REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE SAAD-DAHLEB-BLIDA

N° D'ordre : .....

Faculté des sciences
Département d'informatique

Mémoire Présenté par :

Charlah Malika **Djeziri Nadia**

En vue d'obtenir le diplôme de master

Domaine : mathématique et informatique

Filière : Informatique

Spécialité : Informatique

Option : Ingénierie de logiciel

Analyse des sentiments des tweets afin de trouver des détails concernant un crime donné

M : **Benhabiles Halim**	Président
M : **Ould Aissa Ahmed**	Examinateur
M : **Madani Amina**	Promotrice

Promotion : 2015/2016

Remerciements :

Premièrement et avant toute chose, nous rendons grâce à ***Allah****, le tout puissant, de nous avoir permis de suivre le chemin du savoir, et donné le courage d'achever ce travail.*

Nous tenons, également, à exprimer notre sincère reconnaissance et notre profonde gratitude à tous ceux qui ont contribué de près ou de loin à la réalisation de ce mémoire, notamment notre promotrice, ***Mme. MADANI Amina,*** *qui grâce à elle nous avons eu l'opportunité de découvrir le domaine du l'analyse de sentiment sur Twitter. Ses conseils illuminés et son aide précieux nous ont permis de mener à bien ce modeste travail et on la remercie très chaleureusement.*

Un grand Merci au corps enseignant ainsi qu'à l'administration de l'université de Blida pour tout le savoir qu'ils ont su nous transmettre durant ces cinq dernières années, et aussi d'être toujours là pour nous guider à retrouver le bon chemin par leur sagesse et leurs précieux conseils.

Mes remerciements vont également à nos parents, qui nous ont aidées de prés ou de loin par le fruit de leur connaissance pendant toute la durée de notre parcours éducatif.

Enfin, nous remercions les membres du jury d'avoir accepté d'évaluer notre modeste travail, ainsi que toutes les personnes qui ont contribué de près ou de loin à l'élaboration de ce mémoire.

Dédicaces

Je commence par rendre grâce à dieu et à sa bonté, pour la patience, la compétence et le courage qu'il m'a donné pour arriver à ce stade d'étude.

Je dédie ce modeste travail et ma profonde gratitude à celle qui m'a transmis la vie, l'amour, le courage, A toi chère maman

A Mon très cher père, pour tous ses conseils et pour toute la confiance qu'il a mise en moi et pour son dévouement pour mon bonheur.

A tous mes frères

A Tous mes enseignants

A Tous mes amis

A Tous mes collègues

A Toute ma famille

Djeziri Nadia

Dédicace :

Tous les mots ne sauraient exprimer la gratitude, l'amour, le respect, la reconnaissance, c'est tous simplement que : Je dédie ce travail à :

A Ma tendre Mère Zohra : Tu représente pour moi la source de tendresse et l'exemple de dévouement qui n'a pas cessé de m'encourager. Tu as fait plus qu'une mère puisse faire pour que ses enfants suivent le bon chemin dans leur vie et leurs études.

A Mon très cher Père Abdelkader: Aucune dédicace ne saurait exprimer l'amour, l'estime, le dévouement et le respect que j'ai toujours pour vous. Rien au monde ne vaut les efforts fournis jour et nuit pour mon éducation et mon bien être. Ce travail et le fruit de tes sacrifices que tu as consentis pour mon éducation et ma formation le long de ces années.

A mes cher frère : Oussama, Rachid, Mohamed et Ali

A mes cher sœurs : Zahia, Fatima, Chafika et Aicha.

A toute ma grande famille

Et mes très chers amis

Malika

Résumé

Les réseaux sociaux tels que Twitter et Facebook ne sont pas utilisés seulement pour la communication, mais ils sont aussi utilisés pour prédire le crime. Le crime est défini comme un acte nuisible non seulement à la personne concernée, mais aussi à la communauté dans son ensemble. Dans ce contexte, ce mémoire se concentre sur l'analyse des sentiments des tweets afin de trouver des détails sur un crime donné et plus précisément sur la prévention de suicide via Twitter.

Dans notre travail, nous avons utilisé une méthode d'apprentissage supervisé, la méthode de K plus proche voisin(KNN) pour classer les tweets en deux classes : tweets suspects risque de suicide et tweets suspects non risque de suicide. Nous avons suivi les étapes suivantes : la conceptualisation et l'implémentation d'une base de données ; Définition et stockage dans la base de données d'un vocabulaire lié à la problématique du suicide ; Collecte et stockage des tweets suspects (comportant un élément du vocabulaire défini auparavant) ; Mise en place d'un outil de classification automatique et évaluation du classifieur sur les tweets suspects collectés auparavant.

