République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université AMO de Bouira

Faculté des Sciences et des Sciences Appliquées

Département d'Informatique

Mémoire de Licence

en Informatique

Spécialité : *Votre spécialité*

Thème

Titre de votre projet

Encadré par

— Nom Encadreur

Réalisé par

— Nom Étudiant 1

— Nom Étudiant 2

2015/2016

Remerciements

Avant tout nous remercions notre Dieu le tout puissant d'avoir nous donner la force et le courage de mener à terme le présent travail.

Nous tenons à remercier tout d'abord notre chère promotrice AID AICHA pour sa patience, sa disponibilité, son sérieux, ses remarques, ses conseils, son respect et sa bienveillance. Qu'elle trouve ici le témoignage de notre profonde gratitude.

A nos chers enseignants du département d'informatique, un remerciement particulier et sincère pour tous les efforts que vous avez fournis pour nous encadrer tout au long de ces années, vous nous avez enrichi avec vos connaissances et savoirs, nous avons beaucoup appris avec vous, vos remarques et conseils ont contribué à notre progression et amélioration au cours de notre cursus.

Notre gratitude va également aux membres de jury pour avoir jugé et noté notre modeste travail.

Dédicaces

Au nom du dieu clément et miséricordieux

Je dédie ce mémoire à :

Mon cher père BRAHIM, ma chère mère ASSIA, que nulle dédicace ne puisse exprimer mes sincères sentiments, pour leur patience illimitée, leur encouragements contenu, leur aide, leur tendance inestimable.
Je suis très reconnaissante pour tous les sacrifices que vous n'avez cessé de me donner depuis ma naissance, durant mon enfance et mes études du primaire jusqu'à l'université. Puisse Dieu, le tout puissant, vous garde et vous procure santé, longue vie et bonheur.

Mes chers frères ISMAIL, MOHAMED et ma petite sœur INES pour leur grand amour leur soutien, qu'ils trouvent ici l'expression de ma haute gratitude.

Veuillez accepter, l' expression de mes gratitudes et de ma grande estime.

Ma chère Directrice TALEB LYNDA pour son amour inestimable, son encouragement et son aide.

BELLILI AMIRA

Dédicaces

Au nom du dieu clément et miséricordieux

Je dédie ce mémoire à :

Mes chères parents surtout : a ma très chère mère "que dieu ait son âme" et mon cher père qui n'a jamais cessé de m'encourager tous le long de mon parcours et qui a toujours tout sacrifié pour faire de moi ce que je suis à présent que dieu le protège.

Et bien sur sans oublier Mon frère Taki Eddine et ma petite soeur Maroua.
Tous mes amis et mes amies Ceux qui me sont proches et chers de loin ou de prés.

Veuillez accepter, l' expression de mes gratitudes et de ma grande estime.

BEN MADANI Khaoula

Résumé

Dans beaucoup de discussions autour des catastrophes et des situations d'urgence, on affirme couramment qu'une catastrophe fait émerger l'urgence, et le plus souvent, les deux termes sont utilisés de manière interchangeable.

Les réseaux sociaux sont considérés comme un moyen de communication fiable durant les situations d'urgence en raison de son omniprésence croissante, de sa rapidité de communication, et de son accessibilité multi-plateforme. De plus, les réseaux sociaux comme Twitter ont des caractéristiques spécifiques, comme l'existence des métadonnées telles que les hashtags. Ceci rend notre tâche plus complexe, malgré que ces hashtags ont été définis par Twitter dans le but de regrouper les tweets selon leurs sujets de discussion (Farzindar et Roche, 2013).

Beaucoup d'informations sont maintenant diffusées en cas d'une crise qu'il est impossible pour les humains de bien les trouver, et encore moins les organiser, leur donner un sens et agir en conséquence. Pour filtrer ces informations actionnables et situationnelles, des méthodes de filtrage et de classification doivent être développée et mis en œuvre pour augmenter les efforts humains de compréhension et d'intégration de données. Ces informations partagées doivent être transmises en temps réel aux bonnes personnes et/ou stocker pour des besoins ultérieurs.
La classification et catégorisation d'informations est l'activité du traitement automatique des langues naturelles qui consiste à classer de façon automatique des ressources documentaires, généralement en provenance d'un corpus.

Notre travail, réalisé dans ce mémoire, vise à proposer une solution technique pour améliorer la Situational Awareness dans les situations d'urgence. Cette solution proposée permet de l'utilisation des méthodes d'apprentissage automatique pour classer les tweets liés aux catastrophes afin d'extraire le maximum d'informations pertinentes et actionnables aux décideurs. Nous avons comparé les performances de deux des algorithmes de classification les plus courants, Naïve Bayes et les arbres de décision. L'évaluation de la performance est basée sur la validation des résultats à travers les paramètres d'exactitude, de précision, et le rappel, avec l'application d'outils statistiques.

Mots clés : Situation-awerness, Gestion de crise, Machine Learning, Data Mining, Médias Sociaux, Twitter , Classification d'information, Apprentissage automatique, TAL.

Abstract

In many discussions around disasters and emergencies, it is commonly said that disaster brings urgency, and most often both terms are used interchangeably.

In addition, social media like Twitter have specific characteristics, such as the existence of metadata such as hashtags. This makes our task more complex, despite the fact that these hashtags have been defined by Twitter in order to group tweets according to their topics of discussion (Farzindar and Roche, 2013).

A lot of information is now being released in the event of a crisis that is impossible for humans to find, let alone organize, make sense of and act on. To filter these situational and actionnable information, computational methods must be developed and implemented to increase the human effort of understanding and integrating data. This shared information must be transmitted in real time to the right people and / or stored for future needs.
The classification and categorization of information is the activity of the automatic processing of natural languages, which consists of automatically classifying documentary resources, usually from a corpus.

Our work in this manuscrit aims to provide a technical solution to improve Situational Awareness in emergency situations. This proposed solution allows the use of tweeting preprocessing methods as well as aiming to create a machine learning model to categorize disaster-related tweets to extract the most relevant and actionable information from decision makers. We have compared performance of two of the most common classification algorithms, Naïve Bayes and Decision Trees. The performance evaluation is based on the validation of the results through the parameters of accuracy, precision, error rate and confusion matrix, with the application of statistical tools.

Key words : Situation-awareness, Disaster Management, Machine Learning, Data Mining, Social Media, Twitter, Information Classification, Machine Learning, TAL

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

ONG	Organisation Non Gouvernementales.
SA	Situation-Awareness.
TIC	Technologies de l'Information et de Communication.
IE	Extraction d'Information.
IR	Recherche d'Information.
IF	Filtrage d'Information.
DM	Data Mining.
ML	Machine Learning.
NB	Naïve Bayes
AD	Arbre de Décision.
TAL	Traitement Automatique du Langage naturel.
OMS	Organisation mondiale de la Santé.
PAM	programme alimentaire mondial.
IASC	Comité permanent inter organisations.
GNC	Global Nutrition Cluster.
OIM	Organisation internationale pour les migrations.
ETC	Cluster Télécommunications d'Urgence.
ONU	Organisation des Nations Unies.
UNICEF	United Nations International Children's Emergency Fund.
WHO	World Health Organization.
WFP	World Food Programme.
FAO	Food and Agriculture Organization of the United Nations.

HCR	Haut Commissariat des Nations Unies pour les Réfugiés.
OCHA	Office for the Coordination of Humanitarian Affairs.
ITU	International Télécommunication Union.
NLTK	Natural Language Toolkit.
SAR	Search And Rescue

Introduction générale

Dans beaucoup de discussions autour des catastrophes et des situations d'urgence, on affirme couramment que la catastrophe fait émerger l'urgence, et le plus souvent, les deux termes sont utilisés de manière interchangeable.

Lors de situations de crise telles que les grandes séismes, les tempêtes ou autres types d'incidents, les gens peuvent avoir besoin de nourriture, d'abris et de soins médicaux, etc., ils rapportent et discutent de leurs observations, expériences et opinions dans leur réseaux sociaux. Par conséquent, des informations précieuses utilisées pour les services d'urgence et le grand public est disponible en ligne. Des études récentes montrent que les données des réseaux sociaux et particulièrement Twitter aide à détecter les incidents et leurs sujets ou d'analyser les flux d'informations générés par les personnes sur un sujet.

Donc les réseaux sociaux sont considérés comme un moyen de communication fiable durant les situations d'urgence en raison de son omniprésence croissante, de sa rapidité de communication et de son accessibilité multi-plateforme. Les interactions sur les médias sociaux étant hautement distribuées, décentralisées et se produisant en temps réel, fournissent l'information nécessaire dans les situations d'urgence.

Beaucoup d'informations sont maintenant diffusées en cas d'une crise qu'il est impossible pour les humains de bien les trouver, et encore moins l'organiser, lui donner un sens et agir en conséquence. Pour filtrer des informations utiles, les méthodes de calcul doivent être développée et mis en œuvre pour augmenter les efforts humains de compréhension et d'intégration de données. Ces informations partagées doivent être transmises en temps réel aux bonnes personnes et/ou stocker pour des besoins ultérieurs.

L'analyse sémantique des médias sociaux a ouvert la voie à l'analyse de données volumineuses, discipline émergente inspirée de l'analyse des réseaux sociaux, de l'apprentissage

automatique, de l'exploration de données, de la recherche documentaire, de la traduction automatique (Gotti et al. , 2014), du résumé automatique (Farzindar et Roche, 2015) et du TAL plus globalement.

Par exemple, Twitter qui nous intéresse particulièrement dans ce travail, constitue une source continue et illimitée de données en langage naturel qui est particulièrement difficile à traiter avec les approches classiques de traitement automatique du langage naturel (TAL). Ce type de langage est très éloigné des normes du langage traditionnel, avec ses conventions (telles que les hashtags, les mentions, les retweet, etc.). Son lexique particulier est souvent grossier et contient de abréviations, des émoticons, des acronymes. Sa syntaxe est parcellaire dans le meilleur des cas. Les données extraites de Twitter sont hautement bruitées, non-structurées, et courtes (comportant au maximum 140 caractères par tweet). La classification et catégorisation d'informations est l'activité du traitement automatique des langues naturelles qui consiste à classer de façon automatique des resources documentaires, généralement en provenance d'un corpus (Jaillet et al.,2003).

Dans le cas des tweets, la classification consiste à annoter les différentes phrases d'un tweet avec des classes (exemple : health, food, search&rescue, Education,etc.). Pour chaque classe C, on trouve des termes importants considérés comme des indicatifs pour la classe C (Liu, 2006). Par exemple, les termes dead, kill, bodies et casualtie sont des indicatifs du sujet health (santé). Cependant, les textes courts des tweets ne fournissent pas assez d'occurrences de mots. Ainsi, les méthodes de classification qui utilisent les approches traditionnelles telles que les Sacs de mots sont limitées, car les mots ne se répètent pas assez et génèrent des matrices creuses, ayant des tailles indéterminées. Pour pallier à ce problème, nous proposons l'utilisation des méthodes au prétraitement des tweets ainsi que nous visons à créer un modèle d'apprentissage automatique pour classer les tweets liés aux catastrophes et comparer les performances de deux des algorithmes de classification les plus courants, Naïve Bayes et les arbres de décision. L'évaluation de la performance est basée sur la validation des résultats à travers les paramètres d'exactitude, de précision, le taux d'erreur et la matrice de confusion, avec l'application d'outils statistiques.

Actuellement, il existe un grand intérêt académique et industriel pour le traitement automatique des langues naturelles, l'apprentissage machine, la traduction automatique ou l'extraction d'information telle que les entités nommées. La majorité de ces outils

s'appuient sur des corpus relativement structurés et sans bruits.

De plus, les réseaux sociaux comme Twitter ont des caractéristiques spécifiques, comme l'existence des métadonnées telles que les hashtags. Ceci rend notre tâche plus complexe, malgré que ces hashtag ont été définis par Twitter dans le but de regrouper les tweets selon leurs sujets de discussion (Farzindar et Roche, 2013).

Quelques travaux se sont concentrés sur l'amélioration de leur capacité à produire rapidement des informations analysables et adaptées aux situations d'urgence de masse. Ils introduisent une syntaxe normative basée sur le tweet qui pourrait augmenter l'utilité de l'information générée lors des situations d'urgence , et d'explorer et d'analyser les informations provenant des flux Web sociaux lors d'incidents tels que catastrophes naturelles, incendies ou autres types d'événements d'urgence.
Un deuxième problème qui est concentrés sur la localisation de la bonne information. En outre la diffusion des informations exploitables via Twitter pendant les urgences de masse, envoient également des informations générales qui sont inutiles. Ces informations doivent être filtrées préalablement et délivrées aux utilisateurs sous une forme actionnable, appropriée et personnalisée.

Dans ce travail de recherche, notre objectif est de remédier aux problématiques citées précédemment, et ceci en trouvant et développant une solution dédiée au filtrage et la classification des tweets pertinents à la compréhension de la situation d'urgence. On se base sur le Machine Learning et ses algorithmes afin d'extraire le maximum d'informations pertinentes et actionnables aux décideurs.

La structuration de notre mémoire s'articule autour des chapitres suivants :
Chapitre 1 : présente un état de l'art incluant les concepts de base de la gestion de crise et de la situation-awareness en situation d'urgence. Nous étudions les différentes phases du processus de gestion de crise en décrivant un scénario réel de crise et aussi nous définirons également la situation awareness.
Chapitre 2 : présente un état de l'art sur une synthèse des usages des nouvelles technologies de l'information et de la communication dans la situation awareness en gestion de crise et des nouveaux défis que cette dernière a posé dans le développement des solutions informatiques.

Chapitre 3 : détaille notre solution proposée. Nous présenterons le Machine Learning (ML) dans la gestion de crise avec ses différentes méthodes d'apprentissages supervisées et non supervisées. Et nous citerons quelques travaux connexes au notre.

Chapitre 4 : détaille l'implémentation de la solution proposée. Nous présenterons les outils et les langages de développement utilisés pour concevoir et réaliser cette dernière. Enfin, nous discuterons les résultats obtenus après une comparaison entre les différents algorithmes.

Chapitre 1

Gestion de crise et réponse

1.1 Introduction

« Crise » et « gestion de crise » sont des mots que nous entendons de plus en plus souvent à cause d'une part, du nombre croissant de catastrophes en tout genre (crises financières, catastrophes naturelles...) et d'autre part, de l'importante médiatisation de ces évènements.
Mais sait-on vraiment ce qui se cache derrière ces termes ? Qu'est-ce qu'une crise ? Comment se déroule une gestion de crise ? C'est ce que nous nous proposons d'expliquer dans ce chapitre afin de dresser le contexte de notre travail et de comprendre les tenants et aboutissants de notre étude concernant la gestion de la réponse aux crises.

Même si la notion de gestion de crise est passée dans le langage courant et semble être le quotidien de tout un chacun, elle n'est pas pour autant totalement maîtrisée. Lorsqu'un retour d'expérience est réalisé à la suite d'une crise, il n'est pas rare de constater que des dysfonctionnements ont eu lieu pendant la phase de réponse et que selon toute vraisemblance, il aurait été possible de faire mieux. Certains de ces dysfonctionnements peuvent avoir des conséquences dramatiques, alors même qu'une meilleure gestion aurait pu les limiter, voire les éviter. D'autres ont pu simplement générer des surcoûts plus ou moins critiques.

Ce chapitre est organisé en trois sections : la première décrit un scénario de crise. La deuxième présente gestion de crise et ses différentes phases et la troisième section sera consacrée à la situation-awareness en gestion de crise.

