CrisisTS: Coupling Social Media Textual Data and Meteorological Time Series for Urgency Classification

Romain Meunier[1], Farah Benamara[1,2], Véronique Moriceau [1], Zhongzheng Qiao[3,4,5], Savitha Ramasamy[3,4]

[1] IRIT, Université de Toulouse, CNRS, Toulouse INP, Toulouse, France

[2] IPAL, CNRS-NUS-A*STAR, Singapore

[3] Institute for Infocomm Research (I^{2}R), A*STAR, Singapore

[4] CNRS@CREATE LTD, Singapore

[5] IGP-ERI@N, Nanyang Technological University, Singapore

Correspondence: romain.meunier@irit.fr

Abstract

This paper proposes CRISISTS, the first multimodal and multilingual dataset for urgency classification composed of benchmark crisis datasets that have been mapped with open source geocoded meteorological time series data. This mapping is based on a simple and effective strategy that allows for temporal and location alignment even in the absence of location mention in the text. A set of multimodal experiments have been conducted relying on transformers and LLMs to improve overall performances while ensuring model generalizability. Our results show that modality fusion outperforms text-only models.

1 Introduction

Learning with multiple modalities has become an active line in the research community where current multimodal models have achieved excellent results outperforming their unimodal counterpart in many application scenarios (Xu et al., 2023; Yin et al., 2023). For example, visual/layout information helps to better capture long-range dependencies in automatic summarization (Nguyen et al., 2023) while combining texts with images and/or videos helps over text-based systems in many subjective tasks such as sentiment analysis (Liang et al., 2022) and hate speech detection (Hee et al., 2024).

Multimodal fusion mainly concerns texts, images, videos and audio data, where at most two or three modalities are fused. However, coupling texts with tabular-based time series (hereafter TS) to improve performances of NLP applications has received less attention. Among existing works, Dang et al. (2019) use TS and texts from newspapers to predict which news are the most relevant and Deznabi et al. (2021) clinic notes and TS from medical devices to predict in-hospital mortality. Finally, Conforti et al. (2022) leverage textual and financial signals for stance detection in the financial domain. In this paper, we newly investigate the role of TS in NLP-based crisis management (CM).

During crises such as floods or earthquakes, both urgent (human/infrastructure damages, displaced people, security instructions, etc.) and not-urgent (critics, support, etc.) messages are posted on social media platforms which provide crucial information for various stakeholders like humanitarian and secure organizations to set priorities and decide appropriate actions (Imran et al., 2016; Vieweg et al., 2014; Reuter and Kaufhold, 2018; Sarioglu Kayi et al., 2020; Kozlowski et al., 2020). Urgency detection is either framed as a binary *utility classification* task to filter-out not-relevant from relevant messages to rescue teams, a three-class *urgency classification* (is the message urgent, not-urgent or not-relevant), or a multiclass *humanitarian information types classification* such as caution, advice and people missing. Several unimodal datasets have been developed to address these tasks in (semi-)supervised settings in different languages. Well known datasets in the field include TREC-IS (McCreadie et al., 2019, 2020), HumAID (Alam et al., 2021), and CrisisFACTS (McCreadie and Buntain, 2023).

Urgency detection from short and ill-formed social media content is very challenging due to two main reasons: First, urgent messages are scarce (e.g., about e.g., about 3.87% and 1.93% for infrastructure/human damages respectively in CrisisFACTS), making crisis datasets highly imbalanced towards the not-relevant messages that often contain keywords related to a crisis (e.g., "*This place gonna be on fire tonight*" posted during a music festival), (2) Models need to generalize well to new unseen events that lack annotated data. Three evaluation settings are usually employed to measure portability in real application scenarios: (a) *On-event* where models are trained/tested on data from the same event, e.g., Turkey earthquake, (b) *Out-of-event* that requires training on one or several events

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16082–16099

from various types (e.g., flood, hurricanes) and testing on unseen events of the same type (e.g., flood) that occurred in different localization/dates, and (c) *Out-of-type* where models are trained on events related to different types of crises (e.g., hurricane, earthquake) and tested on a particular different type (e.g., storm). In this last setting, portability to sudden crises (e.g., terror attacks, building collapse, explosions) has shown to be particularly difficult where a significant drop in performances has been observed when compared to expected crises that can be anticipated from e.g., weather forecast (Kersten et al., 2019; Wiegmann et al., 2020; Wang et al., 2021a; Li et al., 2021; Bourgon et al., 2022).

One solution to address these issues is multimodality.[1] Text and image fusion for urgency detection has shown very promising results (Alam et al., 2018; Agarwal et al., 2020; Abavisani et al., 2020; Wu et al., 2022; Basit et al., 2023; Koshy and Elango, 2023; Farah et al., 2024; Bouabid and Farah, 2024). We continue this line of research by exploiting for the first time as far as we know, multivariate TS, i.e that have multiple dimensions recorded over time. TS data (e.g., rainfall, water level) has shown to be reliable sources of information to detect and monitor crises (Zeng and Bertsimas, 2023; de Bruijn et al., 2020). Here we go one step further by first aligning TS with textual data then injecting them into deep learning models to improve urgency classification in out-of-type and out-of-event scenarios. Our contributions are:

(1) CrisisTS, the first multimodal and multilingual dataset for urgency detection composed of benchmark datasets about various expected and sudden crises in French and English that have been mapped with open source geocoded meteorological data.

(2) A simple yet effective temporal and location alignment strategy that allows TS-text mapping even in the absence of location mention in the text, which allows us to cover a huge variety of messages, going beyond existing text-TS alignment strategies that only rely on one location mention within a message.

(3) A set of experiments for multimodal urgency detection relying both on transformers and large language models. To show the portability of our approach, we tested our models on **two languages and 29 crises from 7 types**. Our results show that injecting TS data from various sources improves over text only models across different languages and types of crisis especially on non sudden crisis. The dataset including the alignment strategy are available to the community.[2]

In the following, Section 2 summarizes the related work. Section 3 presents the CrisisTS dataset as well as the alignment strategy. Section 4 presents the experimental settings, models and our results. Finally, we conclude drawing some perspectives for future work.

2 Related work

2.1 NLP for Crisis Management

In crisis situations, real-time textual information can come from the emergency services (Otal and Canbaz, 2024) but also from social media such as X (Twitter) (Reuter et al., 2018) (e.g., more than one million tweets posted during 2023 Turkey–Syria earthquakes (Toraman et al., 2023)). In order to collect, extract or summarize this information, NLP-based crisis management has become a hot research topic in text classification where posted messages are classified into different categories, named entity recognition to detect location mention that helps geolocalise information needs (Suwaileh et al., 2023), and event detection (Rajaby et al., 2022).

Text-based classifiers are mainly trained in a supervised way with either traditional feature-based learning algorithms (Li et al., 2018a; Kaufhold et al., 2020; Alam et al., 2021) or deep learning architectures (Caragea et al., 2016; Castillo, 2016; Neppalli et al., 2018; Kersten et al., 2019; Kozlowski et al., 2020; Chowdhury et al., 2020b; Liu et al., 2021; Wang et al., 2021b; Dusart et al., 2021). More recent works use generative AI (Otal and Canbaz, 2024). Overall, results show that humanitarian categories detection is the most challenging task due to the extremely imbalanced nature of crisis datasets.

2.2 Time Series in Crisis Management

Time series data are ubiquitous in various domains like in finance, energy and public health. TS forecasting is crucial in many applications and consists in predicting future events or trends based on historical data (Benidis et al., 2022). State-of-the-art models range from statistical methods (Huang et al., 2018) to deep learning models such as LSTM (Box

[1] Text-based data augmentation has also been explored (Bayer et al., 2021; Chowdhury et al., 2020a; Meunier et al., 2023) but this is out of the scope of the paper.

[2] https://huggingface.co/datasets/Unknees/CrisisTS

et al., 2015), CNN (Bai et al., 2018) or hybrid models like CNN-LSTM that has been used to predict the effect of a flood (Malik et al., 2024). TS pretrained models have also been largely employed (Fawaz et al., 2018; Wen et al., 2022; Zhang et al., 2024). Large language models applied to TS also start to emerge such as TimeLLM (Jin et al., 2023) and LagLLAMA (Rasul et al., 2023).

In the field of crisis management, time series with meteorological data seems to be a natural ally to predict a large range of crises in order to prepare the population: For example, rainfall data is used to predict daily precipitation during typhoons (Huang et al., 2018), topological data to predict landslides (Yuan and Moayedi, 2020), or weather parameters to predict floods (Sankaranarayanan et al., 2020). In the context of crisis early detection, Li et al. (2018b) trained a GAN with waveform features to mitigate false alerts of earthquake while Moon et al. (2019) use logistic regression for early-detection of heavy rainfalls. Sudden crises have also received a special attention. For instance, Van Le et al. (2021) used geographic information system (GIS) database with deep learning in order to predict the risk of wildfire in tropical climate. Finally, the use of LLMs for TS crisis management is relatively new, see for example Zhu et al. (2024) who developed an LLM enriched with flood knowledge that can interact with a GIS to enhance the public's perception of flood risks.

