A three-level classification of French tweets in ecological crises

Diego Kozlowski, Elisa Lannelongue, Frédéric Saudemont, Farah Benamara, Alda Mari, Véronique Moriceau, Abdelmoumene Boumadane

A Three-level Classification of French Tweets in Ecological Crises

Diego Kozlowski[a], Elisa Lannelongue[b], Frédéric Saudemont[a], Farah Benamara[a,*], Alda Mari[b], Véronique Moriceau[a], Abdelmoumene Boumadane[a]

[a]*IRIT, Université de Toulouse, France*
firstname.lastname@irit.fr
[b]*ENS-INJ, CNRS, France*
firstname.lastname@ens.fr

Abstract

The possibilities that emerge from micro-blogging generated content for crisis-related situations make automatic crisis management using natural language processing techniques a hot research topic. Our aim here is to contribute to this line of research focusing for the first time on French tweets related to ecological crises in order to support the French Civil Security and Crisis Management Department to provide immediate feedback on the expectations of the populations involved in the crisis. We propose a new dataset manually annotated according to three dimensions: relatedness, urgency and intentions to act. We then experiment with binary classification (useful vs. non useful), three-class (non useful vs. urgent vs. non urgent) and multiclass classification (i.e., intention to act categories) relying on traditional feature-based machine learning using both state of the art and new features. We also explore several deep learning models trained with pre-trained word embeddings as well as contextual embeddings. We then investigate three transfer learning strategies to adapt these models to the crisis domain. We finally experiment with multi-input architectures by incorporating different metadata extra-features to the network. Our deep models, evaluated in random sampling, out-of-event and out-of-type configurations, show very good performances outperforming several competitive baselines. Our results define the first contribution to the field of crisis management in French social

*Corresponding author (farah.benamara@irit.fr)

media.

Keywords: Crisis response from social media, Machine learning, Natural Language Processing, Transfer learning

1. Introduction

The rise of social networks changed the way people interact, communicate and access information all over the world. Their use is widespread among individuals, public institutions and companies to convey information about events or products, convey points of view and opinions and share and contrast those opinions with others. All areas of daily life are involved, including civil security and crisis management, the area targeted in this paper.

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, Palen and Liu, 2007). For example, more than 20 million tweets were posted during the superstorm Sandy in 2012 (Castillo, 2016) and around 17,000 during the Notre Dame fire that occurred in France in 2019. When analysing tweets posted by individuals and the media during crisis situations, one can observe that they express their *intentions, desires, plans, goals* and *preferences* to act. For example[1], in the tweet (1), the writer publicly expresses an explicit commitment to provide help after the Irma hurricane tragedy, using an explicit action verb ("*to help*") which is under the scope of an explicit attitude verb ("*want*"). (2) expresses an intention to complain about the absence of assistance without using any explicit intent keywords. Intention to advise, evacuate (cf. (3)) and inform about people's movement (cf. (4)) are other types of actions expressed in crisis situations. It is important to note that such useful messages do not always require an urgent and rapid action from rescue teams: messages like (4), about affected people, or infrastructure damages can be seen as more urgent compared to others types of intention to act (cf. (1) and (2)). Also, both urgent and non urgent messages are drowned in a deluge of off-topic or personal messages that may contain crisis-related keywords, like the word ("*flood*") in (5). Therefore, new tools are needed to early access

[1]All the examples presented in this paragraph are French tweets taken from our corpus and translated into English.

useful information that will allow the emergency units to anticipate actions, coordinate efforts, and share information (Imran et al., 2015, DePaula et al., 2017, Avvenuti et al., 2018, Kim and Hastak, 2018).

(1) #Irma Hurricane: "I want to go there to help."

(2) Irma hurricane: where is disaster assistance one month later?

(3) Emergency heritage at Bordeaux. After the flood, the archaeology lab is looking for volunteers to evacuate collections.

(4) Midnight at Nemours: Surrealist scene: boats, canoes take locals to get their stuff at home #flood

(5) - Darling, I feel that you are not willing to settle the flood in the apartment.

Automatic crisis management using Natural Language Processing (NLP) techniques has recently become a hot topic in the research community (Imran et al., 2015, Qadir et al., 2016). Given a corpus of messages (mainly tweets) collected during a crisis event, the aim is to classify each message according to its relatedness (i.e., useful vs. non useful for emergency responders also known as on-topic vs. off-topic) and/or the type of intentions to act following a predefined taxonomy. Approaches range from unsupervised (Alam et al., 2018, Ning et al., 2017, Zhang and Vucetic, 2016) to supervised learning, using either feature-based (Li et al., 2018, Stowe et al., 2018) or deep learning techniques (Nguyen et al., 2016a,b, Caragea et al., 2016, Neppalli et al., 2018, Aipe et al., 2018). Linguistic resources being the core element in supervised learning, many manually annotated crisis datasets have been built. Most resources are in English (Imran et al., 2016), 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 and Lee, 2019).

Our aim here is to contribute to this line of research focusing for the first time on French tweets related to ecological crises (hurricanes, storms, floods, etc.) in order to support French regional and national alert processing centers to provide immediate feedback on the expectations of the populations involved in crises. Our main contributions are the following:

- **A novel characterization of crisis-related messages according to three dimensions**: relatedness, intentions to act and *urgency*.

While the first two have already been used in the literature, as far as we know, no one has explored urgency as an intermediate level of classification.

- **The first French crisis dataset of around 13k tweets** manually annotated following this new characterization[2]. Annotations focus on crises that occur in metropolitan France and its overseas departments targeting the most discussed crises in the French media. This includes very local crises but also larger crisis events that may affect other countries or continents. In this last case, only messages concerning French territories are relevant. This poses a challenging issue regarding class imbalance as the amount of non useful information that comes with the data is important.

- **A set of experiments for organizing the information gathered on Twitter on crisis situations**. We experiment with binary classification (useful vs. non useful), three classes (non useful vs. urgent vs. non urgent) and multiclass classifications (i.e., intention to act categories). We rely on traditional feature-based machine learning using state of the art features whose efficiency has been empirically proved, and new groups of features. We also explore several deep learning models trained on pre-trained word embeddings as well as contextual multilingual and French embeddings. We then investigate three transfer learning strategies to adapt these models to the crisis domain: language model fine-tuning, domain shift on embeddings and multitask learning. We finally experiment with multi-input architectures by incorporating different metadata extra-features to the network. As far as we know, this is the first study in the field of crisis management that (1) experiments with French transformer-based architectures for crisis management, (2) investigates metadata features in deep architectures, and (3) explores multiple transfer learning models for classification of tweets for natural disasters.

- **Quantitative and qualitative evaluation of our models**. We evaluate in *random sampling*, *out-of-event* and *out-of-type* configura-

[2]A subset of the annotated data is available here https://github.com/DiegoKoz/french_ecological_crisis. The full dataset will be made freely available for the research community upon final acceptance.

tions and provide a detailed error analysis highlighting main causes of misclassification. Out-of-event considers training the model on a set of previous events, and testing it on other temporal non-overlapping events. On the other hand, out-of-type detection aims at training on specific event types (e.g., ecological crises) and testing on other crisis types (e.g., building collapse). Our models achieve very good results outperforming several competitive baselines. These results define the first contribution to the field of crisis management in French social media.

In the next section we provide an overview of the main existing manually annotated resources for crisis management as well as NLP approaches proposed in the literature. Our dataset, the annotation scheme and the quantitative and qualitative analysis are presented in Section 3. Classification models are described in Section 4. Results and error analysis are respectively presented in sections 5 and 6. We end this paper by highlighting the main conclusions of this study and discuss directions for future work.

2. Related Work

2.1. Crisis Datasets

Crisis datasets are mainly tweets extracted using keywords and/or hashtags about the crises names (e.g., #sandy for Sandy superstorm), time period (which usually includes *during* and *after* the crisis event), location (#nyc when Sandy hit New York), or crisis-related keywords (e.g., *flood*, *hurricane*, *storm*). Additional metadata, like user profiles or tweet geolocalisation[3] can additionally be used to better target messages of interest. Tweets are collected via the Twitter Search API or dedicated platforms that aim at crawling tweets in real time. Well known platforms include Artificial Intelligence for Disaster Response[4] (AIDR) (Imran et al., 2014), Twitcident (Abel et al., 2012), Twitris[5] (Purohit and Sheth, 2013) and TweetTracker[6] (Kumar et al., 2011).

[3]Messages with GPS coordinates are a minority.
[4]http://aidr.qcri.org/
[5]http://twitris.knoesis.org/
[6]http://tweettracker.fulton.asu.edu/

The extracted tweets are then manually labelled according to relevant categories[7] that are deemed to fit the information needs of various stakeholders including humanitarian organizations, local police or firefighters. Annotations are performed either by crowd-sourcing workers, humanitarian volunteers or domain experts. Relevance criteria found in the literature can be grouped into the following dimensions:

- *Relatedness* also known as *usefulness* or *informativeness*: Given a message, identify whether its content is useful, adequate and provides valuable information that might be relevant to rescue teams (Jensen, 2012). This dimension is used in almost all state of the art annotation guidelines (Habdank et al., 2017, Kaufhold et al., 2020).

- *Situation awareness*: A message is relevant if it "demonstrates an awareness of the scope of the crisis as well as specific details about the situation" (Verma et al., 2011). Vieweg (2012) further distinguishes between on-topic relevant information that can aid people in making decisions, advise others or offer immediate post-impact help, and on-topic irrelevant including offers, supports and solicitations for donations to charities. Imran et al. (2013) on the other hand, consider personal only, informative direct (post written by an eyewitness) or informative indirect (post written by a person based on information from news, radio or television).

- *Information type* that indicates various types of intention to act categories from a predefined taxonomy such as: caution or advice, donations, people missing, found, or seen and damage infrastructure (Imran et al., 2016, Olteanu et al., 2015).

- *Information source* that identifies the source of the tweet, i.e., who is the author of the message among individuals, media, public and private organizations (Olteanu et al., 2015).

- *Eyewitnesses types*. This dimension was recently proposed by Zahra et al. (2020) who identify three types of eyewitnesses: direct (first-hand knowledge and experience of an event), indirect (messages shar-

[7]Annotations are usually done at the text level. Some studies propose to additionally annotate images within the tweets (see for example (Alam et al., 2018)).

ing valuable information from direct witnesses) and vulnerable direct eyewitness (users reporting warnings and alerts).

Existing datasets are either annotated according to one of the dimensions above or using several dimensions in cascade like relatedness or situation awareness first, then information type for messages that have been identified as relevant. Most annotated datasets are in English. Crisis datasets in other languages include Spanish (Cobo et al., 2015) and Arabic (Alharbi and Lee, 2019) both annotated according to relatedness. For Italian, Cresci et al. (2015) propose a three-level classification following the information type dimension: "damage", "no damage", or "not relevant".