1.2 Définition d'une crise :

Le sens du mot crise a fortement évolué depuis son apparition et diffère d'un domaine à un autre.
Dans les années 1970, la crise est perçue comme « une situation qui menace les objectifs prioritaires des centres de décision, restreint le temps de prise de décision et dont l'occurrence surprend les responsables ». A partir de la fin des années 1980, la notion d'incertitude apparaît dans les définitions proposées. Une crise peut aussi être définie par un « phénomène complexe dynamique, qui constitue une menace pour la survie d'une organisation et de ses membres, qui laisse peu de temps de réaction, et qui entraîne un ajustement du système ». Cette définition, comme les précédentes, met en avant la nécessité de prendre des décisions dans l'urgence. Cependant, elle apporte un élément nouveau : l'aspect dynamique de la crise.[1]

Ces définitions sont plutôt applicables à des crises qui menacent la survie d'une entreprise, or les crises peuvent également menacer la vie de personnes. C'est le cas des crises résultant d'une catastrophe naturelle par exemple, ou des crises humanitaires de manière plus large. Les définitions suivantes couvrent cette acception.
Une crise humanitaire est « toute situation où il y a une menace exceptionnelle et de grande ampleur pour la vie, la santé ou la subsistance de base des individus et d'une communauté ». Van Wassenhove précise ces notions de menace et d'ampleur en décrivant la crise comme étant une évolution défavorable d'une calamité se situant à l'intersection de deux forces :

- la vulnérabilité (par exemple, une forte densité de population).
- un événement déclencheur (distinct des conditions qui créent la vulnérabilité), par exemple un séisme.

Une crise (ou une situation d'urgence, emergency en anglais) est un événement soudain et imprévu menaçant la sécurité d'une population, des propriétés ou de l'environnement et nécessitant à cet effet des actions d'intervention immédiates.[1]

D'après ces définitions on peut donc définir la crise comme un phénomène grave et imprévu menaçant la sécurité d'une population créé par un événement déclencheur, qui plonge le système de départ dans une situation instable, d'urgence et d'incertitude.[1]

1.3 Gestion de crise

La gestion de crise concerne l'ensemble des modes d'organisations, des techniques qui vont permettre à l'organisation de se préparer, de faire face et de tirer un enseignement de la crise afin d'améliorer les procédures et les structures.

1.3.1 Différentes phases de gestion d'une crise

Lors d'une crise, une organisation passe par plusieurs phases, chacune de ces phases possède ses propres actions pour sa résolution. La gestion de crise est conceptualisée par un modèle constitué de quatre phases interdépendantes correspondant au cycle de vie et impliquant des compétences différentes. Ces quatre phases sont : la mitigation, la préparation, la réponse et le rétablissement.

Mitigation (prévention)

Cette phase a pour objectifs de diminuer la probabilité d'apparition des risques liés à la crise et leurs conséquences s'ils surviennent. Si l'on prend le cas d'une catastrophe naturelle, il est possible, par exemple, de gérer les implantations dans les zones à risques ou encore améliorer la résistance des structures susceptibles d'être frappées par ce désastre.[1]

La préparation

Cette phase a lieu également avant une crise. Elle consiste notamment à établir de nouveaux processus de réponse adaptés aux futures crises. Par exemple, des exercices d'entraînements peuvent être réalisés ou encore les secours planifiés.[1]

La réponse

Elle regroupe toutes les actions à réaliser au plus vite après une crise, comme par exemple le déclenchement d'un plan d'opérations de secours ou l'évacuation des populations menacées. L'objectif principal à ce niveau est de mettre en place un ensemble d'actions qui agira sur le système en crise pour qu'il revienne au plus tôt à la normale.

Altay et Green expliquent que la réponse consiste à l'utilisation de ressources et de

procédures d'urgence pour préserver la vie, l'environnement et la structure sociale, politique et économique d'une communauté.

Le rétablissement

Cette étape a lieu une fois que la situation d'urgence a été prise en charge pour faire en sorte que le système perturbé retrouve son régime nominal. Les équipes d'intervention peuvent, par exemple, remettre en état des infrastructures ou prodiguer des soins aux populations déplacées.[1]

La figure 1.1 montre les différentes phase de gestion de crise.

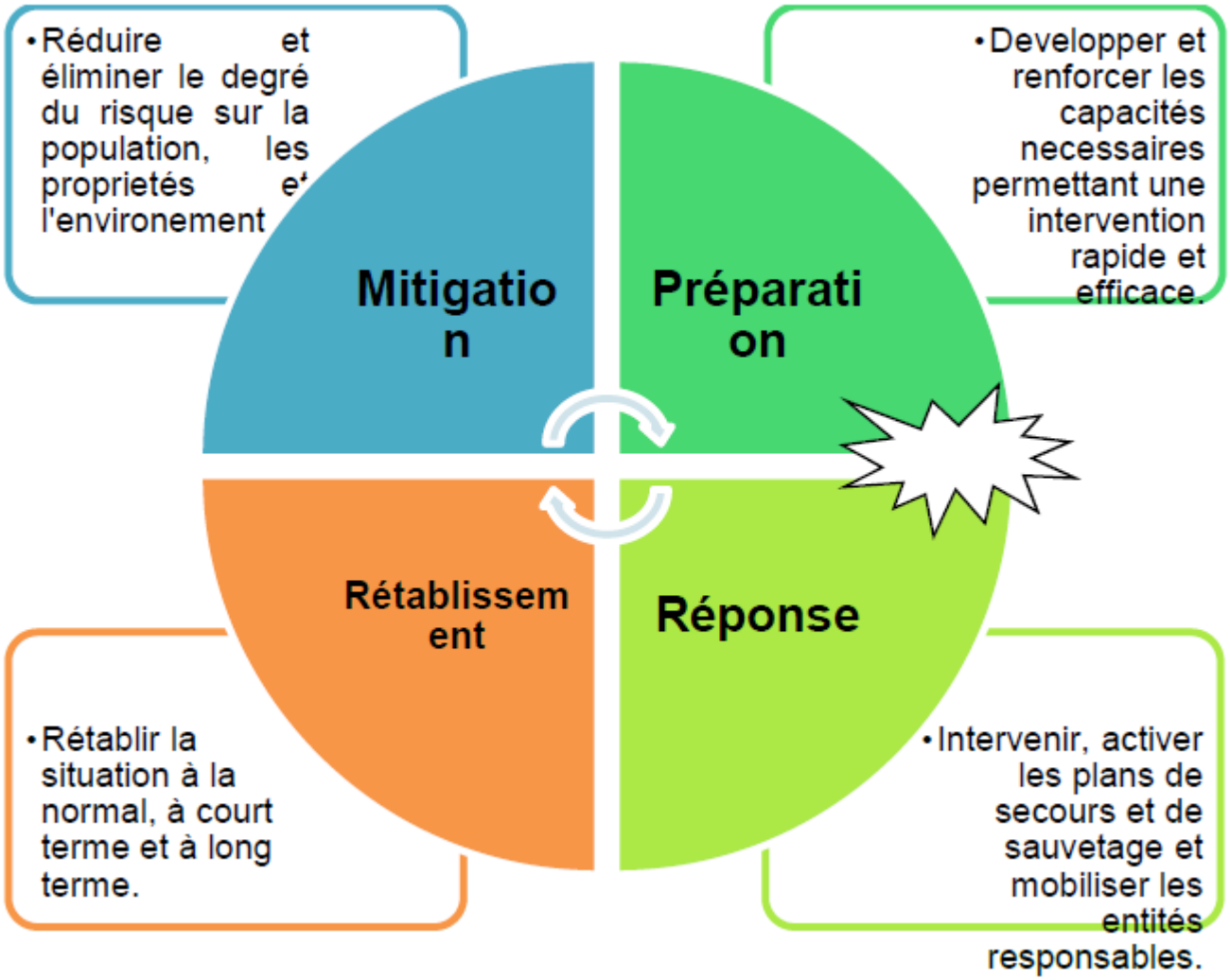

FIGURE 1.1 – Le cycle de gestion de crise.[2]

1.4 Le processus de réponse à une situation de crise

La réponse à une catastrophe est l'une des quatre phases principales de la gestion de crise. Le degré de mobilisation des entités et des acteurs responsables des opérations d'intervention et de secours dépend de l'ampleur de la catastrophe en question. En effet, la réponse face à une catastrophe à grande ampleur nécessite la participation des autorités publiques, des organisations gouvernementales, des organisations non gouvernementales (ONG), des associations caritatives, des médias, les entreprises privées et des citoyens. Toutes ces entités impliquées travaillent conjointement dans le but d'atteindre l'objectif commun, à savoir le sauvetage et l'apport de l'aide à la population et la réparation des biens et des infrastructures endommagés histoire de rétablir la situation à la normale. [2]

Les sociologues experts en la gestion de crise recommandent une organisation décentralisée des opérations de réponse qui favorise la coopération entre ces différents acteurs, tout en considérant le citoyen comme une ressource utile aux opérations. Les acteurs présents sur les lieux de l'incident transmettent aux décideurs les informations situationnelles recueillies décrivant l'état de l'environnement et des ressources pour qu'elles soient analysées et transformées par la suite en décisions d'intervention. Il y a une forte corrélation entre la précision, la temporalité et la fiabilité des informations transmises et la qualité des décisions prises.

1.4.1 Le cycle de réponse

Selon Mehrotra et al , les opérations de réponse et d'intervention peuvent être considérer comme constituées de quatre phases interdépendantes, et ce indépendamment de la nature et de l'ampleur de la crise et des entités impliquées dans ces opérations :

- **Evaluation des dommages** :Dans cette phase, toutes les pertes causées par la crise et leur degré de détérioration sont évalués. Les zones sévèrement touchées, les infrastructures affectées et les autres dégâts nécessitant une intervention prioritaire sont identifiés et le temps nécessaire pour leur réparation est estimé. [2]
- **Evaluation des besoins** :Dans cette phase, les besoins et les situations d'urgence qui requièrent un certain niveau d'intervention sont identifiés puis classés par ordre de priorité en fonction de leur dangerosité. [2]

— **Hiérarchisation des opérations de réponse** :Dans cette phase, chaque besoin et situation d'urgence précédemment identifié est associé avec les ressources d'intervention qu'il lui faut, selon son ordre de priorité déjà établi par les décideurs impliqués.

— **Organisation de la réponse** :Dans cette phase, les ressources d'intervention sont déployées et les décisions de réponse prises par les décideurs (correspondant aux plans d'urgence établi lors de la phase de préparation) sont transmises aux secouristes et autres acteurs impliqués présents sur place. [2]

Ce processus cyclique et continu est répété à chaque disponibilité de nouvelles informations situationnelles. Les besoins et les priorités sont par conséquent réévalués et les décisions révisées puis retransmises le plus rapidement possible.

1.5 Situation-awareness dans la gestion de crise

Pendant les urgences de masse, les populations affectées construisent une compréhension de la situation repose sur des informations incomplètes. Souvent, les victimes potentielles, les membres d'organismes d'intervention officiels et / ou étrangers concernés rassembler les informations disponibles avant de décider quelle action à prendre concernant une situation d'urgence. Ce processus de collecte d'informations ou d'évaluation situationnelle conduit à un état de la conscience situationnelle (SA). [3]

Situational Awareness est un état de savoir ce qui se passe dans votre environnement immédiat et de comprendre ce que cette information signifie pour une situation particulière, y compris la perception des éléments dans l'environnement et comment ces éléments sont liés les uns aux autres.

Atteindre Situational Awareness nécessite de comprendre "les objets dans la région de l'intérêt, "ainsi que" connaître les relations entre les objets qui sont pertinents pour une opération en cours ".

Le point de Endsley indique que situational awareness est un processus complexe qui nécessite la perception et la compréhension des éléments dans son environnement et conduit à prédictions de ce qui va se passer dans un proche avenir. Ce modèle à trois niveaux, représenté dans la figure 1.2 , amène à savoir ce qui se passe dans un environnement donné et comprendre ce qu'une information donnée veut dire dans une

situation donnée, incluant la perception des éléments constituants cet environnement et comment ces éléments sont reliés entre eux.

McGuinness et Foy ont étendu le modèle SA proposé par Endsley en ajoutant un quatrième niveau appelé résolution. Ce niveau fournit la conscience à propos du meilleur chemin à suivre et des meilleures actions à entreprendre pour résoudre les problèmes relatifs à une situation d'urgence.

L'acquisition de situational awareness implique la familiarisation avec les éléments de l'environnement et comprendre la signification de ces éléments. En d'autres termes, dans la situational awareness les gens vont saisir la signification de l'information qu'ils reçoivent. En cas d'urgence, le processus de connaissance des personnes peut inclure l'état d'un agent de danger, les dommages causés aux bâtiments et l'infrastructure, l'emplacement des centres d'évacuation, le nombre et l'emplacement des blessés et / ou animaux. Ces connaissances fournissent aux décideurs informations qui contribuent à la compréhension des situations d'urgence, et peuvent les aider à décider quelles actions prendre.[3]

La réponse à une catastrophe est considérée comme un processus dynamique complexe où les contraintes augmentent en temps réel. Les facteurs qui contribuent à cette complexité incluent : la surprise, la vitesse de développement, l'extension spatiale, le nombre des entités et acteurs d'urgence impliqués, l'incertitude, les écarts de perception, le manque de flexibilité dans la prise de décision, le manque des ressources disponibles, le manque des options d'intervention, incapacité de communiquer et les événements en cascade (l'effet domino).
Sous ces conditions, les participants aux opérations de réponse accumulent deux comportements principaux : un comportement basé-règle et un comportement basé-connaissance. Le premier comportement s'appuie sur les plans de secours et de sauvetage existants, élaborés lors de la phase de préparation et d'entrainement. Le deuxième comportement s'appuie sur les informations contextuelles, les connaissances tacites et l'expérience des secouristes.[3]

Dans leurs recherches sur le comportement humain en cas de catastrophe, Fritz et Marks (1954) déclarent : le problème immédiat dans une situation de catastrophe n'est ni incontrôlé comportement ou réaction émotionnelle intense, mais des déficiences de coordination et d'organisation, compliquées par des personnes agissant sur des définitions individuelles de la situation. Ce qui peut conduire à des résultats problématiques et appelle

à la nécessité d'avoir une commune Situation-Awareness chez les entités responsables et les populations affectées

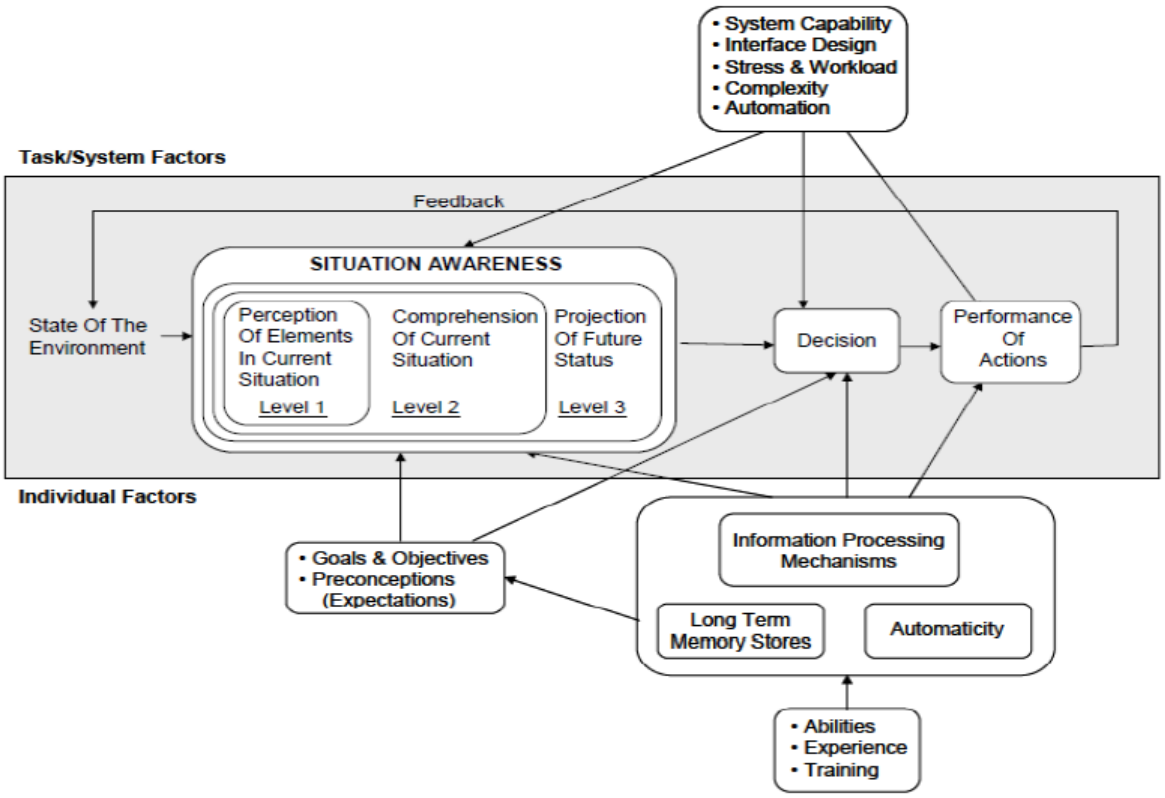

FIGURE 1.2 – Le modèle de Situation-Awareness proposé par Endsley (1995)

1.6 Conclusion

Dans ce chapitre, nous avons présenté les notions de base de la gestion de crise et de catastrophes, ainsi que leurs phases (la mitigation, la préparation, la réponse, et le rétablissement). Nous avons également détaillé la phase de réponse avec une présentation de son processus.