2.3 Time Series and Texts for Crisis Management

Roughly, two major multimodal fusion training exist: *Early fusion* (Iyengar and Nock, 2003) where features from different modalities are merged into a single representation before training (e.g. via embeddings concatenation), and *Late fusion* (Azimi-Sadjadi et al., 2000) where each modality is processed separately (causing a higher computation cost) and the decision from each model are then merged (e.g. mean of the outputs). Vielzeuf et al. (2018) proposed a hybrid fusion mechanism while recent generative models have been fine-tuned with multimodal instructions (Moon et al., 2023).

TS and texts are often temporally aligned and the way this alignment is performed differs according to the task. To predict patient mortality, Deznabi et al. (2021) first train a BERT model and an LSTM on TS then employ a late fusion. Conforti et al. (2022) use a multi-view model where each modality has its own task: stance prediction for texts and financial prediction for TS. Finally, Khadanga et al. (2019) use late fusion to predict patient mortality and the length of stay in intensive care units from clinical notes and TS signals recorded by monitoring instruments. This requires to align discrete text events to continuous TS signals. More recently, LLM have been developed for both time series and texts, with the objective to improve the performance on TS analysis (Chan et al., 2024) or forecasting (Jia et al., 2024; Kim et al., 2024). Our goal here is the opposite: use TS to improve text classification, making existing models not adequate for our task.

When it comes to crisis management, TS and texts can be used to improve crisis prediction from TS data as Cerna et al. (2022) who use NLP techniques to recognize periods with peak interventions in rare events. However, combining these two modalities to improve crisis classification in text messages is new. As far as we know, the only work in this direction has been reported in de Bruijn et al. (2020) who combine rainfall statistics and tweets for utility prediction during floods (i.e., is the tweet about a flood event?). TS and texts are aligned temporally and spatially where a unique location mention within the tweet is mapped with a given latitude and longitude from the TS. This paper extends this initial work by: (1) Considering several types of crises both expected and sudden as annotated in benchmark datasets, (2) Proposing a novel alignment strategy that goes beyond explicit location mentions, (3) Experimenting with early fusion strategies as well as multimodal LLMs, and finally (4) Evaluating models portability to unseen events.

3 The CrisisTS Dataset

3.1 Textual Data

Datasets Selection. We chose the following datasets, with and without annotated location mentions in order to evaluate the feasibility of textual and TS data alignment:

- **KOZLOWSKI**: It is the largest corpus of French tweets annotated for crisis (Kozlowski et al., 2020) and augmented later on by Bourgon et al. (2022). It is composed of 7 types of crisis (Fire, Flood, Storm, Hurricane, Collapse, Explosion, Attack) with several crises which occurred in France such as Notre-Dame fire or flood in the Aude region. It contains tweets, collected 24h before, during (48h) and up to 72h after the crisis, manually annotated for three

urgency categories as well as 6 intent to act categories (similar to humanitarian categories): (1) URGENT that applies to messages mentioning HUMAN/MATERIAL DAMAGES as well as security instructions (ADVICE-WARNING) to limit these damages during crisis events, (2) NOT URGENT that groups SUPPORT messages to the victims, CRITICS or any OTHER messages that do not have an immediate impact on actionability but contribute in raising situational awareness, and finally (3) NOT USEFUL for messages that are not related to the targeted crisis. From the original dataset, we removed 3,628 tweets that are not annotated with intent/humanitarian categories. Note that in this dataset, location mentions are not annotated (see Section 3.3 for the location identification).

- **IDRISI-RE**: This dataset is the largest publicly available for location mention prediction in crisis management (Suwaileh et al., 2023). It is composed of 20,514 tweets in English from the HumAID dataset (Alam et al., 2021), manually annotated for humanitarian categories and location mentions, covering diverse disaster types and geographic areas around the globe. Since IDRISI-RE only contains USEFUL tweets, we have added English NOT USEFUL tweets from HumAID. In the public distribution of HumAID, there are no NOT USEFUL tweets for flood and fire, and only 209 tweets for earthquake. Concerning hurricanes, we have selected all the NOT USEFUL tweets for only Irma hurricane because this crisis is geolocated in only one state in the US (Florida), making possible a direct alignment with TS location.

Datasets Pre-processing. For both datasets, we assigned to each tweet one of the 3 following labels: NOT-CRISIS for tweets labelled as NOT-USEFUL, NOT-SUDDEN for tweets labelled as USEFUL and related to storms, hurricanes and floods, and SUDDEN for tweets labelled as USEFUL and related to fires, collapses, earthquakes and terrorist attacks. It is important to note that in real life, the type of crisis is not known in advance. Therefore, meteorological data are useful to give us information on which kind of crisis (sudden or not) we are facing and then give us a better comprehension of the information conveyed in the tweet which helps in urgency classification.

In addition, and in order to conduct the same experiments on French and English datasets, a unified annotation scheme has been used: We have automatically assigned to the English tweets the URGENT, NOT URGENT and NOT USEFUL labels based on their manually annotated humanitarian categories in order to be able to perform both urgency and utility tasks.

3.2 Time Series Data

Datasets Selection. We created two TS datasets sourced from open government data coming from meteorological stations for two reasons: (a) end users, such as emergency services rely on these types of data, (b) the alignment method can be easily portable to open data from other countries.

- **FRENCHTS**: We have collected meteorological data on Meteo France website.[3] These data are geographically related to one of the 65 French meteorological stations, have a frequency of 3 hours and 39 features, starting on 01/01/2007 00h00m00s and ending at 11/30/2022 21h00m00s for a total of 46,495 timestamps per station.

- **ENGLISHTS**: We have collected meteorological data for geographic areas related to IDRISI-RE crises but only US and New Zealand data from both websites of the National Oceanic and Atmospheric Administration (NOAA)[4] and the National Institute of Water and Atmospheric research (NIWA)[5] could be freely collected. These data are geographically related to a region or state (e.g. Kaikoura for New Zealand, or the state of Maryland for US) and all these data are daily summaries (daily frequency). They start on 01/01/2016 and end at 12/31/2019 for a total of 1,460 timestamps per state.

Datasets Pre-processing. From both datasets, we removed some features that are not relevant for crisis management (e.g. wind direction since it gives no relevant information contrary to wind strength) and features with a high rate of missing values (e.g. 45% for cloud percentage in FRENCHTS). The selected features and their associated metrics are shown in Table 8 in Appendix B. Finally, all features were normalized using the StandardScaler function from Scitkit-learn. A detailed description of these features and their relevance for the task is provided in Appendixes B.1 and B.2.

3.3 Multimodal Alignment Strategy

Textual and time series datasets are aligned geographically and temporally, as shown in Figure 1.

[3] https://donneespubliques.meteofrance.fr/

[4] https://www.ncei.noaa.gov/cdo-web/

[5] https://niwa.co.nz/climate-and-weather

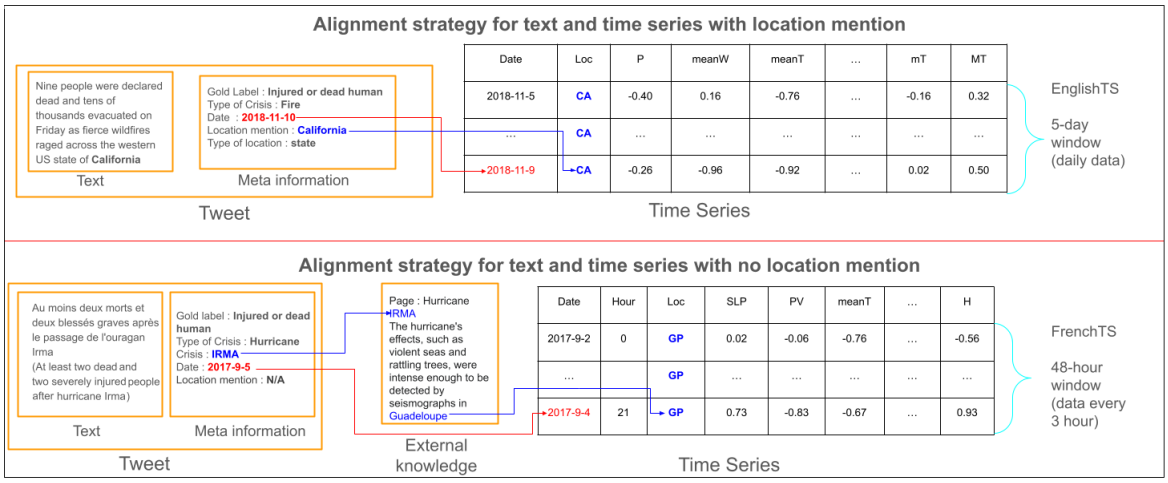

Figure 1: Location and temporal alignment of tweets and time series to create a multimodal dataset. Top (bottom): Alignment in case of known (unknown) tweet location.

