Image-text crisis tweet categorization: a caption-based approach

Badreddine Farah*
University of Orléans, INSA-CVL, LIFO, EA 4022, F45067 Orléans, France
badreddine.farah@univ-orleans.fr

Guillaume Cleuziou
University of Orléans, INSA-CVL, LIFO, EA 4022, F45067 Orléans, France
guillaume.cleuziou@univ-orleans.fr

Cécile Gracianne
BRGM, F45060 Orléans, France
c.gracianne@brgm.fr

Adel Hafiane
INSA-CVL, University of Orléans, PRISME, EA 4229, F18022 Bourges, France
adel.hafiane@insa-cvl.fr

Anaïs Halftermeyer
University of Orléans, INSA-CVL, LIFO, EA 4022, F45067 Orléans, France
anais.halftermeyer@univ-orleans.fr

Raphaël Canals
University of Orléans, INSA-CVL, PRISME, EA 4229, F45072 Orléans, France
raphael.canals@uoMappers.niv-orleans.fr

ABSTRACT

The growth of social media usage this last decade has made available a massive and valuable volume of multimedia data. However, the lack of large multimodal annotated datasets, along with the inherent noise and the diversity of multimodal relations in this type of data presents challenges for machine learning methods. Unlike classic multimodal data, social media data comes with a large diversity of relations between image and text making the interaction between the two modalities more difficult. Previous research concentrated on fusion strategies with separate encoders for each modality. This paper introduces CMB (Caption-based Multimodal BERT), a method of classifying crisis-related social media posts by utilizing information from both images and texts. CMB translates the image modality into a text-compatible space, facilitating intermodal interaction. Furthermore, CMB presents training opportunities to enhance the model's robustness to missing modalities. Experimental results show that CMB is competitive with well-established, costly, and manually crafted multimodal models.

Keywords

Deep Learning, multimodal data, text/image fusion, Crisis data

INTRODUCTION

Proliferation of social media and advancements in internet technology have significantly changed the nature of online content, especially on platforms such as X (previously Twitter[1]), where user-generated data has immense value to a wide range of domains, including public health, economics, and politics (Wu & Mebane Jr, 2022). In times of crisis (natural disasters), this data can be used to gain situational awareness and assist humanitarian organizations in their preparedness, mitigation, response, and recovery efforts. However, the large volume of data present on social media requires the development and implementation of automated systems to effectively categorize and identify informative posts by analyzing information present across modalities.

Many works in this scope focused on processing unimodal data including images (Nguyen et al., 2017) or texts (Kumar et al., 2011). To leverage both modalities, many papers have proposed general-purpose self-supervised

*corresponding author
[1]We will be using Twitter and tweets as a reference for X and X posts

vision-language models (Y.-C. Chen et al., 2020; Kim et al., 2021; Tan & Bansal, 2019). Nevertheless, these large pre-trained (multimodal) models are primarily designed for homogeneous data, such as images and their captions. They therefore do not take into account the heterogeneous informational content that may exist between an image and its associated text, as is the case in social media (Vempala & Preoţiuc-Pietro, 2019), making transfer learning suboptimal (Liang et al., 2022). In this context of multimodal social media tasks, prior works have concentrated their efforts on fusion strategies. The prevailing practice involves using two unimodal encoders, one for each modality, followed by fusing the unimodal embeddings. Many fusion strategies were proposed (Ofli et al., 2020) for instance used the concatenation of the two embeddings, while (Abavisani et al., 2020) introduced a cross-attention mechanism to achieve a more effective fusion. In contrast, (Liang et al., 2022) explored early interaction strategies to enable interactions on low-level features.

However, the lack of large annotated datasets coupled with the variety of the inter-modality relations between image and text presents additional challenges for learning task-oriented fusion. Moreover, previous studies (Fan et al., 2023; Wang et al., 2020) have shown that different modalities tend to overfit and converge at different rates when using two encoders. This indicates that optimizing the same objective for various modalities results in inconsistent learning efficiency. In this paper, we investigate the effectiveness of captions in vision-language tasks for crisis tweets. It is hypothesized that the embedding integration of both texts and images into a common semantic space and the use of a single trainable encoder may thus facilitate the challenging learning problem. For this, we propose to use a caption-based method, which consists of translating a modality (e.g. image) towards another modality (e.g. text). We, therefore, present the CMB method ***C****aption-based* ***M****ultimodal* ***B****ERT*, that leverages captioning models to translate image into text and concatenate resulting captions with the tweet's text before inputting the combined data into a text encoder.

To thoroughly evaluate this approach, we conduct experiments with a variety of captioning models that have been pre-trained on different datasets. We also propose a training strategy that allows to use our method in unimodal settings without significant performance loss. Our experiments focus on crisis context, in particular on the well-studied multimodal CrisisMMD dataset (Alam et al., 2018). The examples in Figures 4 and 5 illustrate the challenges of this dataset in terms of the relationship between modalities (complementarity, heterogeneity or partial redundancy (Vempala & Preoţiuc-Pietro, 2019)). Our contributions are as follows:

- We present CMB, a method to homogenize the input space of image/text data to enable better information processing by BERT-like architecture.
- We compare three captioning models and show that this method can achieve highly competitive results for crisis-related tweets.
- We propose a training strategy to use our models in unimodal settings without significant loss of performance.
- Lastly, we present a qualitative analysis of the success and failure of our method.