Nous avons abordé l'une de ces tâches essentielles permettant de mener à bien les efforts de secours et qui est la situation-awareness. Nous avons également listé les différents défis rencontrés dans le développement des solutions informatiques afin d'appuyer la SA dans les situations d'urgence.

Chapitre 2

TIC et réseaux sociaux dans gestion de crise

2.1 Introduction

Des débuts du télégraphe électrique, au XIXème siècle, jusqu'à l'avènement des connexions Internet par satellite, les technologies de la communication ont connu une suite ininterrompue d'innovations. Conséquemment, les pratiques de communication médiatisée se sont radicalement transformées. La communication de crise ne fait pas exception, et a vu son fonctionnement bouleversé par l'émergence de ces technologies ainsi que par les opportunités qu'elles offrent à leurs utilisateurs.

Au regard des situations d'urgence survenues, on constate que les citoyens soucieux de recueillir ou diffuser de l'information de manière rapide et directe utilisent de plus en plus les médias sociaux. Les autorités ont donc tout intérêt à y être présents, afin de partager également via ces canaux des informations officielles et correctes et de limiter de la sorte la propagation éventuelle de rumeurs.

Afin de mieux appréhender ces nouveaux médias, ce chapitre est organisé en une section qui parle sur les usages des nouvelles technologies de l'information et de la communication dans la situation-awareness en gestion de crise et des nouveaux défis que cette dernière a posé dans le développement des solutions informatique.

2.2 Définition des TIC

Les technologies de l'information et de communication (TIC) regroupent les techniques utilisées dans la transmission des informations, principalement de l'informatique, de l'in-

ternet et des télécommunications. Par extension, elles désignent leur secteur d'activité économique. Cette définition des TIC positionne cette industrie comme support de l'industrie du contenu numérique.[4]
Lorsque nous parlons aujourd'hui des technologies de l'information et de communication, il est particulièrement question d'un rapprochement entre les télécommunications (téléphone, radio, télévision) et l'information. C'est ce rattachement qui a donné naissance au Word Wide Web, c'est à dire au réseau internet que l'on pourrait qualifier de TIC la plus performante dans le sens ou elle réunit tous les supports multimédias en les mettant en réseaux.

Depuis quelques années, avec le développement d'internet, les usages des TIC se sont développés et la plupart des personnes utilisent ces outils pour accéder à l'information.
Un article de Claire Brossaud[5] , intitulé les usages des TIC et rapports à l'incertitude en situation de catastrophes naturelles, annonce que « les prospectivistes s'accordent à penser que les TIC devraient prendre une place croissante et pourraient être à l'origine d'un nouveau paradigme civilisationnel. » C'est pourquoi il est intéressant d'étudier leur influence en gestion de crise. Nous allons discerner leur utilité et la manière dont on peut s'en servir pour mener a bien une communication de crise en cas de problème majeur dans une collectivité.[4]

2.3 L'utilisation des TIC en gestion de crises

L'étude des technologies de l'information et de la communication s'inscrit notamment dans le champ de la sociologie des usages. La notion d'usage vise à décrire tout le spectre de ce que les hommes font avec un dispositif, en incluant les « pratiques déviantes par rapport au mode d'emploi, qui sont autre chose que des erreurs de manipulation». Ce concept reconnaît une certaine autonomie aux usagers, considérés comme capables d'inventer de nouvelles manières d'utiliser les outils dont ils disposent. Ainsi, la sociologie des usages ne s'intéresse pas tant à « ce que les médias font aux individus » qu'à « ce que les individus font des médias » , c'est à dire, à l'influence des usagers des TIC sur le développement et l'évolution des pratiques associées à ces technologies.[6]

Dans le cas des TIC, on constate une grande variété d'usages, certains prévus par les créateurs de l'outil, tandis que d'autres relèvent de pratiques d'appropriation, de braconnage et de détournement à l'initiative des usagers eux-mêmes . Ces nouveaux usages, loin d'être complètement arbitraires, sont guidés par les besoins, l'environnement et les connaissances des usagers, mais aussi par les possibilités techniques et les affordances (ou incitations) de l'outil. Ils sont souvent adossés à des pratiques pré-existantes, adaptées d'un autre outil ou empruntés à une technologie antérieure.[6]

L'usage des TIC a permis l'amélioration de la coordination en temps critique, la collaboration inter et intra-organisationnelle et joue le rôle crucial de médiateur des informations situationnelles entre les multiples acteurs impliqués . De plus, pour Li, Li, Liu, Khan, and Ghani, ces technologies sont utilisées pour : a) alerter efficacement en utilisant plusieurs canaux de communication ; b) intégrer les informations situationnelles depuis des sources hétérogènes ; c) coordonner les différentes opérations d'intervention ; d) encourager les interventions sociales, institutionnelles et publiques ; e) évaluer les dommages causés par la crise.[2]

2.4 Les différentes sources de données au moment de la phase de réponse

Les données sur les catastrophes sont extrêmement hétérogènes, à la fois structurellement et sémantiquement, ce qui crée un besoin d'intégration et d'ingestion de données afin d'aider les responsables de la gestion des situations d'urgence à se rétablir rapidement en cas de catastrophe. Les données sur les catastrophes à collecter et à intégrer pour l'analyse et la gestion peuvent être :

a) Des plans d'action en situation d'urgence.

b) Des rapports situationnels en continu.

c) Des rapports d'analyse des dommages.

d) Des données et des cartes géographiques de la zone touchée.

e) Des informations sur l'état des routes/ponts/aéroports et sur d'autres infrastructures comme l'électricité, le carburant, les hôpitaux, les écoles, etc.

f) Des informations logistiques sur les livraisons de nourriture/eau/médicaments.

g) Des données financières pour gérer les donations.

h) Des images satellitaires de la zone touche après la crise.

En plus de l'unicité du contenu, les données de gestion des catastrophes ont également des caractéristiques temporelles / spatiales différentes et peuvent être classées en trois types différents : Données spatiales ; Données temporelles ; et données spatio-temporelles.

L'analyse de ces données implique l'application de technologies de l'information bien étudiées à ce domaine unique. Les technologies d'analyse de données que nous examinerons dans des situations liées aux catastrophes sont les suivantes :

- **L'extraction d'information (IE) :** Les données situationnelles doivent être extraites depuis des sources hétérogènes et stockées sous un format structuré commun qui permet leur traitement.[7]
- **La recherche d'information (IR) :** Les utilisateurs doivent être en mesure de rechercher et d'accéder aux informations situationnelles pertinentes. Ces besoins sont exprimés en utilisant des requêtes appropriées.[7]
- **Le filtrage d'information (IF) :** Dès que les informations situationnelles arrivent depuis les producteurs, elles doivent être filtrées et redirigées aux consommateurs adéquats.[7]
- **Le data mining (DM) :** Les données situationnelles collectées doivent être analysées pour extraire des modèles et des tendances intéressants.[7]
- **L'aide à la décision :** L'analyse des données situationnelles aide à la prise de décision.[7]

La figure 2.1 montre comment l'extraction, la recherche et le filtrage d'information et le data mining s'inscrivent dans le processus de circulation d'information en gestion de crise, adapté depuis le schéma proposé par Hristidis et al.

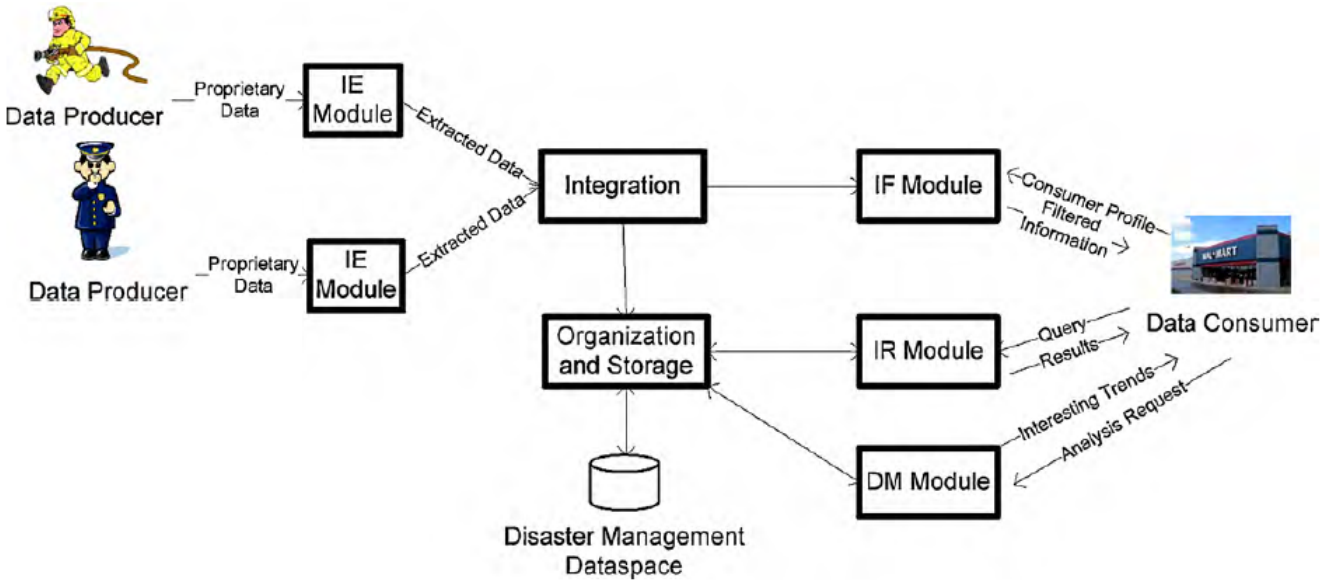

FIGURE 2.1 – Flux de données dans la gestion des catastrophes.

2.5 Qu'est-ce que les médias sociaux ?

Médias Sociaux est une dénomination commune pour les applications internet utilisées pour partager l'information et les opinions. Ce sont les utilisateurs des médias sociaux qui assurent la création du contenu. Celui ci est composé de textes, fragments sonores ou d'images. Le contenu n'est pas seulement créé et utilisé par les autres utilisateurs, mais il est aussi partagé et valorisé en interaction et dialogue.[8]

Communiquer via les médias sociaux se base sur des principes et une philosophie propre :

— Communication en temps réel.

— Communication transparente.

— Communication participative.

Au regard de la planification d'urgence et de la gestion de crise, les médias sociaux peuvent aider à :

— Informer rapidement, par exemple, dans les situations d'urgence, ou démentir les rumeurs .

— Offrir une perspective d'action rapide et améliorer l'autoprotection ;

— Atteindre un public plus large ou groupe cible très spécifique ;
— Faciliter l'interaction avec une participation du citoyen ;
— Être un canal complémentaire aux médias traditionnels ;
— Obtenir rapidement du feedback et rassembler des informations sur la situation, tant pour la communication que pour la gestion.

Qu'il s'agisse d'informer, de réagir ou de rassurer, les médias sociaux contribuent à une communication rapide et ciblée.

2.6 Médias sociaux pendant les situations de crise

L'adoption croissante des médias sociaux pendant les catastrophes a créé des opportunités pour la propagation d'information qui n'existerait pas autrement. Dans la réponse d'urgence les agences publient régulièrement des informations telles que des alertes d'urgence et des conseils par le biais de ces canaux, mais les médias sociaux permettent bien plus que des communications «top-down». Les gens publient des informations sensibles à la situation sur les médias sociaux en rapport avec ce qu'ils vivent, témoignent et / ou entendent d'autres sources. Cette pratique permet à la fois aux populations touchées et à celles qui se trouvent en dehors de la zone d'impact de se renseigner directement et en temps quasi réel sur la situation.[9]

Nous savons que les informations publiées sur les plateformes de médias sociaux dans le temps et la sécurité circonstances peuvent être d'une grande valeur pour ceux qui sont chargés de prendre des décisions dans ces situations tendues. Des recherches antérieures ont montré que l'information qui contribue à SA est signalée via Twitter (et d'autres plateformes de médias sociaux) lors d'urgences de masse.[9]

Maintenant, ceux qui sont chargés des efforts de réponse formelle des services d'incendie locaux aux agences d'aide internationales travaillent à intégrer les informations diffusées sur les plateformes de médias sociaux dans leurs processus et procédures. De nombreux intervenants d'urgence et les responsables humanitaires reconnaissent la valeur de l'information affichée sur les médias sociaux plates-formes par des membres du public (et d'autres), et sont intéressés à trouver des moyens de trouver et organiser rapidement

et facilement l'information qui leur est la plus utile.[9]

2.7 Communication de crise via les médias sociaux

Aujourd'hui, à travers les smartphones, leurs caméras, les réseaux sans-fil et les logiciels de messagerie instantanée, les situations de crises peuvent être documentées et commentées dans le monde entier quelques minutes après leur occurrence. Ce nouveau potentiel en termes d'instantanéité et de portée de la communication de crise est fortement lié aux caractéristiques des systèmes informatiques qui sous-tendent ces échanges, mais aussi des pratiques et usages qui se sont développés parallèlement à l'émergence de ces technologies.

La démocratisation des médias sociaux et des smartphones a également favorisé l'apparition d'un « journalisme citoyen » qui, grâce à sa multitude de participants, devance de plus en plus fréquemment les médias traditionnels pour couvrir les catastrophes. Par exemple, lors du crash d'un avion dans l'Hudson River en 2009, l'information fut publiée en premier lieu sur Twitter, accompagnée d'une photographie prise à l'aide d'un smartphone. (voir figure2.2).

Cet usage s'est rapidement étendu à d'autres plateformes de médias sociaux et à des situations de communication plus complexes que la seule diffusion d'alerte. Ainsi, en 2007, après la fusillade de Virginia Tech, les étudiants et leurs proches ont exploité Facebook et d'autres applications de messagerie pour identifier et recenser les victimes. Le site de partage d'image Flickr a également été utilisé pour documenter les conséquences de différentes situations de crise . Lors d'inondations et d'incendies, les usagers de Twitter ont partagé des informations relatives à la localisation du danger, aux conditions météorologiques ou encore aux procédures d'évacuation, contribuant à la SA et facilitant la gestion de la crise . Après le séisme et le tsunami de la côte Pacifique du Tohoku (11 mars 2011), de nombreuses vidéos ont été publiées sur Youtube, non seulement pour témoigner des destructions mais aussi pour apporter du réconfort et du soutien aux personnes affectées. [6]

FIGURE 2.2 – Premier tweet mentionnant le crash dans l'Hudson River

2.8 Utilisation de Twitter pour atteindre la SA

Twitter est un service de microblogging populaire qui permet aux utilisateurs d'envoyer messages de 140 caractères, connus sous le nom de «tweets», pour tout le monde. Des recherches récentes que les utilisateurs convergent sur Twitter en cas d'urgence de masse .