In this paper, we propose the first French dataset manually annotated according to three dimensions: relatedness, intentions to act and *urgency*. The last dimension is new and was inspired by Vieweg (2012) who defined criteria for situational relevance and non relevance. However, as pointed out by Zade et al. (2018), these criteria are too general and agnostic to the fact that "information relevance may vary across responder role, domain, and other factors". We therefore refine and adapt Vieweg (2012)'s definition to better address the French Civil Security and Crisis Management Department's (hereafter F-CSCM) specifications who perceive actionability in terms of emergency, i.e., the need to rapidly organize the information and set priorities for the human teams that will later read the relevant tweets and decide appropriate rescue actions. We thus propose a three-level taxonomy: First a binary classification of usefulness of the tweet. Then a ternary one for urgency (non useful, useful urgent and non urgent). Finally, further refine useful messages into specific intentions to actionable categories. Actionability is achieved by our three level categorization, which maps fine-grained categories into the needs expressed by the actors on the field.

2.2. NLP approaches for crisis management

This section presents a brief state of the art of NLP-based approaches for crisis management in social media. For a more comprehensive overview of other computational methods in mass emergency, see (Castillo, 2016) and (Imran et al., 2015).

Most approaches are supervised[8] casting the problem either as a binary (i.e., useful vs. non useful) or multiclass (mainly intention to act categories)

[8]Unsupervised techniques (mainly topic modelling) have also been employed either to

classification. Classifiers are evaluated either in on-event (training and testing tweets from the same event) or out-of-event (testing on unseen or sudden events for which no manually annotated data is available at the moment of training) configurations. The results for the former generally outperform the latter.

Besides bag of words representation (Stowe et al., 2018, Palmer et al., 2016, Zahra et al., 2020, Kaufhold et al., 2020), many features have been explored to help the classification task: *linguistic features*, like the sentence structure, lexical density, part-of-speech tags or named entities (Ning et al., 2017, Kaufhold et al., 2020), *temporal features* such as the temporal difference between the post and the event (Kaufhold et al., 2020), *emotional/sentiment features* relying on dedicated lexicons (Ning et al., 2017), *surface-based features* including URL or image presence, punctuation and emoticons (Truong et al., 2014, Li et al., 2018), and finally *user-based features* which refer to message metadata information (e.g., followers, favorites, verified account) (Neppalli et al., 2018).

These features have been used to train different learning models. Li et al. (2018) use different word embedding techniques, like Word2Vec (Mikolov et al., 2013a,b), FastText (Bojanowski et al., 2016) and GloVe (Pennington et al., 2014), both from pre-trained embeddings and built using crisis-related corpora[9], and use statistical metrics to collapse dimensions of the embedding. These representations are then used in a set of traditional machine learning models (Support Vector Machine (SVM) (Cortes and Vapnik, 1995), Random Forest (Breiman, 2001)) and results show that the embeddings trained on crisis datasets outperform the pre-trained embeddings. Palmer et al. (2016) use Naive Bayes (Schneider, 2003), SVM and Maximum Entropy (Berger et al., 1996) to find that the SVM shields the best results. Morstatter et al. (2014) also use Naive Bayes to try to predict if the tweet comes from inside the affected area or not, while Truong et al. (2014) use it for detecting if tweets are informative or not with respect to a specific crisis. Zhang and Vucetic (2016) tackle the problem of the low amount of labelled data in the crisis moment and use a semi-supervised technique where they build clusters based on Brown and K-means applied to word embeddings, and then use a simple

better summarize the data or used as features in supervised learning settings (Alam et al., 2018, Ning et al., 2017, Zhang and Vucetic, 2016).

[9]See for example `https://crisisnlp.qcri.org/`

Logistic Regression to training sizes that go from 20 to 1000 observations. Finally, Li et al. (2018) propose a Naive Bayes classifier together with an iterative self-training strategy: First they start using the pre-trained model from out-of-event information and predict the new unlabelled data. Secondly, they use the observations with the most confident results as new labelled data and retrain the model. This iterative process allows them to shift the domain specificity of the original model.

Deep learning has also been recently employed. Alharbi and Lee (2019) work on Arabic tweets, and find its best results with LSTM and BiLSTM networks for on-event-data, and SVM, LSTM, Convolutional LSTM and BiLSTM for out-of-event data, depending on the crisis. Nguyen et al. (2016b) use a combined model of a CNN with a Multi Layer Perceptron (MLP-CNN) to add TF-IDF features directly into the dense layer. Caragea et al. (2016) explore SVM and CNNs using various n-gram combinations as input vector. Kersten et al. (2019) employ CNN to filter crisis-related tweets and explore model transferability across different types of crisis (flood vs. hurricane events). Alam et al. (2018) propose a neural framework that performs domain adaptation with adversarial training and graph-based semi-supervised learning leveraging both labelled and unlabelled data. Finally, to account for geographical information that is often missing, Hernandez-Suarez et al. (2019) use a toponym (place names within the tweet) extractor relying on BiLSTM with a CRF output layer.

In this paper, we explore both feature-based and deep learning models in three different configurations: First a binary relevance classification, then urgency (non useful vs. urgent vs. non urgent) and finally multiclass intention to act categories. In particular, we propose:

- *A new group of features* showing their effectiveness when combining with state of the art features.

- *Several transfer learning strategies.* Both large general-purpose corpora and crisis-related datasets have been used to generate word embeddings. The problem with the latter is that the amount of data necessary for building a well-behaved embedding is bigger than the size of a domain specific corpus. Using pre-trained embeddings misses the opportunity of extracting information from a domain specific corpus. In this paper, we propose a combination of both approaches. We define a tangent task for which we resort to labelled or unlabelled data,

and then train a deep learning model to predict this task. The training process of the network makes an adaptation of the original word embeddings to better predict the related task. Finally we re-capture the weights inside the internal layer of the network.

Transfer learning has been widely used in many NLP applications (see (Ruder, 2017) for an overview). For example, Ziser and Reichart (2018) use words that are frequent in both source and target domain as pivots and use them to predict surrounding non-pivots. Yang et al. (2017) first learn the source domain representations, and then use a regularized cost function on the target domain that penalize distance of the word vector to the source domain. Kameswara et al. (2018) builds both source and target embeddings, to project them onto the same space by using Canonical Correlation Analysis and then linearly combine them via an optimization formula. However, in the field of crisis management, transfer learning remains under-explored. Pedrood and Purohit (2018) transferred knowledge from past events by representing tweets using sparse coding which learns latent themes in the message from unlabeled data. Singh et al. (2020) employed language model fine-tuning to classify the flood-related feeds in any new location. Nguyen et al. (2016b) used domain adaptation techniques for training a particular event using information from other events. In this paper, we newly: (1) Introduce the concept of *domain shift* on embeddings to refer to a change in the weights of pre-trained, general domain embedding, towards the crisis situation domain using both French and English in-domain datasets, (2) Investigate multitask learning, and (3) Compare its performances with two other transfer learning strategies, namely language model fine-tuning and domain shift.

- *Use of metadata in deep learning approaches:* Although Nguyen et al. (2016b) used an architecture of MLP-CNN, the features concatenated to the dense layer are TF-IDF based features. We propose to use metadata from the tweet (*tweet likes* and *retweets*) and user (*tweets*, *following*, *followers*, *likes* and *lists*) in order to harness not only the textual information, but also the complementary data that a tweet carries.

3. Dataset, annotation methodology and results

3.1. Data Gathering

The first step was to find significant ecological crises that had big repercussions on social networks. To this end, we relied on *catnat.net*[10], a French website that inventories the natural disasters that occurred in the world since 2015. We then selected those crises that seemed most relevant to us according to criteria of impact and importance (i.e., crises the most discussed in the media or that caused the most damage). Although our dataset had as its initial goal to be built around natural disasters (flood, storms, etc.), we decided to build a sub-corpus on a sudden non-ecological crises in order to compare it with the rest of our data.

Our dataset is composed of both labelled and unlabelled tweets collected via the Twitter API[11] as well as the Osirim platform[12] that hosts a Twitter stream, representing 1% of global tweets. Data collection was performed in two steps:

1. First using generic keywords (such as *"flood", "storm"*, etc.) without targeting a specific crisis. It mainly covers minor flood crises that occurred in France between 2017-2018. This dataset, named OTHER, has been used as a starting point for developing the annotation scheme and better understanding F-CSCM's needs.
2. Then using dedicated keywords about crisis names and crisis types targeting tweets posted *before* (24h before), *during* (48h) and *after* the crisis (72h after). The collected tweets concern various crises that affected metropolitan France and the French overseas departments and territories from 2016 to 2019: two floods that occured in the AUDE and CORSICA regions, ten storms (BÉRYL, BERGUITTA, FIONN, ELEANOR, BRUNO, EGON, ULRIKA, SUSANNA, FAKIR and ANA), two hurricanes (IRMA and HARVEY) and two non ecological crises: MARSEILLE BUILDING COLLAPSE and NOTRE-DAME BURNS.

In both steps, we did not rely on geolocalization information as it is often missing. The characteristics of our datasets are presented in Table 1

[10] https://www.catnat.net/

[11] We used *twitterscraper* (https://github.com/taspinar/twitterscraper) which allows scraping tweets beyond the 7-day limit.

[12] https://osirim.irit.fr/site/

Crisis types	Crisis names	Before Crisis	During Crisis	After Crisis	Total	#Labelled Data
Minor flood crises (not defined)	Other (2018-06)				**2,500**	1 696 (13,22%)
Hurricanes	Irma 2017-09-06	7,355	36,009	50,013	**93,377**	1 440 (11.23%)
	Harvey 2017-09-17	654	1,764	2,004	**4,422**	720 (5.61%)
Floods	Aude 2018-10-15	1,139	19,601	7,372	**28,112**	1 770 (13.80%)
	Corse 2018-10-16	8,301	7,592	5,141	**21,034**	720 (5.61%)
Storms	Béryl 2018-07-08	894	1 432	2,195	**4,521**	720 (5.61%)
	Berguitta 2018-01-18	5,227	10,399	8,345	**23,971**	720 (5.61%)
	Fionn 2018-01-17	2,066	5,729	5,734	**13,529**	720 (5.61%)
	Eleanor 2018-01-03	3,569	18,027	6,179	**27,775**	720 (5.61%)
	Bruno 2017-12-27	2,071	5,299	8,105	**15,475**	720 (5.61%)
	Egon 2017-01-12	661	17,094	2,158	**19,913**	720 (5.61%)
	Ulrika 2016-02-13	2,589	5,852	6,539	**14,980**	720 (5.61%)
	Susanna 2016-02-09	5,707	6,502	4,025	**16,234**	720 (5.61%)
	Fakir 2018-04-24	2,122	3,797	8,574	**14,493**	–
	Ana 2017-12-11	1,977	6,970	3,827	**12,774**	–
Non-ecological	Marseille 2018-05-11	6,616	15,931	18,945	**41,492**	720 (5.61%)
	Notre Dame fire 2019-04-15	1,196	13,219	2,643	**17,058**	–
Total		**52,144**	**175,217**	**141,799**	**371,660**	**12 826 (100%)**

Table 1: Distribution of labelled and unlabelled tweets in our dataset.

