Are Sudden Crises Making me Collapse? Measuring Transfer Learning Performances on Urgency Detection

Nils Bourgon, Farah Benamara, Alda Mari, Véronique Moriceau, Gaëtan Chevalier, Laurent Leygue

Are Sudden Crises Making me Collapse? Measuring Transfer Learning Performances on Urgency Detection

Nils Bourgon
IRIT, Université de Toulouse, CNRS,
Toulouse INP, UT3
nils.bourgon@irit.fr

Farah Benamara*
IRIT, Université de Toulouse, CNRS,
Toulouse INP, UT3
farah.benamara@irit.fr

Alda Mari†
IJN, CNRS/ENS/EHESS
PSL University
alda.mari@ens.fr

Véronique Moriceau
IRIT, Université de Toulouse, CNRS,
Toulouse INP, UT3
veronique.moriceau@irit.fr

Gaetan Chevalier
DGSCGC SDAIRS
gaetan.chevalier@interieur.gouv.fr

Laurent Leygue
DGSCGC SDAIRS
laurent.leygue@interieur.gouv.fr

ABSTRACT

This paper aims at measuring transfer learning performances across different types of crises related to *sudden or unexpected events* (like earthquakes, terror attacks, explosions, technological incidents) that cannot be foreseen by emergency services and on the occurrence of which they have virtually no control. Although sudden crises are present in most existing crisis datasets, as far as we are aware, no one studied their impact on classifiers performances when evaluated in an out-of-type scenario in which models are tested on a particular type of crisis unseen during training. Our contribution is threefold: (1) A new dataset of about 3,800 French tweets related to four sudden events that occurred in France annotated for both relatedness (i.e., useful vs. not useful for emergency responders) and urgency (i.e., not useful vs. urgent vs. not urgent), (2) A set of monotask and multitask zero-shot learning experiments to transfer knowledge across events and types, and finally, (3) Experiments involving few-shot learning to measure the amount of sudden events instances needed during training to guarantee good performances. When compared to a cross-event setting, our preliminary results are encouraging and show that transfer from predictable ecological crisis to sudden events is feasible and constitutes a first step towards real-time crisis management systems from social media content.

Keywords

Sudden crises, Transfer learning, Few-shot learning, Zero-shot learning, Social media content.

MOTIVATIONS

Recently, Twitter has been widely used to generate valuable information in crisis situations, showing that traditional means of communication between population and rescue departments (e.g., phone calls) are clearly suboptimal (Vieweg et al. 2014; Olteanu et al. 2015). For example, more than 20 million tweets were posted during the superstorm Sandy in 2012 (Castillo 2016) and the hashtag #NotreDame created during the Notre Dame de Paris fire

*corresponding author
†corresponding author

has been the most used in Twitter in 2019.[1] This information is however not fully accessible because it is not always transmitted to the emergency units. An automatic processing of the on-the-spot citizen-witness communication is a need that we aim to fulfill in this study.

The literature on NLP-based crisis management has been growing fast in the recent years (see (Castillo 2016; Imran, Castillo, et al. 2018) for an overview) and several datasets (mainly composed of tweets) have been proposed in order to account for crisis related phenomena.[2] Messages are annotated according to relevant categories deemed to fit the information needs of various stakeholders like humanitarian organizations, local police and firefighters. Well known dimensions include relatedness (also known as usefulness or informativeness) to identify whether the message content is useful (Jensen 2012), situation awareness (also known as urgency, criticality or priority) to filter out on-topic relevant (e.g., immediate post-impact help) vs. on-topic irrelevant information (e.g. supports and solicitations for donations) (Imran, Elbassuoni, et al. 2013; McCreadie et al. 2019; Sarioglu et al. 2020), and eyewitnesses types to identify direct and indirect eyewitnesses (Zahra et al. 2020). These datasets have been used to train classifiers in a supervised way employing either traditional feature-based learning algorithms (Imran, Mitra, et al. 2016; Li et al. 2018; Kaufhold et al. 2020) or deep learning architectures relying on static or contextual embeddings as input (Caragea et al. 2016; Kersten et al. 2019; Liu et al. 2021; Kozlowski et al. 2020; Neppalli et al. 2018; Chowdhury et al. 2020). Due to the extremely imbalanced nature of crisis datasets where out-of-topic tweets are a majority and to better evaluate the classifiers performance in real scenarios (unseen events lack annotated data), three experimental settings were proposed:

