


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 Master

En Informatique

Spécialité : *Ingénierie des Systèmes d'Informations et du Logiciels*

Thème

Extraction d'information depuis les médias sociaux afin d'améliorer la Situational Awareness dans les situations d'urgence

---

Encadré par

— Dr : Aid Aicha

Réalisé par

— Boudraf Youcef

— Zekhmi Salah

2017/2018

Remerciements

En préambule à ce mémoire nous remercions Allah qui nous aide et nous donne la patience et le courage durant ces préparations de mémoire de fin d'étude. Nous exprimons nos gratitudes les plus sincères aux personnes qui nous ont apporté leur aide et qui ont contribué à l'élaboration de ce mémoire et qui ont accepté de répondre à nos questions avec gentillesse, spécialement le Monsieur Ameur Aissa, On tient à remercier sincèrement Docteur Aid aicha et qui en tant qu'encadreur, s'est toujours montré à l'écoute ainsi que son précieux conseil et son aide durant toute la période du travail. Nos remerciements s'étendent également tous nos professeurs du département informatique pour la richesse et la qualité de leurs enseignements et qui déploient de grands efforts pour assurer à leurs étudiants une formation actualisée. Nous n'oublions pas nos familles pour leur contribution, leur soutien et leur patience. Enfin, nous adressons nos plus sincères remerciements à tous nos proches et amis, qui nous ont toujours soutenu et encouragé au cours de la réalisation de ce mémoire. Merci à tous.

Dédicaces

Je dédie ce travail à mon cher père Mustapha et ma chère mère Samia, que nulle dédicace ne puisse exprimer mes sincères sentiments, pour leur patience illimitée, leurs encouragements contenus, leurs prières pour moi, leurs aides, leurs tendances inestimables. Et à mes chers grands parents Laifa et Messaoud, Aicha et Fatima.

Je suis très reconnaissant 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.

A mon cher frère Nadjib et Ma chère soeur Akila, Mon ancle Said et sa femme Saliha, Notre petit ange Thilleli Chaima et Ouafa. Pour leurs conseils et soutiens. A mes chères collègues de groupe ISIL et de département informatique en particulier Selma, Salah, Amir, Anis, Amine, Lhadj.

Boudraf Youcef.

Je dédie ce mémoire : A ma chère mère et ma sœur Fatiha et Rabia et a mon frère hsane qui a toujours été présente. A ma chère femme qui a toujours été présente, et A mes enfants. A tous mes amis, A tous les étudiants du département informatique.

ZEKHMI salah.

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

ONU	L'organisation des nations unies.
ONG	organisations non gouvernementales.
EVC	L'évaluation de la vulnérabilité et des capacités.
SA	Situational Awareness.
NTIC	nouvelles technologies de l'information et de la communication.
TIC	Technologies de l'information et de la communication.
NLP	Natural Language Processing.
API	application programming interface.
IE	L'extraction d'information.
IR	recherche d'information.
DM	Data Mining.
IF	filtrage d'information.
TAL	Traitement automatique de langage.
NLP	pour Natural Langage Processing.
CBOW	Continuous Bag-of-Words.

Introduction générale

Les médias sociaux et les technologies collaboratives sont des éléments essentiels en matière de recouvrement des différentes activités qui intègrent la technologie, l'interaction sociale, et la création de contenu. Actuellement, ce nouvel outil de média intéresse tout le monde : les industriels, les politiciens, les académiques, les particulier et bien d'autres l'utilisent presque chaque jour pour échanger leurs informations .De ce fait, les médias sociaux constituent des incroyables mines de renseignements disponibles se trouvant désormais sur plusieurs plateformes et sous de multiples formats.
Parmi les intéressés qui s'attirent de plus en plus massivement vers les médias sociaux sont les intervenants en cas de situation d'urgence. Les équipes d'interventions d'urgence s'appuient depuis longtemps sur plusieurs sources de renseignements, notamment les radios et outils de télécommunications dans les véhicules d'urgence, les systèmes de gestion des crises, les systèmes d'information géographique, etc.

En effet, les intervenants commencent à utiliser les médias sociaux pour communiquer et pour recueillir et diffuser de l'information utile et sensible afin d'obtenir une meilleure connaissance de la situation et de prendre des décisions adéquates. La grande quantité de renseignements diffusés et la vitesse à laquelle ils sont transmis peuvent par contre créer une situation où ces renseignements se retrouvent inutilisés ou inutilisables s'il n'est pas possible de les identifier, de les confirmer, de les coordonner, de les agréger ou de les contextualiser. Alors, est-il possible d'établir un mécanisme permettant d'exploiter le mieux possible ces plates-formes d'information et de les filtrer et par la suite de les envoyer en temps réel aux intervenants appropriées ?
Parmi ces réseaux sociaux, Twitter est devenu un instrument de communication incontournable pour tous les acteurs sociaux grâce à son partage rapide des informations, et à accès

public. Les décideurs en situation de catastrophe l'utilisent pour donner la primeur de leur actualité, leurs décisions, et leurs actions à venir. Il constitue également une plate-forme d'échange qui permet à tout un chacun d'exprimer son opinion en réaction à une annonce ou à un évènement. Des informations, parfois très importantes, transitent ainsi dans tous les sens, tous les jours, sans que nous saisissions toute la portée de ce déluge de textes qui, parfois, semblent peu cohérents.

L'analyse des tweets s'inscrit dans le cadre du text mining à bien des égards. Chaque document est un texte rédigé. Nous pouvons appliquer les techniques de fouille de textes usuelles, notamment en passant à la représentation en sac de mots (bag-of-words). Mais les tweets induisent des particularités. Certaines peuvent enrichir l'analyse. Ainsi, leur longueur est calibrée (du moins en ce qui concerne les messages publics), des caractères spéciaux permettent d'identifier les auteurs () et les thématiques (#), les mécanismes de tweet et retweet permettent de suivre la diffusion de l'information. A contrario, d'autres caractéristiques peuvent perturber les analyses. L'espace étant limité, les auteurs utilisent souvent des abréviations, des émoticons pour exprimer des sentiments, et ils ne font pas très attention à l'orthographe. Tout cela engendre du bruit qui peut compliquer notre tâche, qui consiste en l'extraction d'informations situationnelles depuis réseaux sociaux, et notamment Twitter, afin d'augmenter la compréhension des décideurs à la situation d'urgence. L'extraction d'information fait partie du domaine de traitement automatique de langue (TAL). Cette branche de l'informatique traite les différents aspects de la langue (sémantique, syntaxe, morpologie, etc.). Dans ce travail, nous allons tirer profit des tweets pendant une situation de crise pour améliorer la situation awareness, en filtrant et classant les tweets extraits selon des classes prédéfinies.

Ce mémoire est organisé en 4 chapitres :

- **Chapitre 1 :** nous donnerons une définition appropriée des catastrophes tout en décrivant leur types, le processus de gestion d'une catastrophe, l'étape de Réponse et la Situational Awarness. Enfin, quelques problèmes liés à la communication dans la situation d'urgence sont cités.
- **Chapitre 2 :** ce chapitre discute les NTIC et leur rôle dans la situation d'urgence.Nous commencerons par définir ce que sont quoi les NTIC. Ensuite, une vue sur les médias sociaux et Twitter, ainsi que les avantages qu'ils offrent pour qu'ils

soient une source d'informations. Enfin, nous verrons les caractéristiques des informations lors d'une situation de crise, et comment elles sont traitées par les NTIC.

- **Chapitre 3 :** ce chapitre constitue le vif de notre travail. Nous parlerons de l'extraction d'information depuis Twitter permettant de répondre aux problématiques rencontrées par les décideurs durant une situation de crise. Nous présenterons notre solution et les algorithmes utilisés, ainsi que quelque travaux connexes.
- **Chapitre 4 :** Dans le dernier chapitre, nous présenterons les outils utilisés, et les résultats d'évaluation avec leur discussion.

Chapitre 1

La réponse dans la situation d'urgences

1.1 Introduction

L'être humain a toujours été entouré de danger : « Être entouré de danger [...] c'était la situation fondamentale de l'être humain des sociétés primitives » (Beck, 2001). Toutefois, l'évolution de la notion de danger et de sa perception dans le temps, mais aussi dans l'espace, a fait du risque une problématique qui pèse sur les sociétés de nos jours. Ce faisant, le risque devenu incontrôlable impose à la société un nouveau mode de pensée et d'action. Ainsi, la société passe du danger naturel incontrôlable dans un premier temps, et aboutit, enfin, au risque calculable et donc contrôlable grâce au progrès de la science. Par contre, ces mêmes progrès ont enfoncé nos sociétés dans un monde de « risques incontrôlables[9] ».

La volonté de maîtriser les risques soulève la question de l'équité sociale dans la gestion des risques ou des catastrophes. Aussi, et afin d'assurer une sécurité face aux catastrophes, divers programmes et normes ont été conçus et mis en œuvre dans le but d'améliorer tant la qualité des interventions lors d'un sinistre que les procédures de prévention, en visant à responsabiliser, de plus en plus, les acteurs du risque au regard des droits de la personne.

Cependant, bien que des programmes aient été mis en œuvre, on continue à enregistrer des pratiques contraires à toute éthique et une inégalité dans l'application des normes d'un pays à un autre, voire au sein d'un même pays, chaque fois qu'il y a crise. Les acteurs du risque n'assurent malheureusement pas toujours la transparence des procédures et l'égalité des chances pour tous, devant un risque ou un sinistre. L'accès à l'information pour la population devient alors un problème crucial dans un pareil contexte d'urgence et risque d'accentuer la vulnérabilité des populations face à un sinistre. Nous tenterons de montrer

comment le croisement de deux rationalités (celle des populations et celle des institutions) peuvent altérer la communication et, par conséquent, le bon déroulement des plans de gestion de l'urgence.

1.2 Définition de catastrophe :

Définir une catastrophe par ses origines est limité, car une catastrophe peut avoir des effets variés selon le contexte local, également, tenter de définir une catastrophe par le principe de ses effets est limité notamment par le contexte local spécifique [10], donc l'impact de la catastrophe sur l'individu, la communauté, et le pays sert à trouver des définitions plus précises pour la catastrophe.

Une première définition considère que la catastrophe est un événement exceptionnel, dans lequel ni l'organisation ni les ressources normales disponibles peuvent y faire face. Cette définition traite les conséquences de la catastrophe, d'où les effets de la catastrophe selon cette définition sont mortels, dangereux.

L'organisation des nations unies ONU, de son côté définit la catastrophe comme une grave interruption dans le fonctionnement de la société, un événement qui cause de grandes pertes humaines, matérielles, environnementales, dépassant la capacité de société affectée à y faire face avec ses propres ressources[11].

Les conséquences d'une catastrophe peuvent s'étendre à toutes les dimensions de vie de la société en question, tel que le transport, la communication, l'alimentation en gaz et en électricité, etc.

1.3 Les types de catastrophe :

Les catastrophes sont souvent sur la base de la cause qui la déclenche, donc on peut citer les quatre principaux types de catastrophe, et cela n'empêche pas d'y être des sous-types : :

a) **Les catastrophes naturelles.**

b) **Les catastrophes technologiques.**

c) **Les catastrophes sociales.**

d) **Les catastrophes complexes et État défaillant.**

1.3.1 les catastrophe naturelles :

Les catastrophes naturelles, c'est-à-dire les conséquences d'évènements provoqués par des risques naturels qui bouleversent la capacité d'intervention locale et affectent gravement le développement social et économique d'une région, sont traditionnellement ressenties comme étant des situations engendrant des défis et des problèmes de nature essentiellement humanitaire [12]. Elles couvrent les événements géophysiques tels que des éruptions volcaniques ou des tremblements de terre. Ces événements sont fréquemment très localisés, et leurs effets ressentis dans une zone restreinte. La catégorie des catastrophes naturelles couvre également les événements hydrométéorologiques. Ils sont ressentis dans des zones bien plus étendues et inclus[10] :

- **tempêtes (ouragans, typhons, cyclones).**
- **fortes pluies ou chutes de neige.**
- **sécheresse ; etc.**
- **températures excessivement basses ou hautes.**

À leur tour, ces événements d'origine naturelle peuvent déclencher : inondations ; tsunamis ; glissements de terrain et coulées de boue ; avalanches ; érosion excessive ; incendies de forêt ; et mauvaises récoltes.
Les événements biologiques constituent également une catégorie de catastrophe naturelle. Ils comportent : invasion d'insectes ; et épidémies.

1.3.2 les catastrophe technologiques :

La nature n'est pas la seule cause des catastrophes, ces dernières peuvent également trouver leurs origines dans des activités d'origine humaine[10], telles que : des accidents industriels ou technologiques qui entraînent des émissions de radiations, de produits chimiques ou des explosions, accidents survenant durant le transport de produits dangereux, défaillance structurelle de ponts, d'immeubles, de lignes électriques, de barrages ou de mines, accidents de train ou de véhicules ; et engins non explosés.

1.3.3 les catastrophe sociales :

Les catastrophes technologiques résultant de l'activité de l'homme ne sont pas les seules catastrophes que l'homme peut causer mais il peut aussi être derrière des défaillances sociales, quand le comportement d'un groupe se détériorent d'une manière politique, culturelle

et idéologique, avec par exemple (dans un ordre de magnitude croissante) [10] : manifestations ; mouvements de foule ; émeutes ; action terroriste ; conflit ; et guerre.

1.3.4 les catastrophe complexes et les États en faillites :

Une gouvernance adéquate et un État de droit, représente pour une société un moyen de stabilité. Quand ceux-ci sont pas présents, à cause d'un conflit ou d'une catastrophe naturelle de grande ampleur, de complexes défaillances ayant des conséquences économiques, sociales, physiques ou environnementales peuvent se produit, le tout dans un contexte général d'insécurité [10]. La dernière catégorie de catastrophe admet l'importance de deux éléments cités.
Ces types de catastrophe sont catégorisés par leurs causes, mais ils partagent tous le processus de leur gestion.

1.4 Le processus de gestion d'une catastrophe :

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.

a) **La mitigation :**Réduire et éliminer le degré du risque sur la population, les propriétés et l'environnement.
b) **La préparation :** Développer et renforcer les capacités nécessaires permettant une intervention rapide et efficace.
c) **La réponse :** Intervenir, activer les plans de secours et de sauvetage et mobiliser les entités responsables.
d) **Le rétablissement :** Rétablir la situation à la normal, à court terme et à long terme.

1.5 Focus sur l'étape de Réponse :

Cette phase concerne les opérations d'intervention à activer dès que qu'une alerte est communiquée. Ces activités de réponse en cas de catastrophe sont exécutées par des décideurs et des secouristes de diverses organisations gouvernementales, des organisations

humanitaires, des organismes internationaux et nationaux, des entités locales et des particuliers, chacun avec ses rôles et tâches. Leurs objectifs visent à fournir de l'aide et du secours à la population touchée (par exemple, la recherche et le sauvetage, l'évacuation, l'identification des corps, les soins médicaux, installation des camps de refugiés, etc.), minimiser les pertes et le risque d'aggravation, ainsi qu'à stabiliser la situation d'urgence pour un retour à la normale le plus vite possible.

1.5.1 La 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.

1.5.2 La réponse :

La phase de réponse, où sont conduites les activités de recherche et sauvetage, évaluation rapide des dommages et des besoins, et la mise à disposition des aides de premier secours suivies par l'ouverture et la gestion de refuges temporaires pour les individus sans demeure ainsi que la mise à disposition d'assistance humanitaire pour les personnes affectées[13].

Estimation des dommages L'évaluation de la vulnérabilité et des capacités (EVC) recourt à divers outils participatifs pour estimer l'exposition des personnes aux dangers naturels et leur capacité à y résister. Elle fait partie intégrante de la préparation aux catastrophes et contribue à la création de programmes de préparation aux catastrophes orientés vers la communauté au niveau populaire rural et urbain. L'EVC permet d'identifier les priorités locales et de prendre les mesures adéquates pour réduire le risque[14].

Les Besoins précis : La préparation logistique est un élément clé de tout effort de prévention des catastrophes. La planification est à la fois nécessaire et pratique, dans la mesure où il est généralement possible de prévoir les types de catastrophes pouvant affecter un endroit précis ainsi que les besoins que ces catastrophes sont susceptibles d'engendrer. La préparation logistique doit se fonder sur l'évaluation de la vulnérabilité et des ressources[14]. Fondée sur l'évaluation des besoins, l'Unité logistique régionale couvrant la zone affectée par la catastrophe publie une liste d'articles de première nécessité dont les Sociétés nationales et d'autres parties peuvent faire don à une opération (connue sous le nom de tableau de mobilisation). La définition des priorités, la planification du transport, la réception et la distribution des ravitaillements d'urgence constituent un rôle de coordination, essentiel en matière de sauvetage.

Hiérarchisation des opérations de réponse : Les options de traitement du risque peuvent être présentés par ordre de priorité en tenant compte de la gravité du risque, de l'efficacité des contrôles du risque, des coûts et des avantages, des contraintes actuelles. Ces options de traitement, qui forment des recommandations, informeront et seront envisagées dans l'étape de traitement du risque de la gestion du risque ou dans le cycle de la gestion des urgences.

