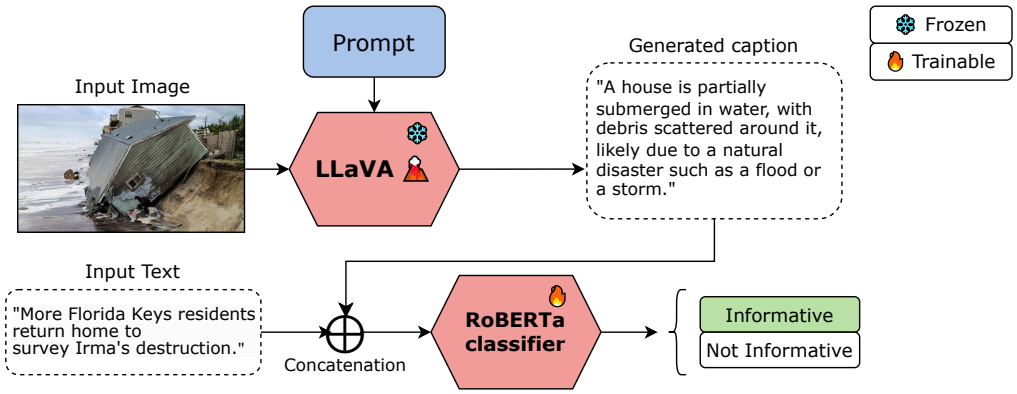

Figure 1. **Architecture of the proposed integration and classification method**

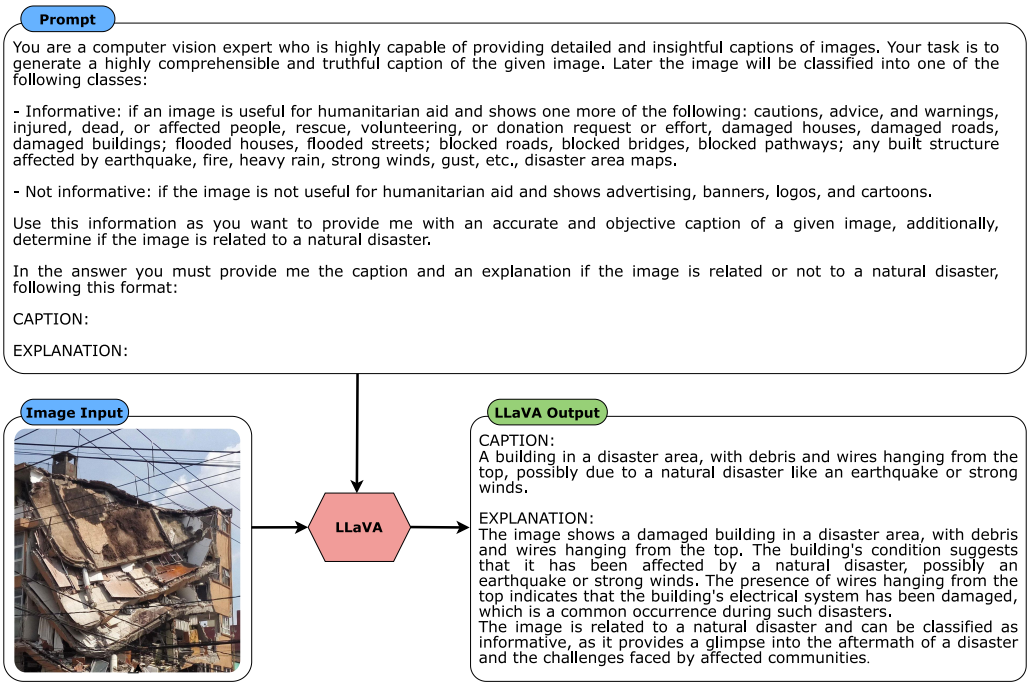

Figure 2. **Information augmentation for the CrisisMMD informativeness task**

As discussed in the previous section, fusion models usually require sophisticated architectures to blend information from text and images, as these data modalities have inherently different features and structures, leading to challenges in model design and optimization. When employing image-only models, the encoded information relies on visual features, mainly disconnected from the semantic features of textual data. The proposed approach enables the use of a text-only classifier, avoiding the need for more complex fusion methods while still achieving state-of-the-art performance.

LLaVA is a large, general-purpose multimodal model combining visual and language understanding that integrates a vision encoder and the Vicuna chat assistant (Zheng et al., 2023). The model is leveraged to generate textual information from images according to a preferred semantic level, such as an image description (image captioning) or task-dependent reasoning. A fundamental feature of the approach is its ability to translate visual features into textual ones, allowing for control over the semantic level and format of the generated content through appropriate prompt engineering.

In the considered setup, multiple text generation tasks are requested, such as visual-only and visual-and-reasoning. The idea is to exploit the language model, through suitable prompts, for different kinds of information amplification, which could in turn improve the downstream classification task. An example is reported in Figure 2, which will be detailed in the Experiments section. Each component of the generated information, derived from pictures or video frames attached to posts, can be regarded as supplementary text, extending the original post's text via concatenation.

Table 1. **Characteristics of the datasets**

	Number of posts	Number of images	*Informative* images (%)
CrisisMMD	11,400	12,708	66.63
Reddit	838	35,551	6.39

Each type of resulting textual representation is then separately fed to a transformer-based model for training and evaluation.