- *On-event*: Training and testing data come from the same event, e.g., Nepal earthquake. This is the easiest setting but lacks practical relevance (Wiegmann et al. 2020).
- *Out-of-event*, also known as *Cross-event*: Training on one or several events from various types (e.g., flood, earthquakes and hurricanes) and testing on unseen events of the same type for which no manually annotated data is available at the moment of training (e.g., flood). For example, train on hurricane Sandy and test on hurricane Irma.
- *Out-of-type*: Training on a pool of events related to different types of crises and testing on a particular unseen type, e.g. train on floods and test on hurricanes or earthquakes.

Lots of attention has been devoted to the on-event setting (Caragea et al. 2016; Gata et al. 2019; Nguyen et al. 2017; Sarioglu et al. 2020). For example, Caragea et al. focused on flooding events and used CNN to classify tweets according to their informativeness (Caragea et al. 2016) while Spiliopoulou et al. proposed an adversarial classifier to remove bias for real-time disaster events observing that mixing data from events of different types (e.g., violent mass attacks and floods) confuses the model (Spiliopoulou et al. 2020). However, measuring models transferability across events and types has been less explored. Nguyen et al. trained relatedness binary classifiers both on the event and out-of-event data, the former outperformed the latter (Nguyen et al. 2017). Wiegmann et al. compared the effectiveness of three deep learning models (CNN, BERT and USE) for relatedness classification in both cross-event and out-of-type settings (Wiegmann et al. 2020). Kozlowski et al. trained a transformer on ecological crises and tested it on a completely different type of event (building collapse) observing a decrease in the macro F-score (Kozlowski et al. 2020). Kersten et al. proposed an in-depth analysis of the performance and transferability of global versus local models when considering floods and hurricanes (Kersten et al. 2019). Finally, Algiriyage et al. investigated relatedness classification in on-event, cross-event and out-of-type settings focusing on nine event types. For the three settings, a Bi-LSTM model with Word2Vec features was the best performing one (Algiriyage et al. 2021).

This paper continues this line of research and aims at measuring transfer learning performances across different types of crises while dealing with *sudden or unexpected events*. Compared to ecological crises like hurricanes and floods, sudden events (like earthquakes, terror attacks, explosions, technological incidents) are difficult to predict (Björck 2016). These events, over which organizations have virtually no control, influence social behavior and the ways the emergency services are organized (James and Wooten 2005; Coombs 2014; Quarantelli et al. 2017). Automatically identifying urgent messages in such crises can therefore be a valuable tool that may potentially reduce their negative consequences. Although sudden crises are present in most existing crisis datasets, few studies consider them in transfer learning settings. Among them, Wang et al. use a seq2seq transformers on two benchmark datasets and evaluate cross-domain adaptation (where the train and test sets do not intersect) (Wang et al. 2021) while Hongmin et al. combine self-training with CNN and BERT models to improve the performance of tweet classifiers for a target

[1] https://blog.twitter.com/en_us/topics/insights/2019/ThisHappened-in-2019

[2] See https://crisisnlp.qcri.org/ for an overview.

disaster, where labeled data is assumed to be unavailable (Hongmin et al. 2021). We want here to go further by newly: (1) Investigate urgency classification going beyond state of the art relatedness binary classification when evaluated in an out-of-type scenario and (2) Measure the amount of data needed to achieve acceptable performances when evaluated in a few-shot learning configuration. When compared to a cross-event setting, our preliminary results are encouraging showing that transfer from ecological and predictable crisis to sudden events is feasible which constitutes a first step towards real-time crisis management systems. Our contributions are as follows:

- A new dataset of about 3,800 French tweets related to four sudden events that occurred in France annotated for both relatedness (i.e., useful vs. not useful for emergency responders) and urgency (i.e., not useful vs. urgent vs. not urgent).
- A set of monotask and multitask zero-shot learning experiments to transfer knowledge across events and types while addressing the particular nature of sudden events. To our knowledge, no one explored out-of-event and cross-event detection for both relatedness and urgency while dealing with class imbalance.
- Development of few-shot learning experiments to measure the amount of sudden events training instances needed to guarantee good performances.