Twitter est un moyen par lequel les gens partagent des informations qui contribuent à SA pendant les périodes d'urgence massive. Sur Twitter, les gens peuvent trouver des informations telles que des mises à jour sur les dommages matériels, des rapports de blessures, des zones en cours d'évacuation, et sites d'hébergement. De plus, les gens peuvent trouver des informations sur où diriger les dons monétaires et matériels, et lire les messages de soutien.

Les gens connaissent les catastrophes différemment, et peuvent nécessiter des informations des différents types pour mieux gérer leur propre situation. L'objectif est que les personnes confrontées à une situation d'urgence massive prennent de bonnes décisions en temps opportun avec le plus d'informations possible. Les informations communiquées via Twitter peuvent aider à atteindre cet objectif.
Additionally et Mileti considère la réponse collective en cas d'urgence de masse et présente

une stratégie en cinq étapes que les gens connaissent lorsqu'ils font face à des dangers :

— Évaluer la vulnérabilité des dangers.

— Examiner les ajustements possibles.

— Déterminer la perception humaine et l'estimation du danger.

— Analyser le processus de prise de décision.

— Identifier les meilleurs ajustements, compte tenu des contraintes sociales, et évaluer leur efficacité.

Ces étapes offrent une compréhension de haut niveau de la façon dont les gens font face à l'immédiat après des événements d'urgence de masse, et nous voyons ces mêmes comportements d'adaptation exposé dans les communications Twitter au cours de masse d'urgence récente .

Nous savons que les gens utilisent Twitter pour offrir des informations, poser des questions et demander aide lors de situations d'urgence massives. De nouvelles questions se posent quant les services de micro-blogging changent ou augmentent le comportement humain en cas de catastrophe situations, et comment l'évaluation situationnelle et l'état final de la situation la sensibilisation change avec l'utilisation de Twitter.[3]
Un exemple d'utilisation de Twitter, après les attentats du marathon de Boston (2013), la police a utilisé Twitter pour demander aux internautes de fournir toutes les images susceptibles d'aider à identifier les terroristes (voir Figure 2.3).

[htbp]

FIGURE 2.3 – Usages de Twitter par les autorités durant des situations de crise (attentat de Boston)

2.9 Conclusion

Nous avons présenté au cours de ce chapitre l'usage des technologies de l'information et de la communication, et plus particulièrement les médias sociaux et Twitter, permettant de contribuer à la mise en place d'une communication de crise. Ce nouveau paradigme de partage d'information suit parfaitement les caractéristiques de la nouvelle informatique dite pervasive.

Dans le chapitre suivant, nous présenterons le Machine Learning (ML) dans la gestion de crise, ses méthodes d'apprentissages ainsi que les défis rencontrés lors du développement d'une solution technologique pour appuyer la SA. Pour finir, nous présenterons notre solution.

Chapitre 3

Approche et solution

3.1 Introduction

Dans le premier chapitre, nous avons discuté et mis en avant les caractéristiques principales à prendre en compte afin d'appuyer la Situational Awareness dans la gestion de catastrophe. Notre objectif est de fournir une solution technique pour améliorer la Situational Awareness dans les situations d'urgence. Dans ce chapitre, nous présentons notre solution en utilisant les techniques du Machine Learning.

D'abord, nous présenterons le Machine Learning (ML) dans la gestion de crise, ses méthodes d'apprentissages supervisées et non supervisées. Ensuite, nous citerons quelques travaux connexes au notre. Enfin, nous détaillerons notre solution proposée.

3.2 Data Mining et Machine Learning

Le Machine Learning - apprentissage automatique - est une branche de l'informatique, plus précisément de l'intelligence artificielle, qui est concernée par le développement de méthodes et d'algorithmes qui apprennent des caractéristiques et des modèles à partir des données disponibles afin de faire des prédictions. Ainsi, l'objectif principal du Machine Learning est de développer des méthodes qui peuvent construire des modèles qui décrivent des données (et de préférence mécanismes) fidèlement. Des exemples pratiques d'application du Machine Learning sont les filtres anti-spam pour les e-mails : ces outils logiciels intégrés aux serveurs de messagerie permettent d'identifier automatiquement les

spams e-mails avec une grande précision. La «logique» derrière ces filtres est apprise automatiquement en analysant le contenu des e-mails et les comportements des utilisateurs.[10]

Frawley et al définissent l'extraction de connaissances à partir de données (ou Data Mining) comme l'extraction non triviale d'informations implicites, précédemment inconnues et potentiellement utiles à partir de données. Le Data Mining est un concept plus large que le Machine Learning. En effet, il est plus préoccupé par les processus analytiques qui conduisent à la découverte de nouvelles connaissances à partir de grands ensembles de données. Il est fortement lié à la gestion et au traitement de données à grande échelle contenues dans les bases de données, les data warehouse, etc. Contrairement au Machine Learning, le Data Mining met l'accent sur les pratiques de la gestion de données et des résultats fournis. Le Data Mining devrait nourrir les entreprises et les systèmes de renseignement et d'aide à la décision avec des informations qui devraient être utile pour les décideurs.

Il y a des chevauchements évidents entre Data Mining et Machine Learning, mais il y a aussi des différences clés. Les chevauchements incluent : les algorithmes partagés et les méthodes utilisées (principalement issues de la statistique). Tous deux s'appuient sur des données et tentent d'en tirer des conclusions[10]. Les différences comprennent :

— ML est plus préoccupé par le processus de découverte des connaissances (algorithmes qui analysent les données), de préférence détaché des caractéristiques données, tandis que DM se concentre sur l'extraction et l'exploitation des connaissances utiles à partir des données disponibles.
— Le but du ML est davantage axé sur la reproduction des connaissances existantes (afin de valider l'utilité et l'exactitude des algorithmes proposés), tandis que le DM doit découvrir et exploiter de nouvelles connaissances à partir de l'ensemble de données fourni.
— Les sources de données : fournir une infrastructure de données efficace sous forme de bases de données et d'algorithmes de récupération efficaces fait partie du DM, tandis que ML suppose généralement des datasets comme entrée des algorithmes (bien que certaines structures de données spécifiques à un algorithme puissent être un partie de ML).

— DM nécessite de grandes quantités de données : l'infrastructure des grandes bases de données est la préoccupation du DM, tandis que le ML peut se concentrer sur des méthodes et des algorithmes explicitement conçus pour les données disponibles limitées.

Le DM et le ML sont des outils reconnus pour soutenir la prise de décision dans de nombreux domaines, notamment les banques, les assurances, l'aérospatiale, et la défense. Exemples d'applications : moteurs de recherche, bioinformatique et séquençage de l'ADN, filtres anti-spam, systèmes de recommandation en ligne, reconnaissance faciale pour les applications de sécurité, détection de la fraude et bien d'autres.[10]

3.3 Le Machine Learning dans la gestion de crises

Comme indiqué précédemment, la gestion des catastrophes et des crises pose un certain nombre de problèmes. La disponibilité de données utiles et complètes est l'un de ces défis. Mais même si les gros volumes de données étaient disponibles, il y aurait d'autres problèmes. Si l'on considère une zone géographique (qui serait sujette à une catastrophe) comme un système dans un état "normal", il y a tellement de façons différentes que ce système peut être sujet à une catastrophe, et il est impossible d'avoir des données raisonnablement complètes sur chacun d'entre eux (il n'est même pas possible d'exprimer tous les scénarios possibles pour les catastrophes)[10].

Actuellement, le concept largement accepté du cycle de gestion des catastrophes Waugh et Tierney divise le processus de gestion des catastrophes en plusieurs phases distinctes, bien qu'il n'y ait pas de consensus clair sur la division en phases :

3.3.1 Mitigation (atténuation)

La phase mitigation du cycle de gestion des catastrophes vise à réduire le risque d'occurrence de la catastrophe et ses conséquences possibles. Les exemples les plus connus de DM et de ML pour l'atténuation des catastrophes sont probablement la prévention des différentes menaces posées par les catastrophes d'origine humaine. Les exemples incluent : la détection des menaces terroristes par l'analyse des réseaux informatiques , les réseaux sociaux , la fusion des données des capteurs pour la détection des menaces nucléaires,

la reconnaissance faciale, etc. Le ML peut être combiné avec des données statiques pour surveiller les conditions changeantes et leur impact sur les caractéristiques statiques de la communauté. Cela renforcerait la capacité de hiérarchiser les actions préventives afin d'éviter un incident[10].

3.3.2 Préparation (Preparedness)

L'un des problèmes les plus difficiles dans la phase de préparation est la planification de l'évacuation . La planification de l'évacuation implique le besoin de combiner les données spatiales et de capturer les comportements des évacués . La recherche principale se concentre sur l'utilisation des méthodes de DM pour identifier les menaces potentielles et les zones de sécurité. Le DM pourrait également être utile pour surveiller les sources d'information, comme les sites Web sur la sécurité publique ; déterminer quelle information est recherchée à la fois comme indicateur de la sensibilisation du public et des ressources qui pourraient être nécessaires en cas d'événement . DM peut également être utilisé pour estimer les conditions changeantes. Yang et al ont montré comment les applications DM permettent d'estimer les changements d'intensité des cyclones tropicaux. Cela pourrait être utile à l'élaboration de plans de préparation[10].

Le ML peut également trouver une application dans les systèmes d'alerte précoce , à la fois pour les catastrophes naturelles et celles causées par l'homme. Les exemples incluent les systèmes d'alerte précoce pour les menaces chimiques et nucléaires, les inondations, les tsunamis et plus encore. Conceptuellement, les systèmes d'alerte précoce reposent sur la détection des anomalies - le manque de données sur les catastrophes est atténué par le fait que tout ce qui n'est pas normal peut être qualifié de menace potentielle et peut être soumis à un examen humain. Une autre méthode, qui a récemment suscité beaucoup d'intérêt, est l'examen humain ou la vérification potentielle de la réalité. L'impact de l'alerte précoce est la récolte de sources de données de réseaux sociaux telles que Facebook et Twitter.

3.3.3 Réponse (Response)

Les progrès récents dans les appareils mobiles capables de communiquer sans fil et dotés d'une puissance de calcul importante ont attiré l'attention des chercheurs et des praticiens . Leur rôle pendant les opérations de réponse peut inclure la communication

(vocale et numérique), ils peuvent servir de capteurs automatisés (généralement équipés de GPS, de détecteurs de mouvement, etc.) et sont capables de produire des images et des vidéos d'une qualité relativement élevée. Ils peuvent être utilisés pour améliorer la connaissance de la situation en glanant des informations provenant de sites pouvant produire des résultats précis . Cependant, si ces capacités de génération de données sont utilisées, les problèmes de fourniture d'informations répétitives et de surcharge d'informations se posent. Des techniques DM pour gérer le contenu et la quantité d'informations présentées aux utilisateurs ont été proposées .Les problèmes de confidentialité associés à l'utilisation des appareils mobiles dans le contexte de DM et une proposition de la solution à ce problème basée sur la collecte de données agrégées, plutôt que liée à des utilisateurs individuels ont été discutés par Meilleur .

L'utilisation de robots dans les opérations de recherche et de sauvetage est une technologie intelligente qui gagne en reconnaissance dans les interventions en cas de catastrophe. ML est étroitement liée à la robotique - l'intelligence des robots autonomes provient généralement d'algorithmes ML. La force des algorithmes ML est exploitée dans la cartographie de nouveaux environnements qui ont été créés dans le résultat de la catastrophe (par exemple des tas de décombres). Les Robots de sauvetage fournissent un exemple clair de la façon dont les algorithmes qui ont la capacité d'apprendre de nouvelles connaissances à partir des données peuvent être appliqués avec succès dans les situations réelles.

3.3.4 Récupération (Recovery)

Cette phase semble recevoir beaucoup moins d'attention de la part des communautés ML et DM, mais elle mérite une attention supplémentaire. Le DM peut être utilisé pour coordonner les différents ordres des décideurs, comme l'a demandé le Groupe de travail sur la reconstruction de l'ouragan Sandy, en recueillant des données auprès de divers organismes .

Mais il existe des possibilités de soutien DM et ML. Le DM peut être utile grâce à l'intégration de médias sociaux qui peuvent être complétés par l'utilisation de données, notamment des photos géolocalisées des dommages transmis par les systèmes sans fil tels que les téléphones intelligents, puis affichées sur des sites Web tels que Facebook et Flickr

. Cette information fournirait non seulement des données spatiales à partir desquelles des données textuelles associées pourraient être extraites, mais si elle était augmentée par l'analyse de la vidéo pour déterminer la gravité de l'impact, elle pourrait contribuer à l'élaboration de plans de rétablissement.

Le DM peut également être utilisé pour aider à estimer les effets sur la récupération économique des zones touchées , tandis que le ML peut être utilisé pour déterminer les stratégies optimales de gestion des débris. Le DM et le ML peuvent tous deux être utilisés dans la planification du rétablissement post-événement pour créer une communauté plus résiliente aux catastrophes [10].

3.4 Les méthodes d'apprentissage automatique

Le but de l'apprentissage machine ("Machine Learning") est d'étudier et d'entrainer des algorithmes afin qu'ils puissent apprendre à apprendre et ainsi être en mesure de faire des prédictions sur une large quantité de données.

Dans l'apprentissage automatique, la classification est l'ensemble des catégories (sous-populations) auxquelles une nouvelle observation appartient, sur la base d'un ensemble de données contenant des observations (ou des instances) dont l'appartenance à une catégorie est connue.

On peut grouper les méthodes classificatoires en deux grandes familles. Cette fois-ci, on prend en considération l'intervention ou non d'un « attribut classe » au fur et à mesure du processus de la classification. Ces deux types sont :
a)supervisé : groupes fixés, exemples d'objets de chaque groupe.
b)non supervisé : on ne connaît pas les groupes au préalable.

3.4.1 Les méthodes de classification supervisées

À cause de la grande quantité de documents échangés et stockés sur les supports électroniques, la classification automatique supervisée est devenue plus que nécessaire pour faciliter l'utilisation et l'analyse des données. À la différence de la classification non supervisée, où l'ordinateur doit trouver automatiquement les classes, la classification su-

pervisée se base principalement sur le fait qu'il existe déjà une classification de documents, c'est-à-dire qu'on dispose d'un ensemble de données déjà classées qu'on appelle «ensemble d'apprentissage» et qu'on l'utilise comme base, pour classer le reste des données. On essaie dans ce type de classification de trouver le maximum d'informations à partir des ensembles d'apprentissage, pour permettre un meilleur groupement des données restant[11].

Parmi les méthodes de classification supervisées les plus populaires, on peut citer par exemple :

Les arbres de décision

Les arbres de décision sont considérés parmi les méthodes les plus populaires pour la classification textuelle. Parmi les algorithmes les plus connus, on peut citer ID3 et C4.5.

Le fonctionnement des arbres de décision se base principalement sur des exemples. En effet, si on veut classer des documents dans des catégories, on doit construire un arbre de décision par catégorie. D'une manière générale, chaque nœud de l'arbre de décision exécute un test If-Then-Else et les feuilles de l'arbre ont les valeurs de décision Oui ou Non. Les tests exécutés, observent les valeurs des attributs de chaque exemple. Pour un texte quelconque par exemple, l'attribut peut être un mot avec une valeur de 0 ou 1 selon que ce mot appartient à ce texte ou non[12].

L'algorithme de l'arbre de décision J48 :

Cet algorithme classifie une instance à l'aide d'un ensemble de valeurs de seuil pour une séquence déterminée de caractéristiques. Lors de la phase d'apprentissage l'algorithme évalue la valeur d'entropie associée à chaque caractéristique. Il choisit en priorité la caractéristique restante ayant la plus petite entropie comme nouveau nœud. Chaque caractéristique est ainsi utilisée pour prendre une décision dans le processus de classification. Lorsqu'on arrive à la fin d'une branche de l'arbre, l'algorithme détecte que toutes les instances font partie de la même classe, il est donc impossible de former un nouveau nœud. Suite à l'entrainement, l'arbre binaire est conservé et chaque nouvelle instance peut être instantanément classifiée suite à quelques décisions binaires sur les seuils[13].