Organisation de la réponse : Dans cette étape, les Données de cycle de réponse sont déployées et les décisions de réponse prises par les responsables sont transmises aux agents spécialiste (pompier, police, etc.) et autres acteurs impliqués présents sur le lieu.
Ce processus cyclique et continu est répété à chaque disponibilité de nouvelles informations situationnelles.

1.6 Facteurs influençant la réponse :

Décider ce qu'est une catastrophe et quand l'aide doit être apportée ne relève pas d'une science exacte. Les facteurs ci-dessous couvrent les enjeux à prendre en compte avant de

déclarer qu'un événement constitue une catastrophe[10].

1.6.1 L'ampleur :

Lorsqu'un individu est grièvement blessé dans un accident cela constitue une catastrophe pour cet individu et ses proches mais cela ne va pas nécessairement impacter la communauté au sens large à moins que cet individu ne soit une personne clé tel qu'un médecin. La communauté peut faire face à cet évènement et ne pas en être affectée négativement. Si un nombre important plus de personnes sont tuées ou blessées, la communauté dans son ensemble pourra être affectée et avoir besoin d'aide. L'ampleur est par conséquent un facteur qui peut déterminer l'intervention lors d'une situation d'urgence.

1.6.2 Les frontières :

Le besoin d'une assistance extérieure nécessitent de prendre en compte ce qui relève de l'intérieur et ce qui relève de l'extérieur. Un feu dans une maison de particulier peut nécessiter l'intervention des voisins ; un important feu en ville peut nécessiter l'aide des villes voisines ; un feu de forêts qui brûle de larges zones peut être considéré comme un problème national. Les catastrophes comme l'épidémie de VIH/SIDA ou une vague de chaleur n'ont pas de frontière, en ce que les populations affectées sont généralement intégrées au sein de la population dans son ensemble.

1.6.3 Ressources :

Un grand pays peut avoir les ressources internes pour faire face à une catastrophe, en y allouant des ressources humaines et matérielles venant d'une autre partie du pays. Un petit pays faisant face à la même catastrophe pourra cependant avoir besoin d'une aide extérieure. De la même façon, si deux communautés de la même taille vivent le même événement, et que l'une a plus de ressources (ou des ressources plus diversifiées – ou les deux) alors cette communauté pourra être plus résistante que celle ayant des ressources limitées ou dépendante d'une seule ressource.

1.6.4 Localisation :

Comme suggéré plus haut, la localisation d'une catastrophe conditionnera son impact. Des catastrophes en zones urbaines densément peuplées auront un impact négatif plus important qu'une catastrophe similaire se produisant dans une zone rurale reculée. A l'inverse il sera plus difficile en zone rurale de faire un diagnostic suite à la catastrophe, d'engendrer une couverture médiatique, de fournir une assistance extérieure et d'accéder aux ressources et compétences locales, notamment si la zone manque déjà d'infrastructures de base et de moyens de communication.

1.6.5 Développement d'infrastructures :

L'intervention post-catastrophe peut être complexifiée du fait du niveau de développement de la zone affectée. Un incendie dans une simple chaumière pourra la détruire entièrement mais il est fort probable que celle-ci ait été construite en matériaux locaux, par de la main d'œuvre locales ayant ces compétences locales. Dans ce cas le besoin de la population affectée consistera en une maison, mais cette maison est simple à remplacer. Un incendie domestique dans un pays industrialisé pourrait ne pas détruire totalement la construction mais il y a moins de chance que le ménage soit dans la capacité de réparer les dommages avec des ressources locales ; le coût de remplacement des habitations sera dans ce cas bien supérieur au coût de reconstruction au sein d'une communauté à faible revenu. La catastrophe peut dont être la même mais son coût économique sera plus élevé dans un pays plus industrialisé.
De bonnes voies de communication peuvent favoriser les opérations de secours alors qu'un manque d'infrastructures de base retardera l'aide et aggravera l'impact de la catastrophe. A l'inverse, lorsque la population dispose de voies de communication fiables, les marchandises de base et la nourriture peuvent être plus longues, rendant les populations moins résilientes au cas où ces infrastructures seraient endommagées.

1.6.6 Gouvernance et capacité :

Le niveau de développement n'est pas uniquement déterminé par les infrastructures physiques, mais aussi par la capacité des gouvernements locaux et nationaux à faire face à une catastrophe, leur niveau de préparation, leurs capacités de réponse à une urgence et leur accès aux services d'urgence.

1.7 Situational Awarness :

Selon [18], la SA est la perception des éléments de l'environnement dans un volume de temps et d'espace, la compréhension de leur signification, et une projection de leur statut dans un avenir proche. Donc la formulation de SA se passe en trois fonctions, la perception des éléments, leur compréhension et leur projection dans le future[15].

1.7.1 Le Modèle SA d'Endsley :

[8] Le modèle Endsley fourni par Mica Endsley (1995b), et largement utilisé dans la littérature désormais, décrit les états de la Situationnal Awereness en fonction de trois niveaux : perception, compréhension, et projection.

Niveau 1 - Perception :

La première étape dans la réalisation de la Situational Awareness consiste à percevoir le statut, les attributs, et la dynamique des éléments pertinents dans l'environnement. Ainsi, ce niveau implique les processus de surveillance, de détection des repères et de reconnaissance simple, qui permettent de prendre conscience des multiples éléments de la situation (objets, événements, personnes, systèmes, facteurs environnementaux) ; états (lieux, conditions, modes, actions).

Niveau 2 – Compréhension :

L'étape suivante de la formation de la SA consiste en une synthèse des éléments disjoints du niveau précédant à travers les processus de reconnaissance des formes, d'interprétation et d'évaluation. Ce deuxième niveau exige l'intégration de ces informations pour comprendre comment cela affectera les buts et les objectifs de l'individu. Cela comprend l'élaboration d'une image complète du monde ou de cette partie du monde qui préoccupe l'individu.

Niveau 3 – Projection :

Il constitue le niveau le plus haut de la SA, et il est associé à la capacité de projeter les éléments de contexte et leur compréhension au future. La précision de la prédiction dépend fortement de la précision du niveau de la perception et du niveau de la compréhension. L'anticipation de la situation future prévue fournit aux décideurs le temps pour résoudre

les conflits et planifier un plan d'action pour atteindre leurs objectifs.

Le modèle de Endsley, illustré par la figure 1.1, montre comment la SA "fournit la base principale pour la prise de décision et la performance ultérieures dans le fonctionnement de systèmes complexes et dynamiques". Seule, elle ne peut garantir une prise de décision réussie, mais la Situationnal Awareness prend en charge les processus de saisie nécessaires (par exemple, reconnaissance des indices, évaluation de la situation, prédiction) sur lesquels reposent les bonnes décisions (Artman, 2000).

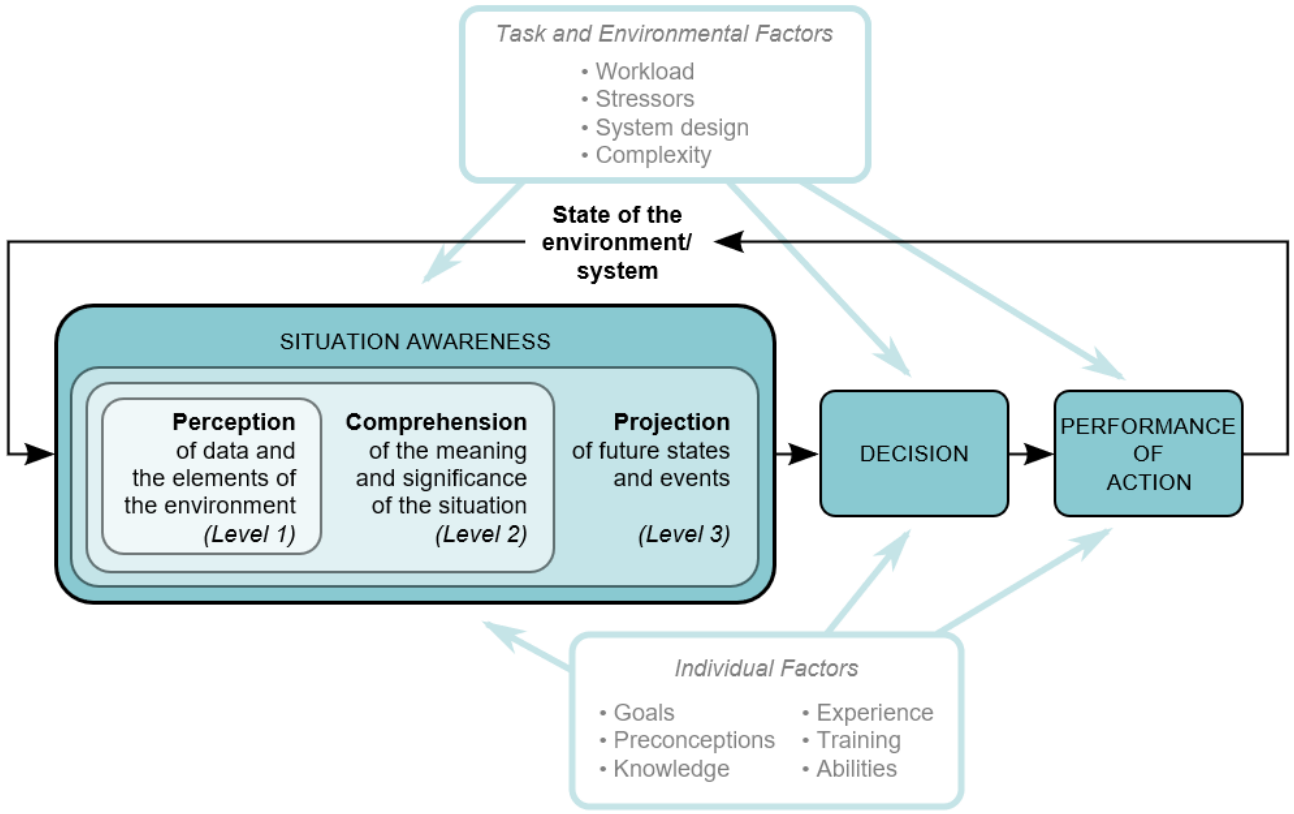

FIGURE 1.1 – modèle Endsley de la Situationnal Awareness.[8]

1.8 Quelques problèmes liés à l'information dans la situation d'urgence :

Les problèmes liés à l'information dans la situation d'urgence consistent donc essentiellement à :

1) Il n'existe presque aucune organisation capable d'assurer pleinement la diffusion

d'information à l'autorité en temps de crise.

2) Les décideurs souffrent d'un manque d'information situationnelle actionnable.
3) Difficile de collecter l'information manuellement et organiser en une façon cohérente.
4) Manque des moyens permettant de filtrer les informations pertinentes par rapport aux informations non pertinentes.
5) la Difficulté d'évaluer et estimer les dommages (région géographique dispersées, vaste région ... etc.).
6) La masse de donnée échangée autour de la crise peut être considérable, d'où la nécessité de la filtrer afin d'extraire des informations qui servent à aider les secouristes et les décideurs à mieux répondre à la situation.

1.9 Conclusion :

Dans ce chapitre nous avons discuté les situations de crise et de catastrophe ainsi que la réponse à ces dernières, qui constitue l'élément clé de la gestion de crise. Afin d'identifier correctement les catastrophes, plusieurs facteurs doivent être pris en compte. Les personnes fournissant de l'aide et les personnes nécessitant de l'aide. Cette relation doit être équilibrée, et la mise en œuvre de l'aide gouvernée par les personnes qui en ont besoin et non par ceux qui l'impose. Il est important de se concentrer sur l'utilisation des NTIC appropriées selon les besoins informationnels.

Les NTIC, qui sont le sujet du prochain chapitre, devraient être une source familière et fiable de communication. En outre, les dirigeants semblent être en mesure de maintenir une certaine SA très élevée par rapport aux autres, contrairement au personnel opérationnel. Ainsi, le rôle du personnel agissant pendant la gestion des catastrophes impose des exigences différentes de l'information.

Chapitre 2

Les NTIC et la réponse dans la situation d'urgences

2.1 Introduction

A la suite d'une catastrophe naturelle, plusieurs besoins émergent : aides de soins, secours et évacuation, nourriture, abri, etc. Cependant, la réponse à ses besoins et la manière dont il faut agir en conséquent nécessitent une bonne compréhension de la situation.
Dans le chapitre précédent, nous avons discuté la Situational Awarness (SA) comme un élément clé dans la réponse à une crise. A son tour, l'élément clé dans la SA est l'information situationnelle. Cette information était jusqu'à un certain temps rarement disponible et étroitement diffusée. La coordination entre les équipes qui actionnent sur le terrain était terriblement difficile par manque de moyens utilisés. Par conséquent, la nécessité de trouver des remèdes à ces problèmes constitue une part majeure des occupations des activistes dans ce domaine.

L'ère actuelle est marquée par l'avènement des nouvelles méthodes de communication qui constituent un élément indispensable dans le comportement quotidien de l'être humain. Ces méthodes de communication sont l'issue d'un rapprochement sans précédent de l'informatique, des télécommunications, et de l'audiovisuel, qui a donné naissance aux NTICs.

2.2 Définition des NTIC :

Les Nouvelles Technologies de l'Information et des Communications (NTIC) désignent généralement ce qui relève des nouvelles technologies utilisées dans le traitement et la transmission des informations et principalement : l'informatique, internet, et la téléphonie mobile.

Les NTIC ont aussi pour définition l'ensemble de technologies utiles à traiter, modifier, et échanger de l'information, plus spécifiquement des données numérisées. La naissance des NTIC est due à la convergence de l'Informatique, des Télécommunication, et de l'Audiovisuel.

L'arrivée du Web 2.0 dans les années 1990 engendre une émergence assez rapide des outils de communication et d'échange de l'information. Une simple recherche sur la toile nous montre le nombre sans précédent d'application informatique qui offre la possibilité de communiquer. Facebook, Twitter et bien d'autres applications connaissant un taux de téléchargement qui dépasse le 500 millions fois sur les stores d'applications mobile sans prise en compte l'utilisation quotidienne de leurs plateformes web.

Les secouristes et les décideurs ne sont pas échappés a cette vague d'utilisation des TIC, ils les ont sans attendre intégrés dans leurs processus de gestion de crise, avec leurs téléphone portables, tablettes, ordinateurs portables, en développent aussi des systèmes d'informations et des outils qui les aident à accomplir leurs travail [16].

2.3 Twitter

2.3.1 Définition :

Twitter est un site de micro-blogging qui est classé onzième mondialement et septième aux Etats Unies, selon le classement du trafic Alexa. Twitter permet aux utilisateurs la communication et la discussion sur divers sujet, en utilisant des messages courts appelés tweets. Selon le site de statistiques en direct, au moment de la rédaction de ce document plus de 500 millions de tweets sont envoyés chaque jour. Grace à sa structure de suiveurs / suiveurs (followers/followers), Twitter devient la plateforme de microblogging la plus populaire, qui permet aux utilisateurs de partager rapidement des informations sur leurs activités personnelles et partager des commentaires récents.

Contrairement aux autres réseaux sociaux tel que Facebook, les relations sociales sur Twitter sont asymétriques et peuvent être assimilées à un réseau social dirigé ou à un réseau de suiveurs (Brzozowski et Romero, 2011). Un utilisateur peut suivre un autre utilisateur sans exiger une approbation ou une connexion réciproque des utilisateurs suivis. Twitter a évolué au fil du temps et il fournit actuellement différentes manières aux utilisateurs de converser et d'interagir en se référant dans un vocabulaire bien défini tel que le retweeting, les mentions d'utilisateurs, et les hashtags.

2.3.2 Hashtags :

Les utilisateurs de Twitter utilisent couramment les hashtags pour mettre en signet le contenu des tweets, participé à un groupe communautaire axé sur le même sujet ou pour relier leurs tweets aux discussions en cours. Ils peuvent être de différentes formes, y compris des mots simples (#winner) ou une combinaison de plusieurs mots (#ripkoufianan). Des études antérieures ont montré que les hashtags peuvent contenir des informations utiles qui peuvent être utilisées pour améliorer diverses tâches NLP sur des tweets tels que l'analyse des sentiments (Wang, Wei, Liu, Zhou et Zhang, 2011) ou la reconnaissance des entités nommées.

2.4 Twitter source d'information :

Au cours des dernières années, Twitter est devenue l'une des sources d'information les plus populaires pour les applications pratiques et la recherche universitaire. Il existe de nombreux exemples d'applications pratiques de données Twitter, allant de la prévision des stocks (Arias, Arratia et Xuriguera, 2013) à la détection d'événements en temps réel (Sakaki et al., 2010), l'analyse des tendances (Mathioudakis & Koudas, 2010), et gestion de crise (Abel et al., 2012). De plus, les données Twitter ont été jugées utiles pour les applications de sécurité publique (Ritterman, Osborne et Klein, 2009). Afin d'aider les chercheurs, les décideurs, et les organisations.