DATASETS

We conducted tests on two distinct datasets to benchmark the effectiveness of the proposed method. The first one, CrisisMMD, is a well-established multimodal dataset collected from Twitter during natural emergencies (Alam et al., 2018b). The second is a novel dataset consisting of video content collected from Reddit over one week at the onset of the 2023 conflict between Israel and Hamas.

CrisisMMD The dataset was collected using event-specific keywords and hashtags during seven natural disasters that occurred in 2017: Hurricane Irma, Hurricane Harvey, Hurricane Maria, the Mexico earthquake, the California wildfires, the Iraq-Iran earthquakes, and the Sri Lanka floods. The dataset has been annotated for three tasks. In the current study, we specifically focus on the *informativeness* task, which aims at recognizing whether a social media post is informative or uninformative for humanitarian aid purposes (see the definition of *informativeness* in Ofli et al., 2020).

In CrisisMMD, text and image pairs, which are annotated separately, could carry different labels. To properly compare the proposed method with previous works (Liang et al., 2022; Ofli et al., 2020), we align with the benchmark guidelines by focusing exclusively on the portion of the dataset where text and images share the same label for the informativeness task. The resulting dataset composition is reported in Table 1.

Reddit Videos To evaluate our method on a recent event, we collected Reddit posts with video attachments during the initial days of the 2023 conflict between Israel and Hamas. We used the Reddit API via the *praw*[1] library, using "*gaza OR israel OR hamas OR palestin**" as a case-insensitive search query, gathering 49,788 posts from October 13th to 19th, 2023. To account for the different nature of this dataset, we adapted the definition of *informativeness* as follows. The post is considered *informative* if it reports or shows direct proof of damage to people, such as injured, dead, or affected persons, or damage to man-made things, such as buildings, infrastructures, or vehicles, including the damaging act itself. Non-primary sources and indirect reporting, including newscasts, interviews performed either in a studio or in the field, reaction videos, edited videos, and non-photographic videos, are considered *not informative*.

Among the collected posts, 2,644 had a video attachment. Of these posts, 1,959 had a text length of at least ten words. This minimum word count was chosen arbitrarily to focus on posts with substantial descriptions. We filter out duplicated videos via video hashing[2].

Videos are then classified as *informative* based on the definition mentioned above, with two annotators using Label Studio[3]. We only keep videos where both annotators agreed on the label. Analogously to the data selection performed in the previous dataset, the annotators performed a second pass to exclude a limited number of videos for which the labels and the original posts' text did not agree. Then, we extracted all the keyframes from the videos and eliminated duplicated frames using image hashing[4]. Each frame was assigned the label and text from its original post, although we recognize that this method might not be perfectly accurate. The final dataset composition is reported in Table 1. This dataset, which includes some level of noise and is related to a human-made emergency, is meant to test the method's effectiveness under challenging conditions[5].

[1] https://github.com/praw-dev/praw
[2] https://pypi.org/project/videohash/
[3] https://labelstud.io/
[4] https://github.com/jgraving/imagehash
[5] The dataset with submission IDs and annotations is available upon request

EXPERIMENTS

LLM Setup

Regarding the text augmentation, we used LLaVA 13B version 1.5 for the CrisisMMD task and LLaVA 7B version 1.5 for the Reddit task because of budget limitations. These models were trained in September 2023 on various datasets, including a set of GPT-generated multimodal instruction-following data. The architecture adopted for the text classifier is RoBERTa base[6], pre-trained on 58M tweets and fine-tuned for sentiment analysis with the TweetEval benchmark.

For the text augmentation on the CrisisMMD dataset, the LLaVA *Temperature* parameter, which controls the randomness of the output, was set to 0.7, while the *Maximum Number of Generated Tokens* was set to 1024. The *Nucleus Sampling* parameter, also known as *Top P*, which determines diversity by only considering a subset of the most probable next words, was set to 0.7. In the case of text augmentation on the Reddit dataset, the settings used were analogous to those for the CrisisMMD dataset, except for the *Temperature* parameter, which was adjusted to 0.2. This is done due to the nature of the dataset, which frequently features video frames with significant motion and complexity. By lowering the *Temperature* setting, the model's range of generative options is somewhat limited, which supposedly helps to generate consistent descriptions of the video frames.

Classifier Setup

LLaVa weights are kept frozen and used only for text generation. The weights of the text classifier are instead fine-tuned while training for the informativeness classification task. We use Adam as optimizer, using a mini-batch size of 64, and weighted cross-entropy as a loss function, with weights accounting for class imbalance. We employ a Cyclical Learning Rate policy (Smith, 2017) for mini-batch Gradient Descent. Cyclical Learning Rate adjusts the learning rate by cyclically changing it between a minimum value of $1e^{-5}$ and a maximum value of $1e^{-4}$. To mitigate overfitting, we use an early-stopping condition on the validation cross-entropy loss, with patience of 5 epochs. Both LLaVa generation and RoBERTa training were performed on an A40 GPU. We did not specifically tune any hyper-parameter; hence, there is likely room for improvements in future studies.

Models are assessed using three metrics: classification accuracy, macro-averaged F1-score (M-F1), and weighted F1-score (W-F1), with weights determined by the number of samples for each class. It is important to note that during emergency events the distribution of samples across categories is usually imbalanced. This motivated the choice of macro-averaged and weighted F1-score to properly synthesize false positives and false negatives.