La classification Naïve de Bayes

Les méthodes Naïve Bayes sont considérées parmi les modèles probabilistes les plus connus. Elles se basent principalement sur le théorème de Bayes (Bayes, 1963).
Les algorithmes Naïves Bayes sont souvent utilisés dans la catégorisation et la classification de documents. Ils permettent d'estimer la probabilité de chaque classe parmi les exemples, étant donné un document, et affectent à ce dernier la classe la plus probable. On appelle ce procédé «Prior probabilities»[12]. Pour classer un ensemble de documents, Naïve Bayes utilise comme entrée les mots qui se trouvent dans ces derniers, ensuite il calcule la fréquence de chaque mot dans les différents documents classés dans une classe donnée.

3.4.2 Les méthodes de classification non supervisées

Les méthodes de classification non supervisées se basent principalement sur la séparation automatique des nuages de points dans un espace, sans le besoin de fournir des données d'apprentissage. Dans ce genre de méthodes, le nombre de classes est fixé au préalable par l'utilisateur.
En général, le fonctionnement des algorithmes de classification non supervisée consiste à partitionner un ensemble d'objets en k sous-ensembles, où k représente le nombre de regroupements attendus par l'utilisateur. Il existe plusieurs stratégies qui permettent de trouver ces regroupements, comme les méthodes se basant sur les densités, sur le partitionnement, de même que des méthodes hiérarchiques et celles utilisant la quantification. Parmi les algorithmes de classifications non supervisée les plus connus, on peut citer par exemple : K-moyen, Single-pass, Suffix tree clustering, Hierarchical Agglomerative Clustering, les cartes auto organisatrices de Kohonen, ART, etc.

3.5 Travaux connexes

Nous discuterons dans cette partie les travaux connexes à notre sujet en les organisant dans des approches qui partagent des traits communs (c.-à-d. propriétés en termes de tweet, en utilisant un sujet probabiliste modélisation et clustering incrémental).

3.5.1 Twicident

C'est un système qui permet aux utilisateurs d'explorer et d'analyser les informations provenant des flux Web sociaux lors d'incidents tels que catastrophes naturelles, incendies ou autres types d'événements d'urgence.(voir figure 3.1)
Son architecture repose sur les modules suivants :

- La fonctionnalité principale est déclenchée par un module de détection d'incident qui détecte les incidents diffusés par les services d'urgence. [14]
- Chaque fois qu'un incident est détecté, Twitcident lance un nouveau fil de discussion sur l'incident et regroupe les médias sociaux et les messages Twitter sur le Web.[14]
- Les messages collectés sont ensuite traités par le module d'enrichissement sémantique qui comprend la reconnaissance d'entité nommée (NER), la classification des messages, le couplage de messages à des ressources Web externes et l'extraction de métadonnées supplémentaires. [14]
- L'enrichissement sémantique est l'un des composants clé du Framework Twitcident. Il se divise en trois sous-composants : Incident Pro-filing qui supporte le filtrage sémantique des messages Twitter pour identifier les tweets pertinents pour un incident donné ; Faceted Search qui permet une recherche à facettes sur les médias filtrés ; et Semantic Enrichment qui donne des moyens pour résumer les informations sur les incidents et fournir des analyses en temps réel. [14]
- Détection d'incidents : Pour détecter les incidents, le système Twitcident s'appuie sur des services de diffusion d'urgence. Les incidents sont immédiatement publiés via le réseau de communication P2000 et décrivent le type d'incident qui a eu lieu, où, et quand cela s'est-il produit et à quelle échelle l'incident est-il classé.[14]

3.5.2 Classification des publications et extraction des informations à partir de messages de micro-blog

Imran et al ont proposé un système automatique d'extraction d'informations liées à une catastrophe qui nécessite deux composantes : Classification des tweets et Extraction à partir des tweets. Premièrement, les messages générés lors d'une catastrophe ayant une valeur très variable, le système doit filtrer les messages qui ne contribuent pas à la SA.

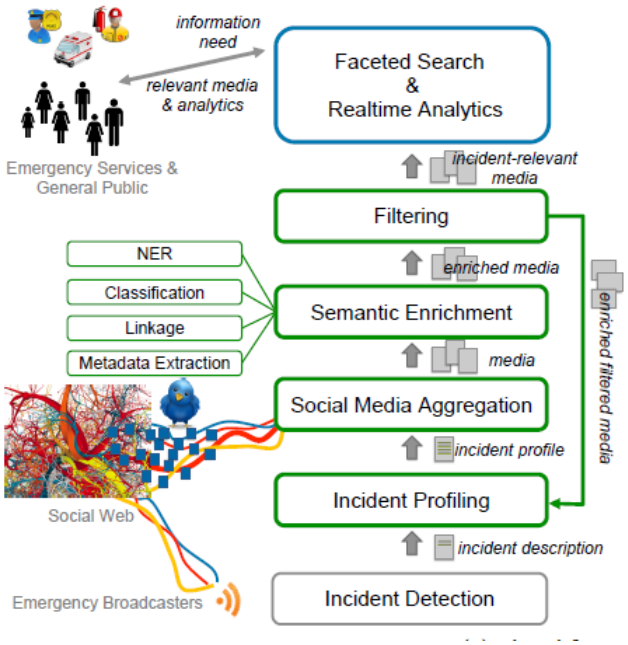

FIGURE 3.1 – Architecture de Twitcident.

Celles-ci incluent celles qui sont de nature personnelle et celles qui ne sont pas pertinentes pour la catastrophe. Par conséquent, ils conçoivent un système de détection des messages informatifs.

Une fois que système a détecté des tweets susceptibles de contenir des informations pertinentes, il doit analyser les tweets candidats pour déterminer le type d'informations à extraire (par exemple, les offres de dons, les rapports de pertes). Le résultat final du système est constitué d'informations brèves et autonomes susceptibles d'accroître la compréhension de la situation.[15]

Classification manuelle et extraction avec crowdsourcing :

- **Tâche 1 :Filtrage des messages informatifs :**La première tâche correspond à une annotation de tweets selon qu'ils sont entièrement de nature personnelle, informative (”directe ”,”indirecte ”, ou ”directe ou indirecte ”), ou autre[15].
- **Tâche 2 :Classer les messages par type :**examiner attentivement un tweet individuel et attribuer une étiquette appropriée parmi les catégories données.[15]
- **Tâche 3 - Classer les messages par sous-type et extraire des informations.**

3.5.3 Tweak the Tweet :

Dans l'article de Starbird,k. et Stamberger,J, les auteurs proposent une solution de basse technologie à utiliser par les microblogueurs sur Twitter, qui pourrait améliorer leur capacité à produire rapidement des informations analysables et adaptées aux situations de crise dans des situations d'urgence de masse. Ils introduisent une syntaxe normative basée sur le tweet qui pourrait augmenter l'utilité de l'information générée lors des situations d'urgence en remodelant doucement la pratique comportementale actuelle. Cette offre repose sur une compréhension des tendances actuelles en matière d'évolution des normes d'utilisation de Twitter, une évolution qui a progressé rapidement mais semble se stabiliser autour de conventions textuelles spécifiques[16].

3.5.4 Les travaux de Vieweg,S et al.

Dans Les travaux de Vieweg,S et al,les auteurs proposent une approche permettant de localiser automatiquement les informations susceptibles de contribuer à la situational awareness dans la multitude de tweets diffusés dans les situations de crise. Leurs objectif primordial est d'aider les populations touchées à collecter et à analyser les informations pertinentes communiquées par les décideurs et la population. Leurs hypothèses est que l'abattage immédiat et dynamique des tweets avec des informations relatives à la connaissance de la situation pourrait être utilisé pour informer et mettre à jour les applications visant à aider les personnes touchées. En utilisant des techniques de TAL et d'apprentissage automatique (ML), les auteurs ont développé une suite de classifieurs pour différencier les tweets sur plusieurs dimensions : la subjectivité, le style personnel ou impersonnel et le registre linguistique (style formel ou informel). Sur la base d'analyses initiales du contenu de tweet, ils postulent que les tweets qui contribuent à la compréhension de la situation sont susceptibles d'être écrits dans un style objectif, impersonnel et formel ; par conséquent, l'identification de la subjectivité, du style personnel et du registre formel pourrait fournir des fonctionnalités utiles pour extraire des tweets contenant des informations pertinentes et actionnables[17].

D'une manière générale, les différents systèmes présentés ci-haut n'offrent aucune solution efficace pour la classification d'information permettant de délivrer aux décideurs et aux secouristes des informations situationnelles personnalisées et pertinentes à leurs

besoins informationnels courants.

Par conséquent, ceci requiert la mise en œuvre de stratégies efficaces en termes de classification des informations brèves et autonomes susceptibles d'accroître la compréhension de la situation et de détecter des informations pertinentes pour la gestion de crises.

3.6 Twitter

Twitter est un outil de microblogging géré par l'entreprise Twitter Inc. Il permet à un utilisateur d'envoyer gratuitement de brefs messages, appelés tweets, sur Internet, par messagerie instantanée ou par SMS. Et comme la taille d'un SMS ne dépasse pas 160 caractères, Twitter a limité la taille d'un tweet à 140 caractères dont 20 caractères réservés au nom de l'expéditeur . Depuis novembre dernier, Twitter a abandonné la limite des 140 caractères pour passer à 280. Twitter fournit notamment une API gratuite pour différents objectifs et pour recueillir les données twitter[18].

3.6.1 Les concepts Twitter

Différents concepts sont définis dans Twitter :

— **Utilisateur :**Un nom précédé d'arobase « @ » et est un lien direct vers un compte twitter. L'utilisateur de ce nom a la permission de voir tous ses tweets, sauf s'ils sont protégés[18].

Les informations suivantes sont stockées pour chaque utilisateur : La langue du tweet, le fuseau horaire de l'emplacement, l'emplacement du Tweet (l'emplacement à partir duquel le tweet a été envoyé). La photo du profil, l'emplacement de l'utilisation, la page web, une brève biographie, et les liens favoris.

— **Tweet :**Un tweet est un message court, limité à 280 caractères. Cette restriction impose aux utilisateurs d'être concis dans ce qu'ils ont à dire. Ceci est également la raison pour laquelle les utilisateurs ont tendance à utiliser les abréviations (par exemple : «fr»-for, «cud»- could). Chose intéressante, est qu'il y a un ensemble riche et bien compris d'abréviations qui est étonnamment cohérent à travers les groupes d'utilisateurs, et même à travers d'autres supports électroniques tels que les SMS et les forumes de discussions.

Comme les utilisateurs veulent transmettre tout ce qu'ils ont à dire en 280 caractères, ils pourraient faire des erreurs d'orthographe et des tweets peuvent être sujet à des erreurs syntaxiques. Cela rend difficile le travail avec Twitter. La plupart du temps, les utilisateurs fournissent des liens vers des ressources externes quand ils ne peuvent pas transmettre l'information complète dans les 280 caractères. Ces liens URL vers des fichiers texte, audio ou vidéo sont appelés artéfacts.[18]

— **Re-tweets :** Si un tweet est convaincant et assez intéressant, les utilisateurs peuvent le republier. Il devient ce qu'on appelle «re-tweeting». Un retweet est similaire au renvoi par courriel. Lorsqu'un utilisateur envoie un re-tweet il est considéré comme ayant approuvé ce contenu et partage son contenu avec ses followers.[18]

— **Les hashtag :**Un hashtag commence toujours par le caractère « # » ; ce qui permet de le repérer très rapidement dans l'analyse des données (lors de la phase de tokenisation). Ces hashtags créent des problèmes durant l'analyse linguistique du texte. En effet, ils sont considérés comme des mots inconnus qui ne se trouvent pas nécessairement dans les dictionnaires. Twitter permet aux utilisateurs d'étiqueter leurs tweets en utilisant les balises de hashtag. Les hashtags aident Twitter à regrouper ensemble les tweets similaires qui ont les mêmes balises de hashtag. Cela rend la recherche sur twitter plus facile et plus rapide.[18]

3.7 Solution proposée

Notre solution consiste à trouver une façon d'exploiter les tweets afin de les classifier selon des classes pertinentes liée à la SA. Pour réaliser notre solution nous avons suivi les étapes suivantes :

— La collecte d'un jeu de données depuis Twitter.

— Prétraitement des tweets.

— Entrainement, évaluation, et construction de modèles de classification en utilisant les algorithmes du Machine Learning expliqués précédemment.

Les modèles de classification construits permettent de résoudre le problème du gros volume d'informations disponibles sur les réseaux sociaux, plus précisément dans Twitter dans notre cas. Les tweets seront ainsi extraits puis classifiés dans des classes afin de facilité leur accès et leur transmission.

3.7.1 Détails sur le jeu de données et sa collecte

La collecte est une étape qui consiste à obtenir les données tweets sur lesquelles nous allons tester notre solution. Nous présentons dans cette section l'approche globale que nous avons suivi, les contrôles de données que nous avons réalisés, les stratégies spécifiques de collecte appliquées, ainsi qu'un échantillon du dataset obtenu.
Pour notre étude préliminaire, nous avons recueilli un échantillon aléatoire d'environ 110 tweets diffusés pendant chacun des événements d'urgence contenant l'un de nos mots clés.
Le dataset se compose de tweets incluant l'un des termes de recherche suivants :
#Typhoon#Lionrock, #Flooding#China#Pakistan, # ItalyEarthquake , #Ecuador_earthquak, # HurricaneMatthew, # EarthquakeMexico, #Hurricane Harvey, #flooding#india, # Hurrica Irma, #tsunami#america, #Vietnam#typhoon, #California#wildfires, #Iran #earthquake, #Bali volcano , #Southern #California #wildfires, #Fuego#Eruption#Guatemala, #ITU.

Dans nos expérimentations, nous avons analysée les tweets qui portent sur la SA. Notre but est d'extraire les attributs, et qui ont été classé dans des groupes sectoriels de l'approche Cluster..

L'approche Cluster a pour objectif de préciser la répartition des tâches entre toutes les organisations humanitaires à l'œuvre dans les différents secteurs, en mieux définissant leurs rôles et leurs responsabilités. Douze groupes sectoriels (« clusters ») identifiés au niveau global selon les responsabilités professionnelles. Il s'agit par exemple du secteur des télécommunications, de la santé, de la sécurité, etc.

Tweets	Kill	Support	Devastate	Evacuate	Rebuild	Classes
#Typhoon #Lionrock slams #Japan, electricity down, at least 11 killed,170000evacuated https ://www.facebook .com/RTvids/videos/ 1296991063644477/	yes	no	no	no	no	health
#DPRK : The @UNhumanrights expert @tojeaquintana calls for increased support for the victims of #Typhoon #Lionrock http ://bit.ly/2cG2XCY	no	yes	no	no	no	donation
RT@LinkTV :#Flooding devastates #Pakistan,northwest #China ; casualties increase in #Afghanistan(Video) http ://bit.ly/bYpszU	no	no	yes	no	no	rescuesearch
#Flooding kills hundreds ;millions evacuated in #China, #Pakistan,&#India	no	no	no	yes	no	protection
#Haiti needs help, esp after #HurricaneMatthew. Every dollar helps rebuild. Join us in supporting http ://aidstillrequired.org/haiti2017/ #AidStillRequired	no	no	no	no	yes	shelter

TABLE 3.1 – Echantillons du jeu de données collectés.

3.7.2 Prétraitement des tweets

Nous nous basons sur une architecture simple du prétraitement du tweets pour réduire le temps de réponse, cette architecture est composé de cinq tâche : Tokenisation, Segmentation des hashtags, Lemmatisation, Détection des mots vides. Chaque une s'exécute dans une phase séparée.