This paper is organized as follows. Section 2 presents our dataset. Section 3 and Section 4 detail, respectively, the zero-shot and few-shot learning experiments as well as the results. Section 5 concludes with a discussion highlighting the main findings of this study.

DATASET

Since our focus is on crises that occur in metropolitan France and its overseas departments, we rely on the only available corpus of French tweets by Kozlowski et al. 2020[3] composed of 12,826 tweets collected using dedicated keywords about ecological crises that occurred in France from 2016 to 2019 and posted 24h before, during (48h) and 72h after the crisis: 2 floods that occurred in Aude and Corsica regions, 8 storms (Béryl, Berguitta, Fionn, Eleanor, Bruno, Egon, Ulrika, Susanna), 2 hurricanes (Irma and Harvey), and 1 sudden crisis (Marseille building collapse).

In this dataset, each tweet is annotated according to an urgency classification composed of three categories: Urgent that applies to messages mentioning human/infrastructure damages as well as security instructions to limit these damages during crisis events, 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 to raise situational awareness, and finally Not useful for messages that are not related to the targeted crisis or information pertaining to events occurring outside the French territories. Among theses three categories, Urgent and Not Urgent messages are considered as useful messages for emergency units.

The collection is extremely imbalanced with 1,442 (11.24%) useful but Not urgent messages, 2,147 (16.74%) Urgent and 9,237 (72.02%) Not useful, which is in line with the proportions reported in other crisis corpora. This dataset has been newly augmented with tweets about four sudden crises posted from the onset of the crisis to 48h after using dedicated keywords possibly coupled with geotag information when available to target tweets posted in the crisis area (please note that geographical information was not used to collect the original dataset): 534 for gas explosion at Sanary-sur-Mer, 1,358 Lubrizol Rouen Plant fire, 520 Notre-Dame de Paris fire and 1,384 concerning the Trèbes supermarket terror attack. The inter-annotator agreements being relatively good in the original dataset (Cohen's Kappa=0.722 for relatedness and 0.658 for urgency (see (Kozlowski et al. 2020)), we decided to annotate the augmented dataset following the same annotation guidelines. Four annotators were asked to label the tweets which results in the distribution shown in Table 1.

We observe that ecological crises share a similar distribution across urgency categories. Indeed the flood crises have similar proportions of not urgent and urgent messages. On the other hand, hurricanes do not follow this pattern. This can be explained by the differences in the impact of these two events in France, which have also affected other countries. In fact, it is not surprising to see more messages of support for France during Irma (and at the same time more informative messages in general), because it had a great impact on France (and specifically on the Saint-Martin and Saint-Barthélemy islands), while Harvey, which was first planned in the French Antilles, had in the end a bigger impact notably on Texas. The eight events in the storms subset contain fewer useful messages overall than other types of crises, fewer not urgent messages, and particularly fewer support messages compared to other crises. Compared to ecological crisis, the Marseille building collapse has a significantly different distribution from the rest of the corpus. Overall, it contains few informative messages but they are evenly distributed. This

[3] https://github.com/DiegoKoz/french_ecological_crisis

Table 1. Dataset statistics.

	Ecological crises	*Building collapse*	*Gas explosion*	*Plant fire*	*Notre Dame fire*	*Terrorist attack*	***Total***
URGENT	1,153	44	363	583	225	398	**2,766**
NOT URGENT	2,343	49	165	637	209	812	**4,215**
NOT USEFUL	8,610	627	6	138	86	174	**9,641**
Total	**12,106**	**720**	**534**	**1,358**	**520**	**1,384**	**16,622**

crisis has many fewer messages of Warning-advice, and this is explained by the fact that it was a sudden event. Hence the media was unable to provide warnings in advance of its arrival.

