M-CATNAT: A Multimodal dataset to analyze French tweets during natural disasters

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

Omar El Bachyr
University of Orléans, INSA-CVL, LIFO, EA 4022, F45067 Orléans, France
omar.elbachyr@usmba.ac.ma

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

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

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

Samuel Auclair
BRGM, F45060 Orléans, France
s.auclair@brgm.fr

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

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

ABSTRACT

The proliferation of social media, especially platforms like X (formerly Twitter), has made available a large volume of real-time data valuable across diverse fields. During natural disasters, such data aids humanitarian efforts by providing crucial insights. However, processing this vast amount of data necessitates automated systems, often relying on annotated datasets for training. While supervised learning dominates this area, multilingual and multimodal annotated datasets are scarce. The present study addresses this gap by introducing M-CATNAT, a multimodal dataset of French tweets about natural disasters. Unlike previous datasets, M-CATNAT integrates annotations for texts, images, and their multimodal combination. Leveraging CrisisMMD guidelines, this work in progress aims to annotate 1,430 tweets, generating over 4,500 labels. The M-CATNAT dataset not only expands resources to non-English languages but also enhances multimodal analysis by furnishing three levels of annotation for each tweet (one per modality plus one for the whole tweet).

Keywords

Deep learning, French multimodal data, Crisis management, CrisisMMD Datatset

INTRODUCTION

The spread of social media use in this last decade has made available massive real time data. Particularly in platforms such as X (previously Twitter) where user-generated content has real value that can be used in a variety of applications including public health, economics and politics (Wu & Mebane Jr, 2022). When a natural disaster

*corresponding author

occurs, tweets can help humanitarian organizations to gather essential information in order to understand the scope of the disaster and prioritize actions to reduce suffering and rebuild communities (Alam et al., 2018). Nevertheless, the need to process the large amount of data available on social media requires the development of automated systems and methods that need annotated data to be trained.

The automation of social media information processing has recently become a hot research topic, with a majority of existing approaches relying heavily on supervised learning methods. Their primary objective is to categorize each post into a specific class based on the targeted task, such as informativeness (Alam et al., 2018) or relatedness (Kozlowski et al., 2020). Annotated datasets being the core element in supervised learning, considerable efforts have been made to manually construct such datasets. While most resources are text-based and predominantly in English, it's worth noting that resources in other languages also exist (Alharbi & Lee, 2019; Cobo et al., 2015; Cresci et al., 2015; Gründer-Fahrer et al., 2018; Kozlowski et al., 2020) (see the following section for more details).

Despite numerous studies indicating that multimodality (e.g. text, images, video) can significantly enhance the performance of these methods (Abavisani et al., 2020; Ofli et al., 2020), there is currently only one extensively studied dataset – CrisisMMD (Alam et al., 2018), which focuses on multimodal English tweets related to crisis (natural disasters). Such a resource in English can be used to fined-tune a multilingual pre-training model (Conneau & Lample, 2019) and thus process French tweets. Although the preliminary results we have obtained with this process are not sufficiently satisfactory, they give reason to hope that the introduction of even a limited number of tweets in French during training would improve performance. Aligning the annotations of this new data (in French) with the annotations of the CrisisMMD dataset is, however, a prerequisite for success.

In this context, this work (in progress) aims to create the M-CATNAT dataset : a multimodal dataset of 1,430 French tweets about various natural disasters with annotations aligned with CrisisMMD tasks. However, the CrisisMMD dataset has the following limitation : it contains annotations for each of the two modalities of a tweet (text and image), but does not propose gold (i.e. manually annotated) annotations for the multimodal tweet as a whole (text+image). Since learning a multimodal model requires such a global label, tweets whose annotations for the two modalities do not match are generally filtered out (Ofli et al., 2020). Unlike CrisisMMD, the M-CATNAT dataset we are preparing includes not only gold annotations for the image modality and the text modality but also the gold label of the tweet as a whole, i.e the multimodal instance (text+image). Interestingly, the full annotation model proposed makes it possible to offer a greater representativeness of tweets in terms of the diversity of relationships between modalities (e.g. redundancy, complementary or contradiction in the information).

In this article, we present the framework defined for the multimodal annotation process of the 1,430 French tweets that will make up the M-CATNAT dataset[1].