— **Tokenisation :** Dans l'analyse lexicale, la tokenisation est le processus de séparation d'un flux de texte en mots, phrases, symboles et d'autres éléments significatifs appelés jetons ou tokens, les jetons peuvent être des mots individuels, des phrases ou même des phrases entières.[18]

— **Suppression des liens et de ponctuation :** Dans cette tâche nous avons supprimé les liens hypertext et la ponctuation comme les virgules et point d'exclamation, qui sont inutile de les traité ou de les utiliser pour classifier les tweets.[18]

— **Segmentation des hashtags :** Un hashtag commence toujours par le caractère « # » ; ce qui permet de le repérer très rapidement dans l'analyse des données (lors de la tokenisation). Ces hashtags créent des problèmes durant l'analyse linguistique. En effet, ils sont considérés comme des mots inconnus et ne se trouvent pas dans les dictionnaires, car les hashtags sont généralement des mots composés inventé par les utilisateurs de Twitter et leur sémantique particulière se perd dans le traitement des textes. Cette tâche va nous aider à extraire le plus de mots possibles à partir des hashtags et le faire relier avec les autres mots récupérés du tweets.[18]

— **Lemmatisation :**La lemmatisation est l'étape qui désigne l'analyse lexicale chargée de faire regrouper les mots d'une même famille qui partagent le même suffixe lexical. Chacun des mots du texte se trouve ainsi réduit en une entité appelée « Lemme ». Ce lemme désigne la forme canonique des mots. La lemmatisation regroupe les différentes formes que peut avoir un mot. Par exemple, un nom en pluriel va être réduit au singulier, un verbe à son infinitif, etc. La lemmatisation aide à regrouper les mots et les faire représenter avec l es lemmes dans le but de réduire la dimension d l'espace des mots. Par conséquence, si les mots partageant un lemme on les considère comme un mot unique.[18]

— **Suppression des mots vide :** Les mots vides (ou stop words) sont des mots qui sont tellement communs qu'il est inutile de les traiter ou de les utiliser dans une recherche d'informations. En Anglais, certains de ces mots sont « the », « is

», « far», etc. Un mot vide est un mot non significatif figurant dans un texte. La signification d'un mot s'évalue à partir de sa distribution (au sens statistique) dans une collection de textes. Un mot dont la distribution est uniforme sur les textes de la collection est dit « vide » et ne permets pas de distinguer les textes les uns par rapport aux autres. En d'autres termes, un mot qui apparaît avec une fréquence semblable dans chacun des textes de la collection n'est pas discriminant, car il ne permet pas de distinguer les textes les uns par rapport aux autres. D'autre part, certains mots grammaticaux sont assez rares pour constituer des mots pleins. La collection des mots vides utilisés dans la classification des tweets est la même collection utilisée dans la recherche d'informations. Elle a pour le but de filtrer les tweets et d'extraire juste les mots pertinents afin de discriminer ces tweets par les mots qu'ils représentent.[18]

3.7.3 Classification des tweets

La classification des tweets consiste à annoter les différentes phrases d'un tweet avec des classes. Pour chaque classe C, on trouve des termes importants considérés comme des indicatifs pour la classe C. Cependant, les textes courts des tweets ne fournissent pas assez d'occurrences de mots. Ainsi, les méthodes de classification qui utilisent les approches traditionnelles telles que les Sacs de mots sont limitées, car les mots ne se répètent pas assez et génèrent des matrices creuses, ayant des tailles indéterminées. Pour pallier à ce problème, nous proposons l'utilisation des méthodes du Machine Learning.

Classes et Attributs

Dans un cas typique d'apprentissage supervisé, les données sont représentées par une table d'exemples ou d'instances. Chaque instance est décrite par un nombre fixe d'attributs dont un label qui dénote la classe à laquelle il appartient.
Le tableau ci-après montre les caractéristiques des bases choisies

	nombre d'instances	nombre d'attributs	Type d'attributs	nombre de classes
Data-set	165	69	Nominal	12

TABLE 3.2 – les caractéristiques de data-set

Les classes

Nous avons utilisés douze classes (les groupes sectoriels de l'approche Cluster) dans la classification des tweets. Cette approche Cluster vise à renforcer la réponse par la prévisibilité, la responsabilité et le partenariat en assurant une meilleure hiérarchisation et en définissant les rôles et responsabilités des organisations humanitaires.

- **Santé**
- **Logistique-transport :**Le cluster logistique assure la coordination et la gestion de l'information pour appuyer la prise de décisions opérationnelles et améliorer la prévisibilité, la rapidité et l'efficacité de l'intervention humanitaire d'urgence. Le cas échéant, le cluster logistique facilite également l'accès aux services logistiques communs. Grâce à son expertise dans le domaine de la logistique humanitaire, le Programme alimentaire mondial a été choisi par l'IASC en tant qu'organisme chef de file du Cluster logistique.[19]
- **Search And Rescue-SAR**
- **Donation**
- **Nourriture**
- **Nutrition :**Le Global Nutrition Cluster (GNC) a été créé en 2006 dans le cadre du processus de réforme humanitaire, qui visait à améliorer l'efficacité des programmes d'intervention humanitaire en assurant une plus grande prévisibilité, une plus grande responsabilité et un meilleur partenariat. La vision du GNC est de sauvegarder et d'améliorer l'état nutritionnel des populations touchées par des situations d'urgence en assurant une réponse appropriée, prévisible, opportune, efficace et à grande échelle. [19]
- **Coordination :**Co-dirigée par l' Organisation internationale pour les migrations (OIM) et des Nations Unies pour l' Agence pour les réfugiés (HCR) pour les catastrophes naturelles et situations de déplacement interne résultant de conflits , respectivement, le Coordination Cluster bénéficie également de l'adhésion de nombreux organismes partenaires, ainsi comme appui de partenaires académiques et du secteur privé. Au niveau des pays, la coordination constituent un mécanisme humanitaire essentiel car ils facilitent la protection et l'assistance aux populations déplacées.[19]
- **Protection-Sécurité :**Le Global Protection Cluster coordonne et fournit des

conseils et des orientations stratégiques interinstitutions au niveau mondial sur la mise en œuvre de l'approche sectorielle des groupes de protection sur le terrain, appuie les mesures de protection dans les protection dans les situations d'urgence humanitaire complexes et liées aux catastrophes naturelles, en particulier en ce qui concerne la protection des personnes déplacées.[19]

— **Telecom/ICT :**Le Cluster Télécommunications d'Urgence (ETC) est un réseau mondial d'organisations qui travaillent ensemble pour fournir des services de communication partagés dans les situations d'urgence humanitaire. Les services de technologies de l'information et de la communication (TIC) fournis en temps opportun, prévisibles et efficaces par l'ETC ont été améliorés : Réponse et coordination entre les organisations humanitaires, Environnement de sécurité opérationnel pour le personnel et les biens, et Prise de décision grâce à un accès rapide aux informations critiques.[19]

— **Water**

— **Shelter :**Shelter Cluster est un mécanisme de coordination du Comité permanent interorganisations (IASC) qui aide les personnes touchées par des catastrophes naturelles et des personnes déplacées affectées par un conflit à vivre dans un abri sûr, digne et approprié. Il permet une meilleure coordination entre tous les acteurs du logement, y compris les gouvernements locaux et nationaux, de sorte que les personnes qui ont besoin d'une assistance en matière de logement reçoivent de l'aide plus rapidement et reçoivent le type d'aide approprié.[19]

— **Education** :Le Cluster Education est un forum formel ouvert pour la coordination et la collaboration en matière d'éducation dans les crises humanitaires. Le Cluster Education rassemble des ONG, des agences de l'ONU, des universitaires et d'autres partenaires dans le but commun d'assurer une éducation prévisible, bien coordonnée et équitable aux populations touchées par les crises humanitaires. Établi en 2007 par l'IASC dans le cadre de l'approche Cluster, le Cluster Éducation s'efforce de défendre l'éducation en tant que droit humain fondamental et composante essentielle de la réponse humanitaire. [19]

Classes (Secteur ou domaine d'activité)	Chef de file sectoriel
Santé	WHO / OMS
Logistique/Transport	PAM / WFP
SAR	-
Nourriture	FAO/WFP
Nutrition	UNICEF
Coordination	HCR (déplacés internes victimes de conflits) ; OIM (déplacés internes victimes de catastrophes naturelles)
Protection/Sécurity	HCR (déplacés internes victimes de conflits) HCR/OHCHR/UNICEF (déplacés internes victimes de catastrophes naturelles et civiles autres que déplacés internes victimes de conflits)
Telecom/ICT	OCHA /PAM /UNICEF
Shelter	HCR (déplacés internes victimes de conflits) ; FICR (animateur) (déplacés internes victimes de catastrophes naturelles)
Education	UNICEF& Save the Children
Donnation	UNICEF
Water/Sanitation	UNICEF

TABLE 3.3 – Liste de classes et leur Chef de file sectoriel.

Les attributs

Nous avons retenu 69 attributs. Les attributs sont regroupés selon les classes présentés ci-haut. Le tableau 3.4 énumère quelques exemples d'attributs pour chacune des classes :

La classe	Les attributs
Santé	Kill,bodies,victim,dead,vaccins, casualtie,patient,medical, vets,injurd...
Logistique/Transport	Boat,trucks,ships,helicopters,deliver, airline....
search/rescue	Evacuate,damage,devastate ,need,missing,affecte,find,Destroy, relief, help,emergency,,search,rescue, humanitarian _assistance,assistance
Nourriture	Food,WFP...
Nutrition	Evacuate,crops
Coordination	Tent,EU..
Protection/Security	Evacuate,armed_forcs ,UNHCR, protect
Telecom/ICT	ITU...
Water	Water ,clean_water
shelter	Rebuild,shelter,repair,shelter_materials, reconstruction, Home,homeless,building
Education	Book,school,student....
donnation	Support,volunteer,donate,UNICEF, job,distribute,supplie,tarps, Blankets,equipement,shipments.....

TABLE 3.4 – Examples d'attributs pour chacune des classes

Après la construction de notre dataset nous divisons généralement nos données en deux sous-ensembles : ensemble d'entrainement et ensemble de test et adaptons notre modèle aux données de train afin de pouvoir effectuer des prédictions sur les données de test. Lorsque nous faisons cela, une chose peut se produire : nous sur-utilisons (overfit) notre modèle. Nous ne souhaitons pas que cette chose se produise, car elle affecte la prévisibilité de notre modèle.

Overfiting

L'overfitting signifie que le modèle que nous avons formé s'est entraîné «trop bien» et qu'il s'adapte désormais bien à la base de données d'entrainement. Cela se produit généralement lorsque le modèle est trop complexe (c'est-à-dire trop de caractéristiques / variables par rapport au nombre d'observations). Ce modèle sera très précis sur les données d'entrainement mais sera probablement très imprécis sur des données non entraînées ou nouvelles. C'est parce que ce modèle n'est pas généralisé, ce qui signifie que vous pouvez généraliser les résultats et ne pas faire d'inférence sur d'autres données, ce qui est en fin de compte ce que vous essayez de faire.[20]
Pour éviter l'overfitting nous avons utilisée la méthode de validation croisée (cross validation) comme solution à ce problème.

Validation croisée

La validation croisée est une procédure de ré échantillonnage utilisée pour évaluer les modèles d'apprentissage automatique sur un échantillon de données limité.

Cela signifie que nous divisons nos données en k différents sous-ensembles (ou plis). Nous utilisons des sous-ensembles k-1 pour former nos données et laisser le dernier sous-ensemble (ou le dernier pli) comme données de test. Nous calculons ensuite la moyenne du modèle par rapport à chacun des plis, puis finalisons notre modèle. Après cela, nous le testons par rapport à l'ensemble de test.[20]

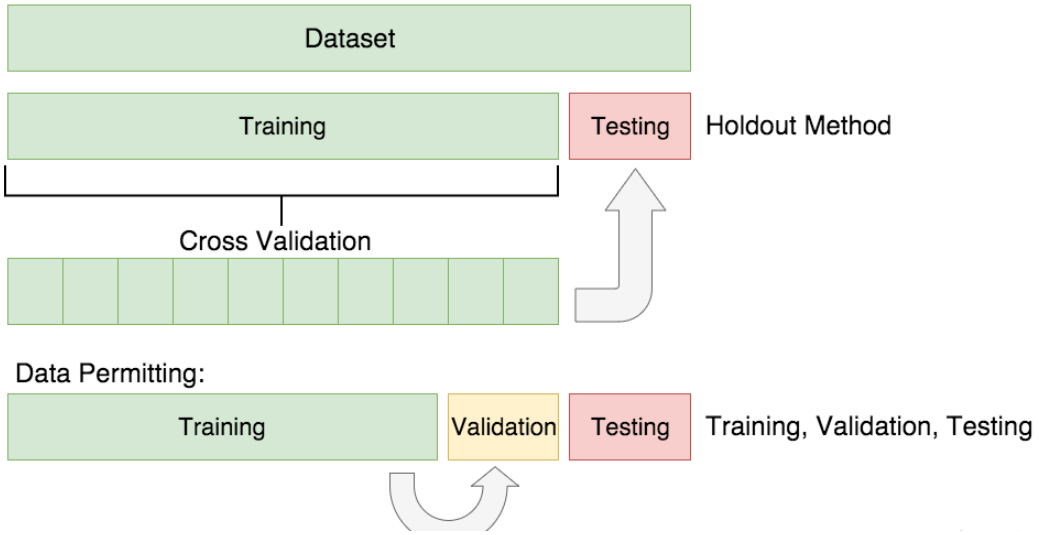

FIGURE 3.2 – Cross Validation

3.7.4 Méthode de classification

La résolution des problèmes de la classification à grande échelle est cruciale dans de nombreuses applications telles que la classification de texte. Dans notre travail, nous avons expérimenté avec deux méthodes de classification, Naïve Bayes (NB) et Arbres de décisions (AD) (voir 3.4). Deux classifieurs ont donc été construits.

Modèle de classifieur

— **Classifieur Arbre de Decision**

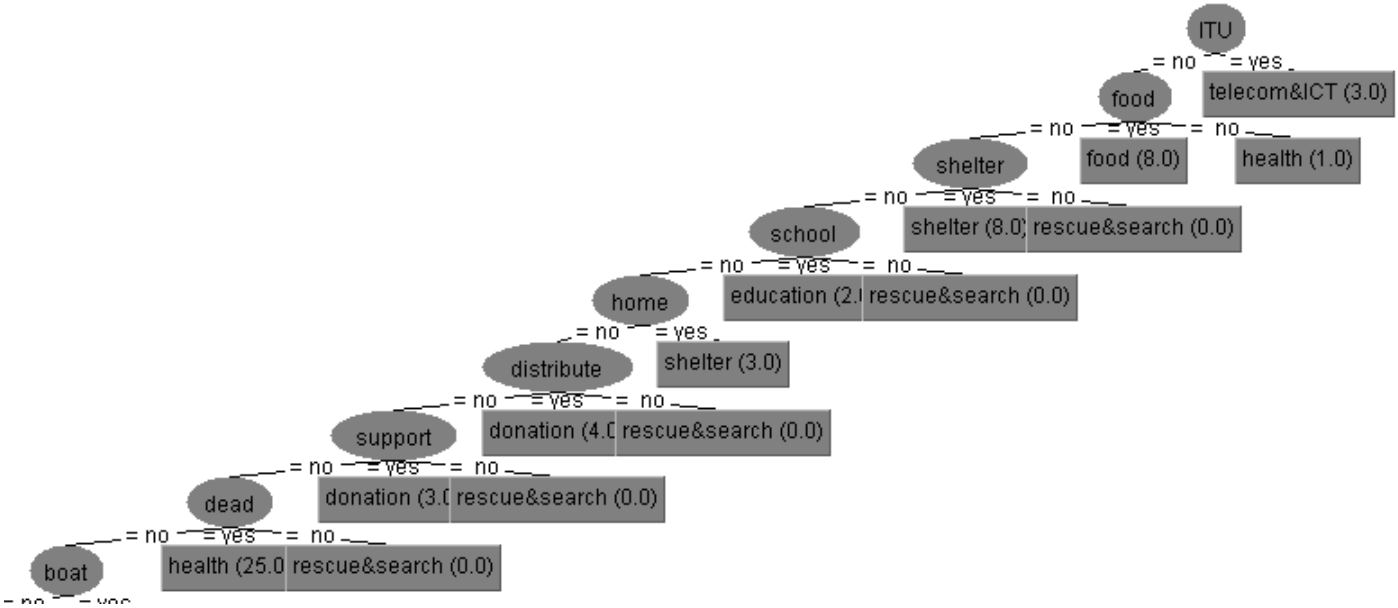

FIGURE 3.3 – Modele de classifieur Arbre de Decision

— **Classifieur Naive Bayes**

Naive Bayes Classifier

Attribute	Class health (0.27)	rescue&search (0.28)	protection (0.04)	donation (0.08)
dead				
no	23.0	50.0	8.0	15.0
yes	26.0	1.0	1.0	1.0
no	2.0	1.0	1.0	1.0
[total]	51.0	52.0	10.0	17.0
medical				
no	47.0	50.0	8.0	15.0
yes	2.0	1.0	1.0	1.0
no	2.0	1.0	1.0	1.0
[total]	51.0	52.0	10.0	17.0
destroy				
no	48.0	47.0	8.0	15.0
yes	1.0	3.0	1.0	1.0
no	2.0	1.0	1.0	1.0
[total]	51.0	51.0	10.0	17.0
affecte				
no	48.0	41.0	8.0	15.0
yes	1.0	10.0	1.0	1.0
no	2.0	1.0	1.0	1.0
[total]	51.0	52.0	10.0	17.0

FIGURE 3.4 – Modele de classificateur Naive Bayes

3.7.5 Architecture de la solution proposée

la figure 3.5 résume tous les étapes de notre solution

FIGURE 3.5 – Architecture de solution

3.8 Conclusion

Nous avons détaillé dans ce chapitre notre solution. Nous avons introduit un processus de collection des tweets qui génère le modèle en se basant sur un détecteur de langue anglaise afin de minimiser le bruit dans les tweets. Notre filtrage basé sur la catégorie

grammaticale a aidé l'amélioration de la précision et de la qualité de la classification.