CrisisMMD

To compare our method against previous works (Liang et al., 2022), we keep the same training, validation, and test data splits provided by the dataset creators (Alam et al., 2018b). Table 2 shows the total number of tweets and attached images for each category, along with their distribution among training, validation, and test sets. The variation in the number of tweets and images in the training set is due to some tweets having multiple image attachments. These tweets are assigned exclusively to the training set.

Table 2. **Labels and data splits for the CrisisMMD informativeness task**

	Train (70%)		**Valid (15%)**		**Test (15%)**		**Total**	
	Text	**Image**	**Text**	**Image**	**Text**	**Image**	**Text**	**Image**
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

We used LLaVA to generate textual descriptions for images using prompt engineering techniques. We instructed the model to act as a computer vision expert and outlined the specific tasks. These tasks included generating a descriptive caption and an explanation for a given image, with information related to the specific classification task. Additionally, we specified the format for the two required outputs: a "*CAPTION*" providing a concise and factual image description and an "*EXPLANATION*" justifying if and why a given image could be *informative* in the current scenario. The prompt utilized is reported in Figure 2.

[6] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

Table 3. **Labels and data splits for the Reddit informativeness task**

	Train & Valid (70% - 15%)		Test (15%)		Total	
	Text	**Frame**	**Text**	**Frame**	**Text**	**Frame**
Informative	61	2,066	11	208	72	2,274
Not informative	651	28,490	115	4,787	766	33,277
Total	712	30,556	126	4,995	838	35,551

We processed the generated texts to craft six distinct inputs for the classifier, namely: the tweet text concatenated with the caption; the tweet text concatenated with the explanation; The tweet text concatenated with both the caption and the explanation; the caption concatenated with the explanation, without the tweet text; the caption alone; the explanation alone.

We compared our approach with traditional networks trained on a single modality (unimodal image and text, Ofli et al., 2020), as well as existing state-of-the-art image-text multimodal classification methods. Studies such as Abavisani et al., 2020 primarily employ a multimodal fusion of global features from each unimodal backbone. Others, like Huang et al., 2020, employ pre-trained multimodal BERT models adapted for image-text classification and inter-modal interaction via the attention mechanisms. We also compare to standard multimodal recognition techniques: score fusion (averaging unimodal classification scores) and feature concatenation (merging global features from each unimodal source). The study from Liang et al., 2022, which proposes a method for integrating information from other modalities into the unimodal models, was considered as a reference for the reported performances of the cited models.

Reddit Videos

Similarly to the CrisisMMD task, we used LLaVA to obtain information about the video frames. We instructed the model to act as a computer vision expert analyzing video frames. The requested output consisted of three text elements: a "*CAPTION*" providing a brief, factual description of the image; an "*EXPLANATION*" offering a justification aligned with the informativeness task; and a "*REASONING*" section, which elucidates the model's "reasoning" concerning the content of the image and its connection to the classification task described. The prompt utilized is reported in Figure 3.

We processed the generated texts to craft six distinct inputs for the classifier, namely: the post text concatenated with the caption; the post text concatenated with the explanation; the post text concatenated with the reasoning; the post text concatenated with both the caption and the explanation; the caption alone; the explanation alone. Finally, the dataset was partitioned into training, validation, and test sets, accounting for 70%, 15%, and 15% of the posts, respectively. We ensured, by leveraging video identifiers, that all frames from the same video appeared only in one partition. When partitioning, we ensured the preservation of class balance among the partitions by stratifying based on class labels. Table 3 displays the total posts and video frames for each category, including their distribution over splits.

To evaluate the robustness and generalization ability of the proposed method on the Reddit dataset, we replicated the same experiment over five random samples. We conducted the same experiment across five distinct pairs of training and validation sets, each generated through random partitioning of non-test data. We ensured that frames from the same video appeared only in one fold, and stratified the sampling on class labels. To evaluate the classifier's performance also at the video level, a video is classified as *informative* if at least 50% of its frames are predicted as *informative*.

RESULTS

CrisisMMD

In Table 4, we report the performance results achieved by different methods, including the proposed one, on the informativeness task.

In Figure 4, we present the test set confusion matrices for the multimodal model developed by Ofli et al., 2020 and for the three best-performing models proposed in this work. Confusion matrices from other studies are not included, as they were not provided in the respective papers.

[†]Reported performances are taken from Liang et al., 2022

[*]Computed from the corresponding confusion matrix reported by Ofli et al., 2020