Compared to ecological crises, the proportions of not useful messages in the augmented sudden crises dataset are low with 1.12%, 10.16%, 16.54% and 12.57% for gas explosion, plant fire, Notre Dame and terrorist attack respectively. On the other hand, urgent messages are a majority for all the crises except the terrorist attack where around 29% of the messages are urgent. These frequencies are however different from the ones observed in the Marseille building collapse where only 6.11% and 6.81% tweets have been annotated as urgent and not urgent respectively. This can be explained by the way this dataset has been collected. Indeed, the use of the "Marseille" keyword to retrieve tweets posted 24h before, 48h during and 72 after the collapse has biased the collection towards the non useful class, as this keyword also refers to a famous football club (Olympique de Marseille) and rap music groups from that city.

ZERO-SHOT LEARNING ACROSS EVENTS AND TYPES

Experimental Settings

Given a tweet, our aim is to develop binary (useful vs. not useful) and three-class (urgent vs. not urgent vs. not useful) deep learning models to deal respectively with relatedness and urgency classifications while measuring models ability to generalize over new unseen sudden events. We experiment with several deep learning models, in particular: CNN with FastText static French embeddings following (Nguyen et al. 2017) and transformer architectures relying on several multilingual (Bert (Devlin et al. 2019)) and French contextualized pre-trained embeddings (FlauBERT (Le et al. 2020) and CamemBERT (Martin et al. 2019)). We only report here the models having the best results.

– **FlauBERT$_{\text{tuned}}$**. It uses the FlauBERT base cased model. We run the HuggingFace's PyTorch implementation of FlauBERT for four epochs and a learning rate of $2e-5$. For better convergence, we use the linear decreasing learning rate during optimization. To avoid exploding gradients, we use a gradient clipping of 1.0. We further fine-tuned FlauBERT$_{\text{base}}$ language models initially trained on a general domain by using 358,834 unlabeled tweets of the dataset. The adaptation consists of training the language models with a masked language model head then use the shifted weights to perform the classification. This process is similar to the initial BERT training. To handle class imbalance, we use either the focal loss (Lin et al. 2017) or the weighted cross entropy loss. Our aim here is to compare with one of the most effective approach for handling imbalanced data (Cui et al. 2019).

– **ML-FlauBERT$_{\text{tuned}}$**. This is a multitask learning version of FlauBERT$_{\text{tuned}}$ that simultaneously learns relatedness (binary) and urgency (three classes). The classifiers for both tasks share and update the same low layers of FlauBERT$_{\text{tuned}}$ except the final task-specific classification layer.

Our dataset being a new extension of an already existing one, we split the data into three non overlapping sets: Eco that groups all ecological crises, $Coll$ considers only the Marseille building collapse and $Sudden$ for the newly added sudden events. We finally add $Sudden_{all} = Coll \cup Sudden$ to group sudden crises all together. Since the number of instances in each event in the original dataset is relatively small (720 for each storm and about 1,500 for some flooding events), we designed three CROSS-EVENT and one OUT-OT-TYPE experimental settings, each one involving training and testing of FlauBERT models on different subsets of Eco, $Coll$ and $Sudden$:

- CROSS-EVENT$_{Eco \rightarrow Eco}$: Train on 80% of Eco and test on the remaining 20% following a random split. This setting can be considered as the most ideal situation where the test is composed of events of the same type (ecological crisis).

- Cross-event$_{Eco+Coll \rightarrow Eco+Coll}$: Train on 80% of $Eco + Coll$ and test on the remaining 20% following a random split. This setting aims to measure the impact of a sudden crisis that has the same distribution of the not useful class than the predictable crises on the overall performances when evaluated on the initial dataset.

- Cross-event$_{Eco+coll \rightarrow Sudden}$: Train on $Eco + Coll$ (77.16%) and test on $Sudden$ (22.84%). This setting makes use of the original data for training where only one event is a sudden crisis and the augmented one for testing.

- Out-of-type$_{Eco \rightarrow Sudden_{all}}$: Train on Eco (72.83%) and test on $Sudden_{all}$ (27.17%).

The first three experiments use a training set composed of crises of different types while the last one is the zero-shot configuration that considers sudden crises only for testing the models. Note that for all these four configurations, the dataset has not been balanced, keeping the original imbalanced nature of the classes in both the train and test sets. Prior to the experiments, a pre-processing step removed URLs and mentions from the tweets and replaced numbers with a tag.