(duplicated tweets have been removed). Our labelled dataset is composed of 12,826 tweets distributed in 14 sub-corpora: 13 specific crisis and one made with general related keywords (cf. OTHER). The unlabelled dataset contains 358,834 tweets and is composed of all the crises mentioned in Table 1 and also three other events, while not including the tweets that have been labelled: ANA and FAKIR storms and a sudden non-ecological event: NOTRE-DAME BURNS. This bigger but unlabelled dataset has been used for training embeddings with domain shift (see Section 4.2.2).

3.2. Annotation guidelines and annotation procedure

The annotation guidelines were designed to address F-CSCM's specifications regarding message urgency classification. To this end, a set of 746 tweets from the OTHER dataset (cf. Table 1) has been used to identify the categorizations that meet their needs while being realistic for constructing machine learning models. We arrived at the following annotation scheme divided hierarchically into three levels, as shown in Figure 1.

First, we have a binary division that determines whether a tweet is useful or not.

- The NOT USEFUL category includes messages that are not related to the targeted crisis. This can concern messages about a non ecological crisis

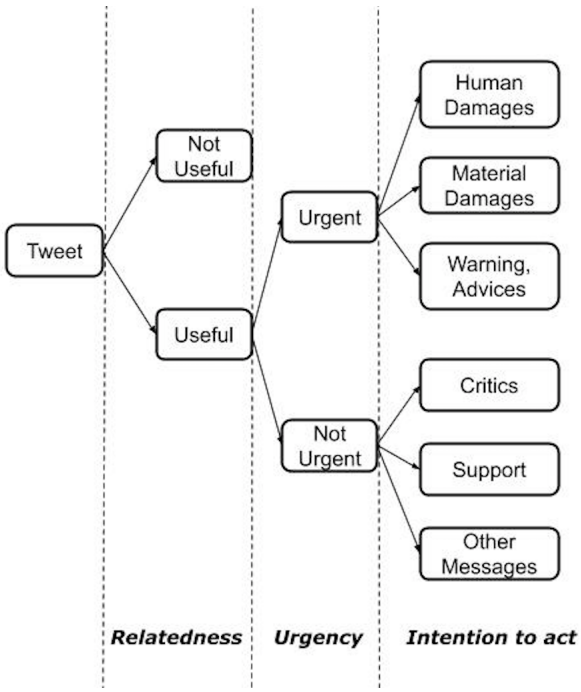

Figure 1: Categories used in our annotation schema

that have been gathered because the period when it occurred overlaps with the crisis we are studying (cf. (6)). Non useful messages can also refer to personal messages unrelated to the topic (cf. (7)), or information pertaining to events occurring outside the French territories (cf. (8)).

(6) Après l'odieux attentat dans l'#Aude, nos pensées vont aux proches des victimes avec une émotion particulière pour le militant CFDT victime de cet acte terroriste. Nous défendrons tjrs les valeurs de liberté, d'égalité et de fraternité contre la barbarie.
(After the shocking terrorist attack in #Aude, our thoughts go to the relatives of the victims with a particular emotion for the CFDT[13] militant victim of this terrorist act. We will always defend the values of freedom, equality and fraternity

[13]CFDT is a French labor union.

against barbarism.)

(7) Le vrai barrage, c'était @NicolasSarkozy. Ils l'ont détruit, il ne faut pas s'étonner de l'inondation...
(The real dam was @NicolasSarkozy. They destroyed it, do not be surprised by the flood ...)

(8) Frappée par l'ouragan Irma, la Floride déplore déjà des victimes - Le Figaro
(Stricken by Hurricane Irma, Florida is already regretting victims - Le Figaro)

- The USEFUL messages category includes those that provide actionable information or raise situational awareness over a crisis that has affected France. Useful messages are divided between urgent and non-urgent depending on the intention-to-act category (or actionability category) they convey.
  - URGENT messages include:
    * HUMAN DAMAGES bring together any message mentioning missing, injured or dead people during crisis events (cf. (9)). The category also concerns messages about displaced populations, evacuations (cf. (10)) or populations isolated or left behind.

      (9) Une automobiliste meurt noyée dans les Hautes-Pyrénées. #Inondation
      (A motorist dies drowned in the Hautes-Pyrénées. #Flood)

      (10) Face au risque d'inondation, 500 personnes ont été évacuées et 1 000 placées sous surveillance, dans les Pyrénées-Orientales...
      (Faced with the risk of flooding, 500 people were evacuated and 1,000 placed under surveillance in the Pyrénées-Orientales).

    * MATERIAL DAMAGES point to any damaged infrastructure that was caused by a crisis (cf. (11)).

      (11) Ouragan Irma a détruit près de 95% de l'île de Saint-Martin!

(Irma hurricane destroyed around 95% of Saint-Martin island!)

* WARNING-ADVICE gives security instructions, tips to limit the damage or weather reports, as in (12).

  (12) #inondation Ne tentez jamais de franchir une rivière en crue respectez signalisation mise en place par les autorités
  (#flood Never try to cross a river in flood respect authorities' signaling)

– NON URGENT category is articulated into:

* SUPPORT messages to the victims (cf. (13)), proposals or requests for donations.

  (13) Courage à tous mes amis héraultais !! #Montpellier #Inondation #Alerterouge
  (Courage to all my héraultais friends !! #Montpellier #Flood #RedAlert)

* CRITICS messages that denounce the lack of effectiveness of rescue services (cf. (14)) or government action.

  (14) Le service travaux ne répond pas, les pompiers ne viennent pas car c'est communal et la police sait pas quoi faire. inondation mons
  (The works service does not reply, firefighters do not come because it is communal and the police does not know what to do. flood mons)

* OTHER MESSAGES that do not have an immediate impact on actionability but contribute in raising situational awareness. This concerns three main cases: (i) messages about animals with a particular focus on flocks of animals caught in the crisis, (ii) messages that aim to provide additional information via external links to URLs, photos or videos (cf. (15)), and (iii) prevention messages that provide general-purpose safety instructions upstream of crisis (cf. (16)).

  (15) Voilà ce que ça donne rue St Pierre #inondation #caen instagram.com/p/**...**
  (Here's what's happening at St Pierre road #flood #caen instagram.com/p/**...**)

(16) Anticipez les intempéries avec AXA. Consultez nos démarches en cas d'inondation → http://go.axa.fr/0P1Mym.... #inondations #CrueSeine
(Anticipate bad weather with AXA. In case of flood, consult our procedures → http://go.axa.fr/0P1Mym... #flood #FloodSeine)

Our data has been manually annotated by two native French speakers, both undergraduate linguistic students using the Brat tool[14]. We performed a three-step annotation where an intermediate analysis of agreement and disagreement between annotators was carried out. Annotators were first trained on 211 tweets in collaboration with the F-CSCM that helped them to better understand the task and adjust the annotation guidelines. Annotators were then asked to separately annotate 1,503 tweets from the OTHER dataset so that inter-annotator agreements could be computed. We got a Cohen's Kappa of 0.722 for relatedness (binary), 0.658 for urgency (three classes) and 0.650 for multiclass (including the non useful category). The main cases of disagreements in all the configurations concern the non useful class.

After adjudication, the final step was the effective annotation. To ensure a similar distribution of tweets regarding the crisis type and the posting period (before, during and after the crisis) in the final labelled dataset, we randomly selected for each crisis 1/6 of tweets before, 3/6 during and 2/6 after, which corresponds to a total of 720 messages per crisis. During the annotation, we observed that IRMA and AUDE contain many more useful messages compared to the other crises. We therefore doubled the number of annotated messages for those crises, as seen in Table 1.

3.3. Quantitative analysis

The distribution over each of the categories of our scheme is presented in Table 2. It consists of 11.24% useful not urgent messages (1 442 tweets) and 16.74% urgent (2 147 labels), as well as 72.02% not useful messages (9 237 tweets).

Figure 2 shows the different proportions of each category for each event, highlighting the type of crises. We observe that the three crises with the highest proportion of informative messages are IRMA (45.14%), OTHERS (41.51%)

[14] http://brat.nlplab.org/

Urgent messages **2,147 (16.74%)**	Human damages	241 (1.88 %)
	Material damages	489 (3.81 %)
	Warning, advice	1 417 (11.05 %)
Not urgent messages **1,442 (11.24%)**	Critics	119 (0.93 %)
	Support	477 (3.72 %)
	Other messages	846 (6.60 %)
Not Useful		**9 237 (72.02 %)**
Total		**12 826**

Table 2: Distribution of labels in our dataset

and AUDE (39.5%). Two flood crises share a similar distribution over the classes with 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, while HARVEY had a bigger impact on other countries. The eight events in the storms subset contain fewer useful messages overall than other types of crises, fewer not urgent messages, and particularly few support messages compared to other crises (0.38% on average for the storm subset versus 5.61% for the rest of the corpus). The non-ecological crisis of MARSEILLE COLLAPSE has a significantly different distribution from the rest of the corpus. Overall, it contains few informative messages but they are evenly distributed. This 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, unlike in the case of natural disasters.

Our dataset is imbalanced with many not useful messages, which is in line with the proportions reported in existing datasets (Vieweg, 2012, Nguyen et al., 2016b, Alam et al., 2018). Nonetheless, this was one of the goals while building the corpus. The demands from the F-CSCM were to prioritize recall over precision because the potential use of automatic tools for processing social media will always involve a human check of the useful messages with the goal of improving actionability. In this sense, it is preferable to err on the side of false positives as opposed to missing messages in order to get as much useful information as possible. For this reason, we used general keywords, and many per crisis, in order to get as many tweets as possible. Moreover the

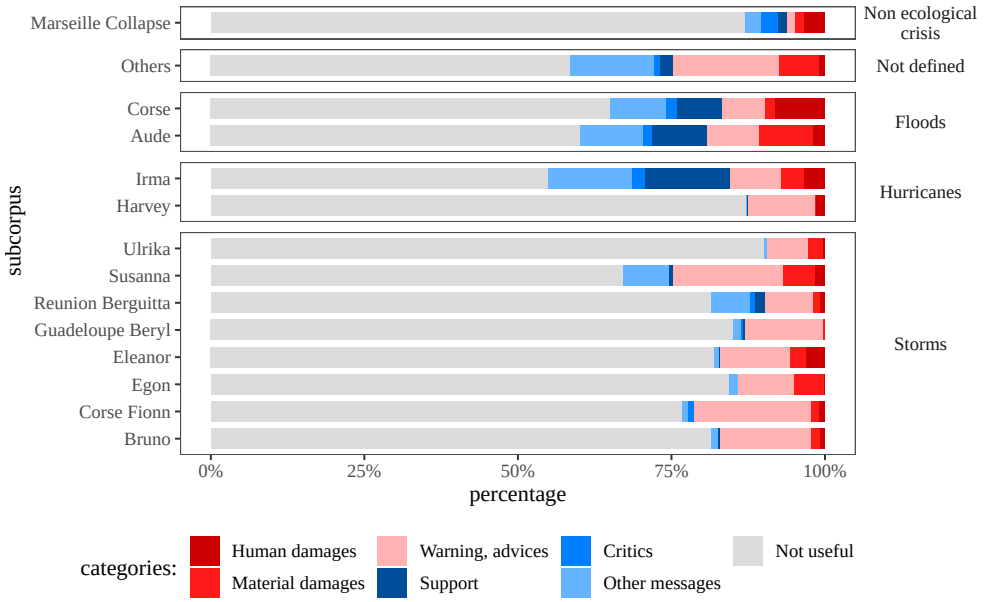

Figure 2: Proportions between classes in respect to the different types of crises. Not useful messages are in light gray, informative not urgent in blue and informative urgent in red.

interest of the F-CSCM resides only on crises occurring in France. For this reason, we have a combination of big crises that occur also outside France, but where messages that make a reference to other regions are not useful, and at the same time small crises happening only in the French region.