Mots clés :

Twitter, Crime, Suicide, Analyse des sentiments, K plus proche voisin, IBK, Classification supervisée, Weka.

Abstract

Social networks such as Twitter, Facebook etc are not only used for networking and communication purposes, but they are also useful for predicting crime; The crime is defined as an act harmful not only to the individual but to the community as a whole. In this subject, this memory tweets sentiment analysis to find details on a particular crime and specifically on Twitter suicide preventing.

In our work, we will use the well-known supervised learning method of K nearest neighbor (KNN) to classify tweets into two classes: tweet suspect suicide risk and not suspect tweet risk of suicide. We follow the following steps: the conceptualization and implementation of a database; Definition and storage in the database of vocabulary related to the problem of suicide; Collection and storage of suspicious tweets (including an element of the vocabulary defined previously); Setting up an automatic classification tool and evaluation of the classifier on suspicious tweets collected before.

Keywords:

Twitter, Crime, Suicide, Sentiment Analysis, K nearest Neighbor, IBK, Supervised classification, Weka.

ملخص

الشبكات الاجتماعية مثل تويتر، الفيسبوك وغيرها لا تستخدم فقط لأغراض التواصل ، ولكنها أيضا مفيدة للتنبؤ بالجريمة. يمكن تعريف الجريمة على أنها فعل ضار ليس فقط للفرد ولكن للمجتمع ككل. في هذا السياق، تظهرلنا هذه المذكرة تحليل المشاعر للعثور على تفاصيل حول جريمة معينة وتحديدا توقع الانتحار عبر تويتر.

في عملنا هذا استخدمنا KNN كطريقة لتصنيف الرسائل التي جمعناها من تويتر، حيث قمنا بتصنيف هذه الرسائل إلى فئتين : فئة خاصة بالرسائل التي لها خطر في الإنتحارو فئة خاصة بالرسائل التي ليس لها خطر في الانتحار حيث قمنا بالتباع الخطوات التالية تصور وتنفيذ قاعدة البيانات تخزين الكلمات المرتبطة بالإنتحار في قاعدة البيانات، جمع و تخزين الرسائل المشتبه فها في قاعدة البيانات(تحتوي هذه الرسائل على كلمة من الكلمات التي تم تخزينها من قبل)،إستعمال أداة تصنيف آلية و تقييم مجموعة من المصنفات بتطبيقها على الرسائل المشتبه فها التي جمعناها من قبل.

Sommaire

CHAPITRE 3: CONCEPTION

CHAPITRE 04 : IMPLIMENTATION ET TEST

Liste des Tableaux

Liste des figures

Liste des équations

Introduction générale :

De nos jours, l'internet est un outil incontournable d'échange d'informations. Il nous offre une quantité considérable d'informations à une vitesse inédite et ses services s'adaptent de plus en plus aux besoins des internautes. Pendant les dernières années, l'Internet a connu encore une plus vaste portée grâce au développement des médias sociaux tels que Facebook, Twitter et LinkedIn, qui comptent des millions de membres. Basés sur des techniques de communication faciles et accessibles pour tous, ces médias favorisent les interactions sociales à travers l'Internet. Les médias sociaux se subviennent aux besoins des individus d'échanger des opinions, de demander des conseils et de communiquer de façon rapide et facile. Offrant un accès libre et gratuit, les médias sociaux ont considérablement favorisé la communication de masse et ils ont déclenché le débat public sur Internet.

Twitter est actuellement la plateforme de microblogage la plus populaire. Elle limite le nombre de caractères utilisés dans un message, appelé tweet, à 140 pouvant contenir également des hyperliens. Plusieurs recherches ont montré que les données publiées par les internautes sur Twitter, reflètent presque en temps réel l'intérêt du public. L'analyse de sentiments et la fouille d'opinion sont devenues les sujets de recherche de beaucoup de scientifiques. Il est utilisé pour décrire l'analyse automatique de texte évaluatif et pour la recherche de valeur prédictive des jugements.
Dans le monde actuel, l'étude de criminologie se concentre sur l'identification des caractéristiques criminelles qui sont de plus technologiquement sophistiquées exprimant souvent leurs émotions sur le web.