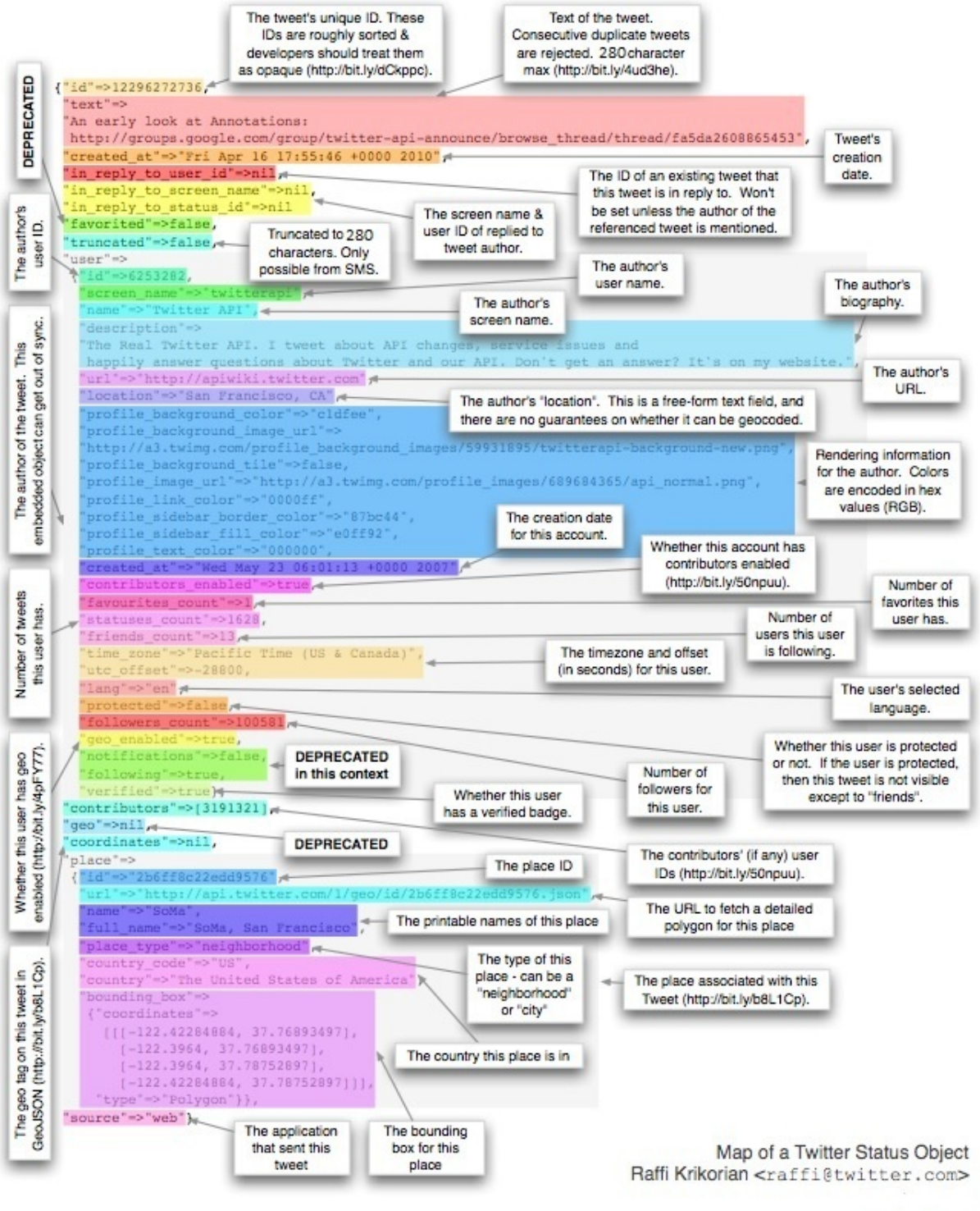

FIGURE 2.1 – Anatomie d'un tweet

à exploiter ses données, Twitter propose une API publique qui facilite son intégration dans les applications externes. L'API Twitter est disponible en tant qu'API de recherche ou de diffusion permettant la collecte de données Twitter à l'aide de différents types de

requêtes, notamment des mots-clés, des hashtags, et des profils d'utilisateurs.

2.5 La communication lors d'une situation de crise :

Aujourd'hui l'usage des NTIC est devenu multitâche, que ce soit pour le grand public, soit pour les décideurs. On désigne par le terme multitâche : les différentes manières de l'utilisation des NTIC, dans le premier chapitre, on a abordé le processus de réponse à une situation d'urgence, et on a expliqué quelles sont les difficultés que l'action humanitaire rencontre durant une situation d'urgence, l'avènement des NTIC résout plus ou moins le problème de la communication lors d'une situation d'urgence.

L'usage des TIC comme on a mentionné pour le cas de Twitter permet d'améliorer la coordination entre les gens en temps critique[16], les réseaux sociaux, les blogs, les sites officiels rendent l'information accessible partout et à tout moment. Selon des études menées sur le comportement des gens lors des situations de crise, la majorité se redirige vers le web pour rapporter et discuter leurs observations, expériences et opinions concernant leurs situations actuelles[17], d'autres trouvent des réseaux sociaux tel que Facebook et Twitter le moyen idéal pour échanger l'information en temps réel. De plus, pour Li, Li, Liu, Khan, and Ghani (2014), 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.

En plus de ça, lors d'une situation de crise les gens manifestent des besoins spécifiques, nourriture, abri, don de sang, don d'argent, certains pays développés ont créé des plateformes qui répondent à ces besoins et elles sont intégrées dans leurs plan de gestion de crise, et conseillent vivement leurs citoyens de se rediriger vers les réseaux sociaux pour avoir des informations et des consignes qui peuvent les aider à surmonter les moment de crise, et agir le plus efficacement.

La production, la distribution, la disponibilité en ligne de la masse d'information externe et interne contenues dans les médias sociaux a été reconnue par les décideurs et les chercheurs [2]. Par conséquent, des informations peuvent être précieuses à la fois pour les services

d'urgence et au grand public.

2.6 Caractéristiques des informations issues de la situation de crise

L'usage des TIC pendant une situation d'urgence, permet d'améliorer la communication, échanger l'information en temps réel entre les décideurs, et par conséquent une meilleure coordination entre eux[16]. Cependant, de nouveaux défis sont apparus. Cette masse d'information peut être traitée comme une masse de données initiale à un traitement spécifique car :

a) Cette masse de données est très volumineuse, vu que le nombre d'utilisateur qui se connecte aux réseaux sociaux s'agrandit chaque jour, donc toute information partagée peut être traitée séparément et peut être porteuse de valeur.

b) Elle est difficile à collecter manuellement et à organiser d'une façon cohérente, sans les intégrer dans des outils adéquats.

c) Certaines informations peuvent être d'une valeur informationnelle critique dans le début de processus de réponse à une situation d'urgence, tel que la détection l'évènement qui est en train de se produire.

d) Ces informations telles qu'elles sont, elles sont brutes, donc leur acquisition et leur traitement peut produire d'autres informations qui sont plus précises et plus efficaces pour la gestion de la situation d'urgence.

e) L'importance de toutes ces informations sur le sujet de la connaissance de la situation critique dans le temps, les victimes de catastrophe, l'évacuation des blessés... etc.

D'autre part, tirer parti des médias sociaux pour la conscience de la situation d'urgence reste un défi, les raisons comprennent :

a) le volume d'information échangée lors de a situation d'urgence reste brut, il comprend toute information en relation avec la situation. Cela peut engendrer une difficulté en matière de recherche de l'information adéquate à un moment donné, que ce soit pour les décideurs, ou pour le grand publique et en particulier, pour les gens ayant une relation directe avec la catastrophe (ex : les victimes).

b) comme mentionné, toute les informations sont traitées, donc un autre besoin émerge et qui est le degré de pertinence d'une information. Une information pertinente est très utiles à la fois pour les systèmes qui traitent ces informations, mais aussi pour les décideurs, pour prendre la bonne décision.

2.7 Les différentes applications des TIC durant une catastrophe :

Les technologies de l'information et de la communication jouent un rôle important dans la prévention des catastrophes, la réponse à l'atténuation et le rétablissement. Les agences gouvernementales et les autres acteurs humanitaires impliqués dans les opérations de sauvetage et les processus de prise de décisions ont grandement besoin d'informations opportunes, prévisibles et efficaces. Selon l'Union internationale des télécommunications les TIC pour la gestion des catastrophes peuvent être résumées en quatre principes : multi-aléas, multi-technologies, multi-phases et multi-parties prenantes :

2.7.1 Multi-danger :

Les risques naturels comprennent les tremblements de terre, les cyclones, les inondations, les coulées de boue, les sécheresses, les tsunamis, les éruptions volcaniques et les incendies. Les risques d'origine humaine tel que les guerres, les insurrections, les révolutions politiques, Pour toutes les catastrophes qui font suite à des risques d'origine naturelle ou humaine, les TIC jouent un rôle essentiel en facilitant la circulation d'informations vitales en temps utile.

2.7.2 Multi-technologies :

En atténuant les effets désastreux des dangers, l'utilisation de différentes technologies et réseaux d'information et de communication, y compris par satellite, radio, réseaux mobiles, internet et ses applications, peut contribuer à renforcer les capacités de réduire la vulnérabilité des populations en communiquant des informations utiles.

2.7.3 Multi-phases :

Les télécommunications sont essentielles à toutes les étapes de la gestion des catastrophes : atténuation, préparation, intervention et secours, redressement et réhabilitation.

2.7.4 Multi-acteurs :

La communauté locale, le gouvernement, le secteur privé, les agences de gestion des catastrophes, les organisations météorologiques, la société civile, les agences humanitaires et les organisations internationales devraient assurer l'accès aux TIC pour mieux coordonner les activités de gestion des catastrophes. Les partenariats sont le meilleur moyen de réaliser cette tâche.
Ces applications montrent l'importance de l'élément clé lors de la situation d'urgence qui est l'information en relation directe avec la situation d'urgence, dite l'information situationnelle, cette information se caractérise par son homogénéité et sa source diverse et sous formats différents. Par conséquent le traitement de cette information situationnelle implique l'application des technologies bien étudiées, et orientées informations.

2.8 Les applications de l'informatique dans le traitement de l'information situationnelle

L'information par son hétérogénéité et diversité représente l'élément clé de la recherche dans le domaine de la gestion de crise. De ce fait, ce domaine attire de plus en plus les acteurs de domaine de l'informatique qui actent dans plusieurs disciplines.[18] les énumèrent comme suit :

2.8.1 L'extraction d'information (IE) :

C'est une sorte de découverte de connaissances. Son but est d'extraire automatiquement des informations structurées - généralement présentées dans un tableau ou visualisées dans des chartes graphiques- de documents non structurés

2.8.2 La recherche d'information (IR) :

La recherche d'information est la science de la recherche de documents pertinents pour une requête d'utilisateur. La requête est généralement exprimée sous la forme d'un ensemble de mots-clés et les documents sont renvoyés sous forme de liste hiérarchisée, classés par pertinence décroissante.

2.8.3 Le filtrage d'information (IF) :

Le filtrage d'informations est une fonction permettant de sélectionner des informations utiles ou intéressantes pour l'utilisateur parmi une grande quantité d'informations. En général, les systèmes FI filtrent les données en fonction de la similarité entre un profil d'utilisateur et le contenu textuel d'un événement et du retour d'information de pertinence d'utilisateur, c'est-à-dire si un utilisateur aime un événement devrait également être envoyé à cet utilisateur.

2.8.4 Le Data Mining (DM) :

Aperçu de l'exploration de données : l'exploration de données ou la découverte de connaissances consiste en l'extraction non triviale d'informations implicites, auparavant inconnues et potentiellement utiles, issues d'une vaste collection de données (Han et Kamber, 2006 ; Tan et al., 2005). En pratique, l'exploration de données fait référence au processus global d'extraction de connaissances de haut niveau à partir de données de bas niveau dans le contexte de bases de données volumineuses.

2.8.5 L'aide à la décision :

Les systèmes d'aide à la décision (DSS), sont une classe spécifique de systèmes d'information qui prend en charge les activités de prise de décisions commerciales et organisationnelles. En particulier, un SAD correctement conçu est un système logiciel interactif destiné à aider les décideurs à compiler des informations utiles à partir de diverses sources de données et / ou modèles commerciaux pour identifier et résoudre les problèmes et prendre des décisions.

Le flux informationnel lors d'une situation de crise intègre plusieurs domaines parmi

les cités auparavant. Ces derniers font parties du processus de circulation de l'information depuis la phase de production jusqu'à la phase de consommation. La figure 1.2 est adaptée de schéma proposé par Hristidis et al. (2010).

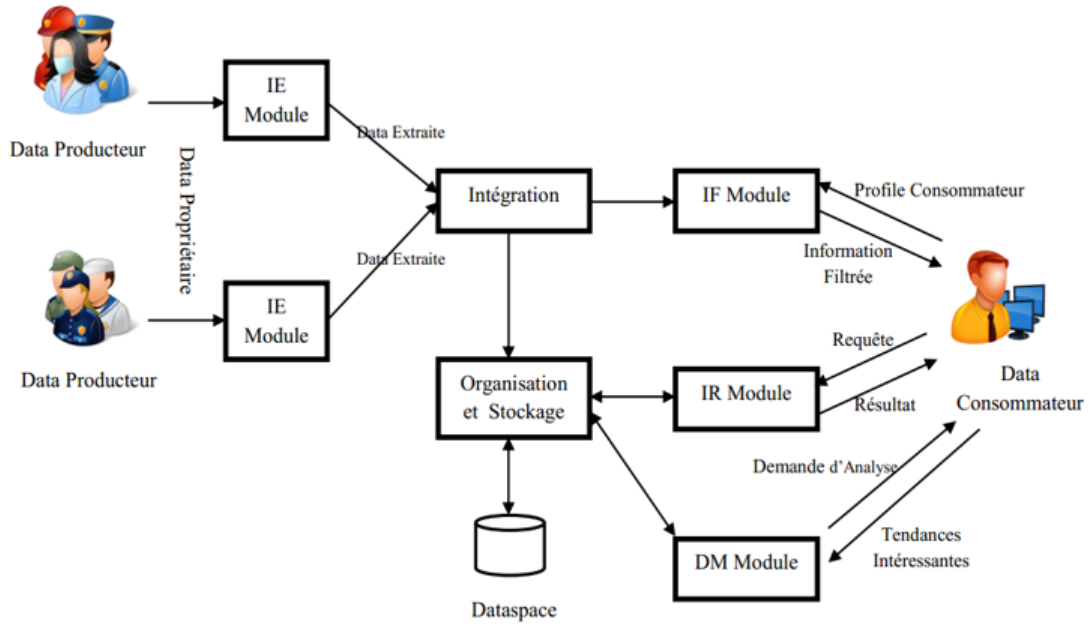

FIGURE 2.2 – Flux d'information en gestion de crise

Dans ce travail nous allons nous intéresser uniquement à l'extraction de l'information, puisque nous allons travailler sur des données textuelles (l'ensemble des informations partagées sur les réseaux sociaux). Nous allons nous appuyer sur le TAL (Traitement automatique de langage), en Anglais NLP pour Natural Langage Processing, qui est d'une manière simpliste le traitement des énoncés linguistiques visant la communication entre les humains à l'aide des machines.

2.9 Conclusion

Nous avons abordé dans ce chapitre les NTICs, qui représentent un moyen efficace pour pallier aux problèmes de communication lors des situations de crise. Elles fournissent des outils et systèmes bien adaptés au besoin des acteurs qui agissent lors de la situation de crise, notamment pour l'accueil et la diffusion de l'information situationnelle en temps réel et la visualisation de ces informations d'une manière simple et facile à comprendre pour ces consommateurs.

Cependant, le volume de ces informations pourrait être considérable, et difficile à interpréter dans une durée de temps courte et sous stress. De plus, ces informations peuvent être à l'origine d'autres informations, qui doivent être extraites. Par conséquent, un autre besoin émerge, et qui est le traitement de ces informations d'une façon à obtenir d'autres informations qui permettent à mieux comprendre la situation de crise.

Dans le chapitre suivant, nous allons aborder le Traitement Automatique des Langues-TAL, ses domaines d'application, ses étapes de traitement, etc. Ensuite, nous allons appliquer certains de ses algorithmes sur un ensemble de données extraites lors d'une situation de crise. Comme nous allons traiter et extraire à partir des données textuelles échangées entre humains, ces données sont produites dans un langage de communication, et cela représente le domaine d'intérêt du TAL.

Chapitre 3

Extraction d'information pour améliorer la situational awarness

3.1 Introduction :

Dans le chapitre précédent, nous avons parlé de Les NTIC et la réponse dans la situation d'urgences et à partir de les Caractéristiques des informations issues de la situation de crise Nous allons fournir dans ce chapitre la méthode d'extraction des informations avec leur traitement. Pour cela, nous nous concentrerons sur le TAL et l'extraction d'information depuis Twitter, ensuite, nous présenterons la problématique qui résume les différents problèmes qui empêchent les intervenants dans une situation de catastrophe à bien agir. ensuite nous présenterons notre approche qui consiste en déférents processus commençant par la collection des informations à partir d'un d'un dataset déjà collecté , puis en enchaine par le nettoyage des tweets et par la suite le traitement des tweets avec l'algorithme word2Vec, qui nous permet de regrouper les mots similaires présents dans les tweets dans des classes sémantiques.