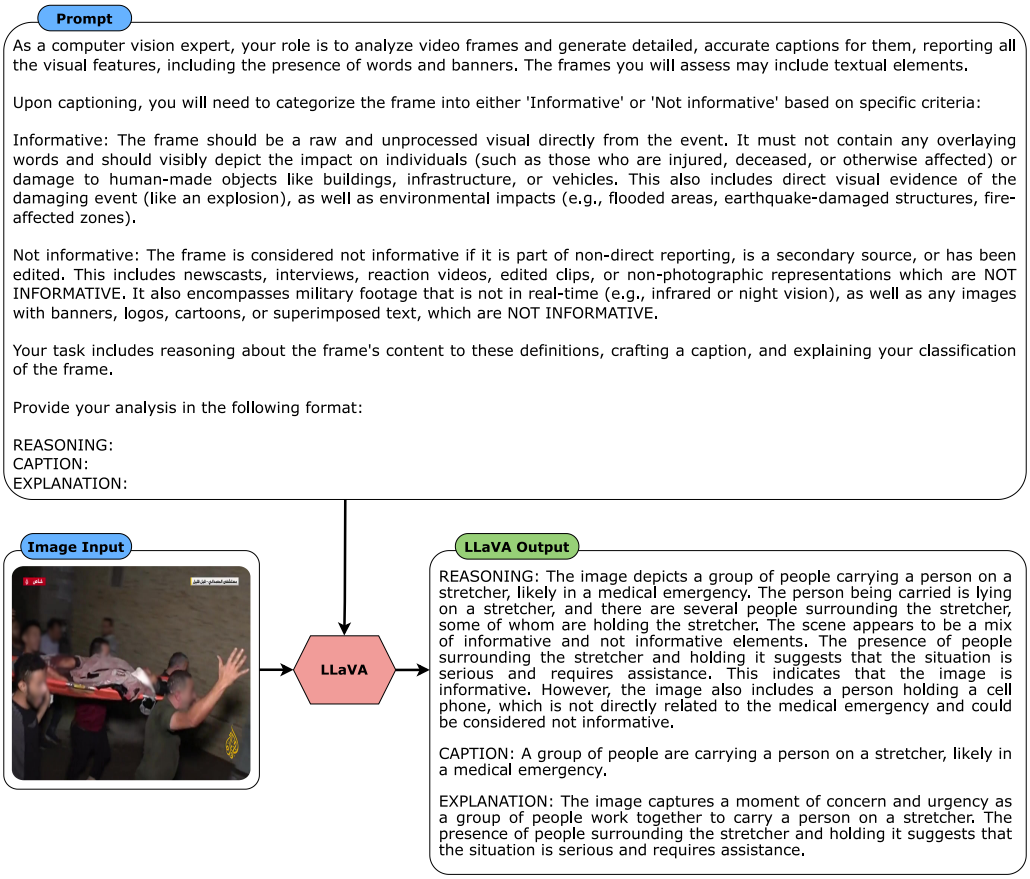

Figure 3. **Information augmentation for the Reddit informativeness task**

Table 4. **Results for CrisisMMD informativeness task**

Method	Acc	M-F1	W-F1
Unimodal Text (Ofli et al., 2020)	80.8	78.5*	80.9
Unimodal Image (Ofli et al., 2020)	83.3	80.5*	83.2
Multimodal Fusion (Ofli et al., 2020)	84.4	81.9*	84.2
Cross-attention (Abavisani et al., 2020)†	88.4	87.6	88.7
CentralNet (Vielzeuf et al., 2018)†	87.8	85.3	86.1
GMU (Ovalle et al., 2017)†	87.2	84.6	85.7
CBP (Fukui et al., 2016)†	87.9	85.6	86.4
CBGP (Kiela et al., 2018)†	88.1	86.7	87.3
MMBT (Kiela et al., 2019)†	86.4	85.3	86.2
VisualBERT (Li et al., 2019)†	88.1	86.7	88.6
PixelBERT (Huang et al., 2020)†	88.7	86.4	87.1
ViLT (Kim et al., 2021)†	87.6	85.1	88.0
ME Feature Concat (Liang et al., 2022)†	90.8	**91.6**	90.3
ME Cross-attention (Liang et al., 2022)†	**92.0**	91.2	91.3
RoBERTa Text Only	86.6	84.8	86.6
LLaVa+RoBERTa Caption Only	85.7	82.8	85.2
LLaVa+RoBERTa Explanation Only	85.5	83.3	85.4
LLaVa+RoBERTa Caption and Explanation	86.8	84.8	86.7
LLaVa+RoBERTa Text & Caption	91.4	90.2	**91.4**
LLaVa+RoBERTa Text & Explanation	88.8	87.6	88.9
LLaVa+RoBERTa Text & Caption & Explanation	90.6	89.3	90.6

