Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Classification de tweets en situation d'urgence pour la gestion de crise

Romain Meunier[1] Leila Moudjari[1] Farah Benamara[1] Véronique Moriceau[1]
Alda Mari[2] Patricia Stolf[1]
[1]IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
[1]prenom.nom@irit.fr
[2]IJN, CNRS/ENS/EHESS, Université Paris Saclay
[2]alda.mari@ens.fr

RÉSUMÉ

Le traitement de données provenant de réseaux sociaux en temps réel est devenu un outil attractif dans les situations d'urgence, mais la surcharge d'informations reste un défi à relever. Dans cet article, nous présentons un nouveau jeu de données en français annoté manuellement pour la gestion de crise. Nous testons également plusieurs modèles d'apprentissage automatique pour classer des tweets en fonction de leur pertinence, de l'urgence et de l'intention qu'ils véhiculent afin d'aider au mieux les services de secours durant les crises selon des méthodes d'évaluation spécifique à la gestion de crise. Nous évaluons également nos modèles lorsqu'ils sont confrontés à de nouvelles crises ou même de nouveaux types de crises, avec des résultats encourageants.

ABSTRACT

Tweets Classification in Emergency Situations for Crises Management

The processing of real-time social media data has become an attractive tool in emergency situations. However, information overload remains a challenge. In this paper, we present a new dataset in French, annotated for crisis management. Relying on evaluation methods specifically designed to crisis management, we evaluate several deep learning models for classification of tweets based on their relevance, urgency and intention they convey to help emergency services during crises. We also evaluate these models on new crises or new types of crises showing interesting results.

MOTS-CLÉS : Gestion de crise, Réseaux sociaux, Apprentissage multi-tâches, Portabilité.

KEYWORDS: Crisis management, Social Media, Multi-task Learning, Portability.

1 Motivations

Les réseaux sociaux ont changé la façon dont les gens interagissent, communiquent ou ont accès à l'information partout dans le monde. Leur utilisation est largement répandue chez les internautes, les institutions publiques et les entreprises pour informer à propos d'événements ou de produits ou encore, et sans être exhaustifs, partager des opinions. Tous les domaines de la vie quotidienne sont concernés, y compris la gestion de crise qui nous intéresse ici. En effet, les catastrophes naturelles représentent une menace permanente pour les populations, les infrastructures et les ressources naturelles. La gestion efficace de ces crises nécessite des outils et des approches innovants pour aider les équipes

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs, pages 204–216.

d'intervention à réagir rapidement et à prendre des décisions éclairées (Vieweg *et al.*, 2014; Olteanu *et al.*, 2015; Palen & Liu, 2007). Dans ce contexte, les réseaux sociaux, en particulier Twitter, offrent une source d'information riche et en temps réel sur les événements en cours (Reuter *et al.*, 2018). Par exemple, plus de 17 000 tweets ont été postés pendant l'incendie de Notre Dame en 2019. Lors du séisme de 2023 en Turquie et en Syrie, des victimes piégées sous les décombres ont appelé à l'aide en publiant des messages sur Twitter (Toraman *et al.*, 2023).

Ainsi, Twitter peut être utilisé pour obtenir des informations cruciales en situation de crise comme l'état des infrastructures ou bien le niveau de préparation de la population face à la crise.

L'une des principales caractéristiques des tweets postés en période de crise est la variété des informations véhiculées qui entraîne des différences de traitement par les services de secours. Par exemple, dans le tweet (a) [1] l'auteur signale un danger qui approche et le message doit donc être traité en priorité. Le tweet (b) exprime quant à lui une critique envers le système de gestion de la crise. Même si ce tweet peut être pris en compte afin d'améliorer le service, il n'a pas de caractère urgent et ne nécessite donc pas d'intervention des secours. Enfin, l'exemple (c) est un avertissement officiel. Il doit être pris en compte par les secours afin qu'ils puissent se préparer au mieux à la catastrophe arrivant.

(a) Le feu de Landiras (au départ à 40km) s'approche de chez moi. Encore 2 villages et c'est à nous d'évacuer. Ce soir on sent vachement le brûlé. On ne panique pas mais le stress monte. Vais mal dormir.

(b) Pourquoi ne pas faire intervenir l'Armée ? Inondations à Villeneuve S/Lot en 59 : les militaires nous apportaient à manger. L'Etat abandonne les sinistrés français !

(c) ALERTE MÉTÉO : Vigilance #orange "orages, pluie-inondation" et #jaune "vagues-submersion". Soyez prudents

Ces dernières années, la littérature sur la gestion de crise a connu une expansion rapide. Elle se focalise surtout sur l'anglais, et malgré la quantité importante de jeux de données élaborés pour la gestion de crise (Imran *et al.*, 2016; McCreadie *et al.*, 2019, 2020; Alam *et al.*, 2018), les ressources en français sont encore rares, peu disponibles et très souvent limitées à une seule crise en particulier. Dans ce cadre, nos objectifs sont multiples. Il s'agit : (1) de détecter automatiquement les tweets pertinents lors d'une crise ; (2) parmi les tweets utiles/pertinents, d'identifier ceux qui sont urgents (au sens où ils doivent être traités de façon urgente par les services de secours) et (3) de caractériser le type d'information véhiculée (conseil, critique, etc.), que nous appellerons *intention*. D'un point de vue linguistique, ce terme est employé pour décrire les postures qu'un individu peut prendre eu égard à une action à entreprendre (Grano, 2017; Giannakidou & Mari, 2021). Dans cette étude, nous employons le terme pour désigner les domaines d'actionnabilité même, à savoir les domaines au sein desquels une action pourrait être entreprise (par exemple les dégâts matériels ou humains pour solliciter de l'aide). Afin de traiter ces trois points, nous proposons :

— Un corpus de tweets en français annoté à la fois selon l'utilité, l'urgence et l'intention du message [2], et qui se caractérise par sa diversité en termes aussi bien de types que de nombre de crises (crises naturelles ou non, prévisibles ou soudaines),

— *Une approche supervisée* pour la classification de tweets en situation de crise. Nous évaluons sur ce corpus un ensemble de modèles qui ont fait leurs preuves dans ce cadre, avec en particulier

1. Les exemples de tweets sont issus de notre corpus.
2. Le corpus est disponible sur demande.

une approche multi-tâches. Notre objectif n'est pas de proposer de nouveaux modèles mais d'évaluer les performances et la portabilité de modèles existants quand ils sont confrontés à des crises de types différents ou à des crises inconnues,

— *Une analyse d'erreurs* montrant les limites des approches proposées et qui permet de proposer des orientations pour améliorer les modèles dans ce domaine.