So far, seven annotators from the research team have fully annotated 837 tweets, out of the 1,430 planned. This ongoing process has resulted in the creation of more than 3,000 manual annotations, as each instance entails three labels (one for the image, one for the text, and one for the multimodal instance with some tweets having multiple images). The CrisisMMD humanitarian classes definition guided our annotation process, ensuring consistency and reliability in the labeled data. In addition, the annotation process was regularly evaluated to check both its consistency (inter-annotator agreement) and its alignment with the CrisisMMD dataset.

RELATED WORKS

The surge in the use of social media has led to an increased interest in processing their content with numerous studies focusing, in particular, on crisis-related tweets. As reported in Table 1, a lot of emphasis has been placed on text-based analysis, given that the majority of annotated datasets primarily feature labels on text alone. These datasets have introduced various tasks, such as assessing the *relatedness* of tweets to a specific disaster (Kozlowski et al., 2020), determining the *informativeness* of tweets (Alam et al., 2021; Olteanu et al., 2015), classifying *crisis types*, and detecting *eyewitnesses* (Zahra et al., 2020). In the domain of image processing for tweets, tasks include estimating the *damage severity* (Alam et al., 2018; Nguyen et al., 2017) and *damage type* (Mouzannar et al., 2018), evaluating *informativeness* (Alam et al., 2018). Most resources are in English, although resources also exist in other languages like Spanish (Cobo et al., 2015), Italian (Cresci et al., 2015), German (Gründer-Fahrer et al., 2018) and Arabic (Alharbi & Lee, 2019). For the French language, (Kozlowski et al., 2020) proposed a text only disaster tweet dataset with three classification levels including *relatedness*, *urgency* and *intent to act*. One should note that the various tasks addressed lead the community to produce datasets more or less compatible with the central resource that is CrisisMMD. Each dataset whose task mentions *humanitarian* is either an implementation of the CriseMMD guidelines or a mapping to a simplified set of classes, keeping it compatible with CrisisMMD[2].

[1]The dataset is available here : https://github.com/badreddineFarah/M-CATNAT

[2]For a substantial survey of existing datasets for English, see (Feng et al., 2022).

Table 1. Crisis related published resources.

	Modal. labels	Tasks	Platform	Size	Language
(Olteanu et al., 2015)	Text	informativeness, humanitarian, source	Twitter	28,000	English
(Alam et al., 2021)	Text	informativeness, humanitarian	Twitter	166,098 (inf.), 141,533 (hum.)	English (94%)
(Kozlowski et al., 2020)	Text	relatedness, urgency, intent to act	Twitter	12,826	French
(Cobo et al., 2015)	Text	relevancy	Twitter	2,187	Spanish
(Cresci et al., 2015)	Text	relatedness, damage	Twitter	5,642	Italian
(Alharbi & Lee, 2019)	Text	informativeness, humanitarian	Twitter	4,037	Arabic
(Zahra et al., 2020)	Text	eyewitness	Twitter	14,000	English
(Alam et al., 2023)	Image	damage severity, humanitarian, disaster type	Twitter, Google, ...	71,198	-
(Nguyen et al., 2017)	Image	damage severity	Twitter, Google	25,758	-
(Alam et al., 2018)	Image, text	informativeness, humanitarian, damage severity	Twitter	16,097	English
(Mouzannar et al., 2018)	Multimodal	damage type	Instagram, Twitter, Google	5,879	English
M-CATNAT (Ours)	**Image, text, multimodal**	**Informativeness, humanitarian**	**Twitter**	**1,430** (in progress)	**French**

Table 2. Disaster events considered.

Crisis type	Affected areas	Search start date	Search end date	# collected tweets	# sampled tweets (images)
Floods	Alpes-Maritimes et Var	10/2/2015 12:00:00 AM	10/4/2015 11:59:00 PM	14846	108 (124)
	Carcassonnais	10/14/2018 12:00:00 AM	10/15/2018 11:59:00 PM	14341	171 (205)
	Secteurs de Béziers / Narbonnais	10/22/2019 12:00:00 AM	10/24/2019 11:59:00 PM	7550	104 (126)
	Pays Basque / Béarn/ Pyrénées	12/12/2019 12:00:00 AM	12/15/2019 11:59:00 PM	6838	132 (158)
	Cote d'Azur 83/06	11/22/2019 12:00:00 AM	11/24/2019 11:59:00 PM	9983	164 (206)
	Cote d'Azur 83/06	12/1/2019 12:00:00 AM	12/1/2019 11:59:00 PM	6353	89 (103)
	Valleraugue, Saumane	9/19/2020 12:00:00 AM	9/20/2020 11:59:00 PM	5721	72 (83)
	Alex storm (south of France)	10/1/2020 12:00:00 AM	10/3/2020 11:59:00 PM	7804	110 (129)
		Date	Time		
Earthquakes	Barcelonnette	07/04/2014	07:26:59 PM	11085	66 (72)
	La Rochelle	28/04/2016	06:46:53 AM	3802	87 (91)
	Thouars	21/06/2019	06:50:57 AM	2668	56 (58)
	Le Teil	11/11/2019	10:52:45 AM	6448	192 (224)
	Strasbourg	12/11/2019	01:38:00 PM	3123	79 (86)
Total				100562	**1430 (1665)**