RELATED WORKS

IMAGE-TEXT MULTIMODAL LEARNING

The image-text machine learning research area encompasses a variety of application domains, such as image captioning (Cho et al., 2022; Mokady et al., 2021), visual question answering (Antol et al., 2015) and visual reasoning (Zellers et al., 2019). Recently, several pre-trained models have been developed to address these tasks : LXMERT (Tan & Bansal, 2019) employs two unimodal streams and a cross-attention encoder to combine information from multiple modalities, while UNITER (Y.-C. Chen et al., 2020) uses a single stream transformer to enable inter-modality attention from the early layers of the model. In contrast, VILT (Kim et al., 2021) allows interaction between text and image at the fine-grained feature level without requiring prior convolution or region detection.

Multimodal learning has been applied to a range of social media projects, including emergency response (Abavisani et al., 2020; Liang et al., 2022; Long & McCreadie, 2022), emotional recognition (Yang et al., 2020), fake news detection (Tuan & Minh, 2021) and political science (Wu & Mebane Jr, 2022). There are two main approaches to multimodal social media post classification: early fusion and late fusion. In early fusion, the interaction between modalities is allowed from the early layers of the model, as shown by the Multimodal Information Injection Plug-in units introduced in (Liang et al., 2022), which preserve the strong intra-modal processing ability of large pre-trained unimodal models. In late fusion, separate encoders are used for each modality, such as DenseNet (Huang et al., 2017) or ResNet (He et al., 2016) for images and BERT (Devlin et al., 2018) or RoBERTa (Liu et al., 2019) for texts.

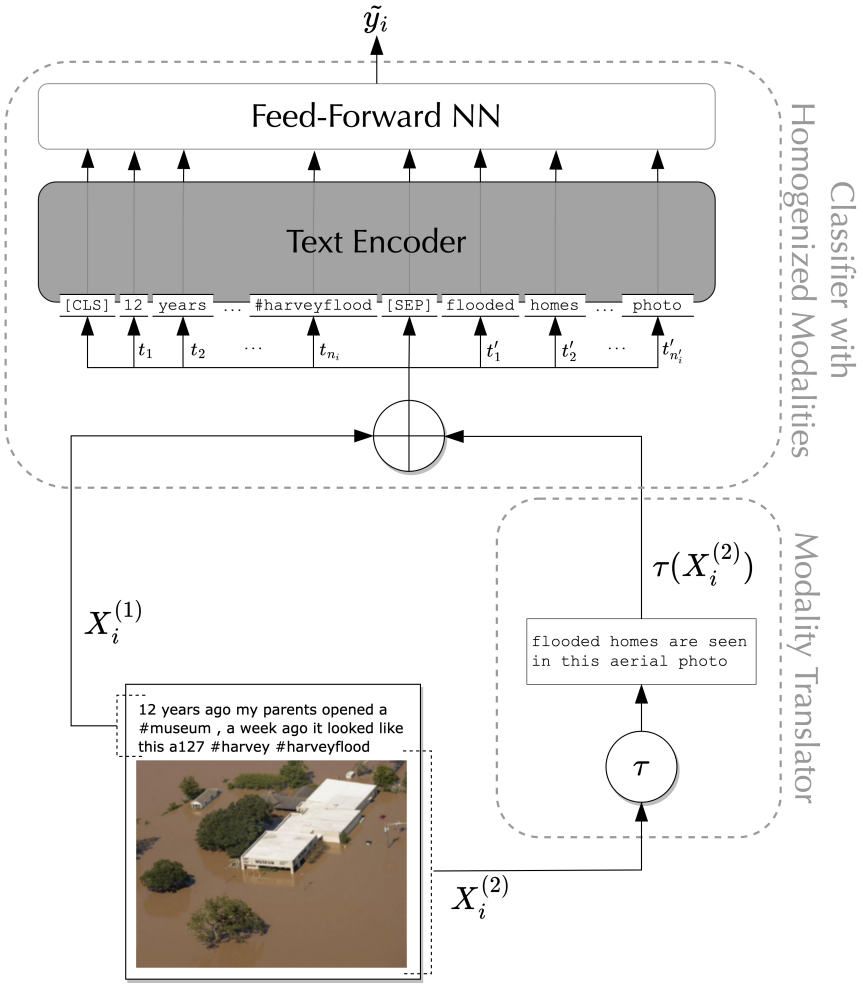

Figure 1. Overview of the CMB method structured in two main parts: a non-learnable process represented by the modality translator (image captioning model) and the homogenized modalities classifier as trainable module.

Interaction between modalities occurs at the high feature level through techniques such as feature concatenation (Long & McCreadie, 2022) or cross-attention (Abavisani et al., 2020).

In this study, however, we explore the effectiveness of image captions, employing pre-trained captioning model to capture the image information. The acquired information is then integrated into a text embedding space to facilitate fusion.