Our methodology has to deal with ambiguity. Indeed, TS we were able to collect do not cover all the crises present in our datasets because the corresponding TS were not freely available (TS could not be collected only for a wildfire in Canada which represents around 4.8% of the whole IDRISI-RE). Ambiguity can also come from location mentions within the tweet (e.g., named entities that refer to different cities). However, our method does not depend on location mention to align the data, which allows us to cover a huge variety of tweets, going beyond existing TS-text alignment strategies that only rely on one location mention.[6] We explain below how we overcome these difficulties.

Location alignment. Since the TS datasets have been collected from meteorological stations, the location of each time series is known (either a city in France, or a state in the rest of the world). For tweets however, there are two alignment methods depending on whether the location mention of the tweet is available or not:

– **(Case a)** *Exact location mention*. If the location mention refers to a unique location in the globe, the tweet is associated to TS data from the state the location belongs to. Otherwise, the tweet is discarded. For example in Figure 1, a tweet mentioning the city *Fontana* is associated with TS from California, whereas a tweet mentioning the city *Florence* is removed since there are 7 cities named *Florence* in the USA and 1 in Italy, so it is impossible to link the tweet to a specific state.

This strategy is applied to geographically align IDRISI-RE with ENGLISHTS. It is important to note that 10% of tweets from this dataset have been removed due to ambiguous location mentions. However, in real life situations, meteorological data can also help in disambiguating these mentions (e.g., a tweet mentioning storm in *Florence* but no TS in Italy indicate an ongoing crisis).

– **(Case b)**: *Unknown in case of the absence of location mention or presence of several mentions.* For example, in *"After having swept across the Atlantic coast, the Bruno storm moves towards Corsica"*, the relevant mention is *Corsica* since the location *Atlantic coast* refers to a past event. A possible solution is to use an automatic named entity extraction which can be quite effective in case of a unique location mention, but this would lead to some difficulties as around 63% of our tweets contain either no or several mentions.

Instead, since each tweet has been associated during scraping to a crisis that can be geolocated, the location mention is manually extracted from the Wikipedia pages of the 29 crisis events and all the tweets for a given event are linked to the corresponding or the closest meteorological station. When the location mention in Wikipedia is a fuzzy area, all the tweets of the crisis event are linked to the meteorological station of the administrative capital of the area. For example, the tweet *wind gusts up to 120km/h recorded in the north of the country* will be linked to the TS from the station Lille-Lesquin.

Temporal alignment. Once a tweet is associated to a location and to geolocated time series, they have then to be temporally filtered out. Given that

[6] A disambiguation module can be useful in these cases and is left for future work.

ENGLISHTS and FRENCHTS do not have the same frequency, this implies fixing a window size to select the appropriate TS.

Considering that the tweets of the French dataset have been scrapped with a 48-hour window before the crisis, we have applied the same window to FRENCHTS resulting in the selection of 16 timestamps (1 timestamp every 3 hours) before the posting date of each tweet. Applying a 48-hour window to ENGLISHTS, which has a daily frequency, would select only 2 timestamps for each tweet which is not enough to be relevant. On the contrary, selecting 16 timestamps as for French would select the time series data for 16 days before the crisis, which may lead to noisy data. After some experiments with different window sizes, we applied a 5-day window for ENGLISHTS as this was the window that gave the best results.

Quality of alignment. Since Case (a) is based on manually annotated and exact location mentions, we considered these alignments as gold alignments. Case (b) however has to be evaluated (this represents 69.46% of tweets (15,368) in our multimodal dataset). To this end, we performed a manual check of the quality of the alignment on a randomly selected subset of 1,150 useful tweets that have been linked to their administrative locations:

–*Location alignment*: We observed 558 tweets with a spatial shift between 0-50 km, 334 tweets between 50 and 150 km, 235 tweets between 150-300 km and 23 tweets beyond 300 km. The maximum shift is 687 km for a mean spatial shift of 114 km.

An outlier analysis (i.e., spatial shift >300 km) shows that outliers were due to a case where two storms were occurring in France at the same time but at different locations. However, thanks to the fact that these two crises were about the same type, these tweets were still associated with meteorological data that describe the same kind of crisis (storms). Finally, when looking into the spatial shift per crisis, we found that all the sudden crisis tweets we verified have a shift between 0-50 km, due to the fact that our sudden crisis tweets are about a local event.

–*Temporal alignment*: An analysis of the impact of a spatial shift on temporal alignment reveals that for a 0-50 km spatial shift, there is no temporal shift ; between 50-150 km an average time shift of 1 time stamp and finally for tweets between 150-300 km, there is an average time shift of 2 timestamps.

CrisisTS Statistics. The distributions of the datasets after alignments are presented in Table 4 and Table 5 (see Appendix A).

4 Multimodal Urgency Classification

4.1 Experimental Settings

We designed two experimental settings to better evaluate the performances in real scenarios (Tables 4 and 5 from Appendix A give the size of each train/test sets, computing infrastructure is detailed in Appendix C):

(1) ***Out-of-event.*** It aims to evaluate if a model can deal with new crisis events with a known type. For this set-up, we defined the three sets of events from KOSLOWSKI (cf. Appendix A). The models are successively tested on each of these three sets while trained on all events that are not present in the chosen test set. To make sure the improvements are not due to randomness or a training artifact, all the models have been run trough three Out-Of-event scenario, for a total 9 experiments. The results are then the mean of the results obtained for each experiments. Due to the low amount of events for each crisis type in IDRISI-RE, we were not able to carry out an *out-of-event* evaluation for the English data.

(2) ***Out-of-type.*** It aims to evaluate if a model can deal with new types of crisis, which is crucial to ensure the portability of the models to unseen events. Thus for each dataset, this second consists in the average of n runs, each run with $n-1$ crisis types for training and the remaining crisis type for testing. All the models have been run through 2 Out-Of-type scenario, for a total 10 experiments. The results are then the mean of the results obtained for each experiments. It is important to note that this second setting is more challenging and is the closest to real world crisis events.

4.2 Models

We designed textual baselines and multimodal models. Since our task is to classify a tweet, a TS unimodal baseline is not relevant. Note that TimeLLM (see below) heavily relies on TS prediction.

(a) Unimodal Text-based Models. We rely on transformers and LLMs, as follows:

- **FlauBERT**$_{FineTuned}$: It is a FlauBERT model (Le et al., 2020) that has been fine-tuned on 358,834 unlabelled tweets posted during crises. Due to an imbalanced dataset, the Focal Loss and Adam optimizer are used. It is reported to be the best

performing model for intent classification on the French KOSLOWSKI dataset (Meunier et al., 2023). We use it as our baseline for the French dataset.

- **FlauBERT**$_{FineTuned-3Tasks}$: It is the current state-of-the-art multitask model on KOSLOWSKI (Bourgon et al., 2022). This model is trained for 3 tasks: detection of utility, urgency and intents.

- **RoBERTa** (Liu et al., 2019): This is a baseline for the English dataset as it is reported to be efficient for crisis management in English (Koshy and Elango, 2023; Rocca et al., 2023; Madichetty et al., 2023). The Focal Loss and Adam optimizer are also used. **RoBERTa**$_{3Tasks}$ is a multitask version of the model trained for 3 tasks: detection of utility, urgency and humanitarian categories.

- **RoBERTa+Sudden** and **FlauBERT**$_{FineTuned}$ **+Sudden.** In order to evaluate the impact of the crisis type (sudden or not) on the global performance, we also trained both FlauBERT and RoBERTA baselines in a multitask architecture, with an auxiliary task to detect the SUDDEN and NOT SUDDEN crises.

- **LLMs.** We rely on Mistral$_{7B}$ (Jiang et al., 2023) and Llama3$_{8B}$ (AI@Meta, 2024) for French and English respectively in a few-shot setting, where 5 examples for each label were randomly chosen from the train set.[7] To avoid any bias, both LLMs use the same prompt (see Appendix D). Each tweet to classify is tested 5 times with different examples. The final predicted label is obtained by a majority vote.

(b) Multimodal Text+TS Models. To evaluate the contribution of time series to the performance of models for all tasks, we trained a multimodal version of each model:

- **RoBERTa + Early Fusion** and **FlauBERT**$_{FineTuned}$ **+ Early Fusion.** Due to the fact that our TS are not labeled, we can only use them in an early fusion configuration. We therefore rely on embeddings concatenation and this fusion strategy is applied to all transformer text-based models including their multitask versions.

- **MM-TimeLLM.** Our goal is to use TS to improve text classification, making existing multimodal models not adequate for our task. A possible alternative could be to train a new multimodal LLM but this requires a huge amount of both textual and TS data. Therefore, we have adapted TimeLLM (Jin et al., 2023), a time series model into a multimodal model, MM-TimeLLM, with the objective to evaluate the impact of multimodal data on a textual classification task. We then compare unimodal LLMs with MM-TimeLLM.