3.2 Définition de traitement automatique du langage(TAL) :

Selon [19], le TAL ne peut pas avoir une définition convenable qui satisfait tout le monde, étant un domaine de recherche et de développement très actif, il combine à la fois un ensemble de théories et un ensemble de technologies pour aboutir à une approche informatisée de l'analyse de texte.définit le traitement du langage naturel comme une gamme de techniques de calcul théoriquement motivées pour analyser et représenter des textes naturels

à un ou plusieurs niveaux d'analyse linguistique dans le but de réaliser un traitement du langage de type humain pour une série de tâches ou d'applications.
Une définition plus concise atteste que le traitement automatique du langage naturel ou des langues (TAL) ou des langues naturelles (TALN) est une discipline à la frontière de la linguistique, de l'informatique et de l'intelligence artificielle, qui concerne l'application de programmes et techniques informatiques à tous les aspects du langage humain. Ainsi parfois appelée ingénierie linguistique.

3.3 Domaines d'application de TAL :

Le traitement du langage naturel fournit à la fois la théorie et les implémentations pour une gamme d'applications. En fait, toute application utilisant du texte est considérée comme application de TAL.[19] énumère les applications utilisant le TAL :

3.3.1 la Recherche d'information :

la recherche d'information est un processus visant à répondre à une requête d'utilisateur formée généralement à l'aide de texte – en interrogeant un moteur de recherche - avec une liste des documents plus ou moins pertinents, qui correspondent aux besoins exprimés dans la requête.

3.3.2 la Réponse aux questions :

Contrairement à la recherche d'informations, qui fournissent une liste de documents potentiellement pertinents en réponse à la requête d'un utilisateur, la réponse aux questions fournit à l'utilisateur soit le texte de la réponse lui-même, soit des passages de réponse.

3.3.3 le résumé des documents :

les niveaux les plus élevés du TAL, en particulier le niveau du discours, peuvent permettre à une implémentation de réduire un texte plus grand en une représentation narrative abrégée du document original, plus courte mais riche.

3.3.4 La traduction automatique :

Considérée la plus ancienne de toutes les applications TAL, différents niveaux de TAL ont été utilisés dans les systèmes de traduction automatique, allant de l'approche «basée sur des mots» à des applications comportant des niveaux d'analyse plus élevés.

3.3.5 les systèmes de dialogue :

peut-être l'application omniprésente du futur, dans les systèmes envisagés par les grands fournisseurs d'applications d'utilisateur final. Les systèmes de dialogue, qui se concentrent généralement sur une application étroitement définie (par exemple, votre réfrigérateur ou votre système audio domestique), utilisent actuellement les niveaux phonétiques et lexicaux du langage.

3.3.6 l'extraction d'informations (IE) :

un domaine d'application plus récent. L'IE se concentre sur la reconnaissance, l'étiquetage et l'extraction dans une représentation structurée de certains éléments clés de l'information, par ex. personnes, entreprises, lieux, organisations, provenant de grandes collections de texte. Ces extractions peuvent ensuite être utilisées pour une gamme d'applications y compris la réponse aux questions, la visualisation et l'exploration de données.
Dans notre travail, nous n'allons pas nous intéresser aux 5 cinq premiers domaines, mais juste au dernier qui est l'extraction d'information.

3.4 Les différents niveaux de langage :

Dans la définition de [19], on constate la présence de certains aspects qu'on doit prendre en compte pour mener à bien un processus TAL. L'analyse du langage nécessite une connaissance de sa structure sur de nombreux niveaux : que sont les mots à utiliser ? Quelle est leurs significations ? Comment avoir une phrase ayant un sens par la combinaison de ces mots ? Comment ces mots contribuent-ils au sens de la phrase ? De la réception des sons (ou leur prononciation) jusqu'à la compréhension approfondie des mots prononcés dans l'environnement ou ils sont prononcés, les linguistes distinguent des niveaux de connaissance lorsque l'on aborde l'analyse automatique de la langue,[19] les détaille comme suit :

3.4.1 la Phonologie ou la phonétique :

Ce niveau traite de l'interprétation des sons de la parole dans et entre les mots. Il existe en fait trois types de règles utilisées en analyse phonologique :

1. les règles phonétiques - pour les sons à l'intérieur des mots.
2. règles phonémiques - pour les variations de prononciation lorsque les mots sont prononcés ensemble.
3. règles prosodiques - pour la fluctuation du stress et de l'intonation sur une phrase.

3.4.2 la morphologie :

Ce niveau traite la forme des mots (de leur flexion – indications de cas, genre, nombre, mode, temps, etc. – de leur dérivation – préfixes, suffixes, infixes – et de leur composition – mots composés). Sous l'appellation de morphosyntaxe, elle représente également l'étude des règles de combinaison des morphèmes (unités minimales de sens) selon la configuration syntaxique de l'énoncé.

3.4.3 le lexique :

À ce niveau, les humains, de même que les systèmes TAL, interprètent la signification des mots individuels. Plusieurs types de traitement contribuent à la compréhension au niveau des mots.

3.4.4 la syntaxe :

Ce niveau se concentre sur l'analyse des mots dans une phrase afin de découvrir la structure grammaticale de la phrase. Cela nécessite à la fois une grammaire et un analyseur. La sortie de ce niveau de traitement est une représentation (éventuellement délimitée) de la phrase qui révèle les relations de dépendance structurelle entre les mots.

3.4.5 la sémantique :

C'est le niveau auquel la plupart des gens pensent que le sens est déterminé, cependant, ce sont tous les niveaux qui contribuent au sens. Le traitement sémantique détermine les significations possibles d'une phrase ou comment les mots font du sens lorsqu'ils sont insérés dans une phrase.

3.4.6 la Pragmatique :

Ce niveau concerne l'utilisation délibérée du langage dans des situations et utilise le contexte au-delà du contenu du texte pour comprendre. Le but est d'expliquer comment le sens supplémentaire est lu dans les textes sans y être réellement encodé. Ou comment les phrases peuvent être interprétées selon leur pragmatique contexte d'énonciation (interlocuteurs, phrases précédentes, connaissance commune du monde, . . .)

3.5 La sémantique :

Un apport sémantique peut servir à réduire les ambiguïtés syntaxiques, à mieux cibler des concepts (par exemple en recherche d'information), mais sa finalité globale est de représenter formellement l'information véhiculée par un énoncé et éventuellement d'en inférer de nouvelles connaissances ou une réponse à la question posée (si l'énoncé est une question). Le dictionnaire français Larousse définit la sémantique comme l'étude qui traite le sens des mots, des phrases ou un ensemble des phrases.

3.5.1 la sémantique distributionnelle (SD) :

La SD comme [20] l'a défini est un champ de recherche en TAL, qui fonde le calcul sémantique sur une représentation vectorielle des contextes des mots, selon cette hypothèse les mots sont sémantiquement proches s'ils apparaissent dans des contextes similaires. L'usage des mots dans une phrase est l'idée fondatrice de cette sémantique, considéré à partir des emplois des mots dans des corpus.
Cette hypothèse remonte à Harris (1954). Ce dernier avait signalé que si des mots peuvent se retrouver dans le même contexte, ça signifie qu'ils sont proches sémantiquement. Ainsi, l'hypothèse démontre que la similarité des sens est liée à la distribution. Si nous prenons l'exemple : « Les topinambours sont dans le panier à légumes » et « J'ai mis des topinambours dans la soupe », même si nous ne connaissons pas le mot « topinambour», nous arrivons tout de même à deviner qu'il s'agit d'un légume[21]. L'idée de la sémantique distributionnelle consiste donc à projeter ces processus cognitifs par ordinateur.
Deux familles de méthodes existent pour traiter la sémantique distributionnelle[20] :

a) **Méthode classique :** modèles explicites fondés sur le décompte des contextes 4 étapes :

- Extraction des contextes.
- Décompte et pondération.
- (éventuellement) Réduction de dimensions.
- Calcul de similarité.

b) **Modèles prédictifs :**

- Procédure d'apprentissage de la représentation vectorielle : le système apprend à assigner des vecteurs similaires à des mots similaires.
- Utilisation d'un réseau de neurones.
- Représentations compactes, de dimension réduite, construites à partir de traits latents Outil word2vec (Mikolov et al. 2013) et (Goldberg and Levy 2014).

3.6 L'extraction d'information :

L'extraction d'information représente l'activité qui consiste à remplir automatiquement une banque de données à partir de textes écrits en langue naturelle. (T. Poibeau). Selon [22] La tâche d'extraction d'informations consiste à identifier les instances d'une classe particulière d'entités prédéfinies, de relations et d'événements dans des textes en langage naturel, et à extraire les propriétés (arguments) pertinentes des entités, relations ou événements identifiés. D'où l'extraction d'informations est une approche guidée par son but d'identifier les occurrences d'événements particuliers ; en extraire les arguments impliqués ; en donner une représentation structurée de résultats[23]. Dans l'extraction d'informations, seule une partie du texte est considérée et traitée, généralement 10% à 20% du texte[23].

3.6.1 le processus d'extraction d'informations

La figure suivante montre l'architecture d'un système d'extraction d'informations simple. Il commence par traiter un document en utilisant plusieurs des procédures de segmentation. Premièrement, le texte brut du document est divisé en phrases à l'aide d'un segmenteur de phrases, et chaque phrase est subdivisée en mots à l'aide d'un tokenizer. Ensuite, chaque phrase est étiquetée avec des balises de partie de discours (POS part of speech), la détection d'une entité nommée. Dans cette étape, nous recherchons des mentions d'entités potentiellement intéressantes dans chaque phrase. Enfin, la détection de relation pour rechercher les relations probables entre différentes entités du texte. Cette architecture est une simple

architecture qui décrit les étapes les plus fréquentes, ça n'empêche pas de rajouter une étape qui nous paraît utile, ou éliminer une étape qui ne semble pas nécessaire dans notre traitement.

3.6.2 les tache de l'extraction d'information :

A partir de ce processus décrit dans la partie précédente on peut extraire les taches de l'extraction d'informations. L'application de l'extraction d'informations sur du texte vise à créer une vue structurée, c'est-à-dire une représentation des informations compréhensibles par une machine. [22] les résume en ce qui suit :

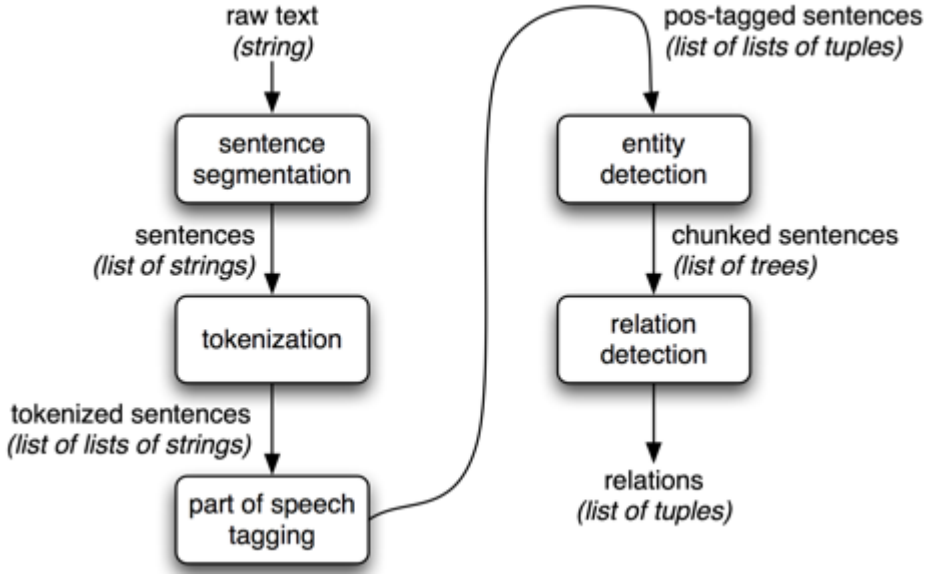

FIGURE 3.1 – processus d'extraction d'information

a) **Reconnaissance d'entité nommée (Named Entities Recognition) :** se focalise sur le problème de l'identification (détection) et de la classification des types prédéfinis d'entités nommées, telles que l'Organisation des Nations Unies, les personnes, les noms de lieux, les expressions temporelles (par exemple, «1er septembre 2011»), les expressions numériques et monétaires (par exemple, «20 millions Euros), etc. La tâche NER peut également inclure l'extrait d'informations descriptives du texte sur les éléments détectés en remplissant une petite échelle modèle.

b) **Résolution de coréférence (CO) :**nécessite l'identification de multiples mentions (co-référentes) de la même entité dans le texte. Les mentions d'entité peuvent être :

- **Nommées :** au cas où une entité est désignée par son nom ; Par exemple, «General Electric» et «GE» peuvent se référer à la même entité du monde réel ;
- **Pronominal :** au cas où une entité est désignée par un pronom ; par exemple, dans «Mohamed a acheté un pantalon. Mais il l'a oublié dans le magasin» .Le pronom « il » se réfère à Mohamed ;
- **Nominal :** dans le cas où une entité est désignée par une phrase nominale ; Par exemple, dans «Sonatrach a révélé ses revenus. La société a également dévoilé ses projets futurs. » la phrase nominative La société fait référence à Sonatrach ;
- **Implicite :** dans certains langages la référence ne nécessite pas une réalisation explicite.

b) **Relation Extraction :** est la tâche de détection et de classification des relations prédéfinies entre les entités identifiées dans le texte.

d) **Extraction d'événement :** fait référence à la tâche d'identification d'événements dans du texte libre et en tirant des informations détaillées et structurées à leur sujet, identifiant idéalement qui a fait quoi à qui, quand, où, par quelles méthodes (instruments), et pourquoi. Généralement, l'extraction d'événement implique l'extraction de plusieurs entités et relations entre elles.

3.7 L'extraction d'information depuis Twitter :

Les tweets liés aux catastrophes sont extraits à l'aide de lexiconbased[24] ou des méthodes basées sur la localisation. L'ancien utilise un ensemble de mots-clés générés par des experts et ce dernier recueille tous les tweets associés à un emplacement spécifique. Tweets filtrés à l'aide de mots clés ne sont qu'une fraction de toutes les catastrophes, et les tweets avec des informations de localisation sont assez rares.

Les deux méthodes manquent de complétude et ont une faible couverture. Une façon d'augmenter la visibilité des tweets liés aux catastrophes étend les ensembles de mots-clés définis par des experts. Par exemple, CrisisLex " est un lexique qui augmente la portion des tweets liés aux catastrophes capturés à partir de Twitter API de streaming.L'effort est de trouver un ensemble des mots-clés qui sont largement utilisés dans différentes catastrophes (ouragan, tornade, bombe, bombardement et explosion).

3.7.1 But :

Le but n'est pas de répondre à une question en particulier, mais de parcourir l'ensemble des hashtags pour découvrir quels types d'informations concernant un sujet ou un domaine sont présents.

- Ne cherche plus à comprendre les textes dans leur ensemble.
- Vise à extraire d'un texte donné des éléments pertinents.
- Identifier les occurrences d'événements particuliers.
- En extraire les arguments impliqués.

3.7.2 Pourquoi nous avons choisi Twitter pour extraire des informations durant la situation de crise ? :

Parmi tous ces médias sociaux, nous avons choisi la plateforme sociale Twitter qui est particulièrement intéressante grâce à son utilisation en forte croissance pour plusieurs raisons :

- il s'agit d'un réseau social très populaire dont le nombre d'utilisateurs a connu une augmentation récente très importante ;
- de par son format de message court, il oblige les rédacteurs à adopter un style très synthétique tout en leur permettant d'inclure des liens vers les sources d'origine.
- il est particulièrement bien adapté à la diffusion et à la propagation d'information.
- Twitter offre des applications pour les développeurs, pour extraire facilement les tweets,
- permet de connaître les sujets qui intéressent les utilisateurs ainsi que leurs réactions.

3.7.3 Pourquoi extraire les Hashtags et pas les Tweets ?

Le hashtag permet de retrouver les posts parlant de ce sujet plus facilement et de découvrir les nouveaux articles liés à cette tendance et catégoriser le contenu. Quand un hashtag est utilisé, le réseau social indexe celui-ci. En cliquant dessus, les autres publications, utilisant le même hashtag, sont affichées. Quand un mot-clé est ainsi beaucoup utilisé, il peut apparaître dans les tendances nationales de Twitter. Notons que ces tendances sont personnalisées pour chaque utilisateur, selon son lieu de résidence et les comptes qu'il suit. Un tweet avec un hashtag a deux fois plus d'engagements que sans hashtag. Il a aussi

droit à plus de 50% de retweets supplémentaires. En pratique, Twitter a un vocabulaire propre, les utilisateurs du service de microblogage créent de nombreux hashtags sous forme d'abréviations.