MACHINE LEARNING FOR EMERGENCY RESPONSE

Using social media information in emergency response is increasingly considered essential for the management of natural disaster crises. Many prior works have focused on unimodal signals: (Kumar et al., 2011) proposes a system to analyze and monitor (text) tweets to help humanitarian organizations better manage crises, (Nguyen et al., 2017) uses images posted on Twitter to determine the level of damage caused by disasters (Little, mild or severe). In a multimodal setting, (Abavisani et al., 2020) applies late fusion with two unimodal encoders to filter informative image-text tweets, while (Liang et al., 2022) employs early fusion to enhance low-level inter-modal feature interactions. Unlike previous approaches, we argue that the inherent complexity within social media data can be challenging due to the verity of relations between modalities, (Vempala & Preoţiuc-Pietro, 2019). To overcome this challenge (Sosea et al., 2021) introduced DisRel, a dataset for learning to distinguish the relationships between images and texts according to their nature (*complementary*, *similar* or *unrelated*). Their work demonstrated that incorporating such tasks can enhance the performance of models trained on CrisisMMD.

In this work, we propose an alternative method leveraging captioning techniques for image-to-text translation, in order to map the image representation space into text facilitating fusion within a shared representation space.

IMAGE CAPTIONING FOR FUSION

In previous works, captions have been used to represent images for multimodal interaction, such as Multimodal Named Entity Recognition (MNER) (S. Chen et al., 2020). It has been argued that models trained for semantic understanding, like image captioning models, may provide better image representations. The use of captions as image representations allows for the incorporation of textual descriptions of the image content, providing additional context and useful information to understand and interpret the image. Moreover, (Wu & Mebane Jr, 2022) uses captions as a translated version of image in addition to the tweet's image and text to improve performance in classifying multimodal tweets. Contrary to (Wu & Mebane Jr, 2022), the method CMB proposed in this paper uses - through captions - modality translation as a substitute and not as a complement to the image to improve inter-modality interactions while keeping the modal simple to train and use in real-world scenarios (with limited resources). Moreover, we argue that the use of one encoder helps to overcome the problem of different convergence rates of each modality encoder and lead to more efficient learning.

METHODOLOGY

PROBLEM STATEMENT

Let $\mathcal{D}$ be a multimodal dataset with N instances, denoted by $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{N}$. Each $X_i \in \mathcal{X}$ represents a multimodal data point with M modalities, represented as $X_i = \{X_i^{(p)}\}_{p=1}^{M}$. Associated with each instance is a label $y_i \in C$, where C is the set of possible labels. The multimodal classification task aims at learning a classifier $h : \mathcal{X} \rightarrow C$ that combines information from different modalities to make accurate predictions. In the following, we consider only bimodal instances ($M = 2$) where the two modalities are different in nature, typically textual and visual modalities. Thus, $X_i^{(1)} = (t_1, \ldots, t_{n_i})$ is a sequence of tokens while $X_i^{(2)} \in \mathbb{R}^{C \times H \times W}$ is an image (where C, H, and W are the number of channels, the height, and the width of the image respectively.

MODEL OVERVIEW

Since modalities of different natures (typically text and image) are hard to combine in multimodal fusion strategies, we propose a two-step categorization approach that leverages modality translation to homogenize the inputs (Figure 1). The first step, *Modality Translator* consists in translating one modality into a space compatible with the space of the other modality using a translator function $\tau(.)$. In our particular context of text-image tweets processing, the image modality $X_i^{(2)}$ is translated into a text (sequence of tokens) by means of an image captioning process $\tau(X_i^{(2)}) = (t'_1, \ldots, t'_{n'_i})$. The second step *Classifier with Homogenized Modalities* takes as input the concatenation of the homogenized modalities $X_i^{(1)} \oplus \tau(X_i^{(2)})$ to train both a text encoder and the final classification layer. Let us notice that the approach allows to process both multimodal ($X_i^{(1)} \oplus \tau(X_i^{(2)})$) or unimodal data ($X_i^{(1)}$ only or $\tau(X_i^{(2)})$ only).

MODALITY TRANSLATORS

In our multimodal approach, the captioning model is viewed as a modality translator. For our investigation, we use three publicly available captioning models, which vary depending on the encoders and the training data they use.

- **CLIP Prefix for Image Captioning** (Mokady et al., 2021). A captioning model using the CLIP (Radford et al., 2021) (Contrastive Language-Image Pre-training) multimodal encoder, which is trained on a large dataset sourced from the Web. It extracts image features using the CLIP encoder and employs them as a prefix for text generation by a language model. Trained with either the MSCOCO dataset (Lin et al., 2014) or the Conceptual Caption dataset (Sharma et al., 2018), the authors argue that the rich semantic features of the CLIP encoder, coupled with a pre-trained language model (GPT2 (Radford et al., 2019)), provide a comprehensive understanding of both visual and textual data, resulting in high quality captions. In the following, these captioning models are referred to as *CLIP_cc* (trained with Conceptual Captions) and *CLIP_coco* (trained with MSCOCO).

- **Fine-grained Image Captioning with CLIP Reward**. In (Cho et al., 2022) the authors use CLIP (Radford et al., 2021) in two ways, first as an image encoder, and then CLIP text and CLIP image are used to compute a multimodal similarity that then serves as a reward function for each generated caption to obtain descriptive and distinctive captions. This captioning model is referred to as *CLIP_reward* in the following.

Table 1. Training and evaluation settings.

Train \ Eval	*Text*	*Image*	*Caption*	*Text + Image*	*Text ⊕ Caption*
Text	T_T	-	-	-	-
Image	-	I_I	-	-	-
Caption	-	-	C_C	-	-
Text + Image	-	-	-	TI_TI	-
Text ⊕ Caption	TC_T	-	TC_C	-	TC_TC

- **Transformer based Captioning**.(Vaswani et al., 2017) uses the Transformer architecture and first process the image using Mask-RCNN (He et al., 2017), the features are then fed to the Transformer in order to generate the caption. This captioning model is referred to as *Trans_cap* in the rest of the article.