Finally, while annotating we found that messages that were scraped for some crises actually belonged to other. This happens because of the overlapping in time of some of the events. For the difficulties that this dataset implies, and the amount of labelled tweets per crisis, we made the following decisions:

- Make a pool of crises for train and test because of the mentioned complexities, and the size of the dataset,
- Design the out-of-event testing as follows: We used the crises ELEANOR and BRUNO for testing, as they did not show the mentioned overlap with other crises and hence there was no information leak from one event to another.

4. Models for automatic classification

We propose several models to automatically classify a tweet according to its relatedness (binary classification), urgency (three classes) and intention to act categories (multiclass). The models range from standard feature-based learning relying on state of the art features as well as new features (cf. Section 4.1), and several deep learning models, including the best performing state of the art models for crisis management. We experiment with both classic pre-trained embeddings (namely FastText (Bojanowski et al., 2017)) and contextual multilingual and French embeddings relying on BERT (Devlin et al., 2019) and FlauBERT (Le et al., 2019) language models (cf. Section 4.2.1). We also experiment with several transfer learning strategies to adapt these models to the crisis domain (cf. Section 4.2.2), as well as multi-input neural network exploring the use of metadata from the tweets (cf. Section 4.2.3).

Prior to learning, we perform standard pre-processing steps: putting all messages in lower case, removing stopwords, punctuation, URLs and user mentions.

4.1. Feature-based models

We mostly used features that have been shown to be efficient in the crisis-management literature, namely bag of words (BOW), linguistic, emotional and surface-based features. Table 3 provides a summary of all the features we used, new features are in bold font. To construct our surface-based features, we used non pre-processed messages. For the other features, we used the TreeTagger tool for French to lemmatize and extract the part-of-speech.

The average polarity of the tweet is determined by averaging the polarity of all the words that compose it relying both on the CASOAR French sentiment lexicon (Chardon, 2013, Benamara et al., 2014) and EMOTAIX (Piolat and Bannour, 2009), a publicly available French emotion and affect lexicon[15]. CASOAR encodes a total of 2,830 entries, among which 297 are expressions composed of at least 2 words[16]. The subjective senses of each entry

[15] https://centrepsycle-amu.fr/outils-recherche/

[16] "*sans doute*" (*without any doubt*) and "*haut de gamme*" (*upmarket*) are examples of such expressions.

is defined according to its polarity (positive, negative, neutral[17]), strength (on a discrete 3-point scale) and semantic category (among reporting verbs, judgement-evaluation, sentiment and advice following (Asher et al., 2008)). In addition to subjective words, the lexicon also contains intensifiers (mainly adverbs such as *too much*, *really*, etc.), negations and modalities (e.g., *certainly*, *maybe*, etc. see (Mari, 2015, Giannakidou and Mari, 2021)). EMOTAIX on the other hand contains 4,921 entries grouped into nine positive and negative emotional categories such as benevolence, surprise and hate.

The other new features are as follows:

- ***Imperative Verbs***. We also relied on a boolean feature for the presence of imperative verbs. In French, the imperative is the mode of injunctive and optative sentences. Therefore, it allows the expression of an order, a request, a restriction or an advice[18]. In the context of crisis related tweets, we have seen during manual annotation that this mode can be used to give advice, as shown by the following example that belongs to the WARNING AND ADVICE category:

  (17) #inondation #crue : savez-vous comment réagir ? **Découvrez** les bons comportements à adopter & les conseils des @PompiersFR #gestesquisauvent
  (#floods #rise: do you know how to react ? **Discover** good behavior to adopt & tips from @PompiersFR (french firefighters) #savinggestures)

- ***Intensifiers***. We observe that urgent messages that belong to the category WARNING-ADVICE tend to employ intensifiers. We therefore check for the presence of intensifiers relying on the CASOAR lexicon.

- ***Number*** measures the presence of numbers in tweets.

- ***Number of following and likes.*** These features give information about the centrality of a user within the social network. This indicates the effect and importance of a user with respect to others. Given that

[17]Neutral expressions are expressions that can convey both positive and negative sentiments like the word *surprise*.

[18]For a typology of imperative sentences in English, see (Condoravdi and Lauer, 2012); the same typology can be extended to French.

the moment of scraping user information differs from the moment of scraping the tweets, there exists some missing cases. For those, we imputed the values by the mean.

Group	Features	Reference
BOW	Bag of words for unigrams and bigrams	[1], [3], [4], [5], [6]
Linguistic & Emotion features	Proper Nouns (i.e., named entities)	[2], [3]
	Verb count	[2], [3]
	Intensifiers	**new**
	Imperative Verbs	**new***
	Average Polarity of the tweet	[6]
Surface-based features	Length of tweet	[8]
	URL	[4], [6]
	Image presence	[9]
	User mention	[7]
	Hashtag	[5], [6]
	Exclamation marks	[10]
	Numbers	**new****
Tweet metadata features	Tweet Likes / Favorites	[6]
	Retweets	[5], [6]
User-based features	Followers count	[6]
	Following count	**new**
	Likes count	**new**
	Tweets count	[6]
	Lists count	[6]

* : [2] used tenses but does not analyse imperative verbs.
** : In addition to phone numbers [6], we account for all other numbers (e.g., number of victims, French regions, etc.)
[1]: (Stowe et al., 2018) ; [2]: (Ning et al., 2017) ; [3]: (Morstatter et al., 2014)
[4]: (Li et al., 2018) ; [5]: (Truong et al., 2014) ; [6]: (Neppalli et al., 2018) ; [7] (Aipe et al., 2018)
[8]: (Imran et al., 2013) ; [9]: (Kaufhold et al., 2020) ; [10]: (Zahra et al., 2020) ; [11]: (Alam et al., 2018)

Table 3: Features used in our experiments. New features are in bold font.

For our experiments, we trained several classifiers: a Naive Bayes, Random Forest, Support Vector Machine (SVM), and Gradient Boosting Machine(GBM)(Chen and Guestrin, 2016). Given the first results, we concluded that the best options are SVM and GBM. Therefore, the remaining experiments were made specifically with these two models. SVM was trained on a linear kernel, while for the GBM, we did a grid search over the parameters using cross validation. We used the Scikit-learn (Pedregosa et al., 2011) implementations of those algorithms. Our models are as follows:

- $\mathbf{BOW_{SVM}}$ and $\mathbf{BOW_{GBM}}$ are both baseline models with a BOW input. We used unigrams and bigrams after removing words less than

3 characters, words appearing in less than 5 tweets of the corpus, and finally, most frequent words in the corpus (in the top 30%).

- **Feat-SOA$_{\text{SVM}}$** and **Feat-SOA$_{\text{GBM}}$** are both a compilation of features from the literature (i.e., those that are not marked *new* in Table 3) that are used respectively with a SVM and a GBM classifier. These two models are also baselines.

- **Feat-(SOA+New)$_{\text{SVM}}$** and **Feat-(SOA+New)$_{\text{GBM}}$** are both SVM and GBM classifiers with all features in Table 3, including the new ones.

4.2. Deep Learning models

4.2.1. Basic configurations

- **CNN$_{\text{basic}}$**. We replicated the methodologies used in (Nguyen et al., 2016b) and (Caragea et al., 2016) and tested other conventional configurations and hyperparameters while changing the architectures and inputs of the models. We tried Fully Connected Neural Network, Long Short Term Memory Network, and Convolutional Neural Network (CNN). We found that CNN was the best architecture using fine-tuned FastText (Bojanowski et al., 2017) pre-trained French embeddings with 300 dimensions. After a grid search over some of the main hyper-parameters of the network, optimum results were obtained with 64 filters, a filter length of four, a pool length of $2,236$ nodes on the dense layer, a dropout of 0.5, using AdaDelta as optimizer. This architecture has been designed with Keras (Chollet and Others, 2015).

- **BERT$_{\text{base}}$**: It relies on the pre-trained BERT multilingual cased model (Devlin et al., 2019). We used the HuggingFace's PyTorch implementation of BERT (Wolf et al., 2019) that we trained for four epochs using a gradient clipping of 1.0.

- **FlauBert$_{\text{base}}$**: It uses the FlauBERT base cased model (Le et al., 2019), the pre-trained French contextual embeddings[19]. 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

[19]We also experimented with CamemBERT (Martin et al., 2019) but the results were lower.

linear decreasing learning rate during optimisation. To avoid exploding gradients, we use a gradient clipping of 1.0.

4.2.2. Transfer learning strategies

The basic idea behind transfer learning is to improve a given target prediction task (e.g. classification) by adapting a network trained on a source task (which can be from the same domain as the target or different) to the target domain task (Pan and Yang, 2010). In this paper, we explore three transfer learning strategies:

(a) Language model fine-tuning. Following the ULMFiT approach (Howard and Ruder, 2018), we fine-tune $\mathbf{BERT_{base}}$ and $\mathbf{FlauBERT_{base}}$ language models (hereafter LM) initially trained on a general domain, for the crisis domain using the set of French unlabeled dataset of 358, 834 tweets. This resulted in two new fine-tuned models: $\mathbf{BERT_{tunedLM}}$ and $\mathbf{FlauBERT_{tunedLM}}$. For both models, adaptation consists of training LM with a masked language model head then use the shifted weights to perform the classification. This process is similar to BERT training (Devlin et al., 2019).

(b) Domain shift. It considers the existence of two corpora. One of them, the target domain, being specifically related with the tasks for which the embeddings will be used, and another, the source domain, that is not related. Using a non-related corpus can be useful when the source domain is a bigger corpus where the representation of off-topic words can be still meaningful. This is exactly the case of crises related tweets, given that the size of the non-labelled corpus of this specific domain is not big enough for building consistent representation of all the vocabulary. We propose here two domain shift models:

- $\mathbf{CNN_{shift}}$. Building directly our own word embeddings based on our non-labelled corpus of 358, 834 tweets result in worst results than using the general domain corpus, given the relatively small size of our corpus. Nonetheless a domain shift of the embedding that uses the unlabelled data was still possible. To this end, we used the information of the period of collection of tweets (before, during and after the crisis) to train a simple CNN model to predict the period (that is a three-class classification). With the fitted model, we capture the fine-tuned weights of the first layer of this model, the embeddings. Given that this matrix has only the size of the vocabulary present in the unlabelled corpus,

we build the final embedding matrix for the labelled dataset as a combination of both word embeddings: Whenever a word was present in the fine-tuned embeddings, we use that embedding, otherwise we keep using the original pre-trained FastText French embeddings. Figure 3 represents this workflow.