Results

Table 2 presents the results of our experiments on relatedness and urgency classifications in terms of precision (P), recall (R) and macro-F1 (F), the latter measure treats all the classes equally which makes it the most suited for measuring the performances of imbalanced classification problems. When testing with sudden events, the results are low regarding our baseline which has been trained and tested only on ecological crisis: introducing new events with a lexical field not present in the training dataset, may explain this decrease.

For all the settings and as expected, binary prediction (in terms of F-scores) outperform urgency classification. Also, multitask models were not productive compared to their monotask counterpart, except for the Cross-event$_{Eco+Coll \rightarrow Sudden}$ setting where the F-scores have been boosted for relatedness (0.474 for monotask vs. 0.488 for multitask).Incorporating $Coll$ in the training set degrades the results when testing on $Sudden$ (e.g., 0.474 vs. 0.488 for relatedness in the monotask and multitask settings) while training only on Eco and testing on $Sudden_{all}$ was the most productive with an F-score=0.640 for binary and F-score=0.554 for three-class classification. Overall, the results of out-of-type indicate that training on expected events is the most appropriate to classify sudden crises.

A closer look into the results obtained by FlauBERT$_{\text{tuned}}$, our best model, per event in the out-of-type configuration (see Table 3) indicates that the Not useful class achieves a good recall whereas Useful a good precision. Indeed, concerning newly collected sudden events, the Useful class is rarely predicted since it is less frequent in the training set (9.5%) than in the test set (45%). This suggests that the zero-shot learning approach allows to (1) filter out irrelevant and off-topic messages and (2) detect useful messages (both urgent and not urgent) with a very high proportion of true positives. This is a very important result in a use case where emergency services monitor social media in order to better set priorities and decide appropriate rescue actions.

The detailed results per crisis in terms of macro F-scores (see Table 4) show that the building collapse obtained the best results followed by the plant fire. The very good scores obtained for the collapse crisis can be explained by (1) the similar distributions of the classes in the training and test sets (see Section 2), and (2) the good precision for the Not Useful class compared to the other sudden events probably because non useful messages are much more frequent for the building collapse (87%) due to the scrapping method we used to collect new sudden events. Although this is closer to a real life scenario where useful tweets need to be found among a lot of data, training our classifiers on this distribution leads to a bias towards the most represented class Not Useful resulting in misclassification such as:

- Une prise d'otage serait en cours dans un hypermarché de #Trèbes près de #Carcassonne. L'homme aurait d'abord agressé un groupe de CRS et blessé l'un d'entre eux par balle [...]
  (An hostage taking is reportedly underway at a hypermarket in #Trèbes near #Carcassonne. The man is said to have first assaulted a group of policemen and shot one of them [...]
  Gold= Urgent, Predicted= Not Useful

Table 2. Zero-shot and few-shot learning results for relatedness and urgency classifications. When present, f means that the corresponding setting has been trained with the focal loss, weighted entropy otherwise.

		FlauBERT$_{\text{tuned}}$			ML-FlauBERT$_{\text{tuned}}$		
		P	R	F	P	R	F
CROSS-EVENT$_{Eco \rightarrow Eco}$	RELATEDNESS	0.840	0.862	0.849	0.841	0.859	0.849
	URGENCY	0.768	0.799	0.782	0.754	0.787	0.769
CROSS-EVENT$^{f}_{Eco+Coll \rightarrow Eco+Coll}$	RELATEDNESS	0.847	0.862	0.854	0.843	0.857	0.850
	URGENCY	0.769	0.773	0.770	0.759	0.779	0.768
CROSS-EVENT$_{Eco+Coll \rightarrow Sudden}$	RELATEDNESS	0.574	0.693	0.474	0.578	0.703	0.488
	URGENCY	0.638	0.602	0.465	0.611	0.586	0.466
OUT-OF-TYPE$_{Eco \rightarrow Sudden_{all}}$	RELATEDNESS	0.676	0.749	0.640	0.665	0.729	0.607
	URGENCY	0.679	0.625	0.554	0.668	0.608	0.549

Table 3. Detailed results per sudden event in the OUT-OF-TYPE configuration for relatedness and urgency classifications obtained by the best model FlauBERT$_{\text{tuned}}$. Best scores per class are in bold font.