Despite recent studies highlighting the effectiveness of multimodal datasets in improving the performance of machine learning models through multimodal training (Abavisani et al., 2020; Liang et al., 2022; Long & McCreadie, 2022; Ofli et al., 2020), limited attention has been given to the creation of new multimodal crisis-related datasets and CrisisMMD remains the most widely used dataset despite its inherent limitations. A first notable limitation is that CrisisMMD independently annotates images and text, necessitating practitioners to filter instances with discordant labels (different labels for image and text) and retain instances with consistent labels for training. This approach does not take into account the various relationships between modalities in social media data (Vempala & Preoţiuc-Pietro, 2019). In response to this gap, (Sosea et al., 2021) introduced Disrel, a dataset for classifying the relationship between image and text. Their work demonstrated that incorporating such tasks can enhance the performance of models trained on CrisisMMD.

This study aims to enhance CrisisMMD in two dimensions. Firstly, we augment the corpus by incorporating French tweets, enabling multilingual and multimodal training. Secondly, we provide three annotations for each instance, hypothesizing that this enriched dataset will enable models to better capture the relationships between images and texts, thereby improving overall system performance. Importantly, we adhered to the CrisisMMD guidelines to ensure alignment between our annotations and the original annotation guide.

DATA COLLECTION

To initiate the annotation process, it is essential to gather data. For this purpose, we used datasets collected as part of the RéSoCIO project by the French Geological Survey (BRGM) via its SURICATE-Nat platform (Auclair et al., 2019). The data collection process relies on the interrogation of Twitter's "academic" API on the basis of a carefully chosen keyword search, exploring French-speaking lexical fields related to floods on the one hand, and earthquakes on the other. Below are the specific queries used for each type of disaster, with the aim of retrieving a maximum number of posts describing the effects of the phenomena, as well as a minimum number of off-topic posts.

Flood key-words The keywords used are the following (taking spelling errors into account) : "*inondation*" (flood), "*inondé*" (flooded / inundated), "*sous l'eau*" (under water), "*rivière en crue*" (swollen river), "*crue*" (flood upward trend / freshet), "*décrue*" (flood downward trend), "*onde de crue*" (flood wave), "*sort de son lit*" (rise above the banks), "*torrentiel*" (torrential), "*emporté par les eaux*" (washed away). The query excludes retweets and tweets with flood lexical fields related to politics, sex or migration, while focusing only on French language tweets by specifying "lang:fr".

Earthquake key-words "*Séisme*" or "*tremblement de terre*" (earthquake), "*magnitude*" (magnitude), and "*terre tremble*" (ground shaking) are the keywords used to retrieve the tweets (taking spelling errors into account as well). The query excludes retweets and tweets with earthquake lexical fields related to sex, migration or politics, while focusing only on French language tweets by specifying "lang:fr".

DATA SAMPLING

As illustrated in Table 2, we sampled tweets from various events of natural disaster, taking one-third from earthquake events and two-thirds from flood events, in order to respect the representativeness of each disaster while keeping the number of tweets sufficient for each disaster. The earthquakes and flash floods selected correspond to significant

events for mainland France, from a phenomenological and/or a crisis management point of view. To maintain data quality and relevance, we applied additional criteria, exclusively selecting tweets with text lengths between 5 to 40 words to ensure sufficient textual content for comprehensive analysis. Additionally, we selectively included tweets with at least one associated image, as the presence of visual content is pivotal for preserving the multimodal nature of the dataset.

The goal is to annotate a total of 1,430 multimodal tweets, with each tweet divided into three instances: text alone, image alone, and the combination of text and image, resulting in approximately 4,775 instances for annotation (because some tweets contain several images). For tweets with multiple images, we annotate the text, each image, and multiple text-image combination instances are created by merging the tweet's text with each associated image.