Dans cet article, nous commençons par présenter les travaux existants sur la gestion de crise pour les médias sociaux, ainsi que les corpus existants en français. La section 3 présente notre corpus annoté. La section 4 présente les modèles et les expérimentations de classification que nous avons menées sur ce corpus, ainsi qu'une analyse d'erreurs permettant d'identifier les principaux défis à relever et pistes d'amélioration pour le futur dans ce domaine. Enfin, nous concluons avec quelques perspectives pour des travaux futurs en section 5.

2 État de l'art

De nombreux travaux d'analyse des tweets en cas de crises ont émergé dans différentes langues, majoritairement l'anglais, que ce soit pendant ou après une crise (Cameron *et al.*, 2012; Imran *et al.*, 2013; McCreadie *et al.*, 2019; Kayi *et al.*, 2020; Seeberger & Riedhammer, 2022; Imran *et al.*, 2014). L'objectif est de proposer un système de classification des messages selon : (1) l'utilité (le message est-il utile ou non pour les services de secours ?), (2) l'urgence (le message est-il urgent ou non pour les services de secours ? Éventuellement, quel degré d'urgence ?) et (3) les intentions véhiculées (déclaration de dégâts, avertissement, critique, etc.). D'autres classifications ont également été proposées comme la détection de messages postés par des témoins directs ou indirects de la crise (Zahra *et al.*, 2020). Des campagnes d'évaluation ont été proposées dans ce domaine, notamment TREC-IS (Incident Streams track) [3] associé à TREC 2019 et 2020 (McCreadie *et al.*, 2019, 2020) et plus récemment CrisisFACTS2022 [4] pour la génération de résumé de situations de crises. Les participants de TREC-IS ont pour objectif le développement de systèmes de veille en temps réel capables de suivre l'évolution d'incidents tels que des catastrophes naturelles, des incidents terroristes ou des crises de santé publique à partir de flux de données textuelles en ligne, tels que des flux Twitter ou des flux de nouvelles en langue anglaise. Les meilleurs systèmes soumis utilisent des approches neuronales. Par exemple, Wang *et al.* (2021) et Dusart *et al.* (2021) utilisent une variation du modèle BERT, qui a été pré-entrainé sur des tweets relatifs à des crises dans une approche multi-tâche. Globalement, les résultats montrent que la détection de l'intention reste la plus complexe en raison du déséquilibre des données d'apprentissage (les messages urgents sont souvent minoritaires, par exemple de l'ordre de 3,87 % pour les dégâts matériels et 1,93 % pour les dégâts humains).

En ce qui concerne les ressources en langue française, on peut citer le projet SURICATE-Nat [5] qui vise à exploiter Twitter pour être informé rapidement de séismes ou d'inondations ayant lieu en France, ou encore le cadre proposé par (Interdonato *et al.*, 2018) pour extraire et regrouper automatiquement des données relatives à une crise sans recourir à une annotation manuelle ou à des catégories prédéfinies. Cependant, les données de cette étude ne sont pas disponibles. Kozlowski *et al.* (2020) ont proposé le premier corpus en français, composé d'environ 13 000 tweets concernant des crises écologiques (catastrophes naturelles) et annotés manuellement selon trois niveaux d'information :

3. https://www.dcs.gla.ac.uk/~richardm/TREC_IS/
4. https://crisisfacts.github.io/
5. http://www.suricatenat.fr/Suricate-Nat/

l'utilité, l'urgence et les intentions. Plus récemment, Diwersy *et al.* (2022) ont utilisé un corpus de discours publics pour analyser les changements dans l'utilisation des mots et expressions pendant la crise sanitaire de la COVID-19 en France, en utilisant des méthodes phonético-textométriques. Caillaut *et al.* (2022) ont construit automatiquement un corpus de 6 023 documents contenant 304 826 entités liées en utilisant la Wikipédia française afin de géolocaliser les publications des réseaux sociaux en temps réel lors de catastrophes naturelles. Enfin, le projet CrisisNLP [6] rassemble une communauté de chercheurs autour de la création de ressources et d'outils de traitement de texte pour l'analyse de données de médias sociaux pendant les crises. Le projet a produit des corpus dans plusieurs langues, dont le français (17 329 tweets relatifs à un événement de glissement de terrain survenu en 2015).

En gestion de crises, les méthodes les plus anciennes utilisées pour classifier les messages s'appuyaient sur l'apprentissage automatique avec notamment l'utilisation de Random Forest (Breiman, 2001) ou des réseaux bayésiens (Schneider, 2003). Les approches utilisées plus récemment sont de deux types : celles qui utilisent des modèles d'apprentissage profond comme les réseaux neuronaux convolutifs (Nguyen *et al.*, 2017) ou les réseaux de neurones récurrents (Alharbi & Lee, 2019), et celles qui utilisent les transformeurs (Zahera *et al.*, 2019) avec des modèles tels que BERT (Devlin *et al.*, 2019). La campagne d'évaluation TREC-IS (Incident Stream) (McCreadie *et al.*, 2020) donne un apercu des méthodes actuelles pour l'identification de tweets en situation de crise et lors de l'édition 2021, les transformeurs sont les modèles qui ont obtenu les meilleurs résultats [7].