		Gas explosion		Plant fire		Notre Dame		Attack		Collapse	
		P	R	P	R	P	R	P	R	P	R
RELATEDNESS	USEFUL	**0.997**	**0.695**	0.967	0.634	0.978	0.313	0.972	0.431	0.652	0.645
	NOT USEFUL	0.03	0.833	0.2	0.812	0.218	**0.965**	0.188	0.914	**0.947**	0.949
URGENCY	URGENT	**0.894**	0.747	0.735	0.667	0.857	0.453	0.814	0.374	0.576	**0.773**
	NOT URGENT	0.758	0.285	0.809	0.265	0.826	0.091	**0.926**	**0.398**	0.515	0.347
	NOT USEFUL	0.03	0.833	0.187	0.841	0.214	0.942	0.188	0.92	**0.951**	**0.952**

Table 4. Detailed results per sudden event in the OUT-OF-TYPE configuration for relatedness and urgency classifications in terms of averaged macro F-scores obtained by the best model FlauBERT$_{\text{tuned}}$. Best scores per task are in bold font.

	Gas explosion	Plant fire	Notre Dame	Attack	Collapse
RELATEDNESS	0.439	0.544	0.415	0.455	**0.798**
URGENCY	0.428	0.468	0.369	0.46	**0.675**

FEW-SHOT LEARNING EXPERIMENTS

Experimental Settings

We have now shown that monotask learning models trained on expected events are the most suitable ones for detecting sudden events in a zero-shot scenario (i.e. the OUT-OF-TYPE$_{Eco \rightarrow Sudden_{all}}$ configuration), the next step is to measure the amount of sudden events needed in the training set to boost the results in a few-shot learning configuration. To this end, we designed two additional CROSS-EVENT experimental settings while randomly varying the number of instances from the sudden crises in the train set. Among the various amounts we tested, we report here the following settings:

- FEW-SHOT 20%: Train on Eco+ (20% of $Coll$) + (20% of each event in $Sudden$) and test on the remaining 80% of ($Coll$+80% $Sudden$).
- FEW-SHOT 5%: Train on Eco+ (5% of $Coll$) + (5% of each event in $Sudden$) and test on the remaining 95% of ($Coll$+95% $Sudden$).

For both settings, we rely on FlauBERT$_{\text{tuned}}$, our best performing model in the zero-shot experiments while keeping the original distribution of the dataset.

Results

Table 5 shows our results in terms of precision (P), recall (R) and macro-F1 (F). We compare few-shot configurations with the zero-shot one where 100% $Sudden_{all}$ has been used for testing (see Table 2). Without any surprise, the more the training set contains instances from $Sudden_{all}$, the better the results. For relatedness, the F1-scores were 0.851 for FEW-SHOT 20% vs. 0.827 for FEW-SHOT 5% while 0.787 vs. 0.750 for urgency. We observe however that only injecting 5% tweets of each crisis in the train improves the results of the classification. To generalize

predictions to unseen events, the model needs few annotated tweets to achieve a good F1-score. For each crisis less than 70 tweets where provided to the train, those messages help the model by introducing keywords and a lexical field specific to the crisis. When looking at Table 6 we can see that Useful and Urgent instances achieve a high recall and F1-score.

Table 5. Few-shot learning vs. zero-shot learning results for relatedness and urgency classifications.

		FlauBERT$_{tuned}$		
		P	R	F
Zero-shot	relatedness	0.676	0.749	0.640
	urgency	0.679	0.625	0.554
Few-shot 20%	relatedness	0.876	0.832	**0.851**
	urgency	0.802	0.779	**0.787**
Few-shot 5%	relatedness	0.836	0.819	0.827
	urgency	0.750	0.756	0.750

Table 6. Micro and macro scores over all the sudden crises in the 5% few-shot learning setting for relatedness and urgency classifications. Zero-shot results are provided for a better comparison