Cette étude s'intéresse à l'analyse des sentiments des tweets afin de trouver des détails concernant un crime donné et plus spécifiquement à la prédiction des suicides à partir des messages publiés sur Twitter. Suicide est l'acte délibéré consistant à mettre fin à sa propre vie. Il révèle de graves problèmes personnels, mais est également souvent le reflet d'une détérioration du contexte social dans lequel vit un individu. Les facteurs de risques sont multiples et complexes (bouleversements dans les relations personnelles, harcèlement, peur, chômage, dépression clinique et bien d'autres formes de maladie mentale, etc.).

Introduction Générale

Le suicide représente un problème majeur dans le monde. De nombreux cas de suicides ont été relatés ces dernières années car les personnes ont fortement interagi via les réseaux sociaux (Facebook, Twitter, …) avant de passer à l'acte.

Il est difficile de traiter un grand corpus de données et prédire la classe la plus probable de chaque tweet et de trouver le taux de suicide dans chaque classe, ce qui nous conduit a un problème de classification de sentiments.

A travers cette étude, nous allons concevoir et mettre en place des méthodes permettant de détecter des risques de passage à l'acte au travers des messages postés par les utilisateurs sur Twitter. Nous allons classer les tweets en deux classes : tweet suspect risque de suicide et tweet suspect non risque de suicide.

Problématiques :

Les crimes sont des problèmes majeurs dans le monde actuel tel que suicide. Donc il est nécessaire de trouver des méthodes qui permettent de prédire et de trouver des détails sur un crime pour diminuer le taux d'un crime. Les réseaux sociaux comme Twitter sont de plus en plus associés à des phénomènes des crimes tels que le suicide. Il est donc très important de détecter les potentielles victimes au plus tôt afin de pouvoir renforcer la prévention d'un crime (suicide) sur le web.

A travers ce travail nous essayons de répondre aux questions suivantes : quelles sont les méthodes utilisées pour trouver des détails sur un crime donné (suicide) en temps réel? Quels sont les algorithmes de classifications plus profondes pour classifier les tweets et avec quel outil ? . Notre travail s'intéresse à l'analyse des sentiments des tweets afin de trouver des détails concernant un crime donné.

L'objectifs:

L'objectif de ce travail est de proposer une nouvelle approche qui permet d'aider les spécialistes pour trouver des détails intéressants concernant un crime donné (suicide) en se basant sur les tweets. À partir d'un vocabulaire associé à la thématique du suicide et des messages récupérés à partir de Twitter, l'application à développer doit permettre de classifier automatiquement les tweets en deux classes: classe de tweets suspects risque de suicide et classe de tweets suspects non risque de suicide.

Organisation du mémoire :

Ce mémoire s'organise comme suit : nous avons commencé par une introduction générale, problématiques et objectifs. Dans le premier chapitre nous allons essayer de comprendre Twitter en général, sa structure, son lexique, son rôle dans la vie politique et social. Ensuite, dans le deuxième chapitre nous proposons une étude approfondie sur l'ensemble de travaux réalisés dans le domaine de l'analyse des sentiments des tweets sur les crimes. Dans le troisième chapitre nous proposons une nouvelle approche de l'analyse des sentiments des tweets afin de trouver des détails sur un crime donné. Et dans le dernier chapitre nous allons faire une implémentation du système et nous détaillons les tests de notre approche.

Chapitre 01 : Généralité sur Twitter

Chapitre 01 : Généralité sur Twitter

1. Introduction :

Les médias sociaux qui ont récemment pu bénéficier d'un considérable essor sont les réseaux sociaux tels que Facebook, LinkedIn, Myspace et Twitter. Ce sont des sites web qui rassemblent des identités sociales telles que des individus, des entreprises et des organisations qui peuvent échanger de l'information à travers des interactions sociales. Grâce à leur caractère maniable et leur accès libre, les réseaux sociaux bénéficient d'un succès croissant auprès du grand public.

Dans les dernières années, Twitter devient plus populaire que Facebook, LinkedIn, Myspace, il repose sur le principe du microblogue qui est un dérivé concis du blog et qui permet de publier un court article.