The CMB approach we propose in this paper allows to independently integrate any of the four captioning models presented above. In the following experiments, we will focus on measuring the contribution of the captions generated by each model in our proposed multimodal framework.

MULTIMODAL CLASSIFIER WITH HOMOGENIZED MODALITIES

Transformer based models (Devlin et al., 2018; Liu et al., 2019; Radford et al., 2019; Vaswani et al., 2017) has achieved state of the art result in large set of natural language processing tasks, by leveraging various self-supervised pre-training techniques. Therefore, without loss of generality, we use BERT (Devlin et al., 2018) as text encoder to extract embeddings from the input $(X_i^{(1)} \oplus \tau(X_i^{(2)}))$, which are then used to perform the classification task using a feed forward neural network layer.

EXPERIMENTAL SETUP

DATASET AND TASKS

The **CrisisMMD** (Alam et al., 2018) benchmark aims to identify crisis events that require emergency response using social media posts as a basis. The dataset for this benchmark consists of tweets (image-text pairs) obtained through searching for specific hashtags on Twitter and labeled for three tasks Informativeness, Humanitarian, and Damage severity assessment. In this work, we only consider the first two multimodal tasks:

- **Informativeness**: a binary classification task identifying whether or not a given tweet (text or image) is informative for humanitarian aid purposes, i.e., useful for providing assistance to people in need.
- **Humanitarian**: Identifying whether a given tweet (text or image) belongs to one of the following seven categories : *infrastructure and utility damage*; *vehicle damage*; *rescue, volunteering, or donation efforts*; *injured or dead people*; *missing or found people*; *other relevant information*; *not humanitarian related*.

This dataset provides a total of 16,058 tweet texts and 18,082 tweet images labelled for these two tasks, resulting into 8,079 multimodal instances for the Humanitarian task and 12,168 instances for the Informativeness task. Unfortunately, the CrisisMMD dataset does not offer one label for each *tweet*, but one label for each *modality* of a tweet. Therefore, it is not directly suited for the (multimodal) classification task we are considering in this work. Thus, we adopt the practice proposed in (Ofli et al., 2020) by using only instances with similar labels on the two modalities (text and image). Although this process excludes a non-negligible number of tweets, it is nevertheless based on the fairly reasonable assumption that a label that matches on both modalities can be used as a label for the entire tweet.

BASELINES

We compare our approach to both uni- and multimodal baselines in the field of image-text classification. Recently, the trend is to use multiple architectures in this research area. To provide a comprehensive evaluation of our approach, we implemented two popular multimodal architectures: the feature concatenation (Long & McCreadie, 2022) and the cross-attention (Abavisani et al., 2020) methods. Additionally, we included the unimodal methods BERT (Devlin

Table 2. Comparisons on CrisisMMD in terms of classification accuracy, Macro F1-score and weighted F1-score.

Models		Informativeness			Humanitarian		
		Acc	F1-m	F1-w	Acc	F1-m	F1-w
Unimodal	BERT	0.8691±0.0031	0.8465±0.0041	0.8667±0.0033	0.8183±0.0068	0.6964±0.0271	0.8179±0.0069
	DenseNet	0.8393±0.0028	0.8186±0.0022	0.8396±0.0024	0.7608±0.0178	0.5366±0.0495	0.7416±0.0272
	CLIP	0.8594±0.0281	0.8407±0.0283	0.8591±0.0265	0.8136±0.0224	0.6105±0.0404	0.8005±0.0325
Multimodal	Features-concat	0.9014±0.0020	0.8870±0.0031	0.9008±0.0023	0.8561±0.0102	0.6772±0.0070	0.8533±0.0098
	Cross-Attention	0.9022±0.0037	0.8885±0.0043	0.9019±0.0037	0.8549±0.0060	0.6718±0.0059	0.8520±0.0058
	CLIP-Concat	0.9020±0.0045	0.8871±0.0054	0.9012±0.0046	0.8608±0.0080	0.7778±0.0146	0.8589±0.0077
CMB	CMB-Trans_cap	0.8920±0.0039	0.8749±0.0045	0.8907±0.0039	0.8527±0.0065	0.7536±0.0274	0.8520±0.0070
	CMB-CLIP-reward	0.9125±0.0033	0.8987±0.0040	0.9115±0.0034	0.8765±0.0080	0.7631±0.0266	0.8755±0.0079
	CMB-CLIP-coco	0.9098±0.0024	0.8954±0.0029	0.9087±0.0025	0.8681±0.0064	0.7156±0.0307	0.8660±0.0066
	CMB-CLIP-cc	**0.9203±0.0036**	**0.9081±0.0043**	**0.9196±0.0037**	**0.8827±0.0033**	**0.7810±0.0176**	**0.8826±0.0032**

et al., 2018) (text only) and DenseNet (Huang et al., 2017) (image only) for comparison. Furthermore, we conduct a comparison between our approach and two other methods: CLIP (Radford et al., 2021) and CLIP-Concat. In the latter, we concatenate features from CLIP image encoder and BERT to accomplish the multimodal classification task. We aim to evaluate the effectiveness of CMB against these well-established methods to illustrate its usefulness in the crisis domain.