		Few-shot 5%			Zero-shot		
		P	R	F	P	R	F
Relatedness	Useful	0.914	0.933	0.923	0.957	0.587	0.728
	Not useful	0.757	0.705	0.730	0.395	0.912	0.551
	Macro scores	0.836	0.819	0.827	0.676	0.749	0.640
Urgency	Urgent	0.757	0.888	0.817	0.763	0.697	0.728
	Not Urgent	0.788	0.671	0.725	0.895	0.255	0.397
	Not useful	0.706	0.708	0.707	0.379	0.922	0.537
	Macro-Average	0.750	0.756	0.750	0.679	0.625	0.554

CONCLUSION

In this paper, we have presented a new dataset of about 3,800 French tweets related to four sudden events that occurred in France annotated for both relatedness and urgency. We also studied the impact of the presence of these sudden crises on classifiers performances when evaluated in an out-of-type scenario. When compared to a cross-event setting, our results show that transfer from ecological crisis to sudden events is feasible, which constitutes a first step towards real-time crisis management systems. We conducted several experiments in either zero-shot or few-shot learning configurations and our main findings are:

- The zero-shot learning approach allows us to (1) filter out irrelevant and off-topic messages and (2) detect useful messages (both urgent and not urgent) with a very high proportion of true positives. This is a very important result in a use case where emergency services monitor social media in order to better set priorities and decide appropriate rescue actions.
- The few-shot learning experiments show that adding only 5% of sudden event data to the training set allows to achieve better results than the zero-shot baseline.

In the future, we plan to continue this preliminary study by first enlarging our sudden crises dataset and second exploring transfer learning across crisis types in a multilingual setting.

ACKNOWLEDGMENT

The work presented in this paper has been supported by the INTACT project funded by FIESP-French Ministère de l'Intérieur (Department of Home Affairs) and the AAP CNRS - INHESJ 2020. Both projects involved the Institut de Recherche en Informatique de Toulouse (IRIT) and Institut Jean Nicod (IJN). The research of Farah Benamara and Véronique Moriceau is also partially supported by DesCartes: the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program. The research of Alda Mari is partially supported by ANR-17-EURE-0017 FrontCog.

REFERENCES

Algiriyage, N., Prasanna, R., Stock, K., and Johnston, D. (2021). "dentifying Disaster-related Tweets: A Large-Scale Detection Model Comparison". In: *Social Media in Crises and Conflicts, Proceedings of the 18th ISCRAM Conference*, pp. 731–743.

Björck, A. (2016). "Crisis Typologies Revisited: An Interdisciplinary Approach". In: *Central European Business Review* 2016.3, pp. 25–37.

Caragea, C., Silvescu, A., and Tapia, A. (Mar. 2016). "Identifying Informative Messages in Disasters using Convolutional Neural Networks". In: *13th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2016)*, pp. 1–7.

Castillo, C. (2016). *Big crisis data: Social media in disasters and time-critical situations*. Cambridge University Press, pp. 1–212.

Chowdhury, J. R., Caragea, C., and Caragea, D. (2020). "On Identifying Hashtags in Disaster Twitter Data". In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, pp. 498–506.

Coombs, W. T. (2014). "Ongoing crisis communication: Planning, managing, and responding". In: *Thousand Oaks, CA: Sage.*

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. (June 2019). "Class-Balanced Loss Based on Effective Number of Samples". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics.

Gata, W., Amsury, F., Wardhani, N. K., Sugiyarto, I., Sulistyowati, D. N., and Saputra, I. (2019). "Informative Tweet Classification of the Earthquake Disaster Situation In Indonesia". In: *2019 5th International Conference on Computing Engineering and Design (ICCED)*, pp. 1–6.

Hongmin, L., Doina, C., and Cornelia, C. (2021). "Combining Self-training with Deep Learning for Disaster Tweet Classification". In: *ISCRAM 2021 Conference Proceedings – 18th International Conference on Information Systems for Crisis Response and Management*, pp. 655–666.

Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (2018). "Processing Social Media Messages in Mass Emergency: Survey Summary". In: *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*. Ed. by P. Champin, F. Gandon, M. Lalmas, and P. G. Ipeirotis. ACM, pp. 507–511.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., and Meier, P. (2013). "Extracting information nuggets from disaster-related messages in social media". In: *Proc. of ISCRAM, Baden-Baden, Germany*.