Pour conclure, un défi important est la capacité des systèmes à identifier des messages urgents pour des crises nouvelles non présentes lors de l'entraînement. Or, d'après le rapport du GIEC 2022 (De Pryck, 2022), le changement climatique va entraîner de nouveaux types de crises et il faut donc avoir un modèle capable de se préparer à l'inconnu pour un déploiement chez les acteurs de cellules de crises. Dans cet article, nous nous intéressons pour la première fois à ce défi sur des données en français en évaluant la portabilité de nos modèles à différents types de crises.

3 Corpus de tweets en français annoté pour la gestion de crise

Le seul corpus pour la gestion de crise existant pour le français et disponible pour la communauté est celui proposé par (Kozlowski *et al.*, 2020) [8]. Ce corpus a l'inconvénient d'être composé quasi exclusivement de tweets en rapport avec des crises météorologiques prévisibles (tempête, ouragan, inondation). Nous avons choisi d'étendre ce corpus à d'autres types de crises, à savoir des crises non prévisibles afin de tester la portabilité des modèles sur ces nouveaux types. Pour cela, nous avons collecté des tweets en français portant sur deux nouveaux types de crise (incendie et attaque terroriste) tout en suivant la même méthodologie de collecte utilisée dans (Kozlowski *et al.*, 2020), c'est-à-dire collecter des tweets postés entre 24h avant la crise et 72h [9] après en utilisant des mots-clés représentatifs de la crise (par exemple, "incendie", "forêt" et "gironde" pour les incendies dans les Landes). Ces nouvelles crises ont la particularité d'être des crises soudaines : l'attaque terroriste de Trèbes en 2018, l'incendie de l'usine Lubrizol en 2019, l'incendie de Notre Dame en 2019, l'explosion/incendie d'un immeuble à Sanary en 2021, les incendies dans les Landes (Landiras

6. https://crisisnlp.qcri.org/
7. https://trecis.github.io/
8. https://github.com/DiegoKoz/french_ecological_crisis
9. Pour le cas des crises s'étalant sur plusieurs jours comme les incendies, la fin de la crise est considérée comme étant le début de résolution de la crise. Par exemple, dans le cas des incendies, il s'agit de l'extinction des premiers feux.

Tweet	Utilité	Urgence	Intention
Quatre départements du sud-est placés en **vigilance** orange aux pluies et inondations	utile	urgent	avert. conseil
Un **bébé** se **noie** dans l'inondation de l'appartement de ses parents	utile	urgent	dégâts humains
Un **immeuble s'effondre** en plein centre ville de Marseille http ://bit.ly/2zxte0S	utile	urgent	dégâts matériels
Inondations de l'Aude : la **solidarité** syndicale s'organise	utile	non urgent	soutien
Étonnant les feux d'artifice dans le #Gard, région d' #Uzes alors que ça brûle en #Gironde #14Juillet2022	utile	non urgent	critiques
#Inondations dans l'Aude : les maires et les députés au plus près des sinistrés	utile	non urgent	autres messages
tu mets le feu	non utile	non utile	non utile

TABLE 1 – Extrait du corpus avec les annotations associées.

et La Teste de Buch) en 2022. À ces crises, nous avons aussi ajouté celle de l'effondrement de 2 immeubles à Lille survenu en 2022 (le type "effondrement" n'étant représenté que par une seule crise dans le corpus initial de (Kozlowski *et al.*, 2020)). Ces tweets ont ensuite été annotés par 2 paires d'annotateurs selon le même schéma d'annotation que celui proposé par (Kozlowski *et al.*, 2020) (cf. Figure 1). Selon ce schéma, les messages ont d'abord été annotés en utilité, puis en urgence et enfin en intentions pour les messages jugés utiles. Les annotateurs se sont d'abord entraînés sur des tweets issus du corpus de (Kozlowski *et al.*, 2020) puis ont annoté les mêmes tweets que ceux utilisés pour calculer l'accord-interannotateur dans (Kozlowski *et al.*, 2020). Les kappas étant alors similaires (autour de 0,70 pour l'utilité, 0,67 pour l'urgence et 0,65 pour l'intention), les annotateurs ont annoté le corpus en entier.

La table 1 donne quelques exemples du corpus ainsi que les labels qui leurs sont associés. Les mots en gras sont des mots qui ont été décisifs pour décider quel label associer à quel message.

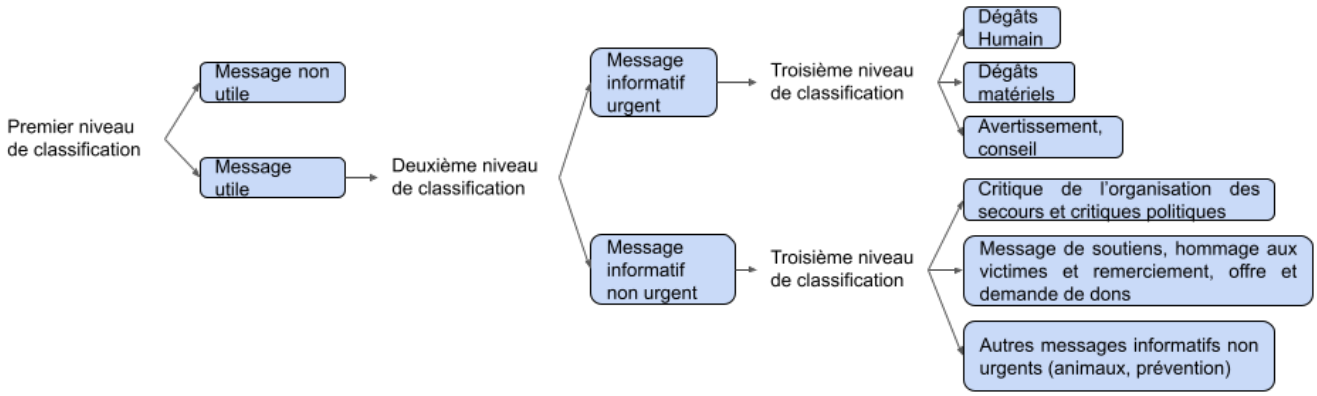

FIGURE 1 – Typologie des messages pour les crises écologiques (Kozlowski *et al.*, 2020).