3.8 Problématique :

Depuis quelques années, le microblogging est devenu une forme de communication très populaire qui attire de plus en plus d'utilisateurs grâce à la facilité et à la rapidité du partage des informations. Chaque jour, des millions de mises à jour sont affichées sous forme de courts messages texte sur des services de microblogging, tels que Twitter. La taille des tweets est limitée par un nombre maximal de caractères (280 caractères). Cette limitation entraîne l'utilisation d'un vocabulaire spécial qui n'est généralement pas utilisé, bruyant et plein de nouveaux mots. En effet, le but est de partager le maximum d'informations en quelques caractères. Il peut donc être difficile de comprendre la signification d'un message textuel court sans connaître le contexte général de sa réalisation. Ce problème de contrainte est, par exemple, un cas fréquent sur la plate-forme de Twitter.

Pour les utilisateurs finaux et les analystes de données, il est difficile de parcourir des millions de tweets contenant beaucoup de bruit et de redondance. De plus, un tweet étant court et dépourvu d'informations contextuelles suffisantes, il est souvent difficile de comprendre les informations associées. Toutes ces difficultés empêchent les utilisateurs de comprendre ou de consommer efficacement des informations.
Dans ce travail, on fait le point sur l'extraction d'information afin d'améliorer la Situational Awareness dans les situations d'urgence. Qui est à l'intersection de plusieurs disciplines : recherche d'information (RI), ingénierie Des connaissances (IC) et traitement automatique des langues (TAL).

Nous avons conclu de nombreux problèmes, notamment :

- les défis associés au volume de renseignements et à l'accès à l'information durant une situation de crise
- la difficulté à trouver des renseignements, des ressources et des actions disponibles immédiates sur le terrain ;
- l'incapacité de déterminer la gravité d'une situation précise en temps réel

- la difficulté à partager l'information entre plusieurs acteurs impliquant dans la situation de crise.
- l'incapacité à agréger, à rechercher, à vérifier les données des médias sociaux, à cause du volume considérable d'information produite.
- L'information non ciblée source d'une mauvaise coordination.

3.9 Notre approche :

Afin de répondre aux problèmes précédemment posés, nous allons nous appuyer sur l'extraction d'information depuis le réseau social Twitter, pour ces caractéristiques avantageuses par rapport aux autres réseaux sociaux, l'unité de base de notre travail sont les tweets extraits pendant une situation de crise et notamment leurs contenus textuels. Par la suite ces tweets seront stockés dans des data-set à fin de les traiter. Le traitement des tweets consiste essentiellement de les classifier selon des classes bien définies. Cette classification dépend de contexte de tweet, et pour déterminer le contexte nous allons nous appuyer sur l'analyse distributionnelle des tweets. L'analyse distributionnelle nous permet de construire des représentations vectorielles qui permettent de capturer la sémantique exprimée par chacune des occurrences. Par la suite, ces représentations peuvent être utilisées pour déterminer le contexte de tweet, ce qui permet le classifier.
La classification des tweets passe par deux étapes :

a) L'extraction des dictionnaires des classes (bag of words) : ici, on construit un sac de mots pour chaque classe à fin de l'utiliser dans l'étape suivante qui est le calcul de similarité. Le sac de mot est un ensemble des mots représentatifs de la classe. Nous allons voir certaines illustrations de sac de mots dans le prochain chapitre.

b) Le calcul de similarité : après l'extraction des dictionnaires de chaque classe, on procède au calcul de la similarité entre le tweet et le dictionnaire de chaque classe.

3.9.1 Dataset :

Pour ce travail nous avons travaillé avec un ensemble de dataset issu de la communication entre tweeters pendant quelques catastrophes qui ont marqué le monde par leurs gravités.

Ces datasets sont disponibles sur le site https ://crisislex.org [25], sous réserve de citer la source.

Détails de dataset :

- **Contenu :** 60K tweets postés lors de 6 événements de crise en 2012 et 2013. Ouragan Sandy en 2012, Inondations en Alberta en 2013, Attentats à la bombe de 2013 à Boston, la tornade 2013 en Oklahoma, Inondations du Queensland en 2013, Explosion du Texas de l'Ouest en 2013.

- **Méthode d'échantillonnage :** 10 millions de tweets au total échantillonnés par mots-clés et régions géographiques ou coordonnées.

- **Étiquettes :** 60 000 tweets (10 000 dans chaque collection) ont été étiquetés par les travailleurs du crowdsourcing en fonction de leur parenté (comme "on-topic" ou "off-topic").

- **Format de données :** fichiers de valeurs séparées par des virgules (.csv) contenant le texte des tweets et des étiquettes pour les libellés, nous avons la changé en valeurs séparées par tabulation (.tsv).
  Dans notre travail, nous allons utiliser les 6 six datasets pour la construction et l'entrainement de modèle, et 1 un pour l'évaluation.

Échantillon de dataset :

Dans notre travail la classification se fait sur la base de l'approche cluster, qui regroupe un ensemble de groupes sectoriels (organisations humanitaires, ONG. . . etc) impliqués dans la gestion de la situation d'urgence, et qui vise à la répartition des tâches et responsabilités selon le domaine d'activité de chacun des groupes. Nous illustrons quelques une des classes via des tweets dans le tableau 2.1 :

Tweet id	Tweet text	Classe
263298821189156865	I don't know how I'm getting back to Jersey since the trains and subways aren't running...	Transport
262767536540643329	@codyfinz my house is creeking... Does that mean she's trying to break in?	Rescue
262679884898381825	I've never waited in a line this long @duanereade but I need milk for my bread pudding! #blameSandy	Food
336932095853215745	RT @EricFrancis: Best stat from #YYCflood is that 100,000+ have been evacuated and only 1,500 needed shelter. The rest were taken in by fel…	Shelter
336823557759856640	Telecommunications Blackout - Flood update 04 - Queensland's telecommunications blackout is now almost... http://t.co/7RFGKgCL	Telecommunication
323895195861151744	2 explosions in Marathon Sports store near finish line. 2 dead and 23 injured so far. Stay safe everyone!	Health
262937571376627714	Help on the way for flood-stricken First Nations: Pickup trucks full of bottled water,and other... http://t.co/32NBdjDJaB	Water
337620411506376704	RT @USATeducation: Thoughts go out to those children in #Oklahoma.	Education
350611174754230272	RT @CBCAlerts: PMO:If guns are seized in High River they must be returned - RCMP should focus on protection of life, property #abflood	Protection
350050730985529345	,"Rutland Park is also coordinating volunteers to help out. Visit: https://t.co/SXXt4PNYA3 for more details. #yycflood #yycW11	Coordination
263118055389925376	RT @redcrosscanada: We are accepting financial donations to assist people affected by flooding in Alberta: http://t.co/Xd9mbM2uj8 #abflood	Donation

TABLE 3.1 – Échantillon de dataset

Autre alternative : (Extraction des tweets directement depuis Twitter) :

Twitter fournit 3 trois variantes d'APIs à fin d'accéder à ces données, API REST pour les données de base de Twitter, API recherche qui permet aux développeurs d'accéder à Twitter Search et une API de streaming qui permet d'accéder au flux publique de Twitter en temps réel. Ce dernier est très intéressant, il permet de collecter les tweets circulant dans Twitter en temps réel et les traquer selon le besoin, souvent nous utilisons le hashtag pour traquer un sujet bien déterminé, donc en cas de situation de crise, il faut juste traquer le hashtag qui parle de celle-ci et par conséquent tout tweet parlant de cette situation sera collecté.

3.9.2 La sélection des tweets étiquetés "on-topic" :

Les dataset contiennent des tweets étiquetés "on-topic" pour les tweets en relation avec la catastrophe, et "off-topic" pour le cas contraire, cet étiquetage a été fait par des être humains, ce qui garantit une grande fiabilité, après il vient la tâche de chargement des tweets ligne par ligne vers un fichier .txt, le chargement ne concerne que le texte de tweet.

3.9.3 La tokenisation et élimination des mots vides :

c'est la méthode de décomposition du texte en composants plus petits (mots, phrases, bigrammes, etc),

- Mot et Ponctuation, divisera le texte en mots et conservera les signes de ponctuation.
  Un exemple. $\longrightarrow$ (un), (exemple),(.)
- Par espace divisera le texte par des espaces uniquement.
  Un exemple . $\longrightarrow$ (Un), (exemple.)
- La phrase divisera le texte par arrêt complet en ne conservant que les phrases complètes.
  Cet exemple. Un autre exemple. $\longrightarrow$ (Cet exemple.), (Un autre exemple.)
- Par expression régulière : divisera le texte par expression fourni. Il se divise par mots uniquement par défaut.

Nous nous s'intéressons à la décomposition par mot et ponctuation.

3.9.4 lemmatisation :

obtention de la forme canonique (le lemme) à partir du mot

-) Pour un verbe : sa forme à l'infinitif.

-) Pour un nom, adjectif, article, ... : sa forme au masculin singulier

3.9.5 la Racinisation (Stemming) :

En linguistique, la racinisation ou désuffixation (anglais : stemming) est un procédé de transformation des flexions en leur radical ou racine (anglais : stem). La racine d'un mot correspond à la partie du mot restante une fois que l'on a supprimé son préfixe et son suffixe, à savoir son radical. Contrairement au lemme qui correspond à un mot réel de la langue, la racine ne correspond généralement pas à un mot réel. Par exemple, le mot « chercher » a pour radical « cherch » qui ne correspond pas à un mot réel. Par contre dans l'exemple de «frontal», le radical est «front» qui lui l'est. Plusieurs variantes d'algorithmes de racinisation existent pour l'Englais. Notre approche utilise le stemmer de Porter. Après ces étapes on obtient un fichier .txt qui contient tout les tweets qui sont prêts au traitement.

3.9.6 Concevoir un modèle word2vec :

C'est un modèle pouvant être classé dans les applications de l'apprentissage pro-fond en traitement automatique de la langue (TAL) (Deng et Yu, 2014). Proposé par Mikolov et al, dans (Mikolovet al., 2013b) et repris par la suite par Goldberg dans une autre variante (Goldberg et Levy, 2014), ce modèle nous permet de construire des représentations continues des mots d'un tweet, en se basant sur la notion de contexte auquel appartient chaque mot contenu dans le tweet. Il permet également de construire des représentations vectorielles permettant de capturer la sémantique exprimée par chacune des occurrences. Par la suite, ces représentations peuvent être utilisées pour comparer les termes entre eux et exprimer, sous forme de distances, les relations entre les termes.

Word2vec fait ce que l'on appelle de l'apprentissage de représentation, plus précisément de

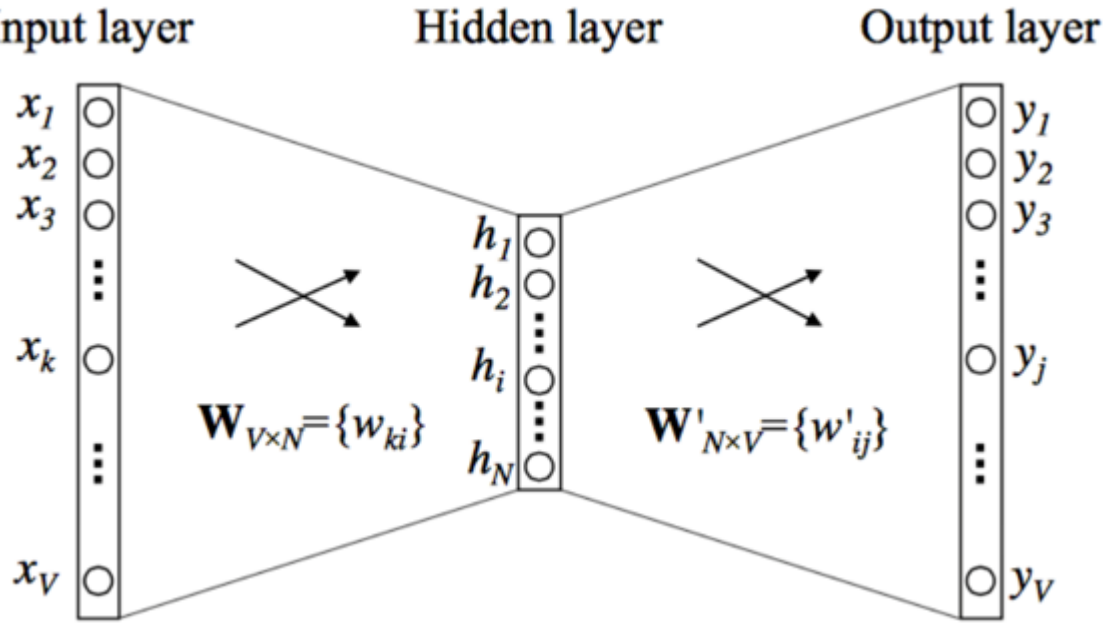

FIGURE 3.2 – architecture générale de word2vec

distribution de mots. Ce n'est pas ce que l'on peut appeler de l'apprentissage profond, il s'agit de réseaux de neurones peu profonds, car il ne possède qu'une seule couche cachée. Cet outil propose deux architectures possibles : nous avons d'une part, celle dite en « sac-de-mots » continus (continuous bag of words, CBOW) et d'autre part l'architecture en Skip-gram [21]. La figure 2.2 illustre une architecture générale de word2vec tel qu'elle décrite dans les travaux de Mikolov et al. (2013a).

Les réseaux de neurones artificiels Word2vec :

Word2vec propose deux réseaux de neurones artificiels simples (i.e. peu profonds) : l'architecture CBOW et l'architecture Skip-gram. Ces architectures ont besoin, pour s'entrainer, de mots centraux (ou mots d'attention) et leurs fenêtres de contexte respectifs. Une fenêtre de contexte correspond aux n mots précédents et n mots suivants un mot central. La valeur de n est à adapter selon la tâche et les données. Chacune de ces architectures est composée de trois couches. Une couche d'entrée, une couche cachée et une couche de sortie. Le figure 2.3 présente l'architecture CBOW et la figure 2.4 l'architecture Skip-gram pour un tweet représenté par : "w2 w1 wc w+1 w+2", et dont le mot central est wc. La couche d'entrée contient un sac-de-mots contenant la fenêtre de contexte pour le CBOW ou le mot central pour le Skip-gram. La couche cachée contient la projection de l'entrée dans la ma-

trice globale des poids. La couche de sortie est la prédiction du modèle. Soit un mot pour l'architecture CBOW, et un contexte pour l'architecture Skip-gram. Cette prédiction est uniquement utilisée pour calculer l'erreur des réseaux et la rétro-propagation du gradient. Cette rétro-propagation permet de corriger la matrice globale en rapprochant dans l'espace multi-dimensionnel les mots de leurs contextes respectifs[25]

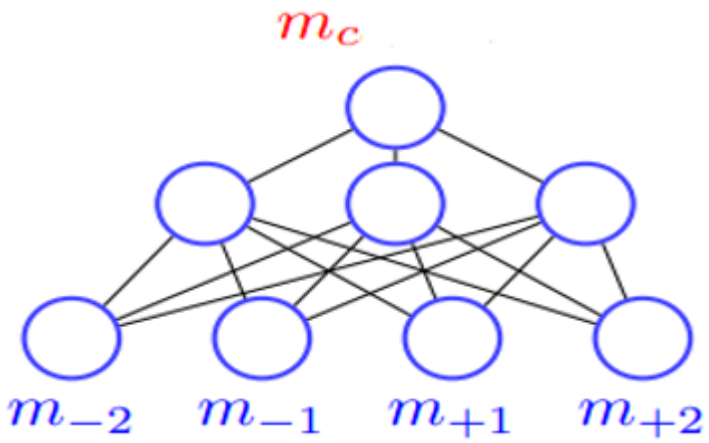

FIGURE 3.3 – Réseau de neurones CBOW

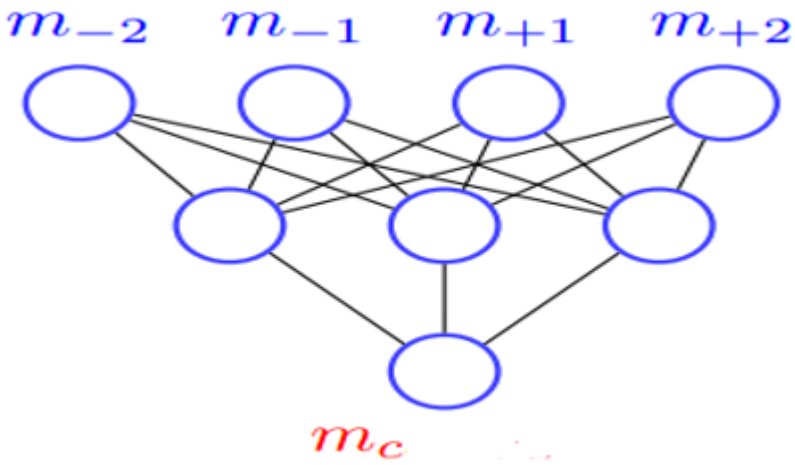

FIGURE 3.4 – Réseau de neurones Skip-gram

Skip gram :

Cette architecture est celle qui retiendra le plus d'attention, car elle fournit les meilleurs résultats, elle est plus adaptée pour les représentations sémantiques et les mots peu fréquents, que l'autre architecture, ce modèle Skip-Gram va en fait prédire le contexte d'où est issu un

mot. Pour ce faire, dans la couche d'entrée, nous avons un vecteur contenant un seul mot. La couche d'entrée du réseau ne contient donc que la représentation en sac-de-mots binaire du mot au cœur du contexte. Ce mot est projeté dans la matrice de poids globale, puis transmis à la couche de sortie qui va prédire un mot. Cette prédiction est ensuite corrigée par rétro-propagation pour chacun des mots de la fenêtre de contexte. Un réseau Skip-gram maximise la vraisemblance suivante :

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j=t-c;j\neq t}^{t+c}\log p(\omega_j|\omega_t)$$

CBOW :

L'architecture du CBOW est un réseau de neurones artificiels simple et log-linéaire. La couche d'entrée de ce réseau de neurones utilise des "sacs-de-mots" binaires représentant une fenêtre de contexte. Dans cette configuration, les vecteurs d'entrée sont des vecteurs de la taille du vocabulaire avec un 1 dans la colonne i si le mot i est présent dans le document, 0 sinon. Chaque mot est alors projeté dans la matrice globale, l'ensemble des représentations est ensuite additionné pour former une unique couche cachée. Cette couche cachée passe par la couche de sortie et les fonctions d'activation type "Softmax" tentent de prédire le mot au coeur de la fenêtre. L'erreur de prédiction est ensuite utilisée pour corriger les matrices de poids via une rétro-propagation de gradient. Cette architecture essaie de maximiser la vraisemblance ci-dessous :

$$\frac{1}{T}\sum_{t=1}^{T}\log\ p(\omega_t|\omega_{t-\frac{c}{2}}\ldots\ldots\omega_{t+\frac{c}{2}})$$

dans laquelle T est la taille des données d'apprentissage et c est la taille maximum de la fenêtre de contexte. L'architecture CBOW est plus efficiente que Skip-gram et capture une meilleure représentation des mots fréquents (Mikolovet al.,2013a).