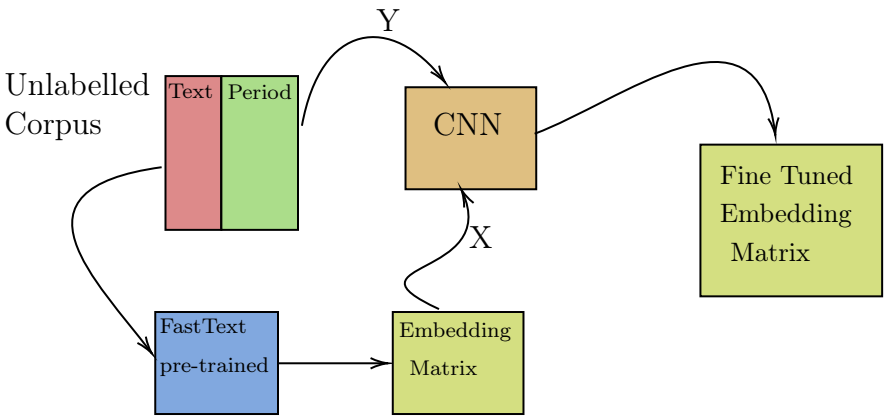

Figure 3: Domain shift on embeddings, here with a CNN architecture.

- **BERT$_{\text{shiftCrisisNLP}}$**: This model aims at fine-tuning **BERT$_{\text{base}}$** on an auxiliary in-domain supervised classification task. We then use this new model to train a second classifier on our dataset, the two tasks being optimized separately. The auxiliary classifier is trained on a subset of the CrisisNLP dataset[20] containing 19,250 manually annotated English tweets about ecological crises[21] which allows to adapt the pre-trained **BERT$_{\text{base}}$** multilingual model to the crisis domain in a cross-lingual fashion. To reduce the bias regarding the crisis location during training (we recall that we are targeting crises that occur in France only), we use Spacy's named entity recognizer and replace all locations (country names, regions, etc.) with the generic tag <Location>. We then align the CrisisNLP categories with our low level categories. However, since there is not a one-to-one mapping between the two annotation schemes,

[20] https://crisisnlp.qcri.org/lrec2016/lrec2016.html

[21] The selected crises are the following: Nepal Earthquake, Chile Earthquake, California Earthquake, Pakistan Earthquake, Cyclone PAM, Typhoon Hagupit, Hurricane Odile, Iceland Volcano, Pakistan Floods, and India Floods.

we group similar categories together when possible[22]. To keep the multilingual property of the model, we freeze the BERT embeddings layer for training to only shift the hidden layers of the model with a learning rate of $2e-5$ and for two epochs using the AdamW optimizer and gradient clipping of 1.0. We afterwards use the learned parameters for training the model on our annotated French corpus for three epochs and with a learning rate of $2e-8$ and AdamW optimizer.

(c) Multitask learning. Following Liu et al. (2019), we fine-tune $\textbf{FlauBert}_{\textbf{base}}$ in a multitask learning framework. The aim is to the share knowledge among the three classifiers (that is binary, three-class and multiclass) when trained jointly by multi-task objectives. In this model, named $\textbf{FlauBert}_{\textbf{multitask}}$, each classifier is then assumed to be a specific task and all tasks share and update the same low layers (that are $\textbf{FlauBert}_{\textbf{base}}$ layers) except the final task-specific classification layer. This is known as hard parameter sharing which has been shown to greatly reduce the risk of overfitting (Baxter, 1997). The learning process works as follows: We retrieve the first's token last hidden state of the shared $\textbf{FlauBert}_{\textbf{base}}$ model followed by a droupout Layer of 0.1 which is then connected to three different layers, one for each classification task. The loss is calculated by summing up the CrossEntropy loss of each task and backprobagating it as a sum through all the model. This allows to share the perception of each task to the problem with the other tasks. We trained the model with a learning rate of $2e-5$ with linear decay and adopting the AdamW optimization algorithm. We increased the size of the clipping gradient to 3.0 because the gradient comes from three different tasks at the same time.

4.2.3. Multi-input models

Following Nguyen et al. (2016b), we also experiment with a multi-input neural network that uses both the text from the tweets as well as other features. Extra-features were added on top of transfer learning models as they achieved better compared to basic models (cf. Section 5). We experiment multi-input with CNN and transformers:

[22]For example: "missing or found people", "injured or dead people" and "displayed people" are aligned with HUMAN DAMAGES while "donation needs, offers or volunteering services", "sympathy and emotional support" with SUPPORT.

- **CNN$_{\text{shift}}$+TF-IDF**. This model is inspired from (Nguyen et al., 2016b) where the additional features are TF-IDF. It is used here as a very strong baseline.

- **CNN$_{\text{shift}}$+Feat$_{\text{meta}}$**. Instead of using the TF-IDF of the vocabulary as additional features, we decided to use only metadata regarding the tweet and user: number of likes and retweets of the tweet, and number of likes, followers, following and lists of the user. This information is not present in any way within the text content, and therefore has the potential of adding new relevant features to the model.

  The multi-input architecture represents a more complicated structure than the traditional deep learning models. The input shape of the text represented as word embeddings is N*D, where N is the number of words and D is the number of dimensions in the embedding. The input shape of the other features is a vector with as many positions as features, in our case six. To concatenate both inputs, the text of shape N*D is first fed as an embedding layer. We then perform the convolutional layers and the max pooling layers. Finally a global max pooling layer flattens the output, which is fed both to an auxiliary fully connected layer, a special dense layer made for optimization purposes, and also to a new vector in which we also concatenate the new features. This vector is then used for a pile of fully connected layers, that produces the main output. Figure 4 shows the structure of this network.

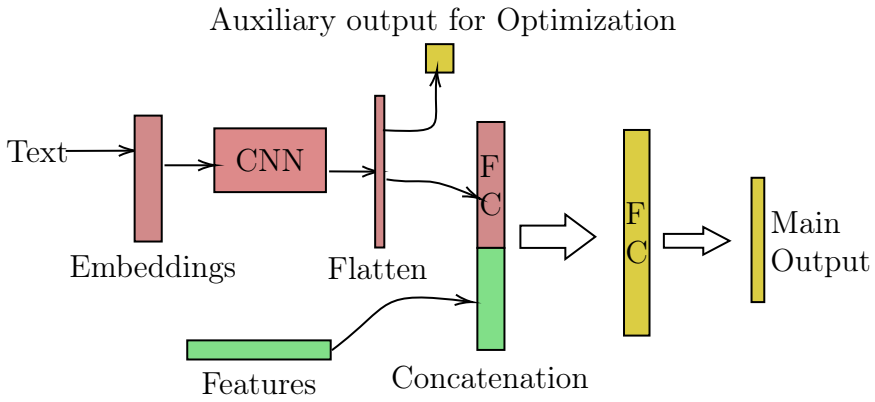

Figure 4: Concatenation of inputs. FC stands for "fully connected".

- **FlauBert$_{\text{base}}$+Feat$_{\text{meta}}$** and **FlauBert$_{\text{tunedLM}}$+Feat$_{\text{meta}}$**. The same idea as above is applied but relying respectively on **FlauBert$_{\text{base}}$** and **FlauBert$_{\text{tunedLM}}$** instead. Our aim is to test whether user metadata can improve over a transformer architecture.

- **FlauBert$_{\text{tunedLM+multitask}}$ +Feat$_{\text{meta}}$**. This final model combines the fine-tuned FlauBert language model trained in a multitask framework with additional metadata features.

5. Results

We experimented with binary classification (useful vs. non useful), three-class (non useful vs. urgent vs. non urgent) and multiclass classifications (i.e., intention to act categories) using both random sampling and out-of-event configurations. In the former, we mixed the tweets for all crises and then randomly select 80% for training and 20% for testing. In the latter, we select those crises that do not show any temporal overlap (as many crises occurred in the same time period) (cf. Section 3.3). Therefore, the train set is composed of 13 crises related to storms, floods and one crisis about building collapse while the test set contains tweets relative to Eleanor and Bruno storms. The distributions of tweets in the train/test sets in both configurations are shown in Table 4 and 5.

Train (10,260)	Urgent messages (1,677)	Human/Material damages	568
		Warning-advice	1,109
	Not Urgent messages (1,192)	Critics	101
		Support	398
		Other messages	693
	Not Useful		7,391
Test (2,566)	Urgent messages (470)	Human/Material damages	162
		Warning-advice	308
	Not Urgent messages (250)	Critics	18
		Support	79
		Other messages	153
	Not Useful		1,846

Table 4: Distribution of labels in the Random sampling configuration (test on 20%).

5.1. Random sampling results

Table 6 shows the results on random sampling configuration.

Train (11,386)	Urgent messages (1,901)	Human/Material damages	673
		Warning-advice	1,228
	Not Urgent messages (1,424)	Critics	119
		Support	474
		Other messages	831
	Not Useful		8 061
Test (1 440)	Urgent messages (246)	Human/Material damages	57
		Warning-advice	189
	Not Urgent messages (18)	Critics	0
		Support	3
		Other messages	15
	Not Useful		1 176

Table 5: Distribution of labels in the out-of-event configuration (test on 2 events: Bruno and Eleanor).

Model	**Binary**	**Three-class**	**Multiclass**
BOW_{SVM}†	0.811	0.729	0.531
BOW_{GBM}†	0.799	0.723	0.521
$\text{Feat-SOA}_{\text{SVM}}$†	0.808	0.721	0.553
$\text{Feat-SOA}_{\text{GBM}}$†	0.818	0.733	0.524
$\text{Feat-(SOA+New)}_{\text{SVM}}$	0.811	0.726	0.572
$\text{Feat-(SOA+New)}_{\text{GBM}}$	0.810	0.711	0.514
$\text{CNN}_{\text{basic}}$†	0.812	0.709	0.489
$\text{BERT}_{\text{base}}\star$	0.824	0.742	0.586
$\text{FlauBert}_{\text{base}}\star$	0.841	0.765	0.617
$\text{CNN}_{\text{shift}}$	0.806	0.701	0.469
$\text{BERT}_{\text{tunedLM}}$	0.846	0.757	0.618
$\text{BERT}_{\text{shiftCrisisNLP}}$	0.822	0.742	0.591
$\text{FlauBert}_{\text{tunedLM}}$	0.853	0.767	**0.654**
$\text{FlauBert}_{\text{multitask}}$	0.847	0.769	0.625
$\text{CNN}_{\text{shift}}+\text{Feat}_{\text{TF-IDF}}$†	0.796	0.705	0.399
$\text{CNN}_{\text{shift}}+\text{Feat}_{\text{meta}}$	0.816	0.702	0.388
$\text{FlauBert}_{\text{base}}+\text{Feat}_{\text{meta}}$	0.834	0.755	0.613
$\text{FlauBert}_{\text{tunedLM}}+\text{Feat}_{\text{meta}}$	**0.854**	**0.778**	0.627
$\text{FlauBert}_{\text{tunedLM+multitask}}+\text{Feat}_{\text{meta}}$	0.854	0.775	0.640

Table 6: Macro F1-score results in the random sampling configuration. † are state of the art baselines while $\star$ are transformers with basic architectures used here as strong baselines.