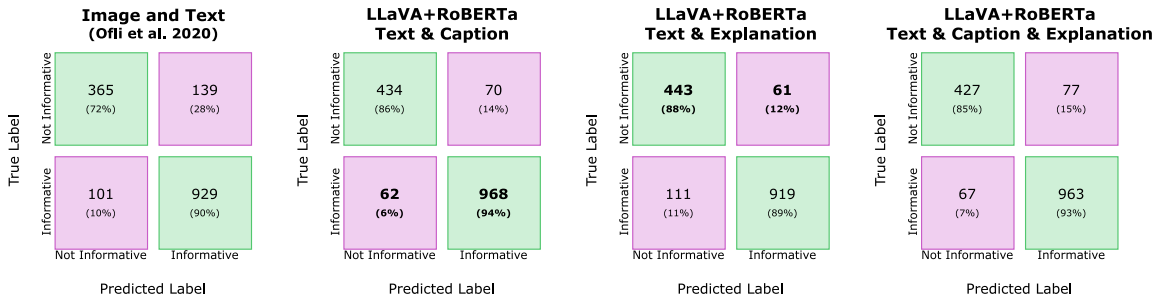

Figure 4. **Test set confusion matrices for the CrisisMMD informativeness task**

Table 5. **Results for Reddit informativeness task**

Method	Acc	W-F1	*Informative* F1
Frame metrics (5 folds average and std)			
LLaVA+RoBERTa Only Text	86.2 (1.0)	88.0 (1.0)	40.0 (6.8)
LLaVA+RoBERTa Text & Caption	96.3 (0.7)	96.3 (0.6)	54.1 (6.7)
LLaVA+RoBERTa Text & Reasoning	96.0 (0.4)	96.0 (0.6)	47.2 (10)
LLaVA+RoBERTa Text & Explanation	96.1 (1.1)	96.0 (0.9)	52.0 (8.8)
LLaVA+RoBERTa Text & Caption & Explanation	95.3 (2.8)	95.6 (2.0)	52.0 (9.0)
LLaVA+RoBERTa Only Caption	89.0 (0.9)	91.4 (0.5)	34.5 (2.0)
LLaVA+RoBERTa Only Explanation	89.4 (1.8)	91.7 (1.0)	34.0 (1.9)
Video metrics - 0.5 threshold (5 folds average and std)			
LLaVA+RoBERTa Text & Caption	90.1 (1.3)	90.0 (1.3)	40.0 (8.0)
LLaVA+RoBERTa Text & Reasoning	90.1 (0.6)	90.0 (0.7)	40.0 (6.0)
LLaVA+RoBERTa Text & Explanation	**90.5** (1.3)	90.1 (0.9)	41.3 (3.0)
LLaVA+RoBERTa Text & Caption & Explanation	89.2 (3.4)	89.2 (2.0)	39.6 (7.8)
LLaVA+RoBERTa Only Caption	90.0 (2.0)	**91.0** (1.3)	**60.0** (3.0)
LLaVA+RoBERTa Only Explanation	90.0 (2.15)	91.0 (1.6)	53.4 (5.6)

Reddit Videos

In Table 5, we report the performance results achieved by the proposed method on the Reddit informativeness task. The table reports the performance metrics computed both at the frame and video levels.

DISCUSSION

CrisisMMD

We can draw the following observations from Table 4. First, the unimodal models generally perform worse than the multimodal approaches. Second, we can see that the late unimodal fusion models (Cross-attention, CentralNet, GMU, CBP, CBGP), as well as the large pre-trained multimodal BERT models (MMBT, VisualBERT, PixelBERT, and ViLT) share similar performances.

The proposed approach, when injecting LLM-generated information, achieves state-of-the-art performances. We observe that the models that consider the original texts of the tweets generally perform better than the ones using only LLM-generated information. Although expected, this highlights the relevance of the tweets' contents for discerning relevance. In addition, we observe that the model combining the original text and captions performs better than combining the original text with other LLM-generated information. This fact may be due to the brief, factual information provided by the captions, which are possibly more helpful to the classifier. In contrast, explanations tend to be more detailed, including descriptions, comments, and evaluations. These additional details could lead the classifier astray, making it possibly more complicated to interpret the overall text accurately.

The confusion matrices presented in Figure 4 show that the proposed model combining tweet text and captions significantly outperforms the other three models in correctly identifying true positives and minimizing false negatives. Although this configuration generates a few more false positives than the model that combines text and explanations, a higher number of false negatives would be far more problematic during an emergency than a slight increase in false positives. The performance looks slightly worse when text is paired with explanations than when it

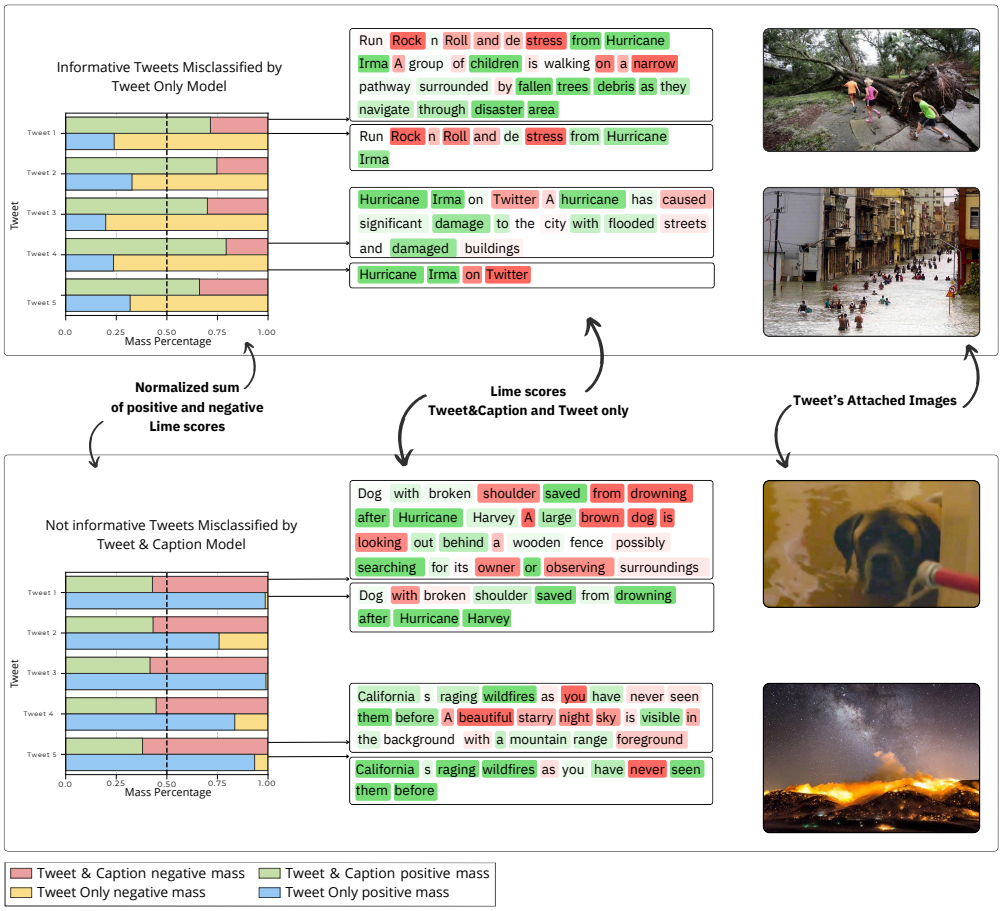

Figure 5. **Information Amplification Analysis in CrisisMMD: On the left, normalized LIME scores are displayed for selected tweets; in the middle, word-level LIME scores are highlighted (green for positive, red for negative); on the right, the images linked to each tweet are reported. *Positive* and *negative* mass percentages refer to the normalized sum of LIME scores within each tweet**