Dans notre travail le vocabulaire d'apprentissage c'est l'ensemble des mots des tweets traités au préalable

3.9.7 Entrainement de modèle :

pour entrainer le modèle, nous avons choisi les deux méthodes CBOW et Skip-gram. Comme tout modèle d'apprentissage, elles nécessitent certain paramètre qui doit être choisis judicieusement pour avoir un résultat satisfaisant. L'entrainement du modèle génère un fichier plus ou moins grand, selon la taille de vocabulaire, qui peut être sauvegardé, pour une éventuelle évaluation.

3.9.8 Classification :

comme on a mentionné précédemment, nous avons conservé un dataset parmi les six pour tester notre modèle. Nous l'avons soumis au même traitement que l'autre dataset à savoir la tokenisation, l'élimination des mots vides, la lemmatisation et le stemming afin de garantir une forte cohérence entre les données d'apprentissage et les données de test.
Nos classes sont issues de l'approche Cluster (les groupes sectoriels de l'approche cluster). L'approche Cluster vise à assurer une réponse stratégique et multi sectorielle au travers de la mobilisation des différents acteurs aux problématiques liées à la situation de crise. Cette approche filtre les activités d'urgence selon les classes suivantes [26] :

- **Rescue And Search**
- **Health**
- **Shelter**
- **Donation**
- **food**
- **transport and logistic**
- **Coordination**
- **Protection - security**
- **Telecom/ICT**
- **Water**
- **Education**

Chaque mot du vocabulaire, c'est-à-dire le texte que nous sommes en train de classer, va être transformé en un vecteur unique pouvant être ajouté, soustrait et manipulé de différentes manières, comme un vecteur dans l'espace. La figure 3.5 représente un vecteur de n dimensions du mot 'health'.

```
[ 0.15215547  0.10465616  0.46609622 -1.1655252  -0.32303596 -0.650485
 -0.27771088  0.04363202 -0.36504525  0.00440468  0.17100409 -0.13546959
 -0.11061854 -0.06453771  0.01972109 -0.72393715 -0.5038698  -0.14289306
  0.6572167   0.13147885  0.55005693  0.20425902 -0.19533464  0.37007114
 -0.8882439  -0.6611291  -0.2886494  -0.34479073  0.11148024 -0.49058157
  0.03616538 -0.23720387 -0.6211895   0.27442363 -0.07722954  0.24752021
 -0.08779252  0.4641448   0.39932302  0.11038842 -0.78841615 -0.42128742
  0.33020893 -0.29365632 -1.1124238  -0.5169877  -0.03829991 -1.0835754
 -0.23039971  0.70913684  0.40504965 -0.19966702  0.15354596  0.4412037
  0.78137845  0.08901598 -0.4885829  -0.311962    0.25172928  0.9677682
 -0.13675103 -0.33137795  0.3144261   0.16901407  0.40023142 -0.05526842
 -0.37032372 -1.063859    0.59039336 -0.13268171  0.08616529  0.0498804
 -0.12973294  0.47787082 -0.00137754  0.14852935  0.16794483  0.7243286
  0.13524786 -0.61629283  0.36917442 -0.00632767  0.42099643  0.34912384...
```

FIGURE 3.5 – Un vecteur de n dimension

Donc pour construire le sac de mots de chaque classe, la similarité des vecteurs (souvent on utilise le cosinus) est un moyen efficace pour déterminer à quel point un mot n est sémantiquement proche de la classe, comme le montre la figure 2.5

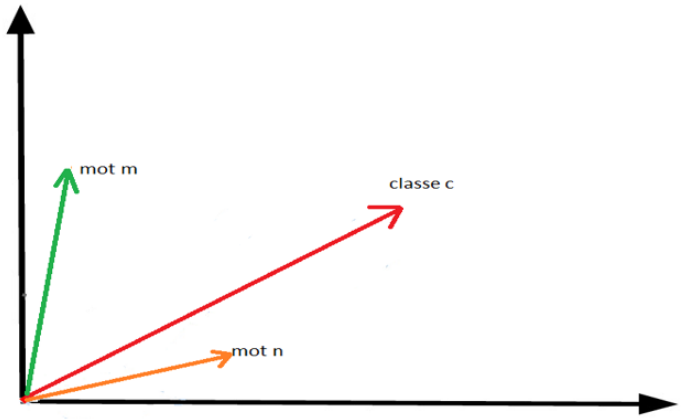

FIGURE 3.6 – Un vecteur de n dimension

Prenons l'exemple de mots n et m et la classe c :
$sim(c,n) = cos(\vec{(c)}, \vec{(n)}) sim(c,m) = cos(\vec{(c)}, \vec{(m)})$ cette figure montre que le vecteur de mot n est plus similaire au vecteur de la classe c que le vecteur de mot m . Par conséquent, la probabilité de retenir le mot n est plus grande que celle de m. Au final, notre sac de mot se compose des mots ayant une grande similarité, par conséquent les plus sémantiquement proches à la classe donnée.

Après la construction des sacs de mots, il vient l'étape de la classification. Le principe reste le même mais les étapes sont différentes.
Soit un tweet T ayant comme mots w1,w2,w3, leurs vecteur respectivement v1,v2,v3 et et deux classe c1 et c2 représentées respectivement par les sacs de mots S1 et S2 contenant trois mots qui sont représentés par leurs vecteurs respectivement,y1,y2,y3,z1,z2,z3, et comme on a mentionné, ces vecteurs peuvent être ajouté les uns aux autres comme le montre la figure 3.7 ou V le vecteur agrégeant v1,v2,v3 , Y le vecteur agrégeant y1,y2,y3, et Z le vecteur agrégeant z1,z2,z3 :

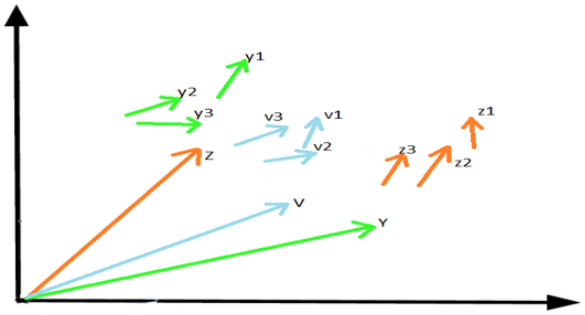

FIGURE 3.7 – Représentations des vecteurs et leurs agrégations

Donc on voit que l'agrégation des vecteurs des mots de tweet qui est le vecteur V est plus similaire au vecteur Y qu'au vecteur Z. Autrement dit, $cos(\vec{(V)}, \vec{(Y)})$ est plus grand que $cos(\vec{(V)}, \vec{(Z)})$ cela signifie que les mots de T ont plus de tendance d'être dans le sac de mot S1 que dans le sac de mot S2. Par conséquent, le tweet T sera classé dans la classe c1 si on parle de classement exclusif, et si on parle de l'ordre de classement la classe c1 sera classée en top suivie par c2.
C'est l'approche qu'on va suivre dans notre travail, et les résultats seront discutés dans le chapitre suivant.

3.10 Architecture de solution proposée :

La formation de la solution proposée comprend plusieurs étapes, depuis la collection des tweets sur lesquels on a travaillé jusqu'à l'aboutissement au résultat final. La figure 3.8 montre les différentes étapes et leurs interactions, les processus numérotés sont décrits dans le tableau 3.2 :

Tableau descriptif

Etapes	Description
01	Collecter les tweets d'une série de catastrophes antérieurs (Dataset)
02	Construction d'un modèle word2vec à partir des dataset
03	Extraction des tweets relevant d'une catastrophe en temps réel
04	Extraction des dictionnaires représentant nos classes (bag of words)
05	Calcul de similarité (similarité cosinus entre les vecteurs de dictionnaires et le vecteur de tweet)
06	Classe similaires (liste ordonnée par similarité entre les dictionnaires et le tweet)

TABLE 3.2 – Tableau descriptif des étapes de notre proposition

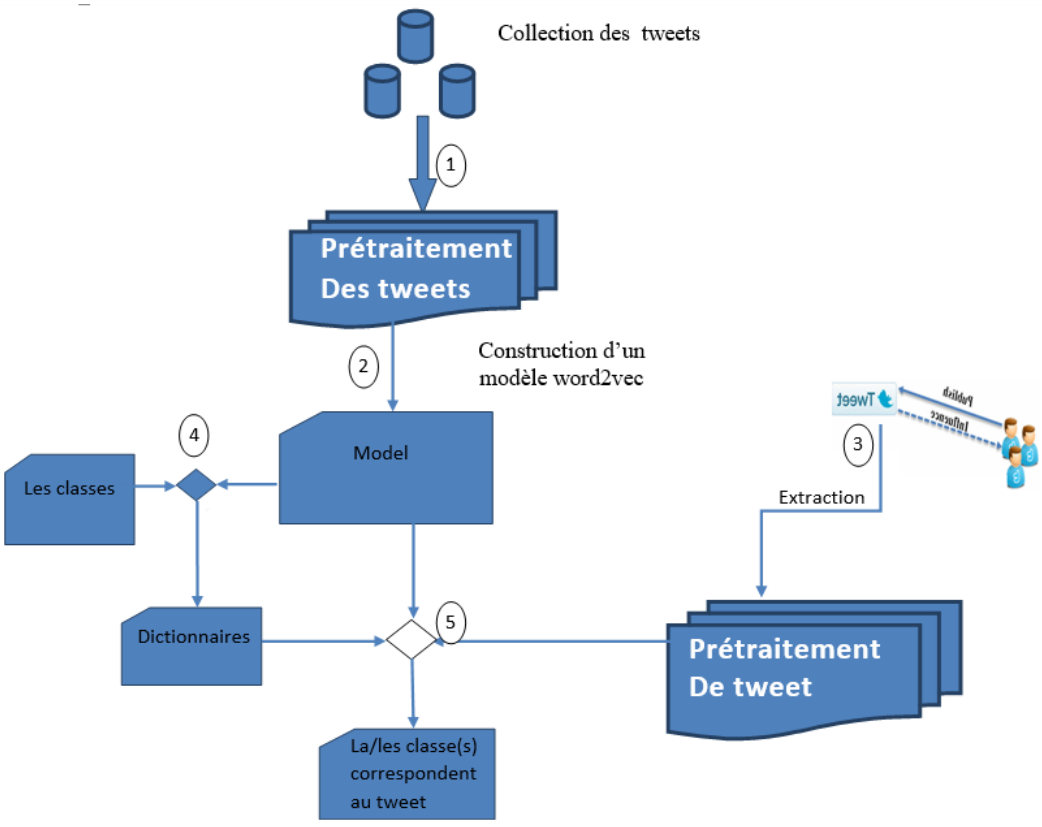

FIGURE 3.8 – Architecture de notre proposition

3.11 Travaux connexes :

L'extraction d'information depuis les réseaux sociaux pour améliorer la situational awarness a été le sujet de plusieurs travaux. Plusieurs systèmes et outils ont été élaborés en réponse à ce sujet, et le point de convergence dans ces travaux sont le sujet à traiter, qui est les informations collectées depuis les media sociaux issues de la situation de crise. La finalité de ces travaux, le point de divergence, consiste en la manière de les traiter. Dans ce qui suit, nous citons quelques travaux :

3.11.1 Twicident [1] :

Twitcident, un frame pour filtrage, la recherche et l'analyse des informations sur les incidents que les internautes publient sur leurs réseaux sociaux. Déclenchée par un module de détection d'incidents qui surveille les services de diffusion d'émergence, cette infrastructure recueille et transfère automatiquement les informations pertinentes sur Twitter. Elle enrichit la sémantique des messages Twitter pour adapter et améliorer l'incident de la recherche et du filtrage au fil du temps. L'enrichissement de la sémantique constitue également la base de l'analyse par facettes de la recherche et de l'altération de l'algorithme fournie par l'infrastructure Twitcident. Dans les évaluations il a été prouvé que l'enrichissement sémantique augmente la performance des messages Twitter pour un incident donné et la recherche d'informations pertinentes à propos d'un incident dans les messages filtrés

3.11.2 SensePlace2 [2]

SensePlace2 est une application Web d'analyse géovisuelle à travers laquelle les médias sociaux peuvent être rassemblés et analysés pour prendre en charge la connaissance de la situation dans la gestion des crises et les domaines d'applications connexes. La version initiale de SensePlace2 est axée sur la recherche de sensations et de sensations avec des informations pertinentes pour la crise extraites de Twitter.

3.11.3 3- Muhammad Imran et al, 2013 [3]

Les chercheurs ont développé un système permettant d'extraire automatiquement des pépites d'information à partir des tweets en temps de catastrophe. Ce système utilise des techniques d'apprentissage automatique à la pointe de la technologie pour classer les mes-

sages dans un ensemble de classes affinées et extraire des informations structurées concises, pouvant être utilisées pour une analyse et une intégration de données complexes au-delà du texte brut.

3.11.4 TweetTracker [4] :

C'est un système de suivi, d'analyse et de compréhension des tweets liés à un sujet bien spécifique. Pour suivre le statut et l'événement, le système collecte les données à l'aide d'un ensemble de critères, notamment les mots-clés, l'emplacement, et les utilisateurs. La source de données peut être choisie parmi les réseaux sociaux les plus utilisés Twitter, Facebook, YouTube, et Instagram. La Modification du nombre total de messages ou la fréquence des messages avec des mots spécifiques peut être tracée pour des périodes de temps différentes. De plus, des mots-clés, des hashtags, des liens, des images et des vidéos avec leurs fréquences sont disponibles pour l'utilisateur.

3.12 Conclusion

Le travail effectué dans ce chapitre couvre deux axes principaux : le premier consistait à l'extraction des informations depuis la plateforme Twitter et les différentes méthodes utilisées pour celle finalité ; et le deuxième axe s'agissait de la présentation de notre approche qui s'appuie sur le TAL et l'analyse distributionnelle en utilisant un algorithme aussi puissant, et largement utilisé dans les processus d'analyse des textes en langages naturelles, on parle alors de word2vec. Cette approche nous a permis à classer selon les tweets selon des classes identifiées au préalable, et les résultats seront discutés dans le prochain chapitre, avec une vue plus détaillée sur les outils qui nous ont servi à fin d'aboutir ce travail.

Chapitre 4

Résultats

4.1 Introduction :

Dans ce chapitre, nous présentons en premier lieu les outils et les langages utilisés pour implémenter notre solution. Ensuite, nous montrons un aperçu de l'évaluation de l'algorithme utilisé, ainsi que les résultats de traitement et d'évaluation des tweets. Pour rappeler, notre solution se base sur la sémantique distributionnelle des mots composant le tweet, à l'aide de l'algorithme word2vec. Cette sémantique nous a permis de déduire le contexte de ce tweet pour qu'il soit classé dans une ou plusieurs classes. Enfin, nous concluons sur quel modèle notre proposition a-t-il été fait.

4.2 Plateformes et outils de développement :

4.2.1 Python :

[5]Python c'est un langage qui peut s'utiliser dans de nombreux contextes et s'adapter à tout type d'utilisation grâce à des bibliothèques spécialisées à chaque traitement. Il est cependant particulièrement utilisé comme langage de script pour automatiser des tâches simples mais fastidieuses.

4.2.2 PyCharm :

[6] PyCharm est un IDE (Integrated Development Environment) spécialisé pour les langages de programmation Python et Django. Il offre de riches et nombreuses fonctionnalités en matière d'édition, l'analyse de code, un débogueur graphique, la gestion des tests unitaires pour vérifier l'intégrité des applications, etc.