Concerning feature-based models, we observe that GBM with the state of the art features is the best model for the binary and three-class configurations of the problem, while the SVM with all features is the best model for multiclass. When analysing the best performing features using the information gain metric, we observe that those related to the user and tweet metadata are the most relevant. Indeed, as shown in Figure 5, the user metadata features are the most informative features, especially the total number of tweets. The new features we proposed are also very informative, namely the number of user likes and number of followings. Other important features include tweet length, the number of retweets and the average polarity. The presence of intensifiers, images and user mentions, on the other hand, are not correlated with the classes, and therefore might not be useful features.

When analysing the differences between feature-based and CNN-based deep learning models, we observe that they have a particularly low performance for all the configurations, the best model being $\text{CNN}_{\text{shift}}$+$\text{Feat}_{\text{meta}}$. When comparing with transformers, we see that all BERT and FlauBert models achieve better results. This is particularly salient with $\text{FlauBert}_{\text{tunedLM}}$ and $\text{FlauBert}_{\text{multitask}}$ which shows that fine-tuning the embeddings to the crisis domain is very productive. Multi-input architectures degrade the results when compared to deep learning baselines, namely CNN_{base}, $\text{FlauBert}_{\text{base}}$ and $\text{BERT}_{\text{base}}$. However, adding meta features on top of $\text{FlauBert}_{\text{tunedLM}}$ and $\text{FlauBert}_{\text{tunedLM+multitask}}$ was very productive, outperforming all the strong baselines.

Finally, we observe that the differences between models increase with the complexity of the problem, and while in binary classification there is a difference up to 5% between the best and the worst, in multiclass there is almost 25% of difference.

5.2. Out-of-event results

In out-of-event testing (cf. Table 7), the deep learning models outperform the feature-based models. The basic architecture (cf. $\text{CNN}_{\text{basic}}$) has the best results for the three-class, while $\text{CNN}_{\text{shift}}$) shows the best performance on the multiclass and binary problems, and $\text{CNN}_{\text{basic}}$+$\text{Feat}_{\text{meta}}$ does so in the binary problem. The multi-input model $\text{CNN}_{\text{basic}}$+$\text{Feat}_{\text{TF-IDF}}$ has a low performance, even when compared with the traditional feature-based models. Again we can see that the variability within the models increases with the number of classes. As already observed in random sampling results, we also see here that the best models are transformers and that

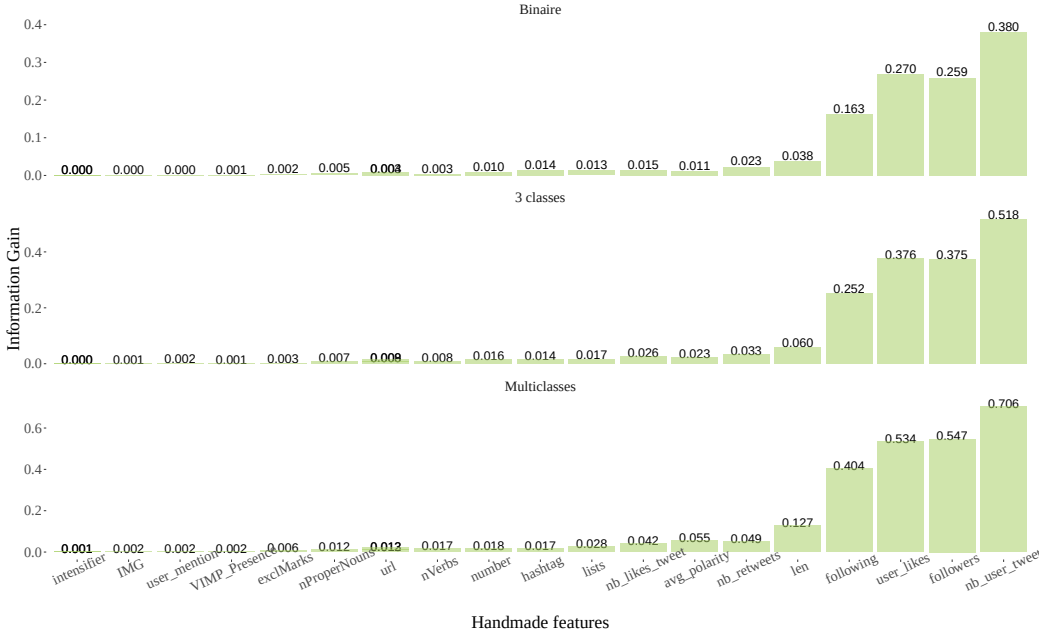

Figure 5: Information gain results for handmade features.

transfer learning models outperform the baselines for all the configurations. FlauBERT$_{\text{multitask}}$ gets the best results for binary and multiclass configurations while BERT$_{\text{shiftCrisisNLP}}$ for the three-class. The domain shift strategy based on cross-lingual transfer learning seems then to be a promising issue for future work which opens the door to crisis management for languages that lack manually annotated resources. Also, meta-features improve over fine-tuned FlauBert for all the configurations but degrades the result when added on top of the same model trained in a multitask framework.

Table 8 shows the percentage difference in the F1-score of each deep learning model when comparing the out-of-event testing and the random sampling. There is a strong correlation between tables 7 and 8, where the models that have worst performance in out-of-event testing are those that suffer the biggest loss of performance with respect to random sampling testing. This means that those models that are taking into consideration the

[22]The same tendency has been observed for BOW and feature-based models.

Model	**Binary**	**Three-class**	**Multiclass**
BOW_{SVM}†	0.813	0.568	0.420
BOW_{GBM}†	0.818	0.555	0.399
$\text{Feat-SOA}_{\text{SVM}}$†	0.818	0.592	0.477
$\text{Feat-SOA}_{\text{GBM}}$†	0.810	0.548	0.406
$\text{Feat-(SOA+New)}_{\text{SVM}}$	0.818	0.594	0.467
$\text{Feat-(SOA+New)}_{\text{GBM}}$	0.797	0.546	0.428
$\text{CNN}_{\text{basic}}$†	0.819	0.630	0.477
$\text{BERT}_{\text{base}}\star$	0.817	0.654	0.549
$\text{FlauBert}_{\text{base}}\star$	0.829	0.653	0.600
$\text{CNN}_{\text{shift}}$	0.823	0.574	0.501
$\text{BERT}_{\text{tunedLM}}$	0.825	0.647	0.572
$\text{BERT}_{\text{shiftCrisisNLP}}$	0.817	**0.665**	0.577
$\text{FlauBert}_{\text{tunedLM}}$	0.803	0.594	0.570
$\text{FlauBert}_{\text{multitask}}$	**0.832**	0.654	**0.631**
$\text{CNN}_{\text{shift}}+\text{Feat}_{\text{TF-IDF}}$†	0.775	0.535	0.383
$\text{CNN}_{\text{shift}}+\text{Feat}_{\text{meta}}$	0.823	0.608	0.468
$\text{FlauBert}_{\text{base}}+\text{Feat}_{\text{meta}}$	0.829	0.618	0.591
$\text{FlauBert}_{\text{tunedLM}}+\text{Feat}_{\text{meta}}$	**0.832**	0.595	0.567
$\text{FlauBert}_{\text{tunedLM+multitask}}+\text{Feat}_{\text{meta}}$	0.810	0.627	0.584

Table 7: Results of out-of-event testing in terms of macro F1-score. † are state of the art baselines while $\star$ are transformers with basic architectures used here as strong baselines.

Model	Binary	Three-class	Multiclass
$\text{CNN}_{\text{basic}}$†	**0.61**	-13.46	-3.44
$\text{BERT}_{\text{base}}$⋆	-0.87	-11.84	-6.16
$\text{FlauBert}_{\text{base}}$⋆	-1.41	-14.53	-2.76
$\text{CNN}_{\text{shift}}$	**0.61**	-20.50	-6.36
$\text{BERT}_{\text{tunedLM}}$	-2.47	-14.55	-7.40
$\text{BERT}_{\text{shiftCrisisNLP}}$	0.57	**-10.33**	-2.37
$\text{FlauBert}_{\text{tunedLM}}$	-5.90	-22.61	-12.86
$\text{FlauBert}_{\text{multitask}}$	-1.78	-14.90	**0.90**
$\text{CNN}_{\text{shift}}$+$\text{Feat}_{\text{TF-IDF}}$†	-2.6	-24.2	-4.2
$\text{CNN}_{\text{shift}}$+$\text{Feat}_{\text{meta}}$	0.37	-17.05	-5.65
$\text{FlauBert}_{\text{base}}$+$\text{Feat}_{\text{meta}}$	-1.47	-18.12	-3.57
$\text{FlauBert}_{\text{tunedLM}}$+$\text{Feat}_{\text{meta}}$	-2.54	-23.64	-9.46
$\text{FlauBert}_{\text{tunedLM+multitask}}$+$\text{Feat}_{\text{meta}}$	-5.14	-19.09	-8.73

Table 8: Loss of generalization power of deep learning models between out of event and random sampling in terms of F1-score. All scores are percentages. † are state of the art baselines while ⋆ are transformers with basic architectures used here as strong baselines.

event-specific information are those that have the worst performance when used for different events.

We provide in Table 9 the precision and recall of the different models. Recall is a particularly important measure given that the priority of the F-CSCM is to retrieve all the useful information, and the output of the models will be ultimately read by humans to make the final decision. Given this, although we do not have an exact weight of how much more important recall is with respect to precision, we need to evaluate our models paying special attention to their recall. Results show that transfer learning models are the best in terms of recall and that the multitask strategy gets 0.750 in the multiclass outperforming all the models by at least 10%. In this table we can also see that most models reached a higher precision than recall and the differences are small in relation to the class imbalance, showing a good compromise of both measures. Overall, the transformer-based transfer learning models seem to be the best option for crisis management in social media as their recall outperform precision for all the configurations.