It is worth noting that since the beginning of the annotation process, around 5% of the data have been deleted from X, leading us to revise the dataset size from 1,500 to 1,430 tweets. In addition, it is still possible that between now and the end of the annotation process, new tweets will be deleted, reducing the total number of tweets available for annotation.

TASKS DESCRIPTIONS

As said before, we use the *humanitarian* and *informative* tasks (from CrisisMMD) to annotate the dataset[3]. However, we also adapt the task to the practises in the community, as its was already done within several works (Abavisani et al., 2020; Liang et al., 2022; Long & McCreadie, 2022; Ofli et al., 2020) which have merged classes relative to damages ("infrastructure_and_utility_damage" and "vehicle_damage") into "infrastructure_and_utility_damage" and classes relative to human casualties ("affected_individuals","injured_or_dead_people","missing_or_found_people") into "affected_individuals", resulting in a five classes task described bellow:

Affected individuals If the tweet/image reports or shows individuals affected by the disaster event, such as people sitting outdoors, individuals standing in lines for assistance, people in need of shelter facilities, missing or found individuals, or deceased individuals.

Infrastructure and utility damage If the tweet/image reports or shows a damaged structure or one whose use is affected by an earthquake, fire, heavy rains, floods, strong winds, gusts, etc., such as damaged houses, roads, buildings; flooded houses, streets, highways; blocked roads, bridges, paths; collapsed bridges, power lines, cars, boats, communication poles, etc.

Rescue volunteering or donation effort If the tweet/image reports or shows any type of rescue, volunteering, or donation effort, such as transporting people to safe locations, evacuating people from the hazardous area, individuals receiving medical or food aid, people in shelter facilities, monetary donations, blood donations, or services, etc.

Other relevant information If the tweet/image does not fit into any of the three categories above but still contains important information useful for humanitarian aid.

Not related or not relevant information If the tweet/image is not useful for humanitarian aid.

As described in the CrisisMMD paper, only informative tweets have been annotated into one of the *humanitarian* classes. Thus, considering other tweets as not informative (not related or not relevant) the annotated dataset can also be used to train models on the task *informativeness*.

ANNOTATION METHODOLOGY

In the annotation methodology, we implemented a four-phase process to ensure high-quality annotation of the dataset. At the end of each phase, an assessment of the annotations was made according to different indicators, followed by a consultation meeting focusing on examples with substantial disagreement occurred. A clarification, accompanied by illustrative examples was then incorporated into the annotation guide. Also, to mitigate the bias in labelling the modalities from each (multimodal) tweet, the order of presentation of the instances to be annotated needs to be carefully considered. An annotator who is first presented with a complete multimodal tweet (text +

[3]The annotation guide is delivered with the resource (https://anonymous.4open.science/r/M-CATNAT-5692/README.md).