Dans le chapitre suivant, nous discuterons les résultats obtenus après la classification afin de choisir la meilleure méthode de classification.

Chapitre 4

Résultats et Discussion

4.1 Introduction

Dans le chapitre 4, nous présenterons l'évaluation de notre solution proposée, à savoir, le filtrage des tweets permettant de contribuer à la SA dans la gestion de crise. Cette solution repose sur la classification et le Machine Learning permettant de déduire un classifieur. L'évaluation de ce classifieur consiste à le valider et à choisir la méthode de classification qui donne le meilleur résultat de précision avec un taux d'erreur minimal.
Dans ce chapitre, nous présentons les outils et les langages utilisés pour implémenter notre solution. Ensuite, nous montrons les résultats obtenus après une comparaison entre les algorithmes de classification.

4.2 Langages d'implémentation

La solution proposée dans notre travail est réalisés avec le langage de programmation Python. Ce langage de programmation présente de nombreuses caractéristiques intéressantes. Il est multiplateforme, c'est-à-dire qu'il fonctionne sur de nombreux systèmes d'exploitation : Windows, Mac OS, Linux, Android, iOS, depuis les mini-ordinateurs Raspberry Pi jusqu'aux supercalculateurs, il est gratuit. C'est un langage de haut niveau. Il est interprété et orienté objet. Enfin, il est très utilisé en bioinformatique et plus généralement en analyse de données.[21]
Nous avons utilisé l'éditeur PyCharm Editor, et pour appliquer les algorithmes d'apprentissage automatique aux profils d'utilisateur, nous avons utilisé Natural Language NLTK

et scikit-learn disponible sous forme de package pour Python.

— **NLTK :**Natural Language Toolkit (NLTK) est une boîte-à-outil permettant la création de programmes pour l'analyse de texte. Cet ensemble a été créé à l'origine par Steven Bird et Edward Loper, en relation avec des cours de linguistique informatique à l'Université de Pennsylvanie en 2001. Il existe un manuel d'apprentissage pour cet ensemble titré Natural Language Processing with Python (en anglais).[22]

— **Scikit-learn :**Librairie Python pour effectuer de l'apprentissage automatique. Il inclut la plupart des méthodes de classification.[23]

4.3 Évaluation des algorithmes d'apprentissage automatique

Dans notre étude accuracy, précision, et le Rappel (Recall) ont été utilisés comme paramètres dans l'évaluation empirique des algorithmes de classification Naive Bayes et les arbres de décision.

Métriques	**Description**
Accuracy	nombre relatif d'exemples correctement classés ou en d'autres termes pourcentage de prédictions correctes.
Précision	Métrique intuitive ,qui représente le rapport entre le nombre d'exemple correctement classés et le nombre total des exemples testé.
Rappel (Recall)	Proportion des solutions pertinentes qui sont trouvées. Mesure la capacité du système à donner toutes les solutions pertinentes.

TABLE 4.1 – Evaluations des métriques.

En référence à la matrice de confusion du tableau4.2, les paramètres d'évaluation peuvent être définis comme suit :

— **TP :**classe positive considérée positive.

— **TN :**classe négative considérée négative.

— **FP :**classe négative considérée positive.

— **FN :**classe positive considérée négative.

- Accuracy $=\frac{TP+TN}{TP+TN+FP+FN}$.
- Precision$=\frac{TP}{TP+FP}$.
- Rappel$=\frac{TP}{TP+FN}$.

4.3.1 Matrice de confusion

Une matrice de confusion (tableau de contingence ou matrice d'erreur) est un tableau cela permet la visualisation d'un algorithme d'apprentissage supervisé.

La précision n'est pas une mesure suffisante pour la performance d'un algorithme, pouvant conduire à des résultats trompeurs. Une matrice de confusion fournit une visualisation plus réaliste d'un algorithme. Chaque colonne de la matrice représente les instances dans une classe prédite, tandis que chaque ligne représente les instances d'une classe réelle.

observation		**Précision**	
		positive	négative
	positive	True Positive(TP)	False Negative(FN)
	négative	False Positive(FP)	True Négative(TN)

TABLE 4.2 – Matrice de confusion.

4.4 Résultat

Nous avons cité que les tweets ont tendance à inclure des symboles et des artefacts qui peuvent confondre le processus de classification. Dans l'exemple suivant nous montrons le résultat d'un tweet avant et après le prétraitement :

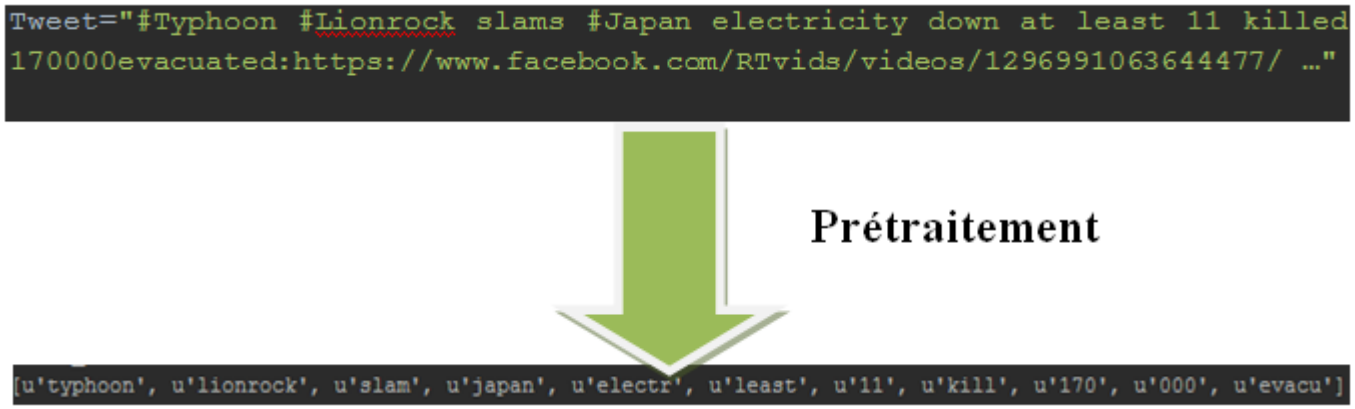

FIGURE 4.1 – Résultat de prétratement d'un Tweet

Alors le résultat de prétraitement consiste à supprimer les hashtag(#), les mots vide, et les liens, mais aussi, extraire les lemmes des mots.

Dans la phase de classification, nous avons utilisé la validation croisée de k=3 ;5 ;7 ;10 ;15 et 20. La performance du modèle de classification est testée à chaque itération. Les tableaux suivant présentent les résultats de la validation croisée des deux méthodes de classification (Naive baise et Arbre de décision) en utilisant les trois mesures cités plus haut.

— **En terme Accuracy**

	Arbre de Decision	Naive Bayes
K=3	67%	56%
K=5	75%	60%
K=7	76%	62%
K=10	77%	63%
K=15	75%	63%
K=20	75%	62%

TABLE 4.3 – Résultats de mesures évaluant les modèle de classification en terme de Accuracy

— **En terme de Précision et Rappel**

	Naive Bayes		**Arbre de Decision**	
	précision	rappel	précision	rappel
Health	72 %	92 %	98%	90 %
Coordination	0.00 %	0.00 %	0.00 %	0.00 %
Food	100 %	22 %	100 %	89 %
donnation	0.00 %	0.00	73 %	57 %
Rescue&search	53 %	100 %	86 %	88 %
education	0.00 %	0.00 %	0.00 %	0.00 %
Nutrition	0.00 %	0.00 %	0.00 %	0.00 %
Telecom&ITC	0.00 %	0.00 %	100 %	100 %
Shelter	100 %	42 %	39 %	89 %
Water	0.00 %	0.00 %	0.00 %	0.00 %
Transport&ligistique	0.00 %	0.00 %	100 %	40 %
protection	0.00 %	0.00%	100 %	29%

TABLE 4.4 – Résultats de mesures évaluant le modèle de classification cas de K=10

Les figures suivantes résument les résultats d'évaluation enregistrés dans chacune des itérations de K en terme de Precision et Rappel.