Finally, when analysing the results of the best performing model (i.e,

Model	Binary		Three-class		Multiclass	
	Prec.	Recall	Prec.	Recall	Prec.	Recall
BOW_{SVM}†	0.865	0.781	0.605	0.543	0.502	0.381
BOW_{GBM}†	**0.888**	0.778	0.588	0.534	0.563	0.354
$\text{Feat-SOA}_{\text{SVM}}$†	0.882	0.623	0.632	0.566	0.558	0.440
$\text{Feat-SOA}_{\text{GBM}}$†	0.868	0.774	0.592	0.522	0.496	0.364
$\text{Feat-(SOA+new)}_{\text{SVM}}$	0.882	0.780	0.636	0.566	0.540	0.434
$\text{Feat-(SOA+new)}_{\text{GBM}}$	0.848	0.765	0.593	0.521	0.522	0.380
$\text{CNN}_{\text{basic}}$†	0.851	0.796	**0.655**	0.611	0.621	0.486
$\text{BERT}_{\text{base}}$⋆	0.805	0.831	0.641	0.671	0.539	0.575
$\text{FlauBert}_{\text{base}}$⋆	0.818	0.842	0.634	0.677	0.560	0.665
$\text{CNN}_{\text{shift}}$	0.851	0.802	0.643	0.540	0.580	0.495
$\text{BERT}_{\text{tunedLM}}$	0.802	0.859	0.614	0.698	0.522	0.651
$\text{BERT}_{\text{shiftCrisisNLP}}$	0.790	0.863	0.640	0.700	0.534	0.646
$\text{FlauBert}_{\text{tunedLM}}$	0.774	0.861	0.568	**0.767**	0.522	0.646
$\text{FlauBert}_{\text{multitask}}$	0.834	0.830	0.619	0.709	0.566	**0.750**
$\text{CNN}_{\text{shift}}\text{+Feat}_{\text{TF-IDF}}$†	0.878	0.730	0.589	0.507	**0.662**	0.361
$\text{CNN}_{\text{shift}}\text{+Feat}_{\text{meta}}$	0.858	0.798	0.624	0.596	0.592	0.463
$\text{FlauBert}_{\text{base}}\text{+Feat}_{\text{meta}}$	0.800	0.853	0.615	0.627	0.546	0.663
$\text{FlauBert}_{\text{tunedLM}}\text{+Feat}_{\text{meta}}$	0.806	**0.873**	0.563	0.670	0.512	0.664
$\text{FlauBert}_{\text{tunedLM+multitask}}\text{+Feat}_{\text{meta}}$	0.780	0.867	0.589	0.711	0.520	0.705

Table 9: Results of the out-of-event testing in terms of precision and recall[24]. † are state of the art baselines while ⋆ are transformers with basic architectures used here as strong baselines.

$\text{FlauBert}_{\text{multitask}}$) per class in the out-of-event configuration, we observe that the results for relatedness classification are quite effective with an F-score of 0.935 for NON USEFUL and 0.732 for USEFUL messages. Concerning urgency classification (i.e., the three-class configuration), the results for the NON USEFUL class remain stable with an F-score of 0.925 while URGENT messages obtain 0.740. The performance of the NON URGENT class is the lowest with an F-score of 0.300 probably because of the very limited number of instances of this class in the test set (18 –cf. Table 5). Finally, for intention to act classification the results are as follows: 0.925 for NON USEFUL, 0.731 for WARNING-ADVICE, 0.636 and 0.656 for respectively for MATERIAL DAMAGES and HUMAN DAMAGES, and 0.6 for SUPPORT.

6. Discussion

6.1. Error analysis

We manually analyse main causes of misclassification in both out-of-event and random sampling configurations. We observe that most errors come from the NON USEFUL class. In particular, we found during corpus annotation that there exist many messages that share a strong resemblance with the useful information, but given that they talk about a country different from France or a non ecological crisis (like terrorist attacks), we considered them as non useful[25]. This can be problematic to the models we trained because what we expected them to learn is not related to this geographical differentiation[26]. In this sense, it is probable that this increased the noise in the dataset and diminished the overall performance.

The following examples illustrate some of these cases. The tweet in (18) is personal and the writer complains about the flood in his apartment and that all his cables are dead. This message has been predicted as MATERIAL DAMAGES because of the presence of words denoting materials (like *cable*) which is not in accordance to the manual annotation that considers it as NON USEFUL since it targets a non ecological crisis. The prediction shows however that our model can easily adapt to new types of crises (we further experiment the out-of-type classification in Section 6.2).

(18) Inondation dans mon appartement super. Je crois que tous mes câbles sont mort. (Xbox, pvr, box etc...) journée de merde.
(Flood in my apartment great. I believe all of my cables are dead. (Xbox, pvr, box etc ...) shity day.)
Gold label: NOT USEFUL
Model label: MATERIAL DAMAGES

Examples (19) and (20) are typical errors of non useful messages annotated as such because the crisis is either out of France or because it occurred

[24]The precision and recall are presented rounded to three digits, while the F1 scores in tables 6, 7 and 8 where calculated before rounding.

[25]We also tried incorporating topic information as features (i.e., as given by LDA) but the results showed that the topics constructed from the corpus were strongly correlated with the particular events, and hence they were not useful for generalization.

[26]Many messages do not contain any information relative to locations (e.g., location name, countries, etc.) but hashtags.

in the past (cf. (19)). Manual annotation of (20) belongs to the NOT USEFUL category whereas the predicted label is HUMAN DAMAGES. Here, the model probably did not learn that the word *Floride (Florida)* implies that the tweet is not informative, which can be complex since the other words of the message taken separately indicate a warning (*hit by Irma hurricane, [location] regrets their victims* where the *location* only changes the label of the whole sentence). Automatic detection of spacial information from the textual content is left for future work.

(19) J'avais 5 ou 6 ans, j'ai faillit mourir noyé après une grande inondation dans notre école J'ai été sauvé in extremis par un lycéen
(I had 5 or 6, I almost died drowned after a great flood in our school I was saved in extremis by a high school student)
Gold label: NOT USEFUL
Model label: HUMAN DAMAGES

(20) Frappée par l'ouragan Irma, la Floride déplore déjà des victimes - Le Figaro.
(Hit by the Irma hurricane, Florida already regrets victims - Le Figaro (French journal))
Gold label : NOT USEFUL MESSAGE
Model label : HUMAN DAMAGES

Finally, we also observed that many misclassification errors come from the presence of irony (cf. (21) where the user criticizes the police) and metaphor, as in (22).

(21) Ne craignez rien, les Parisiens. La Police Nationale arrive pour vous sauver ... #orage #Inondation
(Parisians do not worry, the National Police arrives to save you ...#Storm #Flood)
Gold label: CRITICS
Model label: OTHER MESSAGES

(22) Il y a la tempête dehors, tu sors tu t'envoles !
(There is a storm outside, if you go out you'll fly away!)
Gold label: WARNING-ADVICE
Model label: NOT USEFUL

6.2. Towards out-of-type detection

We define out-of-type classification as the training on a pool of events related to different types of crises, like floods and hurricanes, and testing on a particular different type, like a building collapse. The biggest differences between these types are that while the hurricanes and floods are known with anticipation, the building collapse is a sudden event, and as we saw in Section 3 this difference generates really different imbalances of classes and semantic content (cf. Figure 2).

We therefore performed out-of-type classification by training on all the events but the MARSEILLE COLLAPSE, and testing on this latter. The results on this experiment were surprisingly good, with an F1-score of up to 0.830 for binary classification, 0.641 on three-class and 0.423 on the multiclass. This implies losses in generalization of up to -5.22%, -16.58% and -32.39% for the binary, three-class and multiclass classifications respectively, with respect to the random sampling configuration. This means that the multitask model is relatively stable for binary classification and to a lesser extent for three-class configuration. However, the model is not suitable to capture this type of differences between the training and the test for multiclass in spite of the fact that the MARSEILLE COLLAPSE corpus consists on 720 observations, with 87% of NON USEFUL messages. We also experimented with the random sampling and the out-of-event models removing the Marseille Collapse, and the results were worse than the ones shown, including the event. This means that even when the model is not suitable for out-of-type predictions, it is benefited by the presence of out-of-type data in the training set.

To conclude, out-of-type experiments open the door to a different way of collecting data during sudden event crises. Indeed, when scraping tweets about the Notre-Dame-de-Paris burning, we found that it is possible to easily retrieve the most useful messages from Twitter for sudden event crises, like fires and collapses, given that we can know the exact moment of the crisis. A different framework, but also interesting, could be to do a targeted labelling of the first tweets on a sudden crisis in order to build a model that detects them among the rest of the communications in this social media. The problem of this type of analysis has the particularity of being a really imbalanced dataset, given that from the total number of tweets, only a tiny proportion would be useful. In this sense, besides the tweets labelled by specific search using ex-post information, like time and place, the dataset should be augmented with random non-related tweets, in order to reproduce the original imbalance of classes. Another interesting perspective is to explore the class imbalance

problem in the context of deep learning (Buda et al., 2017, Johnson and Khoshgoftaar, 2019, Koziarski et al., 2019).

7. Conclusions and future lines of work

In this paper, we presented the first crisis-related dataset for French of about 13k tweets and a set of experiments to automatically classify a message according to three dimensions: relatedness, urgency and intention to act categories. This taxonomy has been proposed to better meet the French Civil Security and Crisis Management Department's specifications who perceives actionability in terms of emergency. The models range from feature-based models to deep learning including multi-input architectures and several transfer learning strategies to adapt the deep models to the crisis domain. Our results show that transformer architectures achieved the best results outperforming all the baselines, including state of the art models. In particular, the multitask learning framework trained over the French FlauBert model shows very good performances for random sampling, out-of-event and out-of-type testing. Our results constitute the first study for crisis management in French social media.

We explained that, given the necessity to prefer high recall over high precision, data acquisition process starts with an amount and type of keywords that generates an imbalanced dataset with a lot of not useful information. In order to face this problem, an exact measure of how much more important recall is as compared to precision needs to be made. In the future lines of work, we propose to put an exact cost to false positives and false negatives, and with that information build a weighted F-score. With this new measure of performance we are going to be able to do the resampling of the dataset in order to approach the imbalance of classes in an optimum way.

Acknowledgement

The work presented in this paper has been supported by the Chemi-INTACT project funded by French Ministère de l'Intérieur (Department of Home Affairs) that involved the Institut de Recherche en Informatique de Toulouse (IRIT) and Institut Jean Nicod (IJN). We thank the Colonel Olivier Morin, head of the French Civil Security and Crisis Management Department, for his help and feedback in defining the annotation guidelines, training annotators, and providing actionability specifications.

References

S. Vieweg, C. Castillo, M. Imran, 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.

A. Olteanu, S. Vieweg, C. Castillo, 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.

L. Palen, S. B. Liu, Citizen Communications in Crisis: Anticipating a Future of ICT-supported Public Participation, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI'07, pp. 727–736.

C. Castillo, Big crisis data: Social media in disasters and time-critical situations, Cambridge University Press, 2016.

M. Imran, C. Castillo, F. Diaz, S. Vieweg, Processing Social Media Messages in Mass Emergency: A Survey, ACM Computing Surveys 47 (2015) 67:1–67:38.

N. DePaula, S. Neely, T. Keller, L. Hagen, C. Robert-Cooperman, Crisis Communications in the Age of Social Media, Social Science Computer Review 36 (2017) 523–541.

M. Avvenuti, M. N. La Polla, S. Bellomo, S. Cresci, M. Tesconi, Hybrid Crowdsensing, in: Proceedings of the 26th International Conference on World Wide Web Companion, WWW'17, pp. 1413–1421.

J. Kim, M. Hastak, Social network analysis: Characteristics of online social networks after a disaster, International Journal of Information Management 38 (2018) 86–96.

J. Qadir, A. Ali, R. ur Rasool, A. Zwitter, A. Sathiaseelan, J. Crowcroft, Crisis analytics: big data-driven crisis response, Journal of International Humanitarian Action 1 (2016) 1–21.

F. Alam, F. Ofli, M. Imran, M. Aupetit, A Twitter Tale of Three Hurricanes: Harvey, Irma, and Maria, in: Proceedings of the International Conference on Information Systems for Crisis Response and Management, ISCRAM'18.

X. Ning, L. Yao, X. Wang, B. Benatallah, Calling for Response: Automatically Distinguishing Situation-Aware Tweets During Crises, in: Lecture Notes in Computer Science, LNCS, volume 10604, pp. 195–208.