La table 2 présente la distribution par classe pour toutes les crises disponibles. Il faut noter que certaines crises (incendie de l'usine Lubrizol, incendie d'un immeuble à Sanary, incendie de Notre-Dame et attaque de Trèbes) ont été annotées uniquement selon le premier et le second niveau de classification. La collecte de nouvelles crises a permis d'augmenter le corpus de (Kozlowski *et al.*, 2020) de 52,70 %. Les crises prévisibles (inondations, tempêtes, ouragans) sont les plus représentées avec 12 112 messages contre 7 483 messages pour les crises soudaines (effondrements, incendies, attaque). La collecte de nouvelles données a donc permis de rééquilibrer ces deux types de crises. Logiquement, on remarque aussi que pour les crises soudaines, il y a moins de messages de type AVERTISSEMENT-CONSEIL que pour les crises prévisibles.

Enfin, on note que le corpus est déséquilibré : la classe MESSAGE NON UTILE représente 57,96 % du corpus alors que la classe MESSAGE URGENT 20,26 %. Ceci est dû en grande partie à la méthode de collecte des tweets puisque des tweets postés 24 heures avant les crises ont été récupérés et

	non utile	utile-urgent			utile-non urgent			Total
		avert. conseil	dégâts humains	dégâts matériels	soutien	autres messages	critiques	
Corpus Kozlowski et al.								
Inondation Aude	1 065	150	34	157	157	184	26	**1 773**
Inondation Autre	993	292	35	111	231	16	19	**1 697**
Inondation Corse	468	51	58	12	52	66	13	**720**
Tempête Guadeloupe	612	91	0	2	3	10	2	**720**
Tempête Bruno	586	107	5	11	2	9	0	**720**
Tempête Susanna	484	129	11	38	4	54	0	**720**
Tempête Ulrika	650	47	2	18	0	4	0	**721**
Tempête Réunion	587	56	5	9	12	46	5	**720**
Tempête Fionn	552	138	6	10	0	8	6	**720**
Tempête Egon	609	66	1	35	0	10	0	**721**
Tempête Eleanor	590	82	22	19	1	6	0	**720**
Ouragan Harvey	628	78	10	2	1	1	0	**720**
Ouragan Irma	790	121	47	55	199	199	29	**1 440**
Effondrement Marseille	627	9	24	11	11	19	19	**720**
Nouvelles crises soudaines								
Effondrement Lille	320	2	39	27	12	117	32	**549**
Incendie Gironde Landes	1 394	51	23	93	317	380	165	**2 423**
Incendie Lubrizol	137	583			627			**1 347**
Incendie Sanary	6	363			164			**533**
Incendie Notre-Dame	86	224			209			**519**
Attaque Trèbes	174	398			810			**1 382**
Total	**11 358**	**3 970**			**4 257**			**19 595**

TABLE 2 – Statistiques du corpus annoté.

sont donc très majoritairement non pertinents. En ce qui concerne les intentions, les classes sont aussi déséquilibrées : la classe DÉGÂTS HUMAINS représente seulement 1,93 % du corpus avec 306 messages tandis que la classe AVERTISSEMENT-CONSEIL est constituée de 1 470 tweets soit 9,88 % du corpus. La gestion de crise à partir de tweets s'apparente donc à un problème de détection de signaux faibles.

4 Modèles proposés et résultats

Nous avons évalué ce corpus avec un ensemble de modèles qui ont fait leurs preuves dans ce cadre, avec en particulier une approche multi-tâche. Ont été testés différents modèles tels que BERT, CamemBERT mais nous ne présentons ici que les modèles qui ont obtenu les meilleurs résultats, à savoir :

— **FlauBERT**$_{Fine-Tuned}$: il s'agit d'un modèle FlauBERT (Le *et al.*, 2019) pré-entraîné sur 358 834 tweets non annotés du corpus de (Kozlowski *et al.*, 2020) et qui a montré de meilleures performances comparé au modèle FlauBERT (Le *et al.*, 2019) pre-entraîné sur du texte français provenant de différentes sources (par exemple Wikipédia et livres). Pour entraîner le modèle, nous avons utilisé un optimisateur Adam, avec un taux d'apprentissage de 2^{-5} sur 4 epochs.

— **FlauBERT**$_{Fine-Tuned+MultiTask}$: nous avons également entraîné notre modèle FlauBERT$_{Fine-Tuned}$ avec une architecture multi-tâches. L'objectif est de partager les connais-

sances entre les trois classifieurs de tâche (utilité, urgence et intention) avec un entraînement conjoint. Dans le modèle FlauBERT$_{Fine-Tuned+MultiTask}$, chaque classifieur est entraîné pour une tâche spécifique. Tous les classifieurs partagent les mêmes couches basses (qui sont des couches de **FlauBERT**$_{Fine-Tuned}$) sauf la dernière qui est spécifique à la tâche. Pour entraîner le modèle, nous avons utilisé un optimisateur Adam, avec un taux d'apprentissage de 2^{-5} sur 4 epochs.

Le corpus étant déséquilibré, nous utilisons la fonction de perte *focal loss* (Lin *et al.*, 2016) plus appropriée pour traiter les classes déséquilibrées [10]. Un prétraitement des messages est aussi effectué : les nombres sont remplacés par un token "nombre", les mentions sont supprimées.