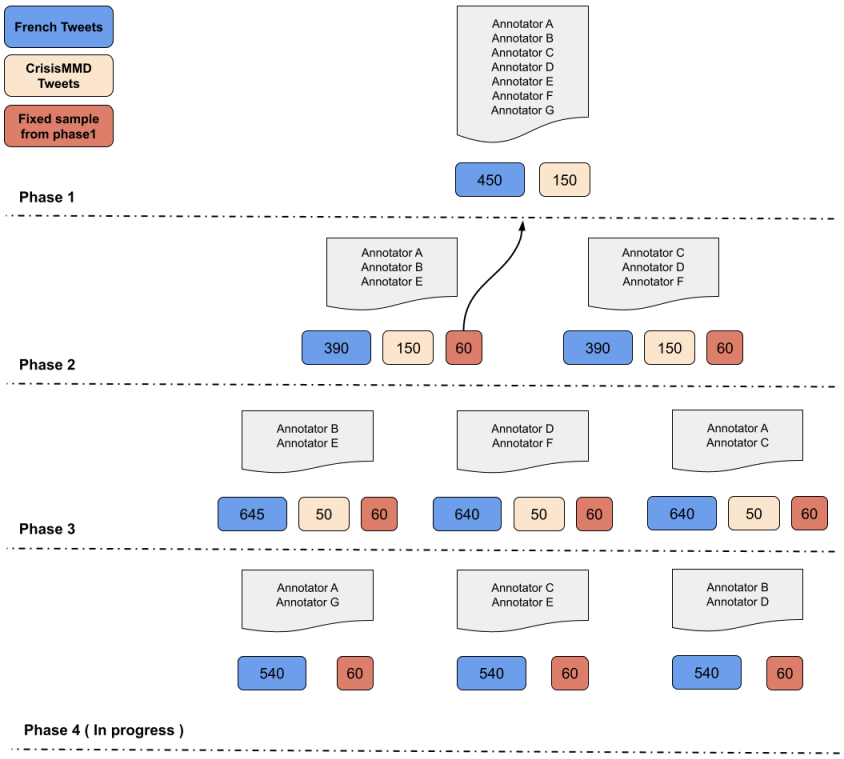

Figure 1. Overview of the annotation protocol. This diagram illustrates our multi-phase annotation methodology. In each phase, we annotate a specific number of new instances (image, text and multimodal instances), highlighted in blue. To maintain alignment with the CrisisMMD, a small portion of this dataset is also annotated. Lastly, to ensure the self-consistency of each annotator, some instances are selected for re-annotation.

image) will undoubtedly be influenced in his annotation (a posteriori) of each of the modalities. We have therefore ensured that the text and image modalities are presented first and in a separate slot (groups of annotators) or phases from the multimodal tweet (text+image).

This sequential presentation aimed to ensure independent labeling decisions for both unimodal and multimodal instances, minimizing potential bias in annotations.

As presented in Figure 1, Phase 1 involved seven annotators collectively annotating one shared dataset of 600 instances constituted of 150 instances from CrisisMMD and 450 instances from new French tweets ($\frac{1}{3}$ images, $\frac{1}{3}$ texts and $\frac{1}{3}$ multimodal instances). In this phase, when an instance is annotated identically by at least 5 annotators (out of 7), this annotation is definitively used as a label. The other instances (for which the annotation is not consensual) are discussed and collectively labelled at the end of the phase during the consultation meeting. This process allowed us to clarify and adjust the annotation guide. This first phase is an agreement (or training) phase for the annotators. Subsequent phases gradually "industrialise" the annotation process, with annotators organised into parallel pools.

Phase 2 comprised two pools, each pool was made up of 3 annotators responsible for annotating 600 instances. Among the 600 instances, we maintained a set of 150 CrisisMMD instances and introduced 60 instances from phase 1 in order to compute the consistency of annotations through the phases for each annotator involved in the process. At the end of this phase, the instances with two identical annotations are definitively labelled, while we proceed to a correction phase during which each disagreement is resolved by an annotator from the other pool.

In phase 3, we decomposed our team into three annotation pairs, for each annotation pool we reduced the number of CrisisMMD instances to 50 while keeping the 60 instances from phase 1. The disagreement is resolved by a third annotator from another pool.

Currently ongoing, Phase 4 involves three new pairs of annotators, each annotating 540 instances, along with the 60 instances from Phase 1. The ultimate goal is to annotate the full set of 1,430 tweets, amounting to approximately 4,775 manual annotations at the conclusion of the process.

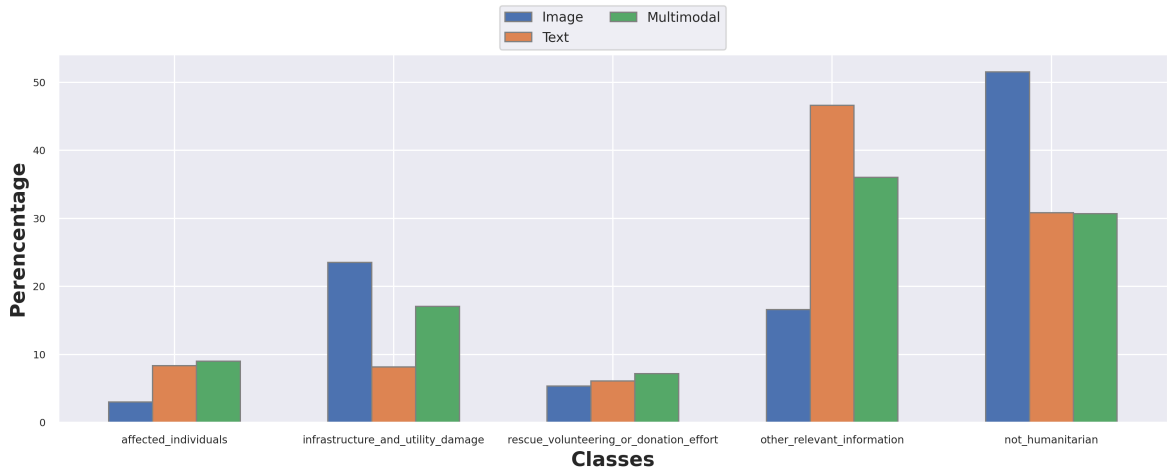

Figure 2. Comparison of label distribution for each modality, the distributions are calculated on a 1,032 fully annotated instances.