As a pioneering approach in multimodal time series analysis, TimeLLM utilizes a frozen LLM backbone for time series forecasting. Like PatchTST (Nie et al., 2023), it follows the channel-independence assumption and segments the time series data into distinct patches. To enable the LLM to understand time series data effectively, the model reprograms these time series patches using pretrained word embeddings from the LLM backbone. To incorporate prior knowledge and textual information, TimeLLM generates a prompt for each time series instance based on designed template. The output embedding of the prompt from the LLM backbone is then used as a prefix to prepend the input embeddings. During training, the LLM remains frozen, and only the reprogramming layer and output layer are updated.

In our adaptation of TimeLLM (see Figure 2), we replace the output projection layer with a classification layer to suit our classification task. In addition, we use our tweet text data for prompting instead of the original template-based prompt (cf. Appendix D). This study utilized Llama 2.

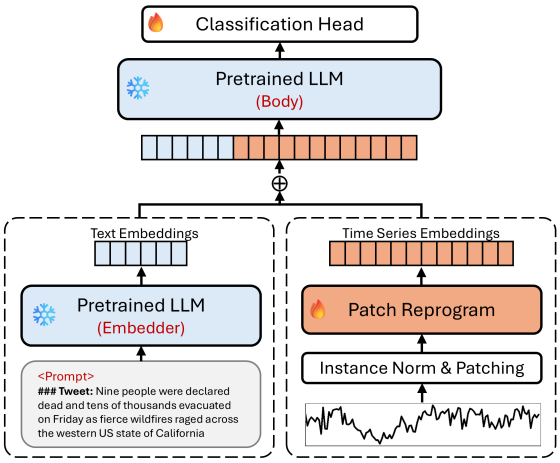

Figure 2: A modelisation of MM-TimeLLM.

4.3 Results

Tables 1 and 2 present the results on the French and English datasets respectively. The best results for each model are in bold and the best result for each task is underlined.

For both English and French, the best results are obtained with a multimodal and multitask model

[7] We also tested other open source LLMs but we only report those that achieved the best scores.

	Model	Out-of-Event			Out-of-Type		
		Main task					
		Utility	Urgency	Humanitarian	Utility	Urgency	Humanitarian
Baselines	(1): FLAUBERT$_{FineTuned}$	59.20	52.39	54.34	**74.68**	63.14	47.34
	(2): FLAUBERT$_{FineTuned-3Tasks}$	**75.35**	**67.11**	**56.24**	74.63	**63.87**	**49.99**
Unimodal models	(3): (1) + SUDDEN TASK	72.73	63.87	51.57	74.44	60.72	44.93
	(4): (2) + SUDDEN TASK	**75.64**	**67.11**	**55.44**	**76.36**	**64.18**	**49.28**
	MISTRAL	69.88	49.35	41.41	69.46	48.16	38.37
Multimodal models	(1) + Early Fusion	74.67	65.23	55.22	74.56	63.23	49.83
	(2) + Early Fusion	74.38	66.13	56.34	75.10	63.73	**51.15**
	(3) + Early Fusion	74.61	65.17	55.56	75.77	64.73	50.52
	(4) + Early Fusion	**75.03**	**67.26**	**56.79**	**75.81**	**64.68**	49.72
	MM-TIMELLM	72.56	28.80	12.05	72.56	56.66	14.72

Table 1: Results on the French dataset in terms of average F-score. All the results are statistically significant at $p < 0.01$ using the McNeymar test.

	Model	Main task		
		Utility	Urgency	Humanitarian
Baselines	(1): ROBERTA	75.55	75.17	72.03
	(2): ROBERTA$_{3Tasks}$	**76.62**	**78.89**	**75.50**
Unimodal models	(3): (1) + SUDDEN TASK	72.77	**77.55**	74.40
	(4): (2) + SUDDEN TASK	**75.73**	76.18	**75.18**
	LLAMA3	48.42	39.19	43.93
Multimodal models	(1) + Early Fusion	73.71	76.58	73.24
	(2) + Early Fusion	77.50	79.46	**74.78**
	(3) + Early Fusion	73.13	76.93	74.56
	(4) + Early Fusion	**77.98**	**79.58**	74.09
	MM-TIMELLM	52.83	41.29	17.47

Table 2: Results on the English dataset in terms of average F-score (*out-of-type* experiment only). All the results are statistically significant at $p < 0.01$ using the McNeymar test.

((4) + Early Fusion) for the urgency task showing that time series data and the SUDDEN vs. NOT SUDDEN knowledge contributes to improve urgency detection. Regarding the performances of the LLMs, the multimodal MM-TimeLLM outperforms the unimodal LLMs (Mistral and Llama3) except on the humanitarian task where its performance drops drastically: In MM-TimeLLM, TS have a more important weight than textual data and this drop suggests that meteorological data are less relevant than texts to detect the humanitarian categories (type of damages, critics, support, etc.). We also note that MM-TimeLLM performs better on French data which have a higher frequency (every 3 hours instead of daily for English).

On the French dataset, we note that all multimodal models using TS data perform better than their unimodal versions for the intent and urgency detection tasks (except FlauBERT$_{FineTuned-3Tasks}$ for urgency). Considering the English dataset, multimodal models outperform only their unimodal version for the RoBERTa baseline and the multitask RoBERTA. The lower impact of TS on the English dataset can be explained by the fact that ENGLISHTS has a lower frequency than FRENCHTS leading to a lower quantity of relevant data. We also note that the results on the utility task are lower than on the urgency and humanitarian tasks. This is due to the lack of NOT USEFUL tweets in the corpus (388 NOT USEFUL tweets vs. 6,356 USEFUL).

Looking at the detailed results per class for the urgency task (see Table 13 in Appendix E), we observe that multimodality increases the performance on the NOT-USEFUL class in English as well as French. In French, an important increase is also noticed for the NOT-URGENT class. Indeed, abnormal meteorological data bring information about the existence of a crisis situation and will help to better detect NOT-USEFUL messages: for example, the tweet *There is a leak in my kitchen, everything is flooded!* would be classified as URGENT by a unimodal model based on text whereas it would be considered as NOT USEFUL by a multimodal model if there is no meteorological anomaly detected. This is confirmed in Table 14 showing better results of the multimodal models for all classes of the utility task in both languages.

Considering our assumption that meteorological data can help to better distinguish not sudden (i.e. meteorological expected conditions) from sudden crises, as expected adding TS data does not increase the performance for sudden crises in both languages (cf. Table 15) whereas performances are improved for not sudden crises in all classes and both languages (cf. Table 16). When looking at the results per crisis (cf. Tables 17 and 18), we notice that multimodal models improve performances for all types of not-sudden crises, except the building collapse. We believe this is due to the fact that meteorological data does not detect any crisis whereas textual information does, which is a typical situation in a sudden event. The only crisis type that shows a loss of performance is fire, probably because in our dataset, some fires are caused by meteorological conditions (e.g. Landes) while others are accidental (e.g. Notre-Dame).

5 Error Analysis

5.1 Impact of the Quality of Alignment

In order to estimate the impact of a spatial shift during location alignment - possibly causing a temporal shift - on classification, we ran a test in an out-of type configuration on the humanitarian task for storm as this is the crisis with the highest shift. For tweets with no temporal shift, 26.66% of tweets were misclassified; with a shift of 1 timestamp 27.92% were misclassified, whereas with 2 timestamps, 31.25% were misclassified. This shows that the performances are quite similar for one timestamp vs. zero shift. Considering that 77.56% of the manually checked tweets have a spatial shift under 150 km, corresponding to a temporal shift of 1 timestamp, alignment errors have a small impact on the classification performances.

5.2 Impact of Time Series Frequency

To better analyze the impact of TS frequency, we experimented by artificially reducing the frequency of FrenchTS to a 6 hour, then 12 hour frequency (instead of a 3 hours). To this end, we follow the out-of type evaluation setting relying on our best model, namely **FlauBERT**$_{FineTuned-3Tasks}$**+Sudden.** The results in terms of F1-scores are presented in Table 3. They show that for the utility and urgency tasks, the lower the frequency is, the lower is the performance.

The humanitarian task however does not follow this scheme where we observed an increase in performances. The results per class show that the increase concerns not-urgent messages (e.g., CRITICS, SUPPORT) while the decrease concerns urgent messages (e.g., HUMAN/MATERIAL DAMAGES). For example, the results for HUMAN DAMAGES are F1=58.80 (resp. 62.31) for 12 hour (resp. 3) frequency while F1=38.54 (resp. 30.39) for 12 hour (resp. 3) frequency for CRITICS. Given that accurate urgent messages detection is crucial from an end user perspective (there is human life at stake), high frequency data is valuable.