IMPLEMENTATION DETAILS

In this work we consider the four modality translators described above. Specifically, for each captioning model, we produced five captions for every image using the authors' provided demo for CLIP-based captioning models (CLIP_cc, CLIP_coco and CLIP_reward) and the demo from a GitHub repository[2] for the Trans_cap model. We keep the most similar caption - from the five generated by each model - to the image in terms of clip similarity score (Cho et al., 2022).

We used the BERT base model from (Wolf et al., 2020) as a text encoder with token type embeddings different for captions ($\tau(X_i^{(2)})$) and texts ($X_i^{(1)}$). To classify, we computed the mean pooling on embeddings provided by BERT and used a classification layer consisting of a ReLU activation function and a dropout layer between two feed-forward layers. The models were trained on a Nvidia Tesla V100 GPU for 3 to 5 epochs with $batch\ size = 128$ and $learning\ rate = 5e^{-5}$. We made our code available on GitHub[3].

UNIMODAL AND MULTIMODAL SETTINGS

To evaluate the robustness of our methodology across diverse scenarios, we conduct an array of experiments involving the training and evaluation of various iterations of our model, each employing distinct modalities such as Text, Image, or Caption. Table 1 uses specific notations to delineate these different experimental setups.

The first three rows of the table focus on unimodal approaches (T_T, I_I, and C_C settings), wherein a single modality is exclusively used for both training and evaluation. For text categorization (T_T, C_C), we employ BERT (Devlin et al., 2018), while for image categorization (I_I), DenseNet (Huang et al., 2017) is utilized.

The fourth row shows multimodal approaches (TI_TI setting), involving the simultaneous use of Text and Image during both training and evaluation. This incorporates a basic feature concatenation strategy, along with the previously mentioned cross-attention architecture (Abavisani et al., 2020) which uses attention mechanisms to merge representations from the two unimodal architectures (BERT and DenseNet).

Finally, the last line corresponds to our multimodal method based on modality translation, whose architecture is generic enough to adapt to multiple settings. We investigate its ability to learn efficient hybrid models for categorizing both multimodal and unimodal data by training it on both text and caption and then evaluate it on different data configurations: Text only (TC_T), Caption only (TC_C) and Text $\oplus$ Caption (TC_TC).

We utilize the same dataset across various configurations. In the TC configurations, both text and caption are used as input during training or testing phases. In the T configuration, only text is used as input to the model, with captions being ignored. Conversely, in the C configuration, only captions are used as input, while text is disregarded. This data split remains consistent across all configurations, encompassing training, validation, and testing phases.

To expand our experiments, we propose a mixed training strategy in which the model is trained on a mixture of data that contains text-only instances ($X_i^{(1)}$) and on instances with Text $\oplus$ Caption ($X_i^{(1)} \oplus \tau(X_i^{(2)})$). The details of our experiments are presented in the following sections.

[2] https://github.com/ruotianluo/self-critical.pytorch.

[3] anonymous

Table 3. Comparison with MARMOT (Wu & Mebane Jr, 2022), a multimodal model integrating caption, text, and image modalities. Two captioning models, namely *Trans_cap* and *CLIP_cc*, are employed. The parameter count (in millions) is reported, excluding parameters associated with captioning models.

	Informativeness			
	Acc	F1-m	F1-w	Parameter count
MARMOT Trans_cap	0.9133±0.0018	0.9012±0.0021	0.9131±0.0018	309
CMB-Trans_cap	0.8920±0.0039	0.8749±0.0045	0.8907±0.0039	112
MARMOT CLIP_cc	0.9220±0.0022	0.9105±0.0024	0.9215±0.0021	309
CMB-CLIP_cc	0.9203±0.0036	0.9081±0.0043	0.9196±0.0037	112

EXPERIMENTAL RESULTS

This section presents the results of our experiments on the proposed method using different captioning models compared with unimodal and multimodal baselines. Additionally, we evaluate the robustness of CMB on unimodal data settings.

CAPTIONING MODELS COMPARISON

Table 2 displays the results of our experiments on the two multimodal tasks of the CrisisMMD dataset (Informativeness and Humanitarian). The first two rows report unimodal classifiers results, while the next two rows present the performances of multimodal baselines. The last four rows show the performances of our approach using each of the four captioning models described before. Firstly, the results obtained confirm that multimodal methods outperform unimodal ones. Additionally, our method is very competitive with other multimodal baselines, regardless of the modality translator used.

Specially, the performance gap between the different captioning models can be significant, as demonstrated by the 3% performance gap on the informativeness task between the method based on Trans_cap and the best-performing method. Importantly, the CLIP_cc captioning model allows the CMB method to outperform the proposed strong baseline of 1-3% in terms of weighted F1-score on both CrisisMMD tasks. Moreover, employing the identical data splits, our experiments yield competitive results with state-of-the-art (Liang et al., 2022) findings as well as results reported on the MVSA dataset. In the tasks of informativeness and humanitarian, we scored 91.96 and 94.84 W-F1, respectively, while their scores were 91.3 and 93.6.