RESOURCE ANALYSIS

In this section, we provide a comprehensive analysis of the M-CATNAT resource at the time of writing (837 unique tweets fully labelled corresponding to 1,032 multimodal pairs since some tweets have multiple images). Firstly, we outline the distribution of labels for each modality and for each type of disaster (earthquakes and floods). Secondly, we present the results of annotation scores, including the Fleiss Kappa score (Fleiss et al., 2013) and the CrisisMMD alignment score. Finally, we show the distribution of class concordance/discordance according to modalities.

LABEL DISTRIBUTION

Figure 2 illustrates the label distribution across different classes. Notably, the class "Affected Individuals" exhibits the lowest representation. The "Not Humanitarian" class, equivalent to "Not informative" in the *informativeness* task

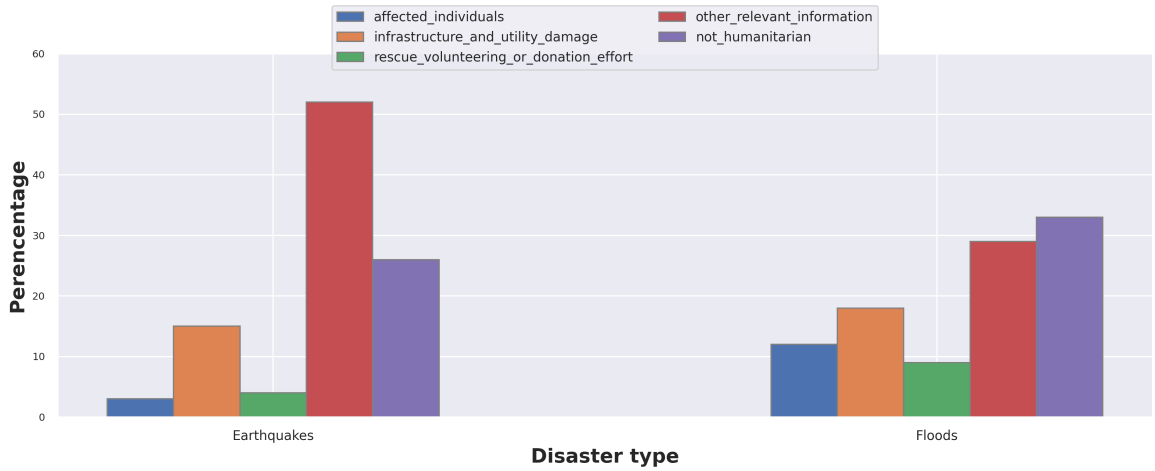

Figure 3. Comparaison of label distribution for earthquakes and floods, the result are on 307 earthquake related tweets and 725 floods related tweets.

Table 3. Fleiss kappa scores on each phase.

Phase	1	2	3
Mean (over pools)	0.64	0.70	0.61
Common examples	0.64	0.68	0.66

Table 4. CrisisMMD alignment scores on each phase.

Phase	1	2	3
CrisisMMD alignment	79%	74%	77%

of CrisisMMD, constitutes 30% in text and multimodal instances but rises to approximately 50% in image labels. Conversely, "Other Relevant Information" represents 46% in texts and 36% in multimodal instances, contrasting with its lower representation at 16% in images. This difference is attributed to the enriching role of textual information in multimodal instances, aiding in disaster characterization regarding intensity and location. The distribution of multimodal labels closely aligns with the text distribution, highlighting the importance of textual information in crisis datasets in general and in the M-CATNAT dataset in particular. The emphasis on textual information stems from its reflective of tweet intentions and provision of more precise details. Classes such as "Rescue Volunteering and Donation Efforts" are more prominent in CrisisMMD (14.5%), indicating that the disasters in our data are comparatively less severe.