is paired with captions. When text, captions, and explanations are combined, the performance falls between the two preceding scenarios, possibly due to the beneficial effect of captions. This supports the hypothesis that explanations make the task more difficult in this scenario.

To understand the effect of the information amplification performed by the LLM-generated text, we performed an explainability analysis via Local Interpretable Model-Agnostic Explanation (LIME, Ribeiro et al., 2016). LIME perturbates the input instances and uses the model's predictions on these instances to fit an interpretable model in the proximity of the selected instance; this second model reveals which features are influential for the specific instance and how much they contribute to the global model's prediction.

Figure 5 illustrates how the LLM augmentation can help the classification task. On the left, each pair of bars represents an *informative* Twitter post from the CrisisMMD dataset, with the different areas representing the normalized sums of positive and negative LIME scores for two configurations, *"Tweet & Caption"* (first row) and *"Tweet only"* (second row).

In the upper section, we observe that positive scores are boosted after augmenting the text with LLM-generated captions, recovering false negative cases. However, the information injection also has a counterproductive effect in some cases, as seen in the bottom section. Here, examples of *informative* tweets are misclassified by the *Tweet & Caption* model due to features with negative scores becoming prevalent. Although the balance between misclassifications remains favorable compared to not using information amplification, as shown in Table 4, this observation highlights the challenges in managing the impact of the additional information introduced.

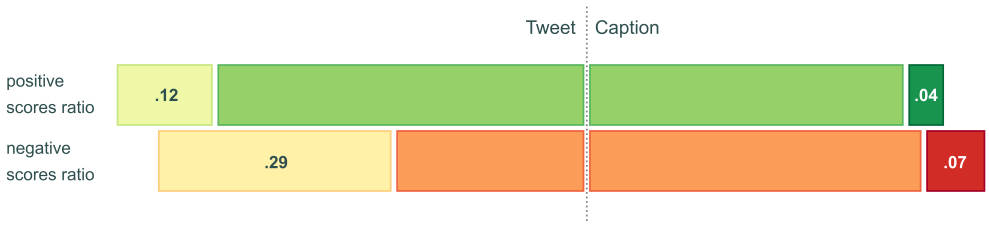

Figure 6. **Normalized LIME score ratios in CrisisMMD test set; Tweet-only lemmata on the left; Caption-only lemmata on the right; lemmata shared by both sections in the middle**

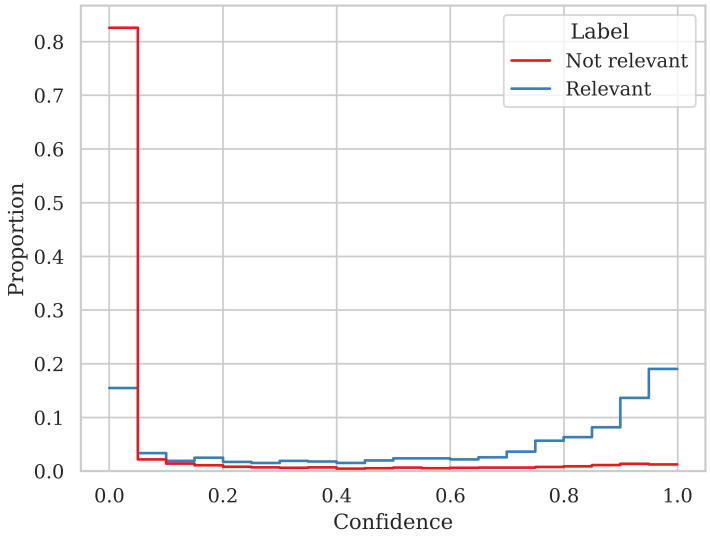

Figure 7. **Test set frame confidence distribution, normalized per class (Reddit Task)**

Figure 6 provides a visual representation of the overall effect of the information introduced by the LLM. The figure refers to the *Tweet & Caption* model and highlights the contribution of each of the two sources – original tweet and caption – to the classification outcome by aggregating and normalizing the LIME scores related to lemmata[7], after excluding English stopwords. In the figure, positive and negative scores are aggregated as rows, while a vertical line partitions tweet-dependent and caption-dependent scores. The three colors highlight the characteristics of the lemmata: the two lighter blocks on the left account for the weight of lemmata seen only in tweets, while the two darker blocks on the right account for the weight of lemmata seen only in captions.

As a general observation, it can be seen that most of the scores originate from lemmata shared between the two sources. The effect of new terms introduced by the LLM is limited, roughly one-fourth of the percentage of scores for tweet-only lemmata. This observation suggests that the tweet's text could consistently describe the images' contents. Overall, the original tweet text accounts for 57% of the positive scores and 52% of the negative scores.