4.1 Protocole d'évaluation

Chaque modèle est évalué pour chaque tâche : (1) la tâche d'utilité qui est un problème binaire (MESSAGE UTILE vs. MESSAGE NON UTILE), (2) la tâche d'urgence qui est un problème ternaire (MESSAGE NON UTILE vs. MESSAGE NON URGENT vs. MESSAGE URGENT), et (3) la tâche de détection d'intention (7 classes). Pour la tâche d'intention, les tweets non annotés en intention ne sont pas pris en compte.

Notre objectif étant d'évaluer les performances des différents modèles lorsqu'ils sont confrontés à différents types de crise, nous les avons testés dans les configurations suivantes :

1. **Généralisation** : Afin d'évaluer la portabilité des modèles aux crises soudaines, nous les avons entraînés sur le corpus initial de (Kozlowski *et al.*, 2020) et les avons testés sur les nouvelles crises que nous avons ajoutées.
2. **Hors-événement (Out-OF-Event)** : Inspirée de (Nguyen *et al.*, 2016) cette évaluation consiste à entraîner un modèle sur un ensemble de crises (par exemple, Ouragan Irma et Tempête Egon) et tester sur d'autres crises du même type (par exemple, Ouragan Harvey et Tempête Fionn). Pour constituer notre corpus d'entraînement, nous avons appliqué la méthode utilisée pour TREC-IS 2018 (McCreadie *et al.*, 2019) et sélectionné pour chaque type de crise, celle qui possède le plus de messages afin d'avoir la meilleure couverture pour chaque type de crise. Ainsi, les crises qui font partie du jeu d'entraînement sont : Inondation Aude, Tempête Egon, Ouragan Irma, Effondrement Marseille, Incendie Gironde Landes et Attaque Trèbes. Ceci donne un total de 8 459 messages pour les tâches de détection d'utilité et d'urgence, et 7 077 messages pour la tâche d'intention. Le jeu de test est composé de 11 136 messages pour les tâches d'utilité et d'urgence et 8 737 messages pour la tâche d'intention et est composé des crises non présentes dans le jeu d'entraînement.
3. **Hors-Type (Out-OF-Type)** : Cette configuration consiste à entraîner le modèle sur des types de crises spécifiques (par exemple, les inondations et les ouragans) puis à les tester sur un autre type de crise (par exemple, les incendies). Le but est de vérifier si le modèle peut s'adapter à de nouvelles crises qu'il ne connaît pas, et donc à de potentielles crises futures. Dans le cas d'une évaluation Hors-Type, pour avoir un résultat représentatif, il faut tester sur chaque type de crise individuellement puis faire la moyenne des F-scores obtenus pour chaque crise. Les résultats présentés ici sont calculés de cette façon. Nous avons fait une expérimentation Hors-Type

10. Nous avons également testé la Weigthed cross-entropy et la cross-entropy mais les résultats étaient moins bons.

sur 5 types de crises (Incendie, Inondation, Ouragan, Tempête, Effondrement) [11]. La méthode d'évaluation hors-type est une évaluation récente et encore peu exploitée : Kersten *et al.* (2019) et Algiriyage *et al.* (2021) utilisent cette stratégie mais sur un nombre de types de crises très réduit. Bourgon *et al.* (2022) ont également utilisé une évaluation similaire. Cependant, la plupart de ces travaux se concentrent uniquement sur la classification des messages en utilité et urgence. Nous allons plus loin ici en abordant, en plus, la tâche plus complexe de classification en intention.

Il est à noter que l'évaluation Hors-Événement est la méthode d'évaluation standard en gestion de crise. Nous proposons dans cet article deux protocoles supplémentaires afin d'évaluer la portabilité des modèles vers de nouvelles crises.

4.2 Résultats

La table 3 présente les résultats obtenus par les deux meilleurs modèles sur les tâches de prédiction de l'utilité, de l'urgence et des intentions dans la configuration *Généralisation*. À titre de comparaison, les macro F1-scores obtenus par (Kozlowski *et al.*, 2020) sur le corpus initial sont de 85,3 pour la tâche d'utilité, de 76,7 pour l'urgence et 65,4 pour les intentions avec le même modèle FlauBERT$_{Fine-Tuned}$ (respectivement, 85,4, 77,5 et 64,0 pour le modèle FlauBERT$_{Fine-Tuned+MultiTask}$). Sans surprise, les résultats obtenus quand les modèles sont testés sur de nouveaux types de crise sont moins bons mais sont tout de même honorables. Globalement, le modèle FlauBERT$_{Fine-Tuned+MultiTask}$ obtient de meilleurs résultats.

Modèles	Corpus de test		**utilité**	**urgence**	**intention**
FlauBERT$_{Fine-Tuned}$	Kozlowski	Macro F1	85,3	76,7	65,4
FlauBERT$_{Fine-Tuned}$	Corpus augmenté	Précision	59,84	**55,87**	**49,82**
		Rappel	60,23	55,88	49,07
		Macro F1	59,92	**55,42**	44,21
FlauBERT$_{Fine-Tuned+MultiTask}$	Corpus augmenté	Précision	**65,34**	54,00	45,12
		Rappel	**65,62**	**62,52**	**54,57**
		Macro F1	**64,66**	55,25	**46,42**

TABLE 3 – Résultats pour l'expérimentation de généralisation.

Les tables 4 et 5 présentent les résultats obtenus sur les 3 tâches dans les configurations *Hors-Événement* et *Hors-Type*. Là aussi, le modèle multi-tâches obtient globalement de meilleurs résultats.