Figure 3 reports the distribution of multimodal annotations according to the type of disaster in our dataset. "Other Relevant Information" exhibits higher representation in earthquake-related tweets, due to the lower magnitude of these earthquakes where tweets often provide basic information about the incident, such as location and magnitude. Conversely, "Affected Individuals" and "Rescue volunteering or donation effort" are more represented in flood-related tweets, reflecting the extensive damage caused by such events.

ANNOTATOR SCORES

In accordance with the annotation methodology outlined in the "Annotation Methodology" section, our process involves four phases, each subdivided into annotator pools. In consequence we propose to measure inter-annotator agreement on each phase using the Fleiss Kappa (Fleiss et al., 2013)[4]. Table 3 reports the results obtained: first row shows the mean Fleiss Kappa score overs pools for each phase, while in the second row Fleiss Kappa scores were calculated specifically for the 60 common examples (highlighted as red samples in Figure 1). The variations observed are mainly due to the changes made to the annotation guide over the phases; it is also significant to note that the inter-annotator agreements remained above 0.60 in each of the three phases. To be complete, we noticed the agreement between annotators varied across modalities. A higher agreement is observed for images (up to 0.85) compared to texts which exhibited a poorer annotator agreement, particularly in Phase 1 (0.56).

[4]In addition, at every phase we made annotations for a sample of 60 fixed instances (in red in Figure 1) in order to monitor the intra-annotator agreement along the whole process. The percentage scores obtained vary between 67% and 78% for the transition from phase 1 to phase 2 and between 82% and 89% for the transition from phase 2 to phase 3.

CRISISMMD ALIGNMENT

Moreover, we assessed alignment with the CrisisMMD dataset (see Table 4). In Phase 1, alignment was computed on 150 examples after disagreement resolution, yielding a score of 79% (with a notable increase to 90% for multimodal instances). In Phase 2, the alignment score was computed on 150 examples for each pool and the mean score obtained is 74%, while it reached 77% in phase 3 on 50 examples for each pool (mean score of the three pools is reported). These results are encouraging given our will to use the proposed resource in addition to CrisisMMD in a multilingual configuration.

MODALITY LABEL ANALYSIS

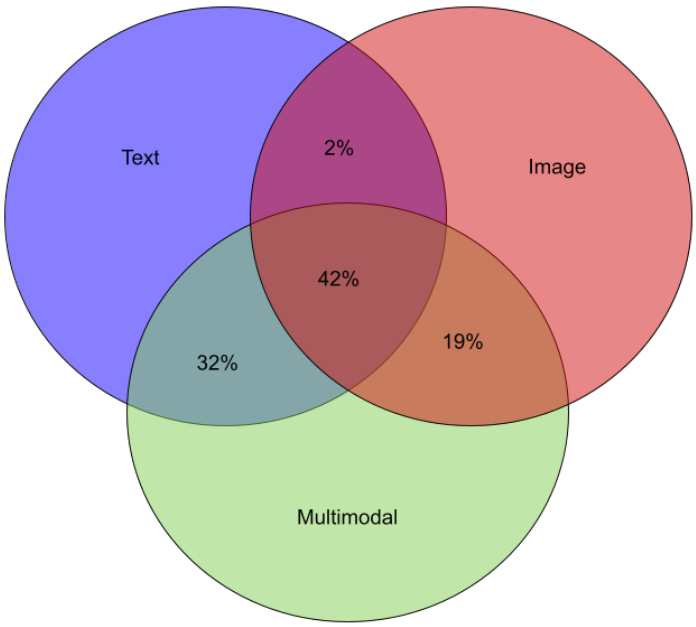

Figure 4. Label Combinations for Image, Text, and Multimodal on 1,032 fully annotated instances.

Figure 4 shows an analysis of the labelling of the 1,032 [5] examples in the current M-CATNAT dataset. It reveals the relation in the distribution of labels across the three labels assigned to one Tweet : the *image* label, the *text* label and the *multimodal* label. A significant portion, accounting for 42% of the examples, demonstrates consistency across all three labels ; this suggests a strong correlation among the different types of data and their respective labels. In 32% of cases, although the text label differs from the image label, it aligns with the multimodal label ; this implies that the textual content plays a crucial role in shaping the combined interpretation. Conversely, 19% of the examples have an image label distinct from the text label but aligning with the multimodal one, highlighting the influence of visual elements in these instances. Following these results, one may consider the need to train models to spot instances where the information looked for may be found in a specific modality. Lastly in a small proportion, just 2%, the image and text labels match but differ from the multimodal label, and lastly, 5% of cases shows complete disparity across all three labels. These findings underscore the variety of relations between different modalities and emphasize the importance of considering multimodal annotation rather than an independent annotation of the text and the image.