Dans ce chapitre, nous allons donner quelques définitions comme celle du blog et du microblogage puis une présentation générale de Twitter ainsi que la structure de ses messages. Ensuite, nous allons décrire un aspect important de Twitter et qui fait de lui un réseau social à part pour les études de recherche d'informations, communément appelée API-Twitter.

2. Définition :

Blog :

Le **Blog**, nommé par contraction des mots Web Log (carnet de bord web en anglais), est un site web personnel dans lequel un ou plusieurs auteurs publient au fil du temps des articles (aussi appelés billets), organisés en catégories et affichés dans l'ordre chronologique inverse. Les visiteurs du blog peuvent ensuite commenter le contenu des articles.[1]

Microblogage :

Microblogage (Microblogging en anglais): est un service en ligne de textes courts. Cette évolution des blogs fonctionne comme un réseau social. Les messages envoyés par une personne sont reçus par une liste d'utilisateurs qui ont souhaité lire les textes de cette personne. Le plus connu de ces services est Twitter.[2]

3. Twitter :

3.1.Définition :

Twitter est un mix de réseau social et de plateforme de micro-blogging. Des informations qui n'excèdent pas 140 caractères appelés tweets sont diffusées sur la plateforme Twitter. Lors de l'écriture d'un tweet (l'information postée), Twitter nous pose la question « What are you doing ? » (Que faites-vous ?). Ce microblogue est donc utilisé pour présenter ce qu'il se passe autour de nous à un moment donné. Il a été utilisé dans diverses campagnes de marque, élections, et en tant que média de nouvelles.

Twitter est un réseau social asymétrique : il n'engage pas une réciprocité. Il est possible pour un utilisateur de restreindre la lecture de ses minimessages en rendant l'accès à son compte privé : en ne le rendant pas public. Les messages sont alors visibles par l'abonné uniquement après validation d'une requête d'ajout à sa liste d'abonnement par l'utilisateur qui a appliqué un accès privé. [3]

La **Figure 1-1** montre une capture d'écran de l'interface de l'utilisateur Twitter. Les mises à jour des statuts peuvent être envoyées via un navigateur Web, SMS, e-mail ou des tierces applications et ils sont affichés sur le profil des utilisateurs.

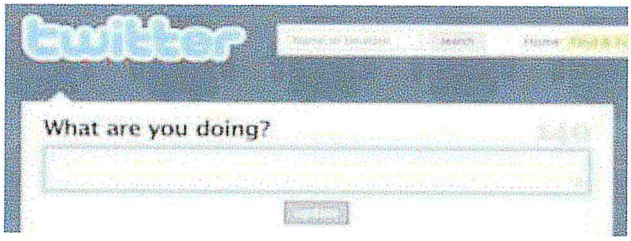

Figure 1-1 : Capture d'écran de l'interface utilisateur de Twitter.[1]

3.2. Historique de Twitter :

Twitter a été créé en 21 mars 2006 à San Francisco au sein de la société américaine startup Odeo fondée par Noah Glass et Evan Williams, et Jack Dorsey. Cette société proposait une plateforme d'hébergement, de diffusion et d'enregistrement de podcast. L'idée de départ lancée par Jack Dorsey était de permettre aux utilisateurs de partager facilement leurs petits moments de vie avec leurs amis. Le 21 mars 2006, M. Dorsey envoyait son premier tweet : « Just setting up my twttr » (« Suis en train d'installer mon twttr »). Le marché du podcast étant déjà très concurrentiel, Jack Dorsey et Noah Glass et Evan Williams furent chargés de développer un nouveau service ouvert au public le 13 juillet 2006, la première version s'intitulait *Stat.us* puis *Twittr*, en référence au site de partage de photos Flickr puis *Twitter*, son nom actuel. Le 25 octobre 2006, les actifs de la société Odeo ont été rachetés par Obvious Corp. Puis en avril 2007, une entité indépendante est créée comme nom Twitter avec Jack Dorsey à sa tête jusqu'en octobre 2008 date à laquelle Evan Williams lui succéda. En mars 2008, Twitter compte un million d'utilisateurs. La société compte 29 employés en février 2009, 300 en octobre 2010 et 900 en avril 2012. En juin 2012, les mots « Twitter » (nom propre), « Twitt » ou « tweet », « Twitteur » ou « Twitteuse », ainsi que « Twitter » ou « Tweeter », font leur apparition dans Le Petit Larousse édition 2013.Twitter dont le prix d'introduction est fixé à 26 dollars entre à la bourse de New York le 31 octobre 2013 sous le symbole « TWTR » avec une première cotation qui s'effectue à 45,10 dollars. L'action atteindra un pic à 73,31 dollars en décembre 2013 avant d'amorcer une chute jusqu'à 31,85 dollars à la fin du lock-up (période durant laquelle un actionnaire ou un investisseur ne peut se défaire de ses actions) le 6 mai 2014. Dick Costolo démissionne de son poste de PDG de Twitter en juin 2015, sur fond de désaveu de sa stratégie. Il est remplacé de façon intérimaire par l'un de ses fondateurs, Jack Dorsey. [4][5]