Les résultats dans les configurations *Hors-Événement* et *Hors-Type* sont relativement similaires pour le modèle FlauBERT$_{Fine-Tuned+MultiTask}$ alors que FlauBERT$_{Fine-Tuned}$ performe mieux dans l'expérimentation Hors-Événement. Il faut noter que les tailles des corpus d'entraînement respectifs sont très différentes. En effet, le jeu d'entraînement de l'expérimentation *Hors-Événement* contient 8 459 messages alors que la taille du corpus d'entraînement *Hors-Type* varie selon le type de crise testé : en moyenne, un jeu d'entraînement pour une expérimentation *Hors-Type* contient 15 017 messages, soit 1,77 fois plus de données d'entraînement que pour l'expérimentation *Hors-Événement*.

11. Afin de garantir un jeu de test homogène pour les 3 tâches qui nous concernent, les données sur l'attaque terroriste de Trèbes n'ont pas été considérées car pas annotées en intention. Pour les données sur les incendies, seule la crise des incendies dans les Landes a été conservée, car elle a la seule annotée dans les 3 tâches.

Modèles		utilité	urgence	intention
FlauBERT$_{Fine-Tuned}$	Précision	65,69	58,85	**56,04**
	Rappel	65,67	60,15	47,37
	Macro F1-Score	65,68	59,39	**50,31**
FlauBERT$_{Fine-Tuned+MultiTask}$	Précision	**70,58**	**62,72**	52,31
	Rappel	**70,28**	**64,47**	**48,92**
	Macro F1-Score	**72,53**	**63,13**	49,65

TABLE 4 – Résultats pour l'expérimentation Hors-Événement.

Modèles		utilité	urgence	intention
FlauBERT$_{Fine-Tuned}$	Précision	73,09	**64,19**	51,26
	Rappel	**75,06**	**65,12**	46,06
	Macro F1-Score	73,19	**63,75**	45,87
FlauBERT$_{Fine-Tuned+MultiTask}$	Précision	**74,88**	63,73	**53,13**
	Rappel	74,58	64,37	**51,49**
	Macro F1-Score	**73,98**	63,04	**50,20**

TABLE 5 – Résultats pour l'expérimentation Hors-Type.

On en déduit donc que, étant donné que FlauBERT$_{Fine-Tuned+MultiTask}$ s'entraîner sur les 3 taches simultanément, il apprend plus vite et a donc besoin de moins de données pour être performant.

Pour l'expérimentation *Hors-Type*, les moins bons résultats sont obtenus sur les crises soudaines : ainsi avec FlauBERT$_{Fine-Tuned}$, on obtient un F1-score de 31,33 quand on teste sur les crises de type effondrement, 40,24 pour les incendies alors qu'on obtient un F1-score de 56,00 quand on teste sur les ouragans. Pour comparer avec les résultats de (Kozlowski *et al.*, 2020), leurs expérimentations *Hors-Type* sur le corpus initial (en testant les modèles sur l'effondrement de Marseille) avaient obtenu un F1-Score moyen de 42,30 avec FlauBERT$_{Fine-Tuned+MultiTask}$, contre 50,20 sur notre corpus augmenté. On peut donc penser que notre nouveau corpus permet de mieux appréhender l'arrivée d'un nouveau type de crise.

Si l'on compare maintenant les performances des deux modèles testés, FlauBERT$_{Fine-Tuned+MultiTask}$ est le plus performant pour la tâche de détection d'utilité. Cependant, en fonction du type d'expérimentation, FlauBERT$_{Fine-Tuned}$, peut être plus efficace, par exemple sur la tâche de prédiction de l'urgence en configuration *Hors-Type*. Pour la tâche de prédiction des intentions, qui est la tâche la plus complexe car 7 classes possibles à prédire, FlauBERT$_{Fine-Tuned+MultiTask}$ obtient de meilleurs résultats dans les configurations *Généralisation* et *Hors-Type*. Il s'agit de deux tâches dont la difficulté est que certains types de crises du jeu de test ne sont pas connus du modèle car absents du jeu d'apprentissage.

4.3 Analyse d'erreurs

Pour analyser les erreurs de classification, nous nous intéressons à la tâche la plus complexe, à savoir la détection des intentions dans un cadre *Hors-Type*. Nous analysons ici en détail les résultats du modèle FlauBERT$_{Fine-Tuned+MultiTask}$ qui a obtenu les meilleurs résultats sur cette tâche.

Dans la table 6, on remarque que les classes pour lesquelles le modèle a les moins bons résultats sont les classes AUTRES MESSAGES et CRITIQUES, tandis que la classe où le modèle est le plus performant est MESSAGES NON UTILES, celle-ci étant la classe comportant le plus de données d'entraînement. De plus, étant donné qu'il s'agit d'une classe commune aux tâches d'intention, d'urgence et d'utilité, le modèle multi-tâches s'entraîne 3 fois plus sur cette classe. Il est donc normal que le modèle soit le plus performant sur cette classe.

	non utile	**avert. conseil**	**autres messages**	**dégâts matériels**	**Soutien**	**dégâts humains**	**critiques**	**Moyenne**
Précision	81,94	46,50	33,14	53,54	59,25	55,07	42,47	53,13
Rappel	83,56	51,34	28,29	52,15	64,76	52,46	27,87	**51,49**

TABLE 6 – Résultats du modèle FlauBERT$_{Fine-Tuned+MultiTask}$ dans la configuration *Hors-Type*.

La classe AUTRES MESSAGES est celle pour laquelle les résultats sont les moins bons. Cependant, d'un point de vue applicatif, on peut relativiser ces résultats car d'après (Kozlowski *et al.*, 2020), "la catégorie AUTRES MESSAGES n'a pas d'impact immédiat sur les actions à mettre en oeuvre mais contribuent à informer les personnes sur la situation. Ils regroupent : (i) les messages à propos d'animaux, (ii) les messages qui ont pour but de donner des informations via un lien, des photos ou des vidéos et (iii) des messages de prévention qui donnent des conseils sur la crise en cours".