Reddit Videos

The proposed approach shows a good separability of the classes already at the frame level, as the distributions in Figure 7 show. Roughly 16% of the *informative* frames are classified as *not informative* with more than 95% confidence. Upon manual inspection of a sample of these frames, the reasons behind the misclassifications were not completely clear, since some relevant frames were present in this subset, together with generic, blank, and indistinct frames isolated from videos labeled as *"informative"*. This observation also suggests that a more fine-grained approach to labeling, such as considering video segments instead of whole videos, could be beneficial for the task, albeit more burdensome as an annotation task. On the other hand, this does not impair the performance of the video classifier, since only individual frames are affected by the misclassification.

Moreover, it should be noted that the reported performance results, based on common-sense thresholds — 0.5 confidence at the frame level and 0.5 proportion of positive frames at the video level — prove to be conservative

[7] "*A form of a word that appears as an entry in a dictionary and is used to represent all the other possible forms*" according to the Cambridge Dictionary

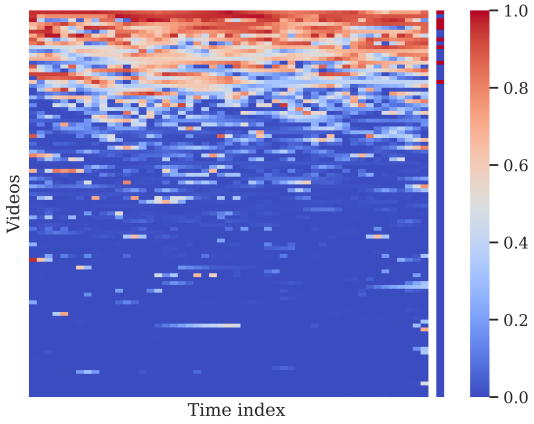

Figure 8. **Confidence for "*informative*" class for each test set video, over normalized time index. True label on the right (Reddit Task)**

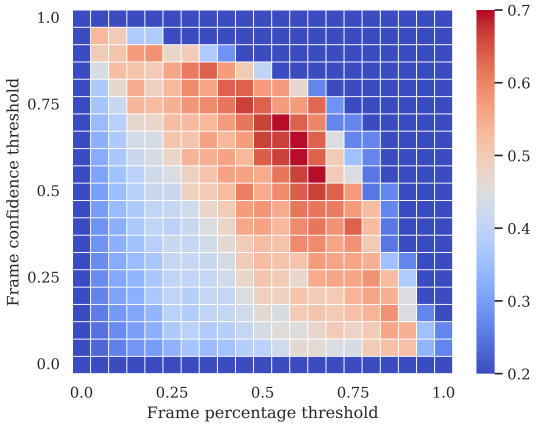

Figure 9. **Response of the test set F1-score for the "*informative*" class varying classification thresholds (Reddit Task)**

estimates of the performance. Figure 9 reports a grid exploration of the thresholds, showing that an increase in the video-level threshold could lead to better performances. We have chosen to evaluate our approach using a common-sense threshold to assess its performance as a general-purpose method. It should also be noted that the *"Caption Only"* model performs better at the video level, even if the F1 score at the frame level is worse. This is due to a higher false positive rate in the majority class (*"Not Informative"*), leading to a better final result since frames predicted as positive are still a minority when considered at the video level.

An understanding of the frame-level labels for the test set videos can be obtained from Figure 8. The figure illustrates how the confidence of being an *informative* frame is distributed along the videos, with red indicating higher confidence. The video durations are normalized for convenience. It can be observed that higher confidence regions, while still present in many of the *not informative* videos (blue region in the bottom part of the graph), are scattered across the video length; these false positives are mostly absorbed by the threshold at the video level. The fact that the *"Caption Only"* performs reasonably well without relying on the original text could also be related to the fact that, often, the text of the Reddit submission in the scenario analyzed is not merely descriptive but contains additional considerations, such as stances and advocacy.

CONCLUSIONS AND FUTURE WORK

We proposed a method for categorizing multimodal posts by extracting text-only representations from multimedia attachments and merging this information with the text of social media posts. One of the advantages of this approach is that it is general-purpose and we showed that it achieves state-of-the-art performances on the informativeness task without the need of fine tuning for a specific case. Further room for improvement is left by performing hyper-parameter tuning on the classifier on the specific case, although this was not the goal of the present study.

Amalgamating text and image information in social media data is crucial, as they often present complementary or conflicting information. Since current multimodal classification methods assume common labels for different modalities, future works should investigate multimodal learning algorithms capable of handling heterogeneous input (e.g., text and image pairs with conflicting labels).

One of the main hindrances to adopting LLMs is their computational cost, which is linked to energy utilization and environmental consequences. We plan to examine the trade-offs offered by alternative smaller models and efficiency-oriented methods such as network quantization, considering the combined effect of performance, execution times, and computational requirements to support a wider choice of use cases.

Finally, we plan to extend the evaluation of the proposed approach to few-shot and continual learning scenarios, since the intrinsically dynamic nature of emergency events poses significant challenges to adopting relevance-oriented classifiers. Moreover, since historical emergency-related data can be limited depending on the kind of emergency, learning informativeness on the fly could be the only viable option in some scenarios.

ACKNOWLEDGEMENTS

This work has been supported by the PNRR-PE-AI "FAIR" project funded by the NextGenerationEU program and by the PRIN 2022 Project "Discount quality for responsible data science: Human-in-the-Loop for quality data".