Furthermore, we can observe the gap between the CLIP_cc and CLIP_coco models despite using the same architecture: we hypothesize that the Conceptual Caption dataset (Sharma et al., 2018) has more coverage and similarity with Twitter images, containing figurative representations with different levels of iconicity such as photographs, drawings, charts, or even non-figurative pieces of digital art for instance. Finally, the clear gap between the CLIP-based models and the simple Transformer model confirms the findings of (Bielawski et al., 2022), indicating that CLIP generalizes better on human-centric tasks (defined by the authors as involving knowledge of cultural, social, aesthetic and/or affective components of the world). In our case, this fits particularly our need on social media, where humans post for other humans.

In order to expand our investigations, we compared our approach with MARMOT (Wu & Mebane Jr, 2022) : a model incorporating images, captions, and textual elements for tweets classification. Employing two distinct captioning models, namely Trans_cap and CLIP_cc, the experiments outcomes are shown in Table 3. Despite the diminished parameter count (approximately one-third) in our method relative to MARMOT, it exhibited competitive performance in both scenarios. Noteworthy disparities were observed, particularly in the Trans_cap scenario, whereas the performance differentials were closely aligned when utilizing CLIP_cc captions. This underlines the influence of employing high-quality captions in this configuration. In the continuation of this paper, we use the best captioning model, namely CLIP_cc.

UNIMODAL AND MULTIMODAL RESULTS

As discussed previously, the CMB method contains two parts : a modality translator and a trainable module (text encoder + classifier). The input of the trainable part is exclusively text-based, enabling the use of both unimodal ($X_i^{(1)}$ or $\tau(X_i^{(2)})$) and multimodal ($X_i^{(1)} \oplus \tau(X_i^{(2)})$) data as input. This section investigates the unimodal and multimodal capabilities of the model by conducting an experiment with three main configurations, TC_T, TC_C and TC_TC. Results of the experiments are reported in Figure 2. Performances of CMB (for the different

[4]We used significance level = 0.05

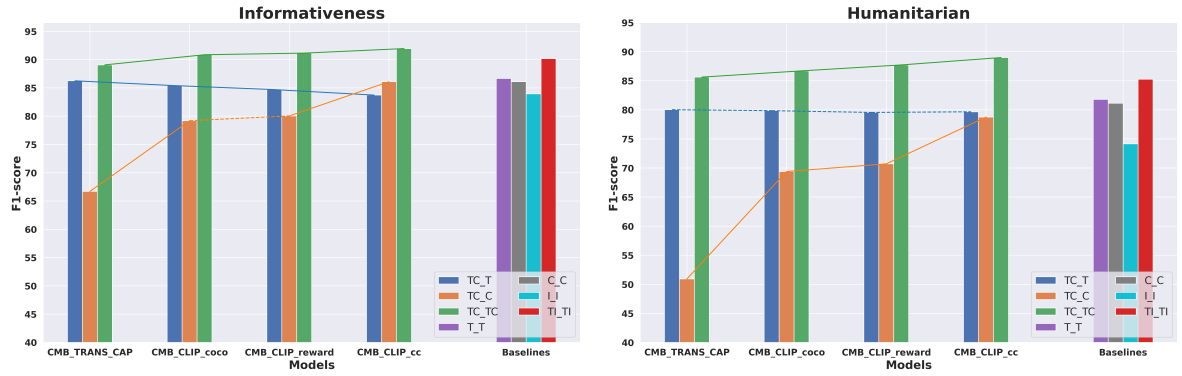

Figure 2. Comparison of modality translators on Unimodal and Multimodal settings (see Table 1). Y-axis shows mean Weighted F1-score results over 10 runs. The X-axis represents the different modality translators used in this experiment. Dotted lines are used for non-significant differences.[4]

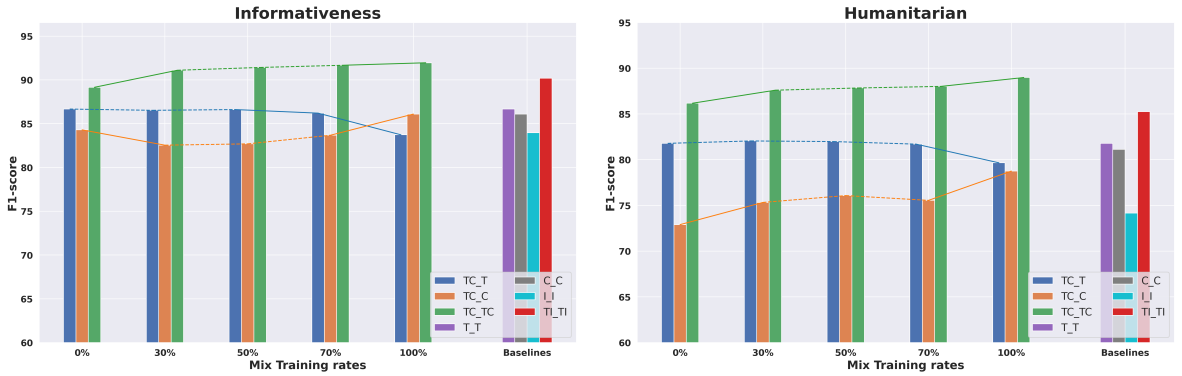

Figure 3. Mix training results. In the X-axis, X% represents the proportion of multimodal instances ($X_i^{(1)} \oplus \tau(X_i^{(2)})$) in training data, the remaining $100 - X$% data are text only ($X_i^{(1)}$). Dotted lines are used for non-significant differences.

captioning models) is compared across the three main configurations as well as to BERT (Devlin et al., 2018) (as T_T configuration), CMB_CLIP_cc (C_C), DenseNet (Huang et al., 2017) (I_I) and the features concatenation model (TI_TI). As mentioned in the previous section, the results clearly illustrate the improved performance when using the two modalities sources for classification. Additionally, these experiments reveal a relationship between the captions-only (TC_C) and multimodal (TC_TC) performances, indicating that the model performance depends on the caption quality. The figures also show that captions-only models can outperform the (I_I) DenseNet model. However, the model in the TC_T setting exhibits poor performance compared to BERT (in T_T configuration) with a 2-3% loss. These results suggest a form of *caption dependency* of the models. To address this issue, we discuss a mix training method in the next section.