Time Frequency	Main task		
	Utility	Urgency	Humanitarian
3 hours (Baseline)	**75.81**	**64.68**	49.72
6 hours	72.04	62.93	52.88
12 hours	71.98	64.16	**53.52**

Table 3: Macro F1-scores on the French dataset with our best model on different frequencies.

6 Conclusion

This paper proposed CrisisTS, the first open source multimodal and multilingual dataset that combines time series and textual data for urgency classification relying on a temporal and spatial alignment strategy that goes beyond explicit location mentions in texts. We also proposed a set of unimodal and multimodal experiments using an early fusion mechanism as well as LLMs that we newly adapted to our task. Results show that coupling TS and texts improves over text-based models in both French and English benchmark datasets while ensuring models portability to unseen events. In the future, we plan to consider other languages as well as design new multimodal fusion strategies.

Acknowledgment

This work has been supported by DesCartes: the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program.

Limitations

Although CrisisTS covers a large amount of crises, it can be extended to more types of crisis and other time series (e.g. the COVID crisis with clinical/health TS data (Suanpang and Jamjuntr, 2021)). The dataset can also be extended to other languages

and easily upgraded with actual public meteorological data.

Ethics Statement

The data used for creating the dataset is composed of texts from datasets benchmarks publicly available to the research community and meterological data publicly available online. The datasets are anonymized and contain no offensive or abusive language. They were collected before Twitter changed to X and conform to the Twitter Developer Agreement and Policy that allows unlimited distribution of either the numeric identification number or the textual content of each tweet.

Finally, analyzing social media is a vital resource in the field of crisis management as shown in many reports (DHS, 2014; Saroj and Pal, 2020). Although using AI models in situations where lives could be at stake, our aim is not to develop a fully automated system but an assistant tool to help rescue teams, with whom we are working in collaboration, to better filter social media urgent messages and anticipate actions.

References

Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. 2020. Multimodal categorization of crisis events in social media. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14679–14689.

Mansi Agarwal, Maitree Leekha, Ramit Sawhney, and Rajiv Ratn Shah. 2020. Crisis-DIAS: Towards Multimodal Damage Analysis - Deployment, Challenges and Assessment. Proceedings of the AAAI Conference on Artificial Intelligence, 34(01):346–353.

AI@Meta. 2024. Llama 3 Model Card.

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. CrisisMMD: Multimodal Twitter Datasets from Natural Disasters. In Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM).

Firoj Alam, Umair Qazi, Muhammad Imran, and Ferda Ofli. 2021. HumAID: human-annotated disaster incidents data from Twitter with deep learning benchmarks. In Proceedings of the International AAAI Conference on Web and Social Media, volume 15, pages 933–942.

Mahmood R Azimi-Sadjadi, De Yao, Qiang Huang, and Gerald J Dobeck. 2000. Underwater target classification using wavelet packets and neural networks. IEEE Transactions on neural networks, 11(3):784–794.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.

Mohammad Basit, Bashir Alam, Zubaida Fatima, and Salman Shaikh. 2023. Natural disaster tweets classification using multimodal data. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 7584–7594, Singapore. Association for Computational Linguistics.

Markus Bayer, Marc-André Kaufhold, Björn Buchhold, Marcel Keller, Jörg Dallmeyer, and Christian Reuter. 2021. Data Augmentation in Natural Language Processing: A Novel Text Generation Approach for Long and Short Text Classifiers. CoRR, abs/2103.14453.

Konstantinos Benidis, Syama Sundar Rangapuram, Valentin Flunkert, Yuyang Wang, Danielle Maddix, Caner Turkmen, Jan Gasthaus, Michael Bohlke-Schneider, David Salinas, Lorenzo Stella, François-Xavier Aubet, Laurent Callot, and Tim Januschowski. 2022. Deep Learning for Time Series Forecasting: Tutorial and Literature Survey. ACM Computing Surveys, 55(6).

Marwen Bouabid and Mohamed Farah. 2024. Detecting Local Crisis Events: A Case Study on the Colorado Wildfires through Social Media and Satellite Imagery. In 2024 IEEE Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS), pages 88–92.

Nils Bourgon, Farah Benamara, Alda Mari, Véronique Moriceau, Gaetan Chevalier, and Laurent Leygue. 2022. Are Sudden Crises Making me Collapse? Measuring Transfer Learning Performances on Urgency Detection. In 19th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2022).

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. 2015. Time series analysis: forecasting and control. John Wiley & Sons.

Cornelia Caragea, Adrian Silvescu, and Andrea Tapia. 2016. Identifying Informative Messages in Disasters Events using Convolutional Neural Networks. In 13th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2016), pages 1–7.

Carlos Castillo. 2016. Big crisis data: social media in disasters and time-critical situations. Cambridge University Press.

Nahuis Cerna, Selene Leya, Guyeux Christophe, and Laiymani David. 2022. The usefulness of NLP techniques for predicting peaks in firefighter interventions due to rare events. Neural Computing and Applications, 34:10117 – 10132.

Nimeesha Chan, Felix Parker, William Bennett, Tianyi Wu, Mung Yao Jia, James Fackler, and Kimia

Ghobadi. 2024. MedtsLLM: Leveraging LLMs for multimodal medical time series analysis. arXiv preprint arXiv:2408.07773.

Jishnu Chowdhury, Cornelia Caragea, and Doina Caragea. 2020a. Cross-Lingual Disaster-related Multi-label Tweet Classification with Manifold Mixup. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 292–298, Online. Association for Computational Linguistics.

Jishnu Ray Chowdhury, Cornelia Caragea, and Doina Caragea. 2020b. On Identifying Hashtags in Disaster Twitter Data. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI).

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2022. Incorporating stock market signals for Twitter stance detection. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4074–4091, Dublin, Ireland. Association for Computational Linguistics.

Xuan-Hong Dang, Syed Yousaf Shah, and Petros Zerfos. 2019. "The Squawk Bot": Joint Learning of Time Series and Text Data Modalities for Automated Financial Information Filtering. arXiv preprint arXiv:1912.10858.

Jens A. de Bruijn, Hans de Moel, Albrecht H. Weerts, Marleen C. de Ruiter, Erkan Basar, Dirk Eilander, and Jeroen C.J.H. Aerts. 2020. Improving the classification of flood tweets with contextual hydrological information in a multimodal neural network. Computers & Geosciences, 140:104485.

Iman Deznabi, Mohit Iyyer, and Madalina Fiterau. 2021. Predicting in-hospital mortality by combining clinical notes with time-series data. In Findings of the association for computational linguistics: ACL-IJCNLP 2021, pages 4026–4031.

DHS. 2014. Using social media for enhanced situational awareness and decision support. Technical report, U.S Departement of homeland security.

Alexis Dusart, Karen Pinel-Sauvagnat, and Gilles Hubert. 2021. ISSumSet: A Tweet Summarization Dataset Hidden in a TREC Track. In Proceedings of the 36th Annual ACM Symposium on Applied Computing, SAC '21, page 665–671, New York, NY, USA. Association for Computing Machinery.

Badreddine Farah, Guillaume Cleuziou, Cécile Gracianne, Adel Hafiane, Anaïs Halftermeyer, and Raphaël Canals. 2024. Image-text crisis tweet categorization:a caption-based approach. ISCRAM Proceedings, 21.

Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2018. Transfer learning for time series classification. In 2018 IEEE international conference on big data (Big Data), pages 1367–1376. IEEE.

Ming Shan Hee, Shivam Sharma, Rui Cao, Palash Nandi, Tanmoy Chakraborty, and Roy Ka-Wei Lee. 2024. Recent Advances in Hate Speech Moderation: Multimodality and the Role of Large Models. Preprint, arXiv:2401.16727.

Ying Huang, Long Jin, Hua-sheng Zhao, and Xiaoyan Huang. 2018. Fuzzy neural network and lle algorithm for forecasting precipitation in tropical cyclones: comparisons with interpolation method by ecmwf and stepwise regression method. Natural Hazards, 91(1):201–220.

Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France. European Language Resources Association (ELRA).

Giridharan Iyengar and Harriet Nock. 2003. Discriminative model fusion for semantic concept detection and annotation in video. pages 255–258.

Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. 2024. GPT4MTS: Prompt-based Large Language Model for Multimodal Time-series Forecasting. Proceedings of the AAAI Conference on Artificial Intelligence, 38(21):23343–23351.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. Preprint, arXiv:2310.06825.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, , and Qingsong Wen. 2023. Time-LLM: Time series forecasting by reprogramming large language models. arXiv preprint arXiv:2310.01728.

Marc-André Kaufhold, Markus Bayer, and Christian Reuter. 2020. Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning. Information Processing & Management, 57(1):102132.

Jens Kersten, Anna Kruspe, Matti Wiegmann, and Friederike Klan. 2019. Robust Filtering of Crisis-related Tweets. In ISCRAM 2019 conference proceedings-16th international conference on information systems for crisis response and management.