4.2.3 API Twitter :

[7] API Twitter fourni aux entreprises, aux développeurs et aux utilisateurs un accès programmatique aux données Twitter via les API (interfaces de programmation d'applications). De manière générale, les API permettent aux programmes informatiques de « se parler » entre eux pour demander et fournir des informations. Cela implique d'autoriser une application logicielle à appeler un point de terminaison : une adresse qui correspond à un type spécifique d'informations que nous fournissons (les points de terminaison sont généralement uniques comme les numéros de téléphone). Twitter donne accès à certaines parties de notre service via des API pour permettre aux utilisateurs de créer des logiciels qui s'intègrent à Twitter, par exemple une solution qui aide une entreprise à répondre aux commentaires des clients sur Twitter. La plateforme d'API fournit ainsi un large accès aux données Twitter publiques que les utilisateurs ont choisi de partager avec le monde. L'API Twitter permet d'accéder à la base de données Twitter et de récupérer/poster plusieurs informations. L'API se décompose en quatre classes :

l'API Stream :

[7] Parmi les API proposé par Twitter ont utilisé l'API Stream. En utilisant l'API Streaming, nous pouvons rechercher simultanément des mots clés, des hashtags, des ID utilisateur et des zones de délimitation géographiques . Jusqu'à 400 mots-clés, 25 zones de délimitation géographique et 5 000 ID utilisateur peuvent être fournis dans une seule demande.

Afin de commencer à utiliser l'API Streaming, plusieurs étapes sont à suivre :

- **Enregistrer une nouvelle application avec Twitter**
- **Authentification API :** La plupart des tâches qu'on peut effectuer avec les données Twitter (en direct) nous obligent à authentifier la demande en s'inscrivant aux clés

API. Lorsque on a enregistré les clés API, on peut les stocker dans un fichier sur l'ordinateur, pu Comme on peut le voir, cela donnera quatre clés distinctes : clé de consommateur, clé de sécurité, jeton d'accès et secret de jeton d'accès.

- **Stocker les clés :**dans un fichier credentials.txt et dans l'ordinateur.
- **L'endroit de stockage (tweets) :**on suppose également que ce dossier est l'endroit où on enregistre tous les fichiers contenant des tweets qu'on collecte. Une fois qu'on a choisi le nom et l'emplacement de ce dossier, on doit définir la TWITTER, variable d'environnement sur cette valeur.
- **Installation Twython :**Le package Twitter NLTK s'appuie sur une bibliothèque tierce appelée Twython. on Installe Twython via pip.
- **La circulation des tweets :** La Twitter classe est conçue comme un moyen d'interagir avec le flux de données Twitter.
  - Collecter tous les tweets actuellement publiés par les utilisateurs.
  - Nous pouvons filtrer le flux public en direct en recherchant des comptes d'utilisateur spécifiques.

4.2.4 NLTK (Natural Language Toolkit) ou Boîte à outils du langage naturel :

[5] NLTK est une plate-forme leader en Python permettant un traitement automatique des langues compatibles avec les données de langage humain. Il fournit des interfaces faciles à utiliser pour plus de 50 corpus et ressources lexicales telles que WordNet, ainsi qu'une suite de bibliothèques de traitement de texte pour la classification, la tokenisation, et catégorisation de texte, l'analyse de la structure linguistique, etc.

4.2.5 Gensim :

[5] Gensim est une librairie spécialisée dans le Topic Modeling. Elle implémente plusieurs algorithmes statistiques de Topic Modeling (Latent Dirichlet Allocation Latent Dirichlet Allocation ou LDA, Latent Semantic Indexing ou LSI, Hierarchical Dirichlet Process ou HDP) et permet également de faire du Word Embedding.

4.2.6 Word2vec :

En intelligence artificielle et en apprentissage machine, Word2vec est un groupe de modèles utilisé pour le plongement lexical) (word embedding). Ces modèles ont été développés par une équipe de chercheurs chez Google sous la direction de Tomas Mikolov. Ce sont des réseaux de neurones artificiels à deux couches entraînés pour reconstruire le contexte linguistique des mots. La méthode est implémentée dans la bibliothèque Python Gensim.

4.3 Evaluation de l'algorithme :

Dans notre travail, nous avons utilisé l'Accuracy et la Précision comme mesures d'évaluation du modèle. Cette évaluation consiste en une étude empirique des différents résultats avec différents jeux de paramètres. Nous allons nous intéresser à la méthode de classification CBOW, une taille variable de la fenêtre.

- **Accuracy :** Pourcentage des tweets correctement classés (1ere classe prédite).
- **Précision :** Proportion d'identifications positives effectivement correcte.

Soit :

TP : classe positive considérée positive.

TN : classe négative considérée négative.

FP : classe négative considérée positive.

FN : classe positive considérée négative.

- $\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$
- $\text{Précision} = \frac{TP}{TP+FP}$

Les résultats sont résumés par le tableau 4.1 :

- **Le nombre d'iterations :** Nombre d'itération d'apprentissage
- **Taille de vecteur :**dimension de chaque vecteur de chaque mot
- **Taille de la fenetre :** le nombre de mots à gauche et à droite de mot cible

Méthode01 : architecture Skip-gram

Paramètre				Nombre des Tweets	Résultat (correctement classés)	Accuracy
Nombre Itérations	Taille vecteurs	Taille fenêtre	Méthode			
300	300	3	Skip-gram	195	148	75.89 %
350	150	2	Skip-gram	195	110	56.41 %
150	100	1	Skip-gram	195	105	53.84 %

TABLE 4.1 – Résultats accuracy Méthode Skip-gram

Précision – cas Taille de fenêtre = 3 :

Classe	**Résultats**		
	True Positive	**False Positive**	**Précision**
Coordination	1	1	50%
Donation	15	2	88.24%
Food	7	4	63.64%
Health	15	3	83.33%
Protection	50	11	81.97%
Rescue	26	2	92.86%
Shelter	7	2	77.78%
Telecommunication	4	7	36.36%
Education	0	0	0%
Transport	3	1	75%
Water	20	14	58.82%

TABLE 4.2 – Tableau de précision pour la taille de la fenêtre = 3 (skip-gram)

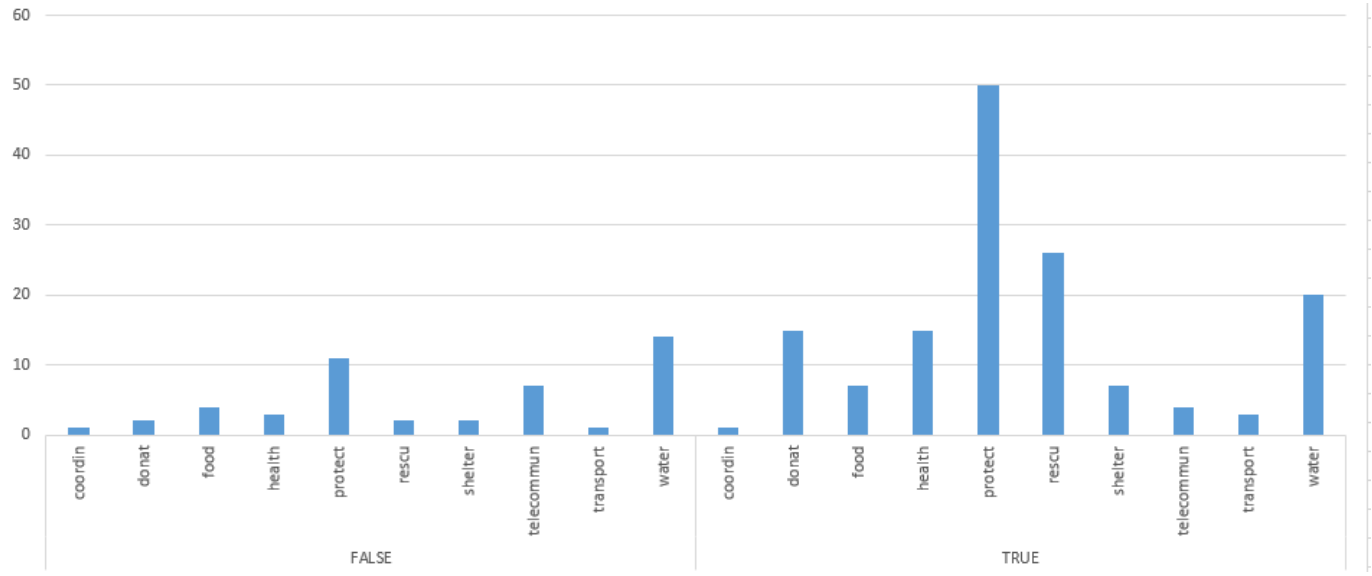

FIGURE 4.1 – variété de la précision pour une taille de fenêtre = 3 (skip-gram)

Précision – cas Taille de fenêtre =2 :

Classe	Résultats		
	True Positive	**False Positive**	Précision
Coordination	5	2	71.43%
Donation	13	15	46.43%
Food	11	6	64.71%
Health	15	12	55.56%
Protection	25	17	59.52%
Rescue	4	6	40%
Shelter	8	9	47.06%
Telecommunication	5	4	55.56%
Transport	5	4	55.56%
Education	0	0	0%
Water	19	10	65.52%

TABLE 4.3 – Tableau de précision pour la taille de la fenêtre = 2

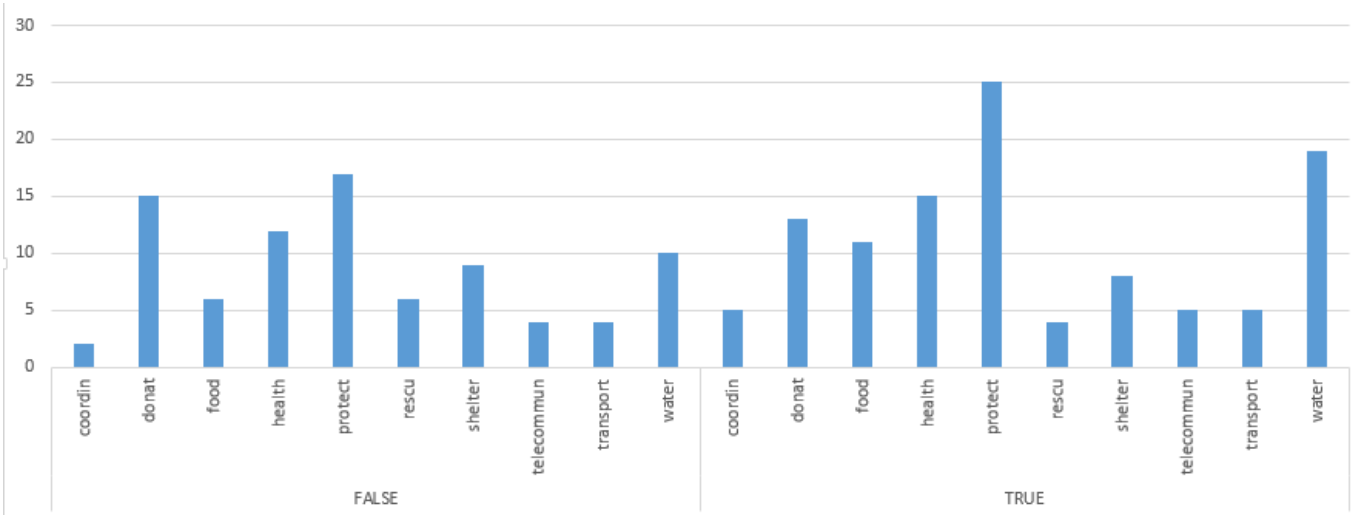

FIGURE 4.2 – variété de la précision pour une taille de fenêtre =2

Méthode 02 : architecture CBOW

Paramètre				Nombre des Tweets	Résultat (correctement classés)	Accuracy
Nombre Itérations	Taille Vecteurs	Taille fenêtre	Méthode			
350	150	2	CBOW	195	131	67.17 %
300	300	3	CBOW	195	109	55.89 %
150	100	1	CBOW	195	98	50.25 %

TABLE 4.4 – Résultats accuracy Méthode CBOW

Précision – cas Taille de fenêtre = 2 :

Classe	Résultats		
	True Positive	**False Positive**	Précision (%)
Coordination	1	3	25%
Donation	26	16	61.90%
Education	1	12	7.70%
Food	1	6	14.29%
Health	1	3	25%
Protection	10	0	100%
Rescue	23	0	100%
Shelter	13	0	100%
Telecommunication	2	4	33.33%
Transport	1	8	11.11%
Water	52	12	81.25%

TABLE 4.5 – Tableau de précision pour la taille de la fenêtre =2 (CBOW)

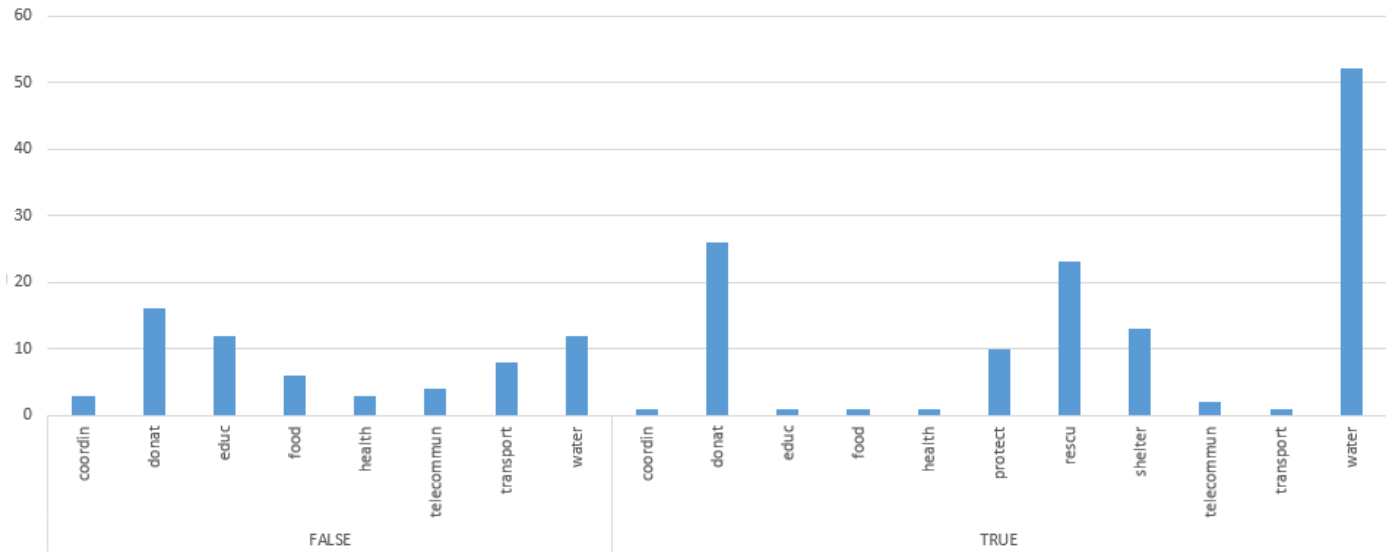

FIGURE 4.3 – variété de la précision pour une taille de fenêtre =2 (CBOW)

Précision – cas Taille de fenêtre = 3 :

Classe	Résultats		
	True Positive	**False Positive**	Précision (%)
Coordination	5	2	71.42%
Donation	13	15	46.42%
Food	11	6	64.71%
Health	15	12	55.56%
Protection	25	18	58.14%
Rescue	4	6	40%
Shelter	8	9	47.06%
Telecommunication	5	4	55.55%
Transport	5	4	55.55%
Education	0	0	0%
Water	18	10	64.29%

TABLE 4.6 – Tableau de précision pour la taille de la fenêtre = 3 (CBOW)

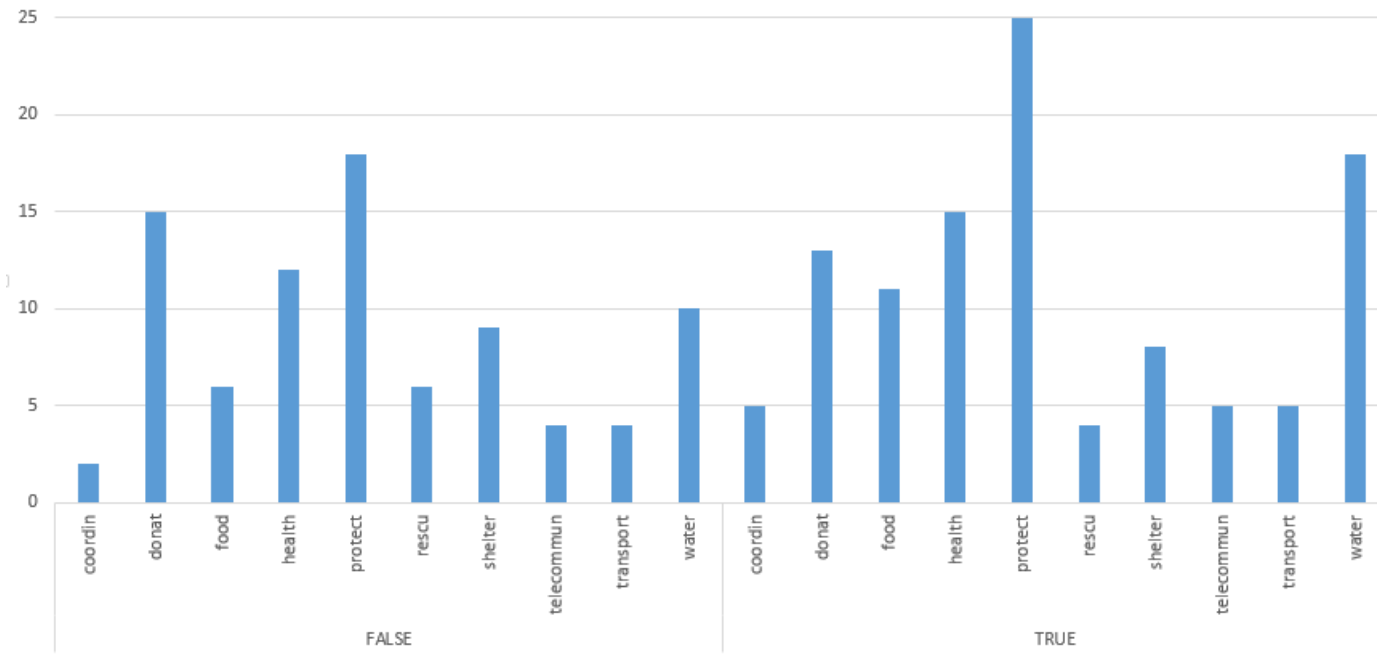

FIGURE 4.4 – variété de la précision pour une taille de fenêtre = 3 (CBOW)

4.4 Résultats

4.4.1 Application de modèle :

Après avoir conçu un modèle word2vec, nous avons construit pour chaque classe un sac de mots qui nous permet de calculer la similarité de tweet à chaque classe.