MIX TRAINING RESULTS

Twitter posts can be either unimodal or multimodal, but in practice, keywords are used as filters to retrieve tweets with the Twitter API, making some tweets containing only text (without image). Rather than using two different models depending on the nature of each tweet (unimodal or multimodal), we suggest that the CMB model can be used to process both unimodal and multimodal tweets. This hybridization requires an appropriate training strategy, called the *mix training* approach. In this study, we propose an experiment to test our method in TC_T and T_T settings. The goal of this mix training approach is to train a hybrid model that can handle both multimodal (*text+caption*) and unimodal (*text-only*) tweets. To this end, we consider training samples with various *text-only*/*text+caption* balances and observe the ability of the derived models to classify *text-only*, or *text+caption* tweets. In this experiment, the captioning model used is CLIP_cc. The proportions of *text+caption* tweets in the training sample is increased from 0% to 100%, with the remaining examples being *text-only* tweets. The results of the mix-training strategy, as presented in Figure 3, provide interesting insights into the performance of the hybrid model in different scenarios.

Figure 4. Examples of caption using success cases when baseline BERT and Cross-attention failed

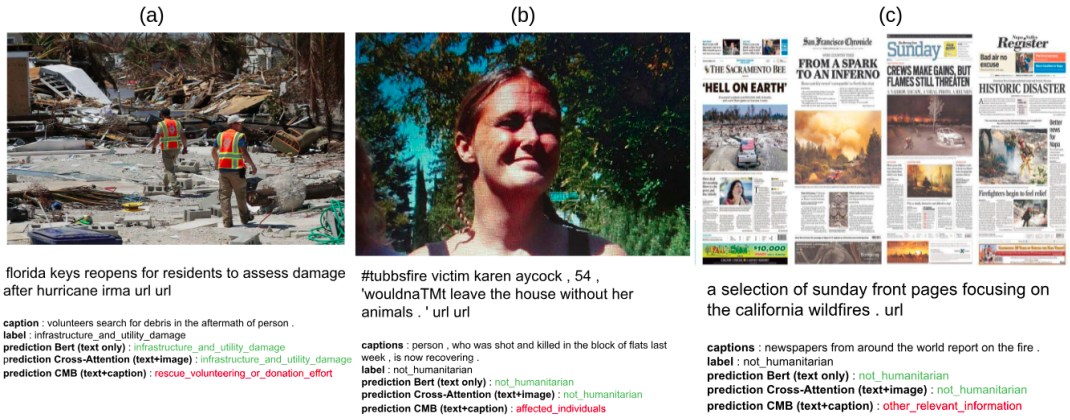

Figure 5. Examples of caption leading to failure cases when baseline BERT and Cross-attention succeed.

When evaluated on text-only tweets (TC_T), the model's performance (F1-score) experiences a slight decrease of approximately 2-3% as the proportion of text+caption tweets in the training sample increases from 50% to 100%. This indicates that the introduction of captions has affected the performance of the model on text-only settings, specially when having more than 70% *text+caption* examples. Additionally, it's evident that the performance in caption-only evaluation (indicated by the orange bar in TC_C) improves as the proportion of *text+caption* examples increases. However, there is one unexplained anomaly observed during the informativeness experiment when transitioning from 0% to 30% of *text+caption* examples.

On the other hand, in the multimodal setting (TC_TC), the same change in the training sample, with an increase in the proportion of text+caption tweets from 50% to 100%, results in an improvement of about 0.5-1% in the model's performance. This suggests that the hybrid model benefits from exposure to a higher proportion of multimodal data during training, leading to better performance in the multimodal tweet classification task. The proposed hybrid model demonstrates its competitiveness with BERT (T_T) when trained on a sample with 50% text+caption tweets, achieving an impressive F1-score of 81.98%. Despite a small loss of performance in the TC_TC setting compared to the text-only scenario, the hybrid model remains clearly superior to the TI_TI baseline, showcasing its capability to effectively handle both unimodal and multimodal tweets. Furthermore, the analysis reveals that the best compromise on caption rates, is achieved at 50%. This suggests that a balanced mixture of text-only and text+caption tweets in the training sample is optimal for achieving peak performance in the hybrid model. Finally, this experiment indicates that the mix training strategy makes the learned models more robust and usable for classifying both unimodal and multimodal tweets in real time.