Swaraj Khadanga, Karan Aggarwal, Shafiq Joty, and Jaideep Srivastava. 2019. Using clinical notes with time series data for ICU management. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing

and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6432–6437, Hong Kong, China. Association for Computational Linguistics.

Kai Kim, Howard Tsai, Rajat Sen, Abhimanyu Das, Zihao Zhou, Abhishek Tanpure, Mathew Luo, and Rose Yu. 2024. Multi-Modal Forecaster: Jointly Predicting Time Series and Textual Data. arXiv preprint arXiv:2411.06735.

Rani Koshy and Sivasankar Elango. 2023. Multimodal tweet classification in disaster response systems using transformer-based bidirectional attention model. Neural Computing and Applications, 35(2):1607–1627.

Diego Kozlowski, Elisa Lannelongue, Frédéric Saudemont, Farah Benamara, Alda Mari, Véronique Moriceau, and Abdelmoumene Boumadane. 2020. A three-level classification of French tweets in ecological crises. Information Processing & Management, 57(5):102284.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 2479–2490, Marseille, France. European Language Resources Association.

Hongmin Li, Doina Caragea, and Cornelia Caragea. 2021. Combining self-training with deep learning for disaster tweet classification. In ISCRAM, pages 719–730. ISCRAM Digital Library.

Hongmin Li, Doina Caragea, Cornelia Caragea, and Nic Herndon. 2018a. Disaster response aided by tweet classification with a domain adaptation approach. Journal of Contingencies and Crisis Management, 26(1):16–27.

Zefeng Li, Men-Andrin Meier, Egill Hauksson, Zhongwen Zhan, and Jennifer Andrews. 2018b. Machine learning seismic wave discrimination: Application to earthquake early warning. Geophysical Research Letters, 45(10):4773–4779.

Yunlong Liang, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. 2022. MSCTD: A multimodal sentiment chat translation dataset. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2601–2613, Dublin, Ireland. Association for Computational Linguistics.

Junhua Liu, Trisha Singhal, Lucienne T.M. Blessing, Kristin L. Wood, and Kwan Hui Lim. 2021. CrisisBERT: A Robust Transformer for Crisis Classification and Contextual Crisis Embedding. In Proceedings of the 32nd ACM Conference on Hypertext and Social Media, HT '21, page 133–141, New York, NY, USA. Association for Computing Machinery.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Sreenivasulu Madichetty, Sridevi Muthukumarasamy, and Sreekanth Madisetty. 2023. A RoBERTa based model for identifying the multi-modal informative tweets during disaster. Multimedia Tools and Applications.

Haider Malik, Jun Feng, Pingping Shao, and Zaid Ameen Abduljabbar. 2024. Improving flood forecasting using time-distributed cnn-lstm model: a time-distributed spatiotemporal method. Earth Science Informatics, pages 1–20.

Richard McCreadie and Cody Buntain. 2023. CrisisFACTS: Building and Evaluating Crisis Timelines. Proceedings of the 31st Text Retrieval Conference (TREC 2022).

Richard McCreadie, Cody Buntain, and Ian Soboroff. 2019. TREC Incident Streams: Finding Actionable Information on Social Media. In Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management, València, Spain, May 19-22, 2019. ISCRAM Association.

Richard McCreadie, Cody Buntain, and Ian Soboroff. 2020. Incident streams 2019: Actionable insights and how to find them. In Proceedings of the 17th ISCRAM Conference.

Romain Meunier, Farah Benamara, Véronique Moriceau, and Patricia Stolf. 2023. Image and Text: Fighting the same Battle? Super Resolution Learning for Imbalanced Text Classification. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 10707–10720, Singapore. Association for Computational Linguistics.

Seung-Hyun Moon, Yong-Hyuk Kim, Yong Hee Lee, and Byung-Ro Moon. 2019. Application of machine learning to an early warning system for very short-term heavy rainfall. Journal of Hydrology, 568:1042–1054.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, et al. 2023. Anymal: An efficient and scalable any-modality augmented language model. arXiv preprint arXiv:2309.16058.

Venkata Kishore Neppalli, Cornelia Caragea, and Doina Caragea. 2018. Deep Neural Networks versus Naive Bayes Classifiers for Identifying Informative Tweets during Disasters. In Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management, ISCRAM'2018.

Laura Nguyen, Thomas Scialom, Benjamin Piwowarski, and Jacopo Staiano. 2023. LoRaLay: A multilingual and multimodal dataset for long range and layout-aware summarization. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 636–651, Dubrovnik, Croatia. Association for Computational Linguistics.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In International Conference on Learning Representations.

Hakan T Otal and M Abdullah Canbaz. 2024. LLM-Assisted Crisis Management: Building Advanced LLM Platforms for Effective Emergency Response and Public Collaboration. arXiv preprint arXiv:2402.10908.

Faghihi Hossein Rajaby, Bashar Alhafni, Ke Zhang, Shihao Ran, Joel Tetreault, and Alejandro Jaimes. 2022. CrisisLTLSum: A benchmark for local crisis event timeline extraction and summarization. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5455–5477, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Biloš, Hena Ghonia, Nadhir Vincent Hassen, Anderson Schneider, et al. 2023. Lag-llama: Towards foundation models for time series forecasting. arXiv preprint arXiv:2310.08278.

Christian Reuter, Amanda Lee Hughes, and Marc-André Kaufhold. 2018. Social Media in Crisis Management: An Evaluation and Analysis of Crisis Informatics Research. International Journal of Human–Computer Interaction, 34(4):280–294.

Christian Reuter and Marc-André Kaufhold. 2018. Fifteen years of social media in emergencies: A retrospective review and future directions for crisis informatics. Journal of Contingencies and Crisis Management (JCCM), 26(1):41–57. Special Issue: Human-Computer-Interaction and Social Media in Safety-Critical Systems.

Roberta Rocca, Nicolò Tamagnone, Selim Fekih, Ximena Contla, and Navid Rekabsaz. 2023. Natural language processing for humanitarian action: Opportunities, challenges, and the path toward humanitarian NLP. Frontiers in Big Data, 6.

Suresh Sankaranarayanan, Malavika Prabhakar, Sreesta Satish, Prerna Jain, Anjali Ramprasad, and Aiswarya Krishnan. 2020. Flood prediction based on weather parameters using deep learning. Journal of Water and Climate Change, 11(4):1766–1783.

Efsun Sarioglu Kayi, Linyong Nan, Bohan Qu, Mona Diab, and Kathleen McKeown. 2020. Detecting Urgency Status of Crisis Tweets: A Transfer Learning Approach for Low Resource Languages. In Proceedings of the 28th International Conference on Computational Linguistics, pages 4693–4703, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Anita Saroj and Sukomal Pal. 2020. Use of social media in crisis management: A survey. International Journal of Disaster Risk Reduction, 48:101584.

Pannee Suanpang and Pitchaya Jamjuntr. 2021. A comparative study of deep learning methods for time-series forecasting tourism business recovery from the COVID-19 pandemic crisis. Journal of Management Information and Decision Sciences, 24:1–10.

Reem Suwaileh, Tamer Elsayed, and Muhammad Imran. 2023. IDRISI-RE: A generalizable dataset with benchmarks for location mention recognition on disaster tweets. Information Processing & Management, 60(3):103340.

Cagri Toraman, Izzet Emre Kucukkaya, Oguzhan Ozcelik, and Umitcan Sahin. 2023. Tweets Under the Rubble: Detection of Messages Calling for Help in Earthquake Disaster. arXiv preprint arXiv:2302.13403.

Hung Van Le, Duc Anh Hoang, Chuyen Trung Tran, Phi Quoc Nguyen, Nhat Duc Hoang, Mahdis Amiri, Thao Phuong Thi Ngo, Ha Viet Nhu, Thong Van Hoang, Dieu Tien Bui, et al. 2021. A new approach of deep neural computing for spatial prediction of wildfire danger at tropical climate areas. Ecological Informatics, 63:101300.

Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. 2018. Centralnet: a multilayer approach for multimodal fusion. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pages 0–0.

Sarah Vieweg, Carlos Castillo, and Muhammad Imran. 2014. Integrating social media communications into the rapid assessment of sudden onset disasters. In Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings 6, pages 444–461. Springer.

Congcong Wang, Paul Nulty, and David Lillis. 2021a. Crisis domain adaptation using sequence-to-sequence transformers. In ISCRAM, pages 655–666. ISCRAM Digital Library.

Congcong Wang, Paul Nulty, and David Lillis. 2021b. Transformer-based Multi-task Learning for Disaster Tweet Categorisation. In International Conference on Information Systems for Crisis Response and Management.

Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. 2022. Transformers in time series: A survey. arXiv preprint arXiv:2202.07125.

M Wiegmann, J Kersten, F Klan, M. Potthast, and B. Stein. 2020. Analysis of Detection Models for Disaster-Related Tweets. In Proceedings of

the 17th International Conference on Information Systems for Crisis Response and Management, ISCRAM'2020.