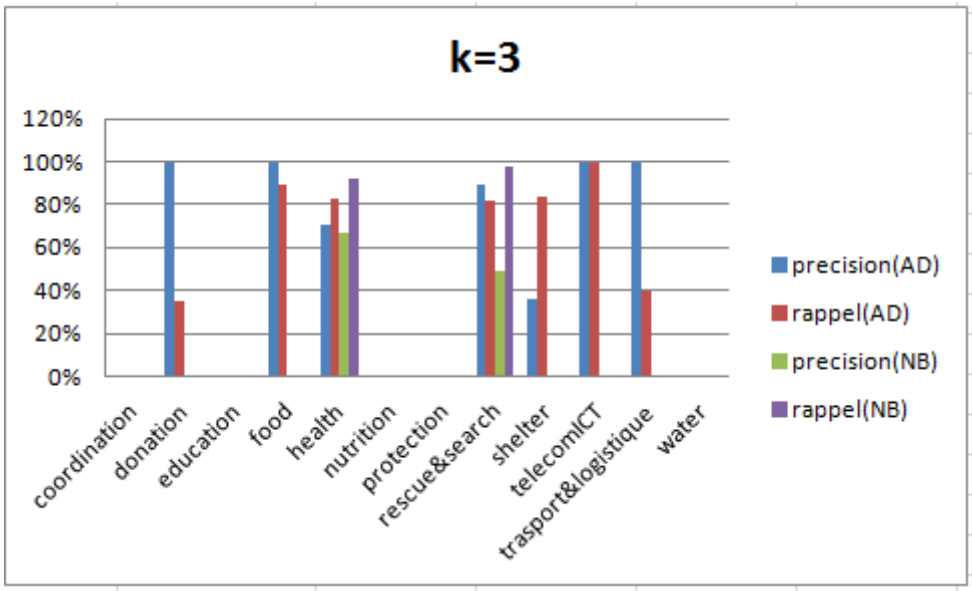

FIGURE 4.2 – Résultat d'évaluation cas de K=3

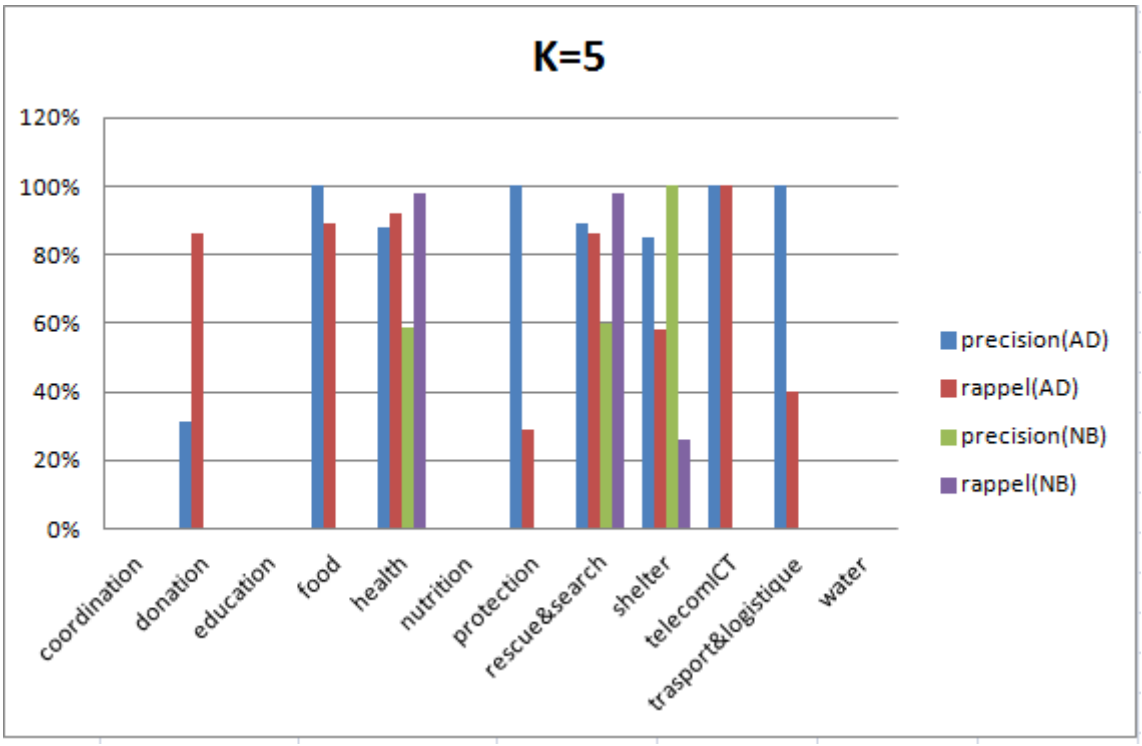

FIGURE 4.3 – Résultat d'évaluation cas de K=5

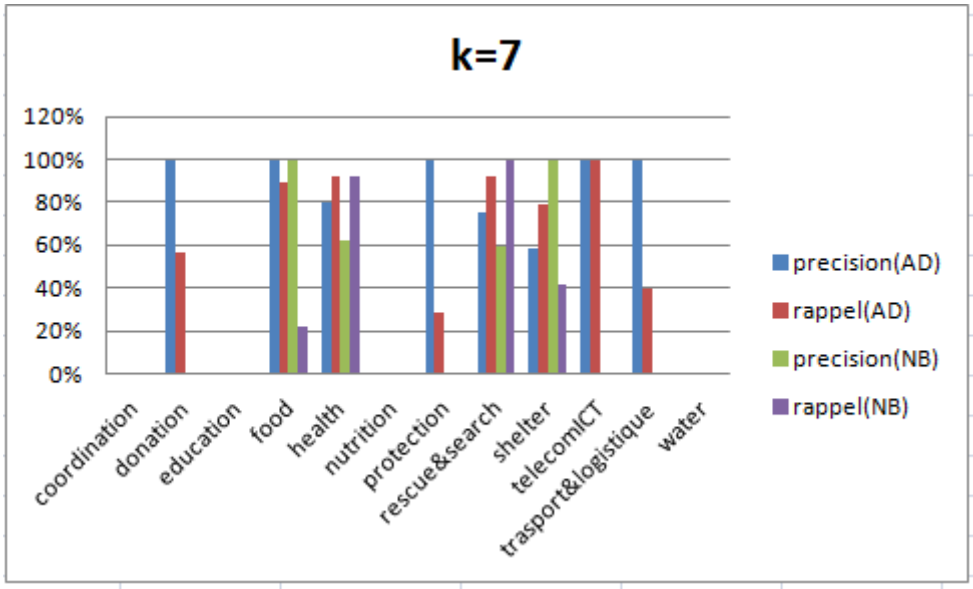

FIGURE 4.4 – Résultat d'évaluation cas de K=7

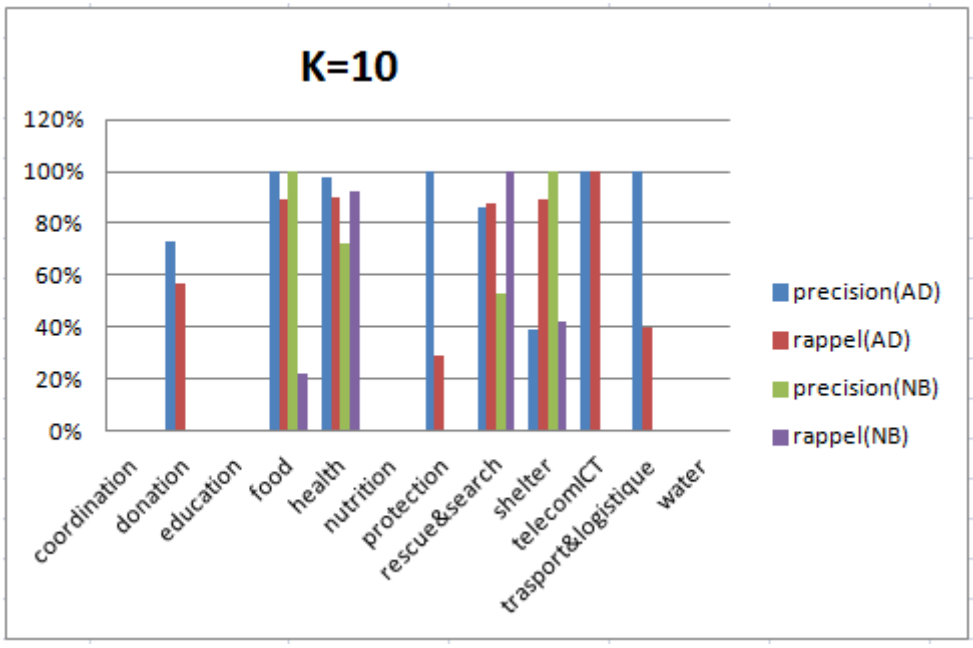

FIGURE 4.5 – Résultat d'évaluation cas de K=10

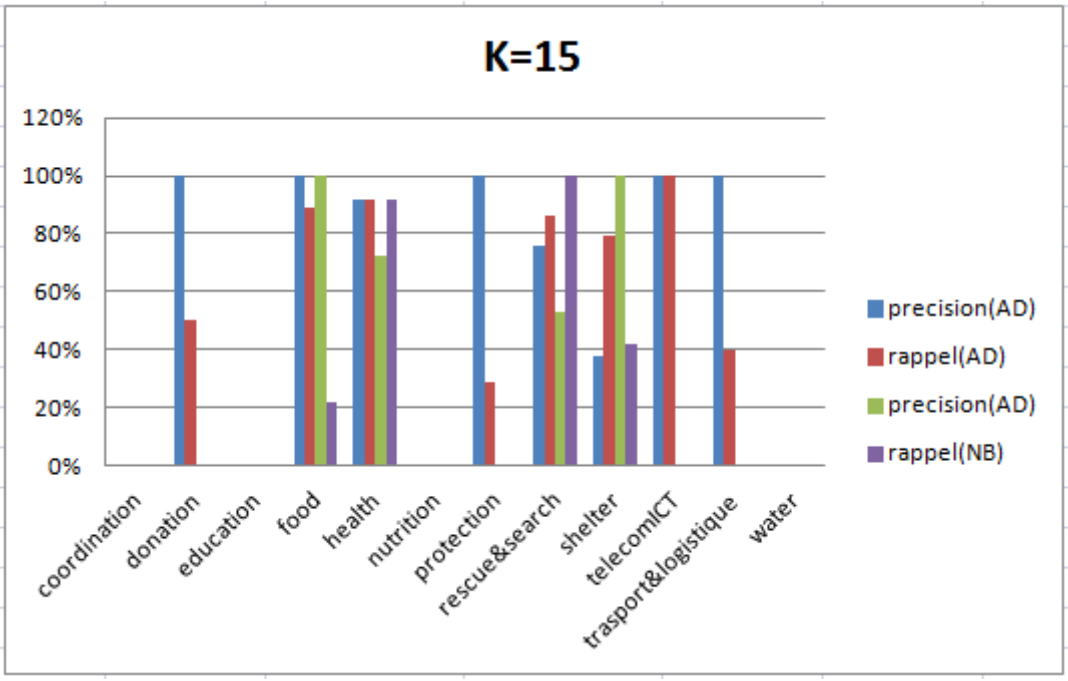

FIGURE 4.6 – Résultat d'évaluation cas de K=15

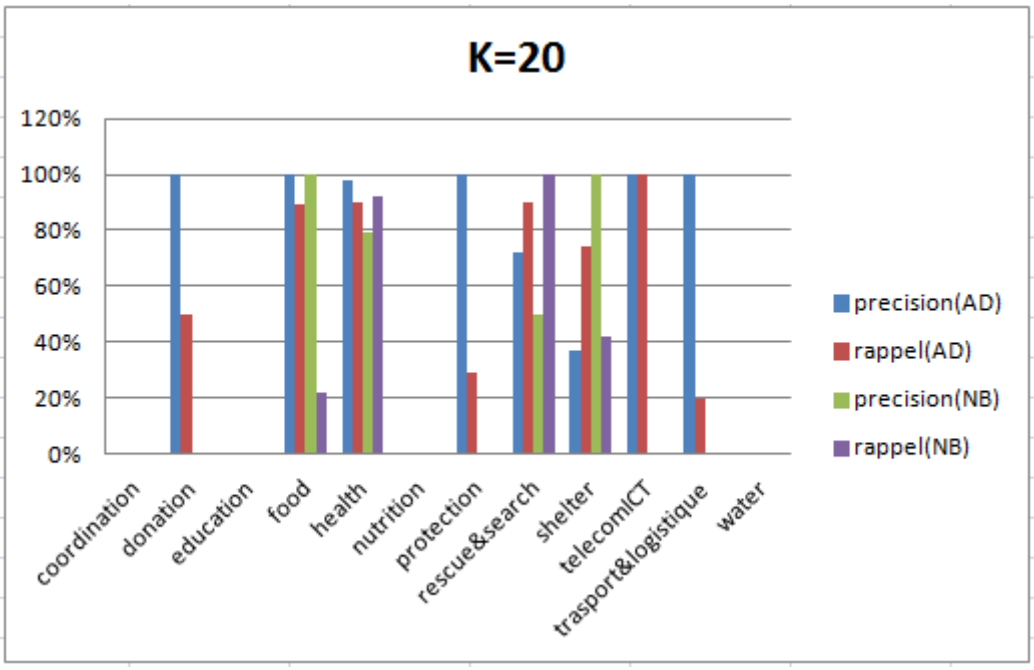

FIGURE 4.7 – Résultat d'évaluation cas de K=20

Les résultats du test démontrent qu'il y a une différence significative entre Naive Baise et Arbre de décision Ceci est vrai pour les trois mesures. Les arbres de décision sont plus élevés que la classification naïve bayésienne en termes de précision , Accuracy et Rappel. Cela signifie que l'apprentissage des Arbres de décision est meilleur.

4.5 Discussion

Nous démontrons qu'un classifieur basé sur les fonctionnalités linguistiques fonctionne bien pour identifier les classe des tweets contribuer à la SA. En plus nous montrons que les caractéristiques linguistiques, y compris la Tokenization, la lemmatisation et la suppression des mots vides, améliorent les performances du classificateur.
Ces résultats suggèrent que l'indentification des principales caractéristique d'un tweet peut aider à prédire si un tweet appartient à l'une des classes. Les liens entre la SA et ces caractéristiques nous permettre d'enrichir notre modèle de classification, et nous améliorons également notre proposition de filtrage de volume d'informations diffusé en cas d'urgence.
Pour l'évaluation des classifieurs, nous avons testé un échantillon de tweets annotés manuellement. En tant qu'une solution proposée, il classera continuellement les tweets entrants en fonction des modèles construits sur les données des événements similaires.
La solution proposée permet aux décideurs de trouver des réponses à ces besoins dans un événement de crise et permet aux services d'urgence d'analyser les informations que les gens publient sur Twitter.

4.6 Conclusion

Nous avons présenté dans ce dernier chapitre les différents langages et outils de développement que nous avons utilisé afin d'implémenter notre solution proposé. Dans le but de tester son fonctionnement et sa faisabilité avec une évaluation de différentes méthodes et des algorithmes de classification. Les résultats d'évaluation obtenus montrent en général l'efficacité de ces derniers.

Conclusion générale et perspectives

Pendant les urgences de masse, les populations affectées construisent une compréhension de la situation reposant sur des informations incomplètes (Harrald et Jefferson, 2007). Souvent, les victimes potentielles, les membres d'organismes d'intervention officiels et / ou étrangers concernés collectent les informations disponibles avant de décider quelle action à prendre concernant une situation d'urgence. Ce processus de collecte d'informations ou d'évaluation situationnelle conduit à un état de la conscience situationnelle (SA).

La situational awareness est un processus complexe qui nécessite la perception et la compréhension des éléments dans son environnement et conduit à prédictions de ce qui va se passer dans un proche avenir (Endsley, 1995 ; Endsley ect. . . .2003). Ainsi, pour satisfaire la SA, une collecte, une communication, et un partage d'informations situationnelles et de décisions décrivant ce qui se passe et ce qui se fait sur le terrain et en back-end doivent être effectués. Une solution technologique soutenant cette nécessité a été proposée et développée dans le cadre de ce travail. Nous avons présenté et décris les plus pertinents et connexes à notre travail dans le chapitre 3.
Twitter est un média utilisé par les utilisateurs dans le but de diffuser des tweets liés à une catastrophe. Lors d'un incident, les utilisateurs utilisent ce moyen pour diffuser à la fois les tweets informatif et non informatifs. Dans la majorité des cas les tweets informatifs dépassent les non informatifs.

Comme indiqué précédemment, la gestion de crises pose un certain nombre de problèmes. Parmi eux, la disponibilité de données pertinentes, actionnables, et complètes. Cependant, même si les gros volumes de données étaient disponibles, il y aurait de nouveaux problèmes

qui en découlent. Pour pallier à ces problèmes, nous avons proposé une solution technique liée à Twitter qui permet l'utilisation des méthodes de Machine Learning pour classer les tweets. Nous avons testé et comparé les performances de deux de ces algorithmes de classification les plus courants, Naïve Bayes et les arbres de décision.
Dans le cas des tweets, la classification consiste à annoter les différentes phrases d'un tweet avec des classes. Pour chaque classe, on trouve des termes importants considérés comme des indicatifs pour la classe. L'apprentissage automatique consiste simplement à comprendre les données et les statistiques. C'est un processus où les algorithmes dévoilent la pertinence des données, puis prédisent des résultats probables.
L'évaluation de la performance de ces données est basée sur la validation des résultats à travers les paramètres d'exactitude, de précision et le rappel avec l'application d'outils statistiques.
D'âpres la comparaison entre les deux méthodes de classification (NA et AD) et l évaluations des résultats obtenu, nous avons remarqué que le classifieur Arbre de Décision donne des meilleurs résultats en terme de précision.

Comme perspectives, il serait intéressant d'explorer d'autres caractéristiques et d'étudier leurs effets sur d'autres paramètres d'évaluation. Aussi, l'améliorer de la sélection, l'optimisation des paramètres et la sémantique seront un autre axe de la recherche permettant d'améliorer les résultats de classification.

Bibliographie

[1] Carine Rongier. *Gestion de la réponse à une crise par la performance : vers un outil d'aide à la décision. Application à l'humanitaire.* PhD thesis, INPT, 2012.

[2] Aicha Aid. *Formulation d'un environnement générique d'un service dans un système pervasif public en cas de situation d'urgence.* PhD thesis, Université Mouloud Mammeri, 2016.

[3] Sarah Elizabeth Vieweg. *Situational awareness in mass emergency : A behavioral and linguistic analysis of microblogged communications.* PhD thesis, University of Colorado at Boulder, 2012.

[4] Calameo. Tic en gestion de crise.

[5] Claire Brossaud. Usages des tic et rapports a l'incertitude en situation de catastrophes naturelles. *Développement durable et territoires. Économie, géographie, politique, droit, sociologie*, (Dossier 11), 2008.

[6] Antonin Segault. *Communication de crise en phase post-accidentelle nucléaire : organisation et partage des connaissances sur le Web.* PhD thesis, Bourgogne Franche-Comté, 2017.

[7] Vagelis Hristidis, Shu-Ching Chen, Tao Li, Steven Luis, and Yi Deng. Survey of data management and analysis in disaster situations. *Journal of Systems and Software*, 83(10) :1701–1714, 2010.

[8] Service Public fédéral. Les media sociaux en communication de crisen.

[9] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in mass emergency : A survey. *ACM Computing Surveys (CSUR)*, 47(4) :67, 2015.

[10] Adam T Zagorecki, David EA Johnson, and Jozef Ristvej. Data mining and machine learning in the context of disaster and crisis management. *International Journal of Emergency Management*, 9(4) :351–365, 2013.

[11] Lamri Laouamer. *Approche exploratoire sur la classification appliquée aux images*. PhD thesis, Université du Québec à Trois-Rivières, 2006.

[12] Hassane Hilali. *Application de la classification textuelle pour l'extraction des règles d'association maximales*. PhD thesis, Université du Québec à Trois-Rivières, 2009.

[13] Landry Jacques-André and Jonathan Bouchard. Rapport final.

[14] Fabian Abel, Claudia Hauff, Geert-Jan Houben, Richard Stronkman, and Ke Tao. Semantics+ filtering+ search= twitcident. exploring information in social web streams. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, pages 285–294. ACM, 2012.

[15] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Extracting information nuggets from disaster-related messages in social media. In *Iscram*, 2013.

[16] Kate Starbird and Jeannie Stamberger. Tweak the tweet : Leveraging microblogging proliferation with a prescriptive syntax to support citizen reporting. In *Proceedings of the 7th International ISCRAM Conference*, volume 1, pages 1–5. Information Systems for Crisis Response and Management Seattle, WA, 2010.

[17] Sudha Verma, Sarah Vieweg, William J Corvey, Leysia Palen, James H Martin, Martha Palmer, Aaron Schram, and Kenneth Mark Anderson. Natural language processing to the rescue ? extracting" situational awareness" tweets during mass emergency. In *ICWSM*, pages 385–392. Barcelona, 2011.

[18] Billal Belainine. Classification supervisée de textes courts et bruités : application au domaine des médias sociaux. 2017.

[19] Marianne Jahre and Leif-Magnus Jensen. Coordination in humanitarian logistics through clusters. *International Journal of Physical Distribution & Logistics Management*, 40(8/9) :657–674, 2010.

[20] Adi Bronshtein. Train/test split and cross validation in python, may 2017.

[21] Patrick Fuchs and Pierre Poulain. Cours de python, Septembre 2018.

[22] Edward Loper and Steven Bird. Nltk : The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, pages 63–70. Association for Computational Linguistics, 2002.

[23] Christian Gagne. Scikit-learn apprentissage et reconnaissance, Septembre 2016.