Texte	**Label**	**Prédiction**
(1) Abritez vous réfugiez vous sur les étages ne sortez pas Signalez votre présence au numero numero Aude Inondation Catastrophe VigilanceRouge	Autres messages	Avert-Conseil
(2) C est nul Dans l Aude numero morts numero disparus le chef des pompiers de Carcassonne a constaté avec colère un afflux de touristes voyeurs dans les communes sinistrées J invite les gens à venir armés de pelles et de raclettes pas de leur portable pour aider les habitants	Autres messages	Critiques
(3) Inondations dans l Aude les maires et les députés au plus près des sinistrés	Autres messages	Soutiens
(4) Certaines personnes commencent déjà à s installer pour dormir la journée a été longue Trebes Aude	Autres messages	Message non utile
(5) INONDATIONS À Carcasonne l eau continue de monter la préfecture invite les habitants qui vivent au bord de l Aude à monter dans les étages L hôpital est inondé intemperies	Avert-Conseil	Dégâts Matériels
(6) Si c est aussi rapide et efficace qu à St Martin et à St Barth je souhaite bon courage aux pauvres victimes de ces inondations	Critiques	Soutiens
(7) Courage à nos compatriotes victimes des intempéries dans l Aude Malheureusement plusieurs victimes à déplorer Nos services de sécurité et de secours sont remarquables	Dégâts Humains	Soutiens
(8) Inondations l aéroclub de narbonne est en danger de mort intemperies	Dégâts Matériels	Dégâts Humains
(9) C est la tempête dehors je sais pas comment j ai survécu	Message Non Utile	Autres messages
(10) Comme c est désolant de voir des immeubles s effondrer suite à un gros manque d entretien De tout coeur avec les habitants de cette avenue et ceux qui y travaillent tout les jours Aux collectivités de réagir maintenant	Soutiens	Critiques

TABLE 7 – Exemples d'erreurs de classification.

La table 7 présente quelques exemples d'erreur de classification. Dans le premier message, on constate que le cas (iii) de la définition ci-dessus pose problème car il est très proche de la définition de la catégorie AVERTISSEMENT - CONSEIL.

En dehors de ce type d'erreur spécifique à la catégorie AUTRES MESSAGES, les erreurs dans les autres catégories sont plutôt similaires. Par exemple, pour le deuxième message, l'erreur de classification s'explique par le fait que l'intention du message est de critiquer la population et pas le système de secours, ce qui le différencie d'un message de type CRITIQUES. Mais comme le terme "chefs des pompiers" est utilisé, cela peut induire le modèle en erreur.

Il existe aussi le problème des intentions multiples : certains messages véhiculent bien l'intention indiquée dans le label mais l'intention prédite par le modèle est tout aussi vraie (exemples (3) et (7)).

Le manque de contexte peut aussi expliquer certaines erreurs de classification (exemples (4) et (9)).

Certaines erreurs peuvent aussi s'expliquer par des usages figuratifs (métaphores, ironie, etc.) : par exemple dans le message (8), l'utilisation du mot "mort" va entraîner une classification en DÉGÂTS HUMAINS alors qu'ici c'est un aéroclub qui est concerné. Le message (6) est quant à lui porteur d'ironie, couramment utilisée pour exprimer des critiques mais très difficile à détecter automatiquement (Zeng & Li, 2022). Ceci explique peut être pourquoi la classe CRITIQUES a les deuxièmes moins bons résultats.

5 Conclusions et perspectives

Nous avons présenté dans cet article un nouveau corpus en français annoté pour la gestion de crise. Ce corpus est varié en types et en nombre de crises : crises prévisibles (tempête, ouragan, inondation) et crises soudaines (effondrement, incendie, attaque terroriste). Cette diversité permet aux intervenants de se préparer à faire face aux situations imprévues et à mettre en place des stratégies de gestion de crise adaptées. Nous avons mené des expérimentations pour évaluer l'efficacité de modèles d'apprentissage supervisé pour la détection automatique des tweets pertinents et urgents liés à une crise. Nous avons montré que les modèles obtenaient des résultats encourageants lorsqu'ils sont testés sur de nouvelles crises ou même de nouveaux types de crises. Au vu des résultats, nous constatons que le problème reste encore déséquilibré en termes de classes comme en termes de représentativité des crises. Aussi, nous envisageons d'augmenter automatiquement le corpus sur les classes et les crises les moins représentées. Une autre piste envisagée est de chercher d'autres sources d'information liées aux crises et de les combiner avec les tweets dans une perspective de détection multi-modale.

Remerciements

Ce travail a été réalisé dans le cadre du projet CNRS-prématuration INTACT impliquant l'Institut de Recherche en Informatique de Toulouse (IRIT) et l'Institut Jean Nicod (IJN).

Références

ALAM F., OFLI F. & IMRAN M. (2018). CrisisMMD : Multimodal Twitter Datasets from Natural Disasters. In *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM)*.

ALGIRIYAGE N., SAMPATH R., PRASANNA R., DOYLE E. E., STOCK K. & JOHNSTON D. (2021). Identifying disaster-related tweets : a large-scale detection model comparison. In *Social Media in Crises and Conflicts, Proceedings of the 18th ISCRAM Conference*, p. 731–743.

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

BOURGON N., BENAMARA F., MARI A., MORICEAU V., CHEVALIER G. & LEYGUE L. (2022). Are Sudden Crises Making me Collapse ? Measuring Transfer Learning Performances on Urgency Detection. In *19th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2022)*.

BREIMAN L. (2001). Random forests. *Machine learning*, **45**, 5–32.

CAILLAUT G., GRACIANNE C., ABADIE N., TOUYA G. & AUCLAIR S. (2022). Automated construction of a French Entity Linking dataset to geolocate social network posts in the context of natural disasters. In I. D. LIBRARY, Éd., *19th International Conference on Information Systems for Crisis Response and Management*, 19 th ISCRAM 2022 Conference Proceedings, Tarbes, France. HAL : hal-03631387.