S. Zhang, S. Vucetic, Semi-supervised Discovery of Informative Tweets During the Emerging Disasters, 2016.

H. Li, D. Caragea, C. Caragea, N. Herndon, Disaster Response Aided by Tweet Classification with a Domain Adaptation Approach, Journal of Contingencies and Crisis Management 26 (2018) 16–27.

K. Stowe, J. Anderson, M. Palmer, L. Palen, K. Anderson, Improving Classification of Twitter Behavior During Hurricane Events, the 6th International Workshop on Natural Language Processing for Social Media (2018) 67–75.

D. T. Nguyen, S. Joty, M. Imran, H. Sajjad, P. Mitra, Applications of Online Deep Learning for Crisis Response Using Social Media Information, in: 4th International Workshop on Social Web for Disaster Management (SWDM).

D. T. Nguyen, K. A. A. Mannai, S. Joty, H. Sajjad, M. Imran, P. Mitra, Rapid Classification of Crisis-Related Data on Social Networks using Convolutional Neural Networks, in: Eleventh International AAAI Conference on Web and Social Media, ICWSM'16.

C. Caragea, A. Silvescu, A. Tapia, 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.

V. K. Neppalli, C. Caragea, D. Caragea, 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.

A. Aipe, A. Ekbal, M. NS, S. Kurohashi, Linguistic feature assisted deep learning approach towards multi-label classification of crisis related tweets, in: Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management, ISCRAM'2018.

M. Imran, P. Mitra, C. Castillo, 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), European Language Resources Association (ELRA), 2016.

A. Cobo, D. Parra, J. Navón, Identifying Relevant Messages in a Twitter-based Citizen Channel for Natural Disaster Situations, in: Proceedings of the 24th International Conference on World Wide Web, WWW'15, pp. 1189–1194.

S. Cresci, M. Tesconi, A. Cimino, F. Dell'Orletta, A Linguistically-driven Approach to Cross-Event Damage Assessment of Natural Disasters from Social Media Messages, in: Proceedings of the 24th International Conference on World Wide Web, WWW'15, pp. 1195–1200.

S. Gründer-Fahrer, A. Schlaf, G. Wiedemann, G. Heyer, Topics and topical phases in German social media communication during a disaster, Natural Language Engineering 24 (2018) 221–264.

A. Alharbi, M. Lee, Crisis Detection from Arabic Tweets, in: Proceedings of the 3rd Workshop on Arabic Corpus Linguistics, pp. 72–79.

M. Imran, C. Castillo, J. Lucas, P. Meier, S. Vieweg, Aidr: Artificial intelligence for disaster response, in: Proceedings of the 23rd International Conference on World Wide Web, WWW'2014, p. 159–162.

F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, K. Tao, Twitcident: Fighting Fire with Information from Social Web Streams, in: Proceedings of the 21st International Conference on World Wide Web, WWW'2012 Companion, pp. 305–308.

H. Purohit, A. P. Sheth, Twitris v3: From Citizen Sensing to Analysis, Coordination and Action, in: proceedings of The International Conference on Weblogs and Social Media, ICWSM'2013.

S. Kumar, G. Barbier, M. A. Abbasi, H. Liu, TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief, in: Proceedings of the Fifth International Conference on Weblogs and Social Media, ICWSM'2011.

F. Alam, F. Ofli, M. Imran, CrisisMMD: Multimodal Twitter Datasets from Natural Disasters, in: Proceedings of the International AAAI Conference on Web and Social Media, ICWSM'2018.

G. E. Jensen, Key criteria for information quality in the use of online social media for emergency management in New Zealand., Master's thesis, 2012.

M. Habdank, N. Rodehutskors, R. Koch, Relevancy assessment of tweets using supervised learning techniques: Mining emergency related tweets for automated relevancy classification, in: Proceedings of the 4th International Conference on Information and Communication Technologies for Disaster Management, ICT-DM'2017, pp. 1–8.

M.-A. Kaufhold, M. Bayer, C. Reuter, 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 (2020) 102–132.

S. Verma, S. Vieweg, W. J. Corvey, L. Palen, J. H. Martin, M. Palmer, A. Schram, K. M. Anderson, Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency, in: Proceedings of the Fifth International Conference on Weblogs and Social Media, ICWSM'2011.

S. E. Vieweg, Situational Awareness in Mass Emergency: A Behavioral and Linguistic Analysis of Microblogged Communications, Phd dissertation, University of Colorado at Boulder, 2012.

M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, P. Meier, Extracting information nuggets from disaster-related messages in social media, in: Proceedings of the 10th Proceedings of the International Conference on Information Systems for Crisis Response and Management, ISCRAM'2013.

A. Olteanu, I. Weber, D. Gatica-Perez, Characterizing the Demographics Behind the #BlackLivesMatter Movement, CoRR abs/1512.0 (2015).

K. Zahra, M. Imran, F. O. Ostermann, Automatic identification of eyewitness messages on twitter during disasters, Information Processing & Management 57 (2020) 102–107.

H. Zade, K. Shah, V. Rangarajan, P. Kshirsagar, M. Imran, K. Starbird, From situational awareness to actionability: Towards improving the utility of social media data for crisis response, ACM Transactions on Computer-Human Interaction 2 (2018) 195:1–195:18.

M. Palmer, L. Palen, K. Anderson, M. J. Paul, K. Stowe, Identifying and Categorizing Disaster-Related Tweets, In Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media (2016) 1–6.

B. Truong, C. Caragea, A. Squicciarini, A. H. Tapia, Identifying valuable information from Twitter during natural disasters, in: Proceedings of the ASIST Annual Meeting, volume 51, pp. 1–4.

H. Li, X. Li, D. Caragea, C. Caragea, Comparison of Word Embeddings and Sentence Encodings as Generalized Representations for Crisis Tweet Classification Tasks, in: Proceedings of the ISCRAM Asian Pacific 2018 Conference, ISCRAM Asian Pacific 2018, pp. 1–13.

T. Mikolov, K. Chen, G. S. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, arXiv preprint cs.CL (2013a) 1–12.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed Representations of Words and Phrases and their Compositionality, in: C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 26, 2013b, pp. 3111–3119.

P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching Word Vectors with Subword Information, CoRR 1607 (2016).

J. Pennington, R. Socher, C. Manning, Glove: Global Vectors for Word Representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.

C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (1995) 273–297.

L. Breiman, Random Forests LEO, Machine Learning 45 (2001) 5–32.

K.-M. Schneider, A comparison of event models for Naive Bayes anti-spam e-mail filtering, in: AAAI-98 workshop on learning for text categorization, 1.

A. L. Berger, V. J. Della Pietra, S. A. Della Pietra, A Maximum Entropy Approach to Natural Language Processing, Computational Linguistics 22 (1996) 39–68.

F. Morstatter, N. Lubold, H. Pon-Barry, J. Pfeffer, H. Liu, Finding Eyewitness Tweets During Crises, in: ACL Workshop on Language Technology and Computational Social Science.

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

A. Hernandez-Suarez, G. Sanchez-Perez, K. Toscano-Medina, H. Perez-Meana, J. Portillo-Portillo, V. Luis, L. García Villalba, Using Twitter Data to Monitor Natural Disaster Social Dynamics: A Recurrent Neural Network Approach with Word Embeddings and Kernel Density Estimation, Sensors 19 (2019) 1746.

S. Ruder, An overview of multi-task learning in deep neural networks, CoRR abs/1706.05098 (2017).

Y. Ziser, R. Reichart, Pivot based language modeling for improved neural domain adaptation, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1241–1251.

W. Yang, W. Lu, V. Zheng, A simple regularization-based algorithm for learning cross-domain word embeddings, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2898–2904.

P. Kameswara, Sarma, Y. Liang, B. Sethares, Domain adapted word embeddings for improved sentiment classification, in: Proceedings of the

Workshop on Deep Learning Approaches for Low-Resource NLP, pp. 51–59.

B. Pedrood, H. Purohit, Mining help intent on twitter during disasters via transfer learning with sparse coding, in: R. Thomson, C. Dancy, A. Hyder, H. Bisgin (Eds.), Social, Cultural, and Behavioral Modeling, pp. 141–153.

N. Singh, N. Roy, A. Gangopadhyay, Localized flood detectionwith minimal labeled social media data using transfer learning, CoRR abs/2003.04973 (2020).

P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.

J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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 and Short Papers), pp. 4171–4186.

H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, D. Schwab, Flaubert: Unsupervised language model pre-training for french, arXiv preprint arXiv:1912.05372 (2019).

B. Chardon, Chaîne de traitement pour une approche discursive de l'analyse d'opinion, Phd thesis, Université Paul Sabatier, Toulouse, 2013.

F. Benamara, V. Moriceau, Y. Y. Mathieu, Catégorisation sémantique fine des expressions d'opinion pour la détection de consensus, in: TALN-RECITAL 2014 Workshop DEFT 2014: DEfi Fouille de Textes (DEFT 2014 Workshop: Text Mining Challenge), pp. 36–44.

A. Piolat, R. Bannour, An example of text analysis software EMOTAIX-Trope use: The influence of anxiety on expressive writing , Current psychology letters 25 (2009).

N. Asher, F. Benamara, Y. Y. Mathieu, Distilling Opinion in Discourse: {A} Preliminary Study, in: In proceddings of the 22nd International Conference on Computational Linguistics, COLING'2008, pp. 7–10.

A. Mari, Modalitś et Temps. Des modl̀es aux données., Peter Lang, 2015.

A. Giannakidou, A. Mari, (Non)Veridicality in Grammar and Thought: mood, modality, and propositional attitudes, The University of Chicago Press, 2021.

C. Condoravdi, S. Lauer, Imperatives: meaning and illocutionary force, in: C. Pinon (Ed.), Empirical Issues in Syntax and Semantics, volume 9, pp. 37–58.

T. Chen, C. Guestrin, XGBoost, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, New York, NY, USA, 2016, pp. 785–794.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching Word Vectors with Subword Information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.

F. Chollet, Others, Keras, https://keras.io, 2015.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew, Huggingface's transformers: State-of-the-art natural language processing, ArXiv abs/1910.03771 (2019).

L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, B. Sagot, Camembert: a tasty french language model, arXiv preprint arXiv:1911.03894 (2019).

S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (2010) 1345–1359.

J. Howard, S. Ruder, Fine-tuned language models for text classification, CoRR abs/1801.06146 (2018).

X. Liu, P. He, W. Chen, J. Gao, Multi-task deep neural networks for natural language understanding, in: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pp. 4487–4496.

J. Baxter, A bayesian/information theoretic model of learning to learn via-multiple task sampling, Mach. Learn. 28 (1997) 7–39.

M. Buda, A. Maki, M. A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks, CoRR abs/1710.05381 (2017).

J. M. Johnson, T. M. Khoshgoftaar, Survey on deep learning with class imbalance, Journal of Big Data 6 (2019).

M. Koziarski, B. Krawczyk, M. Wozniak, Radial-based oversampling for noisy imbalanced data classification, Neurocomputing 343 (2019) 19–33.