It's worth noting that crisisMMD use cases typically focus only on instances with identical labels across modalities. However, our study overlooks more than 50% of the data, disregarding the diverse relationships between images and text present in the multimodal tweets dataset.

CONCLUSION

Leveraging social media information during natural disasters can significantly help humanitarian aid organizations. Deep learning models play a crucial role in filtering and classifying this data efficiently. However, the effectiveness of such models relies heavily on labeled data for training. Research has highlighted the potential of multimodal

[5]We can get 1,032 image-text instances from the 837 annotated tweets as a tweet can have multiple images.

data in enhancing performance, yet available datasets are predominantly unimodal and in English. Although one multimodal dataset exists, its separate labeling for images and text poses limitations. This work addresses this gap by introducing M-CATNAT, a French multimodal dataset aligned with CrisisMMD, featuring three labels for each instance. This alignment facilitates multilingual training, thereby broadening the applicability of deep learning techniques. We provide annotator scores and other analyses to support deep learning practitioners in utilizing the dataset effectively. In future work, our aim is to further expand the annotated data and train models on this dataset.

ACKNOWLEDGEMENT

This work was supported by the French national research agency (ANR) under the program IA.iO (ANR-20-THIA-0017-01), as well as within the RéSoCIO project (ANR-20-CE39-001). Opinions expressed in this paper solely reflect the authors' view; the ANR is not responsible for any use that may be made of information it contains. 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., Alam, T., Hasan, M. A., Hasnat, A., Imran, M., & Ofli, F. (2023). Medic: A multi-task learning dataset for disaster image classification. *Neural Computing and Applications*, *35*(3), 2609–2632.

Alam, F., Ofli, F., & Imran, M. (2018). Crisismmd: Multimodal twitter datasets from natural disasters. *Twelfth international AAAI conference on web and social media*.

Alam, F., Sajjad, H., Imran, M., & Ofli, F. (2021). Crisisbench: Benchmarking crisis-related social media datasets for humanitarian information processing. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*, 923–932.

Alharbi, A., & Lee, M. (2019). Crisis detection from arabic tweets. *Proceedings of the 3rd workshop on Arabic corpus linguistics*, 72–79.

Auclair, S., Boulahya, F., Birregah, B., Quique, R., Ouaret, R., & Soulier, E. (2019). Suricate-nat: Innovative citizen centered platform for twitter based natural disaster monitoring. *2019 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, 1–8.

Cobo, A., Parra, D., & Navón, J. (2015). Identifying relevant messages in a twitter-based citizen channel for natural disaster situations. *Proceedings of the 24th international conference on world wide web*, 1189–1194.

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. *Advances in neural information processing systems*, *32*.

Cresci, S., Tesconi, M., Cimino, A., & Dell'Orletta, F. (2015). A linguistically-driven approach to cross-event damage assessment of natural disasters from social media messages. *Proceedings of the 24th International Conference on World Wide Web*, 1195–1200.

Feng, Y., Huang, X., & Sester, M. (2022). Extraction and analysis of natural disaster-related vgi from social media: Review, opportunities and challenges. *International Journal of Geographical Information Science*, *36*(7), 1275–1316.

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions*. john wiley & sons.

Gründer-Fahrer, S., Schlaf, A., Wiedemann, G., & Heyer, G. (2018). Topics and topical phases in german social media communication during a disaster. *Natural language engineering*, *24*(2), 221–264.

Kozlowski, D., Lannelongue, E., Saudemont, F., Benamara, F., Mari, A., Moriceau, V., & Boumadane, A. (2020). A three-level classification of french tweets in ecological crises. *Information Processing & Management*, *57*(5), 102284.

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.

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.

Mouzannar, H., Rizk, Y., & Awad, M. (2018). Damage identification in social media posts using multimodal deep learning. *ISCRAM*.

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.

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

Olteanu, A., Vieweg, S., & Castillo, C. (2015). What to expect when the unexpected happens: Social media communications across crises. *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, 994–1009.

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)*.

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.

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).

Zahra, K., Imran, M., & Ostermann, F. O. (2020). Automatic identification of eyewitness messages on twitter during disasters. *Information processing & management*, *57*(1), 102107.