CAMERON M. A., POWER R., ROBINSON B. & YIN J. (2012). Emergency situation awareness from twitter for crisis management. In *Proceedings of the 21st international conference on world wide web*, p. 695–698.

DE PRYCK K. (2022). *GIEC. La voix du climat*. Presses de Sciences Po.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186.

DIWERSY S., DIDIRKOVÁ I., DODANE C. & HIRSCH F. (2022). Les temps de la crise sanitaire au prisme d'une série chronologique : une étude phonético-textométrique. In *16th International Conference on Statistical Analysis of Textual Data*.

DUSART A., PINEL-SAUVAGNAT K. & HUBERT G. (2021). ISSumSet : a tweet summarization dataset hidden in a TREC track. In *Proceedings of the 36th annual ACM symposium on applied computing*, p. 665–671.

GIANNAKIDOU A. & MARI A. (2021). *Truth and veridicality in grammar and thought : Mood, modality, and propositional attitudes*. University of Chicago Press.

GRANO T. (2017). The logic of intention reports. *Journal of Semantics*, **34**(4), 587–632.

IMRAN M., CASTILLO C., LUCAS J., MEIER P. & VIEWEG S. (2014). Aidr : Artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW'2014, p. 159–162 : ACM. DOI : 10.1145/2567948.2577034.

IMRAN M., ELBASSUONI S., CASTILLO C., DIAZ F. & MEIER P. (2013). Extracting information nuggets from disaster-related messages in social media. *Iscram*, **201**(3), 791–801.

IMRAN M., MITRA P. & CASTILLO C. (2016). Twitter as a Lifeline : Human-annotated Twitter Corpora for NLP of Crisis-related Messages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, (LREC'2016) : European Language Resources Association (ELRA).

INTERDONATO R., DOUCET A. & GUILLAUME J.-L. (2018). Unsupervised Crisis Information Extraction from Twitter Data. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, p. 579–580. DOI : 10.1109/ASONAM.2018.8508261.

KAYI E. S., NAN L., QU B., DIAB M. & MCKEOWN K. (2020). Detecting urgency status of crisis tweets : A transfer learning approach for low resource languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 4693–4703.

KERSTEN J., KRUSPE A., WIEGMANN M. & KLAN F. (2019). Robust filtering of crisis-related tweets. In *ISCRAM 2019 conference proceedings-16th international conference on information systems for crisis response and management*.

KOZLOWSKI D., LANNELONGUE E., SAUDEMONT F., BENAMARA F., MARI A., MORICEAU V. & BOUMADANE A. (2020). A three-level classification of french tweets in ecological crises. *Information Processing & Management*, **57**(5), 102284.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2019). FlauBERT : Unsupervised Language Model Pre-training for French. *arXiv preprint arXiv :1912.05372*.

LIN Y., SHEN S., LIU Z., LUAN H. & SUN M. (2016). Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2124–2133, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1200.

MCCREADIE R., BUNTAIN C. & SOBOROFF I. (2019). TREC incident streams : Finding actionable information on social media. In *Proceedings of the 16th ISCRAM Conference*.

MCCREADIE R., BUNTAIN C. & SOBOROFF I. (2020). Incident streams 2019 : Actionable insights and how to find them. In *Proceedings of the 17th ISCRAM Conference*.

NGUYEN D., AL MANNAI K. A., JOTY S., SAJJAD H., IMRAN M. & MITRA P. (2017). Robust classification of crisis-related data on social networks using convolutional neural networks. In *Proceedings of the international AAAI conference on web and social media*, volume 11, p. 632–635.

NGUYEN D. T., MANNAI K. A. A., JOTY S., SAJJAD H., IMRAN M. & MITRA P. (2016). Rapid Classification of Crisis-Related Data on Social Networks using Convolutional Neural Networks. *arXiv preprint arXiv :1608.03902*.

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

PALEN L. & LIU S. B. (2007). 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, p. 727–736.

REUTER C., HUGHES A. L. & KAUFHOLD M.-A. (2018). Social media in crisis management : An evaluation and analysis of crisis informatics research. *International Journal of Human–Computer Interaction*, **34**(4), 280–294.

SCHNEIDER K.-M. (2003). A comparison of event models for naive bayes anti-spam e-mail filtering. In *10th Conference of the European Chapter of the Association for Computational Linguistics*.

SEEBERGER P. & RIEDHAMMER K. (2022). Enhancing crisis-related tweet classification with entity-masked language modeling and multi-task learning. *arXiv preprint arXiv :2211.11468*.

TORAMAN C., KUCUKKAYA I. E., OZCELIK O. & SAHIN U. (2023). Tweets under the rubble : Detection of messages calling for help in earthquake disaster. *arXiv preprint arXiv :2302.13403*.

VIEWEG S., CASTILLO C. & IMRAN M. (2014). Integrating Social Media Communications into the Rapid Assessment of Sudden Onset Disasters. In *Proceedings of the 6th International Conference of Social Informatics*, SocInfo'14, p. 444–461.

WANG C., NULTY P. & LILLIS D. (2021). Transformer-based multi-task learning for disaster tweet categorisation. *arXiv preprint arXiv :2110.08010*.

ZAHERA H. M., ELGENDY I. A., JALOTA R., SHERIF M. A. & VOORHEES E. (2019). Fine-tuned bert model for multi-label tweets classification. In *TREC*, p. 1–7.

ZAHRA K., IMRAN M. & OSTERMANN F. O. (2020). Automatic identification of eyewitness messages on Twitter during disasters. *Information Processing & Management*, **57**(1), 102–107.

ZENG Q. & LI A.-R. (2022). A survey in automatic irony processing : Linguistic, cognitive, and multi-X perspectives. In *Proceedings of the 29th International Conference on Computational Linguistics*, p. 824–836, Gyeongju, Republic of Korea : International Committee on Computational Linguistics.