Imran, M., Mitra, P., and Castillo, C. (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).

James, E. and Wooten, L. (2005). "Leadership as (un)usual: How to display competence in times of crisis". In: *Organizational Dynamics* 34.2, pp. 141–152.

Jensen, G. E. (2012). "Key criteria for information quality in the use of online social media for emergency management in New Zealand." MA thesis. Master thesis, Victoria University of Wellington.

Kaufhold, M.-A., Bayer, M., and Reuter, C. (2020). "Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning". In: *Information Processing & Management* 57.1, pp. 102–132.

Kersten, J., Kruspe, A., Wiegmann, M., and Klan, F. (2019). "Robust Filtering of Crisis-related Tweets". In: *Social Media in Crises and Conflicts, Proceedings of the 16th ISCRAM Conference*, pp. 814–824.

Kozlowski, D., Lannelongue, E., Saudemont, F., Benamara, F., Mari, A., Moriceau, V., and Boumadane, A. (2020). "A three-level classification of French tweets in ecological crises". In: *Information Processing & Management* 57.5, p. 102284.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (May 2020). "FlauBERT: Unsupervised Language Model Pre-training for French". English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 2479–2490.

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

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988.

Liu, J., Singhal, T., Blessing, L. T. M., Wood, K. L., and Lim, K. H. (2021). "CrisisBERT: A Robust Transformer for Crisis Classification and Contextual Crisis Embedding". In: *HT*, pp. 133–141.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., Villemonte de la Clergerie, É., Seddah, D., and Sagot, B. (Nov. 2019). "CamemBERT: a Tasty French Language Model". In: *arXiv e-prints*, arXiv:1911.03894, arXiv:1911.03894. arXiv: 1911.03894 [cs.CL].

McCreadie, R., Buntain, C., and Soboroff, I. (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*. Ed. by Z. Franco, J. J. González, and J. H. Canós. ISCRAM Association.

Neppalli, V. K., Caragea, C., and Caragea, D. (2018). "Deep Neural Networks versus Naive Bayes Classifiers for Identifying Informative Tweets during Disasters". In: *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management*. ISCRAM'2018.

Nguyen, D. T., Al Mannai, A., Joty, S., Sajjad, H., Imran, M., and Mitra, P. (2017). "Robust classification of crisis-related data on social networks using convolutional neural networks". In: *Eleventh International AAAI Conference on Web and Social Media*.

Olteanu, A., Vieweg, S., and Castillo, C. (2015). "What to Expect When the Unexpected Happens: Social Media Communications Across Crises". In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW '15, pp. 994–1009.

Quarantelli, E., Boin, A., and Lagadec, P. (2017). "Studying Future Disasters and Crises: A Heuristic Approach." In: *Handbook of Disaster Research*, pp. 61–83.

Sarioglu, E., Nan, L., Qu, B., Diab, M., and McKeown, K. (Dec. 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*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 4693–4703.

Spiliopoulou, E., Maza, S., Hovy, E., and Hauptmann, A. (2020). "Event-Related Bias Removal for Real-time Disaster Events". In: *Proceedings of Findings at EMNLP 2020.*

Vieweg, S., Castillo, C., and Imran, M. (2014). "Integrating Social Media Communications into the Rapid Assessment of Sudden Onset Disasters". In: *Proceedings of the 6th International Conference of Social Informatics*. SocInfo'14, pp. 444–461.

Wang, C., Nulty, P., and Lillis, D. (2021). "Crisis Domain Adaptation Using Sequence-to-Sequence Transformers". In: *ISCRAM 2021 Conference Proceedings – 18th International Conference on Information Systems for Crisis Response and Management*. Ed. by A. Adrot, R. Grace, K. Moore, and C. W. Zobel. Blacksburg, VA (USA): Virginia Tech, pp. 655–666.

Wiegmann, M., Kersten, J., Klan, F., Potthast, M., and Stein, B. (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.

Zahra, K., Imran, M., and Ostermann, F. O. (2020). "Automatic identification of eyewitness messages on twitter during disasters". In: *Information Processing & Management* 57.1, pp. 102–107.