Xuehua Wu, Jin Mao, Hao Xie, and Gang Li. 2022. Identifying humanitarian information for emergency response by modeling the correlation and independence between text and images. Information Processing & Management, 59(4):102977.

Peng Xu, Xiatian Zhu, and David A. Clifton. 2023. Multimodal Learning With Transformers: A Survey. IEEE Trans. Pattern Anal. Mach. Intell., 45(10):12113–12132.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. arXiv preprint arXiv:2306.13549.

Chao Yuan and Hossein Moayedi. 2020. Evaluation and comparison of the advanced metaheuristic and conventional machine learning methods for the prediction of landslide occurrence. Engineering with Computers, 36(4):1801–1811.

Cynthia Zeng and Dimitris Bertsimas. 2023. Global flood prediction: a multimodal machine learning approach. arXiv preprint arXiv:2301.12548.

Kexin Zhang, Qingsong Wen, Chaoli Zhang, Rongyao Cai, Ming Jin, Yong Liu, James Y Zhang, Yuxuan Liang, Guansong Pang, Dongjin Song, et al. 2024. Self-supervised learning for time series analysis: Taxonomy, progress, and prospects. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Jun Zhu, Pei Dang, Yungang Cao, Jianbo Lai, Yukun Guo, Ping Wang, and Weilian Li. 2024. A flood knowledge-constrained large language model interactable with gis: enhancing public risk perception of floods. International Journal of Geographical Information Science, 38(4):603–625.

A Dataset distribution

Tables 4 and 5 show the distribution for French and English textual data respectively.

The testsets used in the Out-of-event experiments are as follows:

- **Set 1:** Flood in Aude, Attack in Trebes, Beryl storm, Collapse in Lille, Explosion in Sanary, Fire in Landes (6,939 tweets).
- **Set 2:** Flood in Corsica, Fionn storm, Collapse in Marseille, Explosion Lubrizol, Fire Notre-Dame (4,036 tweets).
- **Set 3:** Flood Autre, Explosion Lubrizol, Irma hurricane, Fire Notre-Dame, Collapse in Lille (5,406 tweets).

Regarding the Out-of-type, testsets are shown in the Tables (i.e., for earthquake in the English data, the testset is composed of 1,105 tweets).

B Time Series Data Quality

B.1 Standard Data Analysis

An important factor when using meteorological time series data is their quality. We provide here standard data analysis (such as the mean, standard deviation, maximum, minimum, number of missing values, abnormal values, etc.) of the TS selected features in both FrenchTS (cf. Figure 3 and Table 6) and EnglishTS (cf. Figure 4 and Table 7). This analysis allows to check whether each feature has the correct metric (e.g., a MEAN TEMPERATURE of 287.82 shows that the measure is in Kelvin and not in Celsius or Fahrenheit) but also detect regular anomaly (e.g., MEAN WIND SPEED can not be negative).

From the features we chose (we decided to remove some features, such as dew point in FrenchTS, as we considered that feature is not useful for crisis management), we observe that the EnglishTS features from the US and New Zealand TS do not contain missing data (Figure 4) whereas in FrenchTS, the feature with the most missing data is MAXIMUM WIND GUST SPEED with a rate of 16.8% (Figure 3).

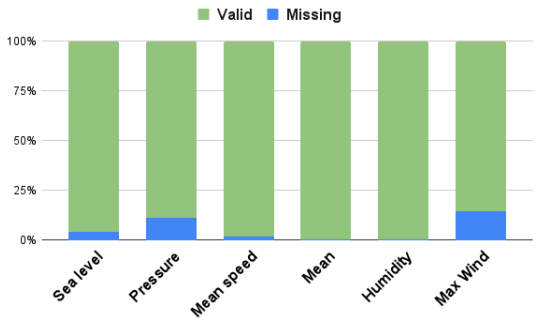

Figure 3: Percentage of missing data in FrenchTS dataset.

When looking into Table 6 for FrenchTS, we notice that the Max(Maximum Gust Wind Speed) is lower than Max(Mean Wind Speed). This is due to the number of missing data that is higher for MAXIMUM GUST WIND SPEED (see Figure 3).

B.2 Selected Features and Relevance for Crisis Classification

Table 8 shows the final list of features selected for each time series dataset.

Crisis (# events / # tweets)	Urgent			Not Urgent			Not Useful
	Hmn-Dmg	Mat-Dmg	Adv_Warn	Support	Critics	Other	
Not Sudden (13 / 11,513)							Not Crisis
Flood (3 / 3,593)	102	237	431	198	53	405	2,167
Hurricane (2 / 2,160)	57	57	199	200	29	200	1,418
Storm (8 / 5,760)	52	142	716	22	13	147	4,668
Sudden (7 / 3,855)							Not Crisis
Fire (2 / 2,458)	23	94	51	340	170	385	1,395
Attack (1 / 61)	14	0	0	40	3	2	2
Collapse (2 / 1,269)	63	38	11	23	51	136	947
Explosion (2 / 67)	1	7	0	53	2	4	0
Total (20 / 15,368)	312	575	1,408	876	321	1,279	10,597

Table 4: Distribution of textual data in the multimodal French dataset.

Crisis (# events / # tweets)	Urgent						Not Urg	Not Useful
	Hmn-Dmg	Hmn-Miss	Evac	Mat-Dmg	Need	Warn	Volunteer	
Not Sudden (7 / 4,552)								Not Crisis
Flood (1 / 431)	45	133	2	82	3	65	101	N/A
Hurricane (6 / 4,300)	318	9	478	699	309	936	1,372	179
Sudden (2 / 2,192)								Not Crisis
Fire (1 / 1,087)	514	49	82	102	49	28	263	N/A
Earthquake (1 / 1,105)	102	3	66	252	16	361	96	209
Total (9 / 6,923)	979	194	628	1,135	377	1,390	1,832	388

Table 5: Distribution of textual data in the multimodal English dataset.

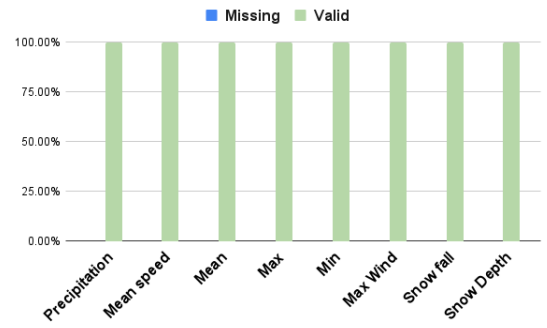

Figure 4: Percentage of missing data in EnglishTS dataset.

Feature	Mean	SD	Max	Min
Sea Press.	101,532.83	914.82	109,160	93,850
Press. Var.	2.30	124.07	2,17	- 2,190
Mean Wind Spd.	3.19	4.08	96	0
Mean Temp.	287.82	9.38	340.85	238.75
Humidity	75.75	16.86	100	1
Max G. Wind Spd.	7.37	5.08	77.33	0

Table 6: Data Quality Analysis of the FrenchTS dataset in terms of mean, standard deviation, maximum and minimum values.

After a quantitative analysis of our features, the next step is to analyse their relevance to our crisis management task. To this end, given a feature F, we compute the value P that represents the percentage of quartile that belongs to the values range of F, following the formula below:

Feature	Mean	SD	Max	Min
Precipitation	2.29	5.26	52.62	0
Mean spd. wind	4.74	1.54	11.65	1.67
Mean temp.	13.40	10.62	31.95	-16.99
Max temp.	19.97	11.06	39.05	-11.26
Min temp.	6.83	10.54	24.95	-24.14
Max G. Wind Spd.	9.88	2.55	19.74	4.11
Snow fall	0.99	6.13	101	0
Snow Depth	2.61	10.1	105	0

Table 7: Data Quality Analysis of the EnglishTS dataset in terms of mean, standard deviation, maximum and minimum values.

Feature	FrenchTS	EnglishTS
Precipitation	N/A	mm
Sea level Pressure	Pa	N/A
Pressure variation	Pa	N/A
Mean speed Wind	m.s-1	m.s-1
Mean Temperature	K	°C
Humidity	%	N/A
Maximum Wind Gust speed	m.s-1	m.s-1
Max Temperature	N/A	°C
Min Temperature	N/A	°C
Snow fall	N/A	mm
Snow depth	N/A	mm

Table 8: Features with their associated metrics in time series datasets. N/A means that the corresponding features from the source data are either not available or contain a high rate of missing values.

$$P(Q_i, F_{d,j}) = \frac{Q_i - \min(F_{dj})}{\max(F_{dj}) - \min(F_{dj})} * 100$$

where $F_{d,j}$ is a feature j in a given dataset d, Q_1 is the first quartile (i.e. 25% of values

of the feature j are lower than Q_1), Q_2 is the second quartile (or median) and Q_3 the third quartile (i.e. 75% of values are lower than Q_3). For example, for the HUMIDITY feature in the FrenchTS dataset, we get 64.65 which means that the first quartile of the humidity feature is equal to 64.65% of the range of humidity:

$$P(Q_1, F_{Fr,Humidity}) = \frac{65-1}{100-1} * 100 = 64.65$$

Computing these percentages for each feature allows for anomaly detection. Indeed, a linear data would have a first quartile close to 25% of the value range, a median close to 50% and third quartile close to 75%. Hence, a huge difference between these linear values for first quartile implies that there is also a huge difference between 25% of the lowest data and the rest of the data. Similarly, if $P(Q_3, F_{d,j}) << 75\%$, it means that there is a gap between 25% of the highest values and the others. Therefore, 25% of data can easily be detected as a crisis situation since it can correspond to an extreme meteorological situation.