REFERENCES

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

Alam, F., Ofli, F., & Imran, M. (2018a). Processing social media images by combining human and machine computing during crises. *International Journal of Human–Computer Interaction*, *34*(4), 311–327.

Alam, F., Ofli, F., & Imran, M. (2018b). CrisisMMD: Multimodal Twitter datasets from natural disasters. *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM)*.

Avvenuti, M., Cimino, M. G. C. A., Cresci, S., Marchetti, A., & Tesconi, M. (2016). A framework for detecting unfolding emergencies using humans as sensors. *SpringerPlus*, *5:43*.

Bono, C., Pernici, B., Fernandez-Marquez, J. L., Shankar, A. R., Mülâyim, M. O., & Nemni, E. (2022). TriggerCit: Early Flood Alerting using Twitter and Geolocation–a comparison with alternative sources. *Proc. ISCRAM 2022, Tarbes, France*.

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. In J. Su, X. Carreras, & K. Duh (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), Austin, Texas, USA* (pp. 457–468). The Association for Computational Linguistics.

Havas, C., Resch, B., Francalanci, C., Pernici, B., Scalia, G., Fernandez-Marquez, J. L., Van Achte, T., Zeug, G., Mondardini, R., Grandoni, D., Kirsch, B., Kalas, M., Lorini, V., & Rüping, S. (2017). E2mC: Improving emergency management service practice through social media and crowdsourcing analysis in near real time. *Sensors*, *17*(12). https://doi.org/10.3390/s17122766

Huang, Z., Zeng, Z., Liu, B., Fu, D., & Fu, J. (2020). Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*.

Imran, M., Castillo, C., Lucas, J., Meier, P., & Vieweg, S. (2014). AIDR: Artificial. *Proceedings of the 23rd International Conference on World Wide Web*, 159–162. https://doi.org/10.1145/2567948.2577034

Imran, M., Ofli, F., Caragea, D., & Torralba, A. (2020). Using AI and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions. *Information Processing & Management*, *57*(5), 102261.

Kiela, D., Bhooshan, S., Firooz, H., & Testuggine, D. (2019). Supervised multimodal bitransformers for classifying images and text. *Visually Grounded Interaction and Language (ViGIL), NeurIPS 2019 Workshop, Vancouver, Canada, December 13, 2019*.

Kiela, D., Grave, E., Joulin, A., & Mikolov, T. (2018). Efficient large-scale multi-modal classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), 5198–5204.

Kim, W., Son, B., & Kim, I. (2021). ViLT: Vision-and-Language Transformer without convolution or region supervision. *International Conference on Machine Learning*, 5583–5594.

Kruspe, A., Kersten, J., & Klan, F. (2020). Detecting event-related tweets by example using few-shot models. *ISCRAM 2019 Conference Proceedings – 16th International Conference on Information Systems for Crisis Response and Management*, 825–835.

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

Liang, T., Lin, G., Wan, M., Li, T., Ma, G., & Lv, F. (2022). Expanding large pre-trained unimodal models with multimodal information injection for image-text multimodal classification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15492–15501.

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *NeurIPS*.

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

Mozannar, H., Rizk, Y., & Awad, M. (2018). Damage identification in social media posts using multimodal deep learning. *International Conference on Information Systems for Crisis Response and Management*.

Nguyen, D. T., Ofli, F., Imran, M., & Mitra, P. (2017). Damage assessment from social media imagery data during disasters. *2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 569–576.

Ofli, F., Alam, F., & Imran, M. (2020, May). Analysis of social media data using multimodal deep learning for disaster response. In A. L. Hughes, F. McNeill, & C. W. Zobel (Eds.), *17th International Conference on Information Systems for Crisis Response and Management (ISCRAM)* (pp. 802–811). ISCRAM Digital Library.

Ovalle, J. E. A., Solorio, T., Montes-y-Gómez, M., & González, F. A. (2017). Gated multimodal units for information fusion. *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*.

Peters, R., & De Albuquerque, J. P. (2015). Investigating images as indicators for relevant social media messages in disaster management. *12th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*.

Ramachandram, D., & Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, *34*(6), 96–108. https://doi.org/10.1109/MSP.2017.2738401

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Sakaki, T., Okazaki, M., & Matsuo, Y. (2013). Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Trans. Knowl. Data Eng.*, *25*(4), 919–931. https://doi.org/10.1109/TKDE.2012.29

Smith, L. N. (2017). Cyclical learning rates for training neural networks. *2017 IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 464–472.

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, *39*, 156–168. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2017.12.002

Stollberg, B., & De Groeve, T. (2012). The use of social media within the global disaster alert and coordination system (GDACS). *Proc. 21st Intl. WWW Conf.*, 703–706.

Vielzeuf, V., Lechervy, A., Pateux, S., & Jurie, F. (2018). CentralNet: A multilayer approach for multimodal fusion. *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 575–589.

Wang, Y., Xu, X., Yu, W., Xu, R., Cao, Z., & Shen, H. T. (2021). Combine early and late fusion together: A hybrid fusion framework for image-text matching. *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. https://doi.org/10.1109/ICME51207.2021.9428201

Wiegmann, M., Kersten, J., Klan, F., Potthast, M., & Stein, B. (2020). Analysis of detection models for disaster-related tweets. *ISCRAM 2020 Conference Proceedings – 17th International Conference on Information Systems for Crisis Response and Management*, 872–880.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*.