4.4.2 La distribution de sac de mot :

Word2vec permet de construire des sacs de mots propre à un contexte donné, en se basant sur la méthode de calcul de similarité en cosinus entre une moyenne simple des vecteurs des mots donnés et les vecteurs de chaque mot du modèle. Les figures suivantes illustrent la distribution de sac de mots de 'health' et shelter selon le corpus généré de dataset :

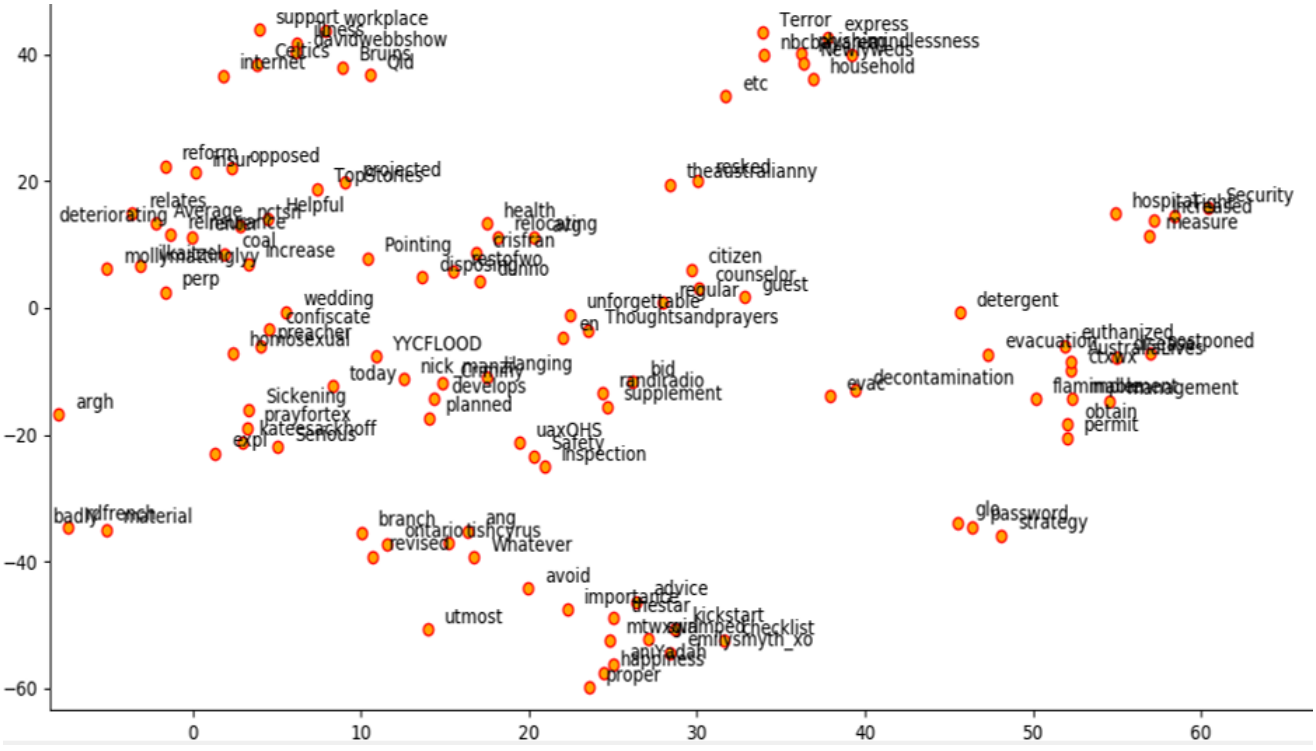

FIGURE 4.5 – Sac de mot de 'health'

Nous les avons appliqué ces sacs de mots sur un ensemble de tweets. Ces tweets ont été soumis au même prétraitement, à savoir la tokenization, la suppression des mots vide, la lemmatisation et le stemming. Les résultats sont résumés dans ce tableau :

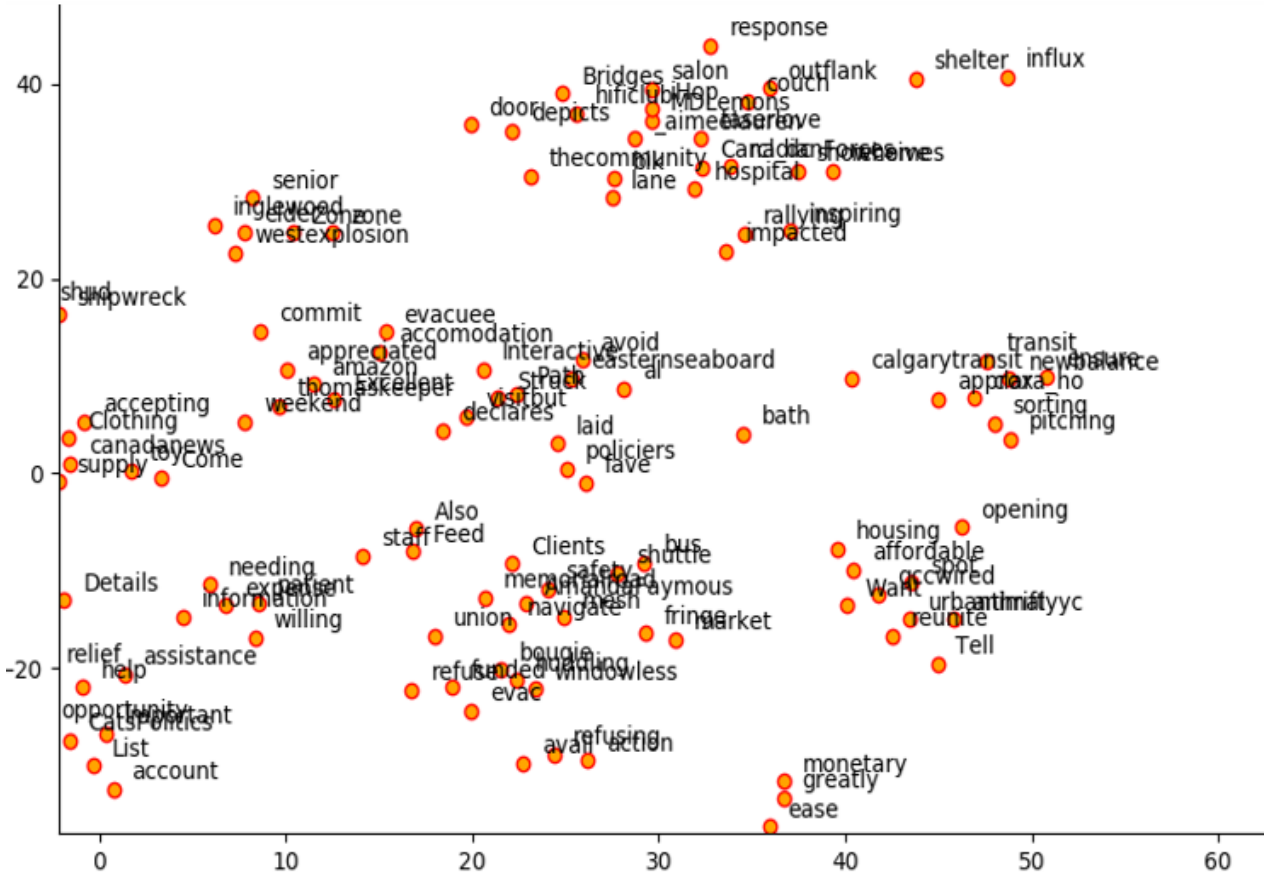

FIGURE 4.6 – Sac de mot de 'shelter'

On constate que nous avons admis qu'un tweet peut être classé dans une ou plusieurs classe, selon le besoin ; soit nous récupèreront juste la première classe, soit nous fixerons un seuil de similarité, ou alors nous fixerons un nombre de classe à retourner.

4.5 Discussion

Nous avons démontré que la sémantique distributionnelle est une solution efficace pour identifier les classes des tweets à fin de contribuer à l'amélioration de la situation-awarness. En traitant les tweets, nous avons démontré que le processus commun de TAL améliore cette distribution, et par conséquent améliore les résultats.

Nous avons démarré de l'hypothèse selon laquelle un ensemble de mots sont sémantiquement proches s'ils apparaissent dans des contextes similaires. Cette sémantique est fondée sur l'usage des mots à partir de leurs emplois dans des corpus différents.

A partir de cette hypothèse, nous sommes arrivés à construire un modèle qui sert à classifier

Tweet avant traitement	**Tweet après traitement**	**Résultats**		
		Classe 1	Classe 2	Classe 3
Oklahoma tornado - in pictures http://t.co/xWPLjF0r0i ? ??????? @guardian	'tornado', 'picture', 'guardian'	telecommunication 0.5526272	Rescue 0.5489382	/
RT @acarvin: My #okc #tornado Storify from last night. RT @nprnews: Tweets Capture 'Shock And Awe' At Tornado's Deadly Power http://t.co/mS…	'acarvin', 'okc', 'tornado', 'last', 'night', 'nprnews', 'tweets'	shelter 0.7076663	protection 0.69037145	/
RT @ThunderObsessed: Kevin Durant has donated $1 million dollars to the Moore relief efforts.	'thunderobsessed', 'donated', 'dollar', 'relief', 'effort'	donation 0.7769535	Coordination 0.6895181	food 0.6476439
RT @SimpleSpacesAB: Red Cross Challenge. Rebuild after #yycfloods. Match our donation, make a difference. Important details http://t.co/JZ…	'yycfloods', 'match', 'donation', 'make', 'difference', 'important', 'detail'	donation 0.75055873	coordination 0.7062174	shelter 0.68633044

TABLE 4.7 – Tableau des résultats

les tweets d'une manière proche au processus d'interprétation humain, qui représente un gain de temps et d'efforts.Cette solution sert à aider les décideurs et les secouristes à mieux trouver des réponses à leurs besoins d'information pendant une situation de crise, ainsi qu'elle permet de tirer profit des informations circulant dans les réseaux sociaux.

4.6 Conclusion :

Nous avons présenté dans ce chapitre les différents langages et outils de développement que nous avons utilisé afin d'implémenter notre solution proposée. Nous avons démontré que la sémantique distributionnelle représente une très bonne solution pour comprendre le contexte d'un tweet pour en mieux tirer profit dans l'amélioration de situation-awarness des décideurs durant les situations d'urgence.

Conclusion générale et perspectives

Les plateformes des media sociaux sont devenues un moyen important de partager des informations sur le Web, en particulier les événements critiques tels que les catastrophes naturelles, industrielles et celles causées par l'homme. Actuellement, Twitter comme étant un concurrent principal dans les réseaux sociaux est utilisé pour diffuser des nouvelles sur des incidents et des victimes.

Dans ce sens notre thème présenté dans ce mémoire aborde justement la détermination des tweets dont leurs contenu concerne des catastrophes, et par la suite leurs envoyer vers les décideurs appropriés en temps réel.

Afin d'atteindre cet objectif, nous avons concevoir une approche basé sur le Framework word2vec qui est un outil leader dans le domaine de traitement automatique des langues (TAL) ; cette approche est constituée de 3 étapes à savoir l'extraction des informations à partir des tweets puis les nettoyer (tokinisation, et lemmatisation) afin d'obtenir que des mots pertinent en leur forme rationnelle ; et finalement procéder à leur classification en fonction des domaines d'activités adéquates.

Nous pouvons confirmer que les résultats obtenus concernons la correspondance des tweets par rapport à leur classe étaient presque parfaite et par conséquent nous avons assuré que les tweets vont bel et bien être envoyé vers les destinataires correspondant dont l'objectif de notre thème. En réalité la réalisation de ce mémoire et le développement de l'application nous ont apporté plusieurs avantages sur nos compétences techniques et analytiques à savoir :

- Toucher profondément des concepts principaux d'un domaine de recherche très pro-

metteur dans le traitement sémantique des textes.

- Maitrise quelques outils de l'intelligence artificielle et de l'apprentissage machine tel que word2vec sequancematcher.
- Maitrise le langage python et ces nombreux avantages en particulier son adaptation à tout type d'utilisation grâce à ses bibliothèques spécialisées à chaque traitement.

Finalement nous suggestions quelque idées nous apparaissant utile et avantageuse bien pour notre thème que autre thèmes correspondant à savoir : Elargir les sources de données par l'intégration d'autres réseaux sociaux au moins au moment des situations catastrophes.

Bibliographie

[1] Semantics + filtering + search = twitcident exploring information in social web streams.

[2] Anthony C. Robinson Scott Pezanowski Alexander Savelyev Prasenjit Mitra Xiao Zhang Alan M. MacEachren, Anuj Jaiswal and Justine Blanford. Geotwitter analytics support for situational awareness.

[3] Extracting information nuggets from disaster- related messages in social media.

[4] Intelligent disaster response, via social media analysis.

[5] Natural language processing with python.

[6] Site official de pycharm, https ://www.jetbrains.com/pycharm/documentation.

[7] Site official twitter (help.twitter.com).

[8] Endsley. Tward atheory of situational awarness in dynamic system. 1995.

[9] Giddens et le sociologue allemand Beck. résument dans le concept de « société de modernité réflexive. 2002-2003.

[10] Hélène Juillard. Catastrophe et situation d'urgence définition impact et réponse. page 7, 2014.

[11] United national international strategy for disaster reduction, living with risk : a global review of disaster reduction initiative]. 2004.

[12] Brookings-Bern. La protection des personnes affectÉes par des catastrophes naturelles. 1999.

[13] Bekele Geleta. Gestion des catastrophes et des risque. 2011.

[14] Fédération internationale des Sociétés de la Croix-Rouge et du Croissant-Rouge. Qu'est-ce que l'evc ? introduction à l'évaluation de la vulnérabilité et des capacités. (2006-2010).

[15] Jorgen Ernstsen and Daniela Villange. Situation awareness in disaster management. May 2014.

[16] Aid Aicha. Formulation d'un environnement générique d'un service dans un système pervasif public en cas de situation d'urgence. 2015.

[17] William J. Corvey-Leysia Palen James H. Martin Martha Palmer Aaron Schram et Kenneth M. Anderson Sudha Verma, Sarah Vieweg. Natural language processing to the rescue ? : Extracting "situational awareness" tweets during mass emergency.

[18] Tao Li-Steven Luis Yi Deng Vagelis Hristidis, Shu-Ching Chen. Survey of data management and analysis in disaster situations. 2010.

[19] Elizabeth D. Liddy. Natural language processing. 2001.

[20] Cécile Fabre et Nabil Hathout. Apports de la sémantique distributionnelle dans les recherches en linguistique. 2016.

[21] Emmanuelle Dusserre. Utilisation de la méthode distributionnelle pour la constitution de classes sémantiques d'une liste de formes du lexique scientifique transdisciplinaire. 2016.

[22] Jakub Piskorski and Roman Yangarber. Information extraction : Past, present and future. 2016 22.

[23] Xavier Tannier Xavier.Tannier@limsi.fr. Analyse de textes et extraction d'information.

[24] Practical extraction of disaster-relevant information from social media. in proceedings of the 22nd international conference on world wide web,.

[25] Réseaux de neurones pour la représentation de contextes continus des mots.

[26] Note d'orientation sur la mise en oeuvre de l'approche de responsabilite sectorielle (« cluster approach ») pour renforcer l'action humanitaire.

[27] Lexicon for collecting and filtering microblogged communications in crises. in proceedings of the aaai conference on weblogs and social media (icwsm'14). aaai press, ann arbor, mi, usa.

[28] Word2vec et semvue dÉtection et visualisation de termes pertinents de maintenance edf r/d.