QUALITATIVE ANALYSIS

Finally, we provide several examples to illustrate the success and failure cases of our proposed method. Figure 4 shows three examples in which the cross-attention baseline and BERT failed, but our method made the correct prediction on the Humanitarian task. In particular, Figure 4(a) is an example of the success of using a caption, as the synthetic image is difficult to be processed by a conventional image model pre-trained on ImageNet (Deng et al., 2009). In figure 4(b) the successful generation of the words "blocks the road" may have contributed to the prediction of the correct label, highlighting the model's ability to capture discriminative information. Finally, Figure 4 (c) is a multimodal example in which the text alone fails to predict the label; in this case, the caption captures information that is merged with the text to achieve the right prediction.

In Figure 5, we present three examples where the caption model produced inaccurate results. The image in Figure 5(a) illustrates the potential of captions to be misleading and result in incorrect predictions; the model captured some information from the image in a poorly structured sentence, resulting in an incorrect prediction. The image in Figure 5(b) demonstrates a failure due to an inaccurate caption generated by the captioning model, highlighting the sensitivity of our approach to the modality translator. Finally, the image in Figure 5(c) shows a situation in which both the label and the prediction fall into a gray area, where either could be considered a valid label. These examples provide insight into the successes and failures of our proposed method, specifying potential areas for improvement.

CONCLUSION

In conclusion, our multimodal classification method based on a modality translation process was found to be competitive with strong baselines and state-of-the-art results on the CrisisMMD dataset. Our analysis confirms that multimodal approaches generally outperform unimodal ones and that the performance gap between different captioning models can be significant. We also found that the caption quality is important for the overall performance of the model, and we proposed a mixed training method to create a hybrid model that can handle both unimodal and multimodal sources. However, our method has several limitations that should be addressed in future work, including the sensitivity on caption quality and the focus on the crisis domain. Overall, CMB is a promising approach for multimodal classification and highlights the importance of considering multiple modalities in tasks involving social media data.

ACKNOWLEDGEMENT

This work was supported by the National Research Agency under the program IA.iO (ANR-20-THIA-0017-01). The computation was performed using Leto resources from CASciModOT federation.

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. (2018). Crisismmd: Multimodal twitter datasets from natural disasters. *Twelfth international AAAI conference on web and social media*.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). Vqa: Visual question answering. *Proceedings of the IEEE international conf. on computer vision*, 2425–2433.

Bielawski, R., Devillers, B., Van De Cruys, T., & Vanrullen, R. (2022). When does CLIP generalize better than unimodal models? when judging human-centric concepts. *Proceedings of the 7th Workshop on Representation Learning for NLP*, 29–38. https://doi.org/10.18653/v1/2022.repl4nlp-1.4

Chen, S., Aguilar, G., Neves, L., & Solorio, T. (2020). Can images help recognize entities? a study of the role of images for multimodal ner. *arXiv preprint arXiv:2010.12712*.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2020). Uniter: Universal image-text representation learning. *Computer Vision–ECCV 2020: 16th European Conf., Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, 104–120.

Cho, J., Yoon, S., Kale, A., Dernoncourt, F., Bui, T., & Bansal, M. (2022). Fine-grained image captioning with clip reward. *arXiv preprint arXiv:2205.13115*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

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

Fan, Y., Xu, W., Wang, H., Wang, J., & Guo, S. (2023). Pmr: Prototypical modal rebalance for multimodal learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20029–20038.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. *Proc. of the IEEE int. conf. on computer vision*, 2961–2969.

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

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

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

Kumar, S., Barbier, G., Abbasi, M. A., & Liu, H. (2011). Tweettracker: An analysis tool for humanitarian and disaster relief. *Proceedings of the International AAAI Conference on Web and Social Media*.

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*, 15492–15501.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *European conf. on computer vision*, 740–755.

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. *CoRR*, *abs/1907.11692*.

Long, Z., & McCreadie, R. (2022). Is multi-modal data key for crisis content categorization on social media? *ISCRAM 2022 Conference Proceedings 226 19th International Conference on Information Systems for Crisis Response and Management*, 1068–1080.

Mokady, R., Hertz, A., & Bermano, A. H. (2021). Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Nguyen, D. T., Ofli, F., Imran, M., & Mitra, P. (2017). Damage assessment from social media imagery data during disasters. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 569–576. https://doi.org/10.1145/3110025.3110109

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

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 8748–8763.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics*, 2556–2565.

Sosea, T., Sirbu, I., Caragea, C., Caragea, D., & Rebedea, T. (2021). Using the image-text relationship to improve multimodal disaster tweet classification. *The 18th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2021)*.

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

Tuan, N. M. D., & Minh, P. Q. N. (2021). Multimodal fusion with bert and attention mechanism for fake news detection. *2021 RIVF Int. Conference on Computing and Communication Technologies (RIVF)*, 1–6.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vempala, A., & Preoţiuc-Pietro, D. (2019). Categorizing and inferring the relationship between the text and image of twitter posts. *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, 2830–2840.

Wang, W., Tran, D., & Feiszli, M. (2020). What makes training multi-modal classification networks hard? *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12695–12705.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., . . . Rush, A. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Wu, P. Y., & Mebane Jr, W. R. (2022). Marmot: A deep learning framework for constructing multimodal representations for vision-and-language tasks. *Computational Communication Research*, *4*(1).

Yang, X., Feng, S., Wang, D., & Zhang, Y. (2020). Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia*, *23*, 4014–4026.

Zellers, R., Bisk, Y., Farhadi, A., & Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. *Proceedings of the IEEE/CVF conf. on computer vision and pattern recognition*, 6720–6731.