Table 9 (resp. Table 10) shows the values of P for each quartile Q_i in FrenchTS and EnglishTS respectively. For example for FrenchTs, we can see that the third quartile is at 12.29% of the values range of MAXIMUM WIND GUST SPEED which implies that there is an important gap between all the data below the third quartile and the data after which can be useful in crisis detection.

Feature	Q_1	Q_2	Q_3
Sea level pressure	**47.55**	50.36	**53.63**
Pressure variation	**48.62**	50.23	**51.83**
Mean speed wind	1.98	3.44	**5.83**
Mean temperature	**41.63**	47.70	**56.02**
Humidity	**64.65**	78.79	89.90
Max Wind Gust Speed	4.91	8.02	**12.29**

Table 9: Features relevance in the FrenchTS dataset as given by P for each quartile. The most relevant percentages are in bold font.

C Computing Infrastructure

In order to improve the reproducibility of the experiments, we describe here the computing infrastructure we used:

- 2 CPU AMD Milan EPYC 7543 (32 core 2,80 GHz)
- 512 Go of memory
- 8 GPU Nvidia A100 SXM4 80 Go

Feature	Q_1	Q_2	Q_3
Precipitation	0.00	0.08	**3.69**
Mean speed wind	19.54	28.26	**38.78**
Mean temperature	**44.99**	62.69	81.94
Max temperature	**45.50**	64.76	80.70
Min temperature	**45.00**	63.25	83.66
Max Wind Gust Speed	24.50	35.00	**46.77**
Snow fall	0.00	0.00	**0.00**
Snow Depth	0.00	0.00	**0.00**

Table 10: Features relevance in the EnglishTS dataset as given by P for each quartile. The most relevant percentages are in bold font.

D Prompts and LLMs Hyper-parameters

We describe here the prompts and the parameters we used during our experiments in order to ensure reproducibility of the results. We recall that we used a five-shot prompt-tuning for Mistral and Llama3.

In Figure 5, we provide an example of a prompt for Llama3, using the standard three role format: The System role defines the task and how the model will answer; the User represents the input and the assistant is the response of the model. For Mistral, since the tweet is in French, the system input is a direct translation of the Llama system input with an adaptation to fit for the French label. Figure 6 gives an example of a prompt for MM-TimeLLM.

Input:

System:

You are a crisis expert who must classify tweet between the 9 following category: caution_and_advice, rescue_volunteering_or_donation_effort, sympathy_and_support, not_humanitarian, infrastructure_and_utility_damage, injured_or_dead_people, displaced_people_and_evacuations, requests_or_urgent_needs, missing_or_found_people, other_relevant_information. You must only return the label. You must classify the following tweet:

User:

Nine people were declared dead and tens of thousands evacuated on Friday as fierce wildfires raged across the western US state of California

Input:

Assistant:

injured_or_dead_people

Figure 5: Example of prompt used in Llama3.

Mistral and LLAMA3 used the same hyper parameters as shown in Table 11, while those in MM-

Input:

You are a crisis expert, you will receive a tweet and meteorological data and you must classify the tweet to one of the 9 following labels: caution_and_advice, rescue_volunteering_or_donation_effort, sympathy_and_support, not_humanitarian, infrastructure_and_utility_damage, injured_or_dead_people, displaced_people_and_evacuations, requests_or_urgent_needs, missing_or_found_people, other_relevant_information. You must only return the label. You must classify the following tweet: Nine people were declared dead and tens of thousands evacuated on Friday as fierce wildfires raged across the western US state of California

Figure 6: Example of prompt used in MM-TimeLLM.

TimeLLM are shown in Table 12.

Parameter	Value
max_new_token	100
temperature	1.0
do_sample	True
epochs	10
learning rate	(2e-5)
batch size	6

Table 11: Parameters used for testing LLMs unimodal models.

E Detailed Results

Parameter	Value
dropout	0.2
learning rate	0.001
epochs	10

Table 12: Parameters used for testing the adapted multimodal version of TimeLLM.

To measure the impact of time series on the fine-grained detection of urgency categories, we further detail our results per class for each dataset in urgency (cf. Table 13) and utility tasks (cf. Table 14). In both tables, the best unimodal and multimodal models are compared, showing the improvement of early fusion when trained in a multitask configuration, sudden crisis detection being the most productive secondary task.

We further analyze the results per crisis type. Table 15 (resp. Table 16) shows the performances of our best models in urgency classification when only trained on sudden (resp. expected) crises. Tables 17 and 18 give more detailed results per crisis type in order to see the impact of multi-modality for different kind of crisis. Multimodal models improve performances for all types of not-sudden crises, except the building collapse.

Model	F-score per class			Macro F-score
	Not-Urgent	Urgent	Not-Useful	
RoBERTa$_{3Tasks}$	86.56	93.56	56.50	78.89
RoBERTa$_{4Tasks}$ + Early Fusion	86.26	**93.69**	**58.79**	**79.58**
FlauBERT$_{3Tasks}$	47.54	62.85	81.21	63.87
FlauBERT$_{4Tasks}$ + Early Fusion	**49.41**	62.03	**82.86**	**64.68**

Table 13: Results per class for the three-class urgency task of our best unimodal and multimodal models in the English (first two lines) and French (last two lines) datasets in the out-of-type experiments.

Model	F-score per class		Macro F-score
	Useful	Not-Useful	
RoBERTa$_{3Tasks}$	97.37	57.28	77.50
RoBERTa$_{4Tasks}$ + Early Fusion	**97.64**	**58.33**	**77.98**
FlauBERT$_{3Tasks}$	67.78	81.47	74.63
FlauBERT$_{4Tasks}$ + Early Fusion	**69.27**	**82.34**	**75.81**

Table 14: Results per class for the binary utility task of our best unimodal and multimodal models in the English (first two lines) and French (last two lines) datasets in the out-of-type experiments.

Model	F-score per class			Macro F-score
	Not-Urgent	Urgent	Not-Useful	
RoBERTa$_{3Tasks}$	87.35	93.75	63.16	86.57
RoBERTa$_{4Tasks}$ + Early Fusion	86.34	**93.94**	**65.42**	86.54
FlauBERT$_{3Tasks}$	50.7	51.9	76.92	59.84
FlauBERT$_{3Tasks}$ + Early Fusion	**54.37**	48.69	**78.18**	**60.41**

Table 15: Results per class for the three-class urgency task of our best unimodal and multimodal models in the English (first two lines) and French (last two lines) in the out-of-type experiments **when only considering sudden crises** in the test set.

Model	F-score per class			Macro F-score
	Not-Urgent	Urgent	Not-Useful	
RoBERTa$_{3Tasks}$	85.83	93.35	49.85	82.92
RoBERTa$_{4Tasks}$ + Early Fusion	**86.18**	**93.43**	**52.17**	**83.57**
FlauBERT$_{3Tasks}$	45.44	70.15	84.07	66.56
FlauBERT$_{3Tasks}$ + Early Fusion	**46.11**	**70.92**	**85.52**	**67.52**

Table 16: Results per class for the three-class urgency task of our best unimodal and multimodal models in the English (first two lines) and French (last two lines) in the out-of-type experiments **when only considering not sudden crises** in the test set.

Model	F-score per Crisis Type				
	Sudden crisis		Not sudden crisis		
	Fire	Collapse	Hurricane	Storm	Flood
FlauBERT$_{FineTuned-3Tasks}$	**38.36**	45.36	60.87	50.98	54.34
FlauBERT$_{FineTuned-3Tasks}$ + Early Fusion	37.19	**46.69**	**62.18**	**54.19**	**55.48**

Table 17: Results per crisis of our best unimodal and multimodal models in French for the multi-class humanitarian task in the out-of-type experiments.

Model	F-score per Crisis Type			
	Sudden crisis		Not sudden crisis	
	Fire	Collapse	Hurricane	Flood
RoBERTa$_{3Tasks}$	**94.03**	79.12	89.29	76.55
RoBERTa$_{4Tasks}$ + Early Fusion	93.26	**79.83**	**90.02**	**77.13**

Table 18: Results per crisis of our best unimodal and multimodal models in English for the three-class urgency task in the out-of-type experiments