[1] https://twitter.com/?lang=fr

3.3. Les Followers :

Twitter a mis en place un concept de followers (suivre les gens). Donc, on a des followers (des personnes qui nous suivent) et ont suit les gens (on est leur follower), c'est-à-dire que l'on suit les informations qu'ils postent et dès qu'un certain utilisateur met à jour son statut, tous les followers sont informés. Ce résultat est obtenu en ajoutant la nouvelle entrée à leur page personnelle, un aperçu est représenté sur la **Figure 1-2.**

Cette opération est réalisée en cliquant sur le bouton **suivre** ou (Follow) sur une page Twitter. On peut suivre tous les autres utilisateurs à moins que cet utilisateur à mis son profil en mode privée. Dans ce cas, une demande d'approbation doit être envoyée en premier.

Figure 1-2 : Capture d'écran de la page personnelle d'ensemble Twitter[2]

[2] https://twitter.com/?lang=fr

3.4. Type des tweets :

Il existe plusieurs types de tweets sont :

- **Tweet normal** : tout message de 140 caractères maximum publié sur Twitter.
- **Réponses :** Tweet qui commence par le @nomdutilisateur d'un autre utilisateur et qui répond à l'un des Tweets de celui-ci, par exemple : @Assistance Je n'arrive pas à croire que tu n'as pas aimé ce film ! [7]
- **Mention :** Tweet contenant le nom d'utilisateur d'un autre utilisateur de Twitter précédé du symbole @, par exemple : Bonjour @Assistance ! Quoi de neuf ? [7]
- **Message direct (DM) :** Un tweet privé envoyé à une personne qui vous suit, vous ne pouvez pas envoyer un message direct à quelqu'un qui vous ne suit pas.[7]

3.4.L'API-Twitter :

L'API Twitter est une passerelle ou interface de programmation permettant de se connecter aux données Twitter de façon automatisée. Elle peut être utilisée pour afficher automatiquement des tweets sur un site web ou pour extraire des données à des fins de veille sur les réseaux sociaux.[9]

Les utilisateurs de Twitter génèrent plus de 400 millions de Tweets tous les jours. Certains de ces tweets sont disponibles pour les chercheurs à travers des API publiques .Il y a plusieurs types d'informations à extraire à partir de Twitter et qui sont les suivants:

- Information sur un utilisateur.
- Tweets publiés par un utilisateur, et
- Les résultats de la recherche sur Twitter.

Pour accéder aux données de Twitter les APIs peuvent être classés en trois types en fonction de leur méthode de conception et d'accès :

- **API REST** sont basés sur l'architecture, REST maintenant couramment utilisés pour la conception des API Web. Ces API utilisent la stratégie d'attraction pour la récupération de données historiques. La réponse à une requête ressemble plus ou moins à ce que l'on obtiendrait en tapant un mot clé dans le moteur de recherche de Twitter. Le nombre de requêtes est limité à 450 demandes toutes les 15 min.[8]
- **Streaming API** fournit un flux continu de l'information publique de Twitter. Ces API utilisent la stratégie de pression pour la récupération de données. Une fois la demande de

renseignements est faite, l'API streaming fournit un flux continu de mises à jour sans autre intervention de l'utilisateur. La limite de l'API est difficilement atteignable, il faudrait atteindre un volume équivalent à 1% des messages publiés sur Twitter à un instant t[8]. Le Streaming API a trois types de paramètres :

- **flux public (Public streams)** : Ce sont des courants contenant les tweets publics sur Twitter.
- **Les flux de l'utilisateur (User streams)** : Ce sont les flux mono-utilisateur.
- **Site flux (Site streams)** : Ce sont des flux multi-utilisateurs et destinés à des applications qui accèdent aux tweets de plusieurs utilisateurs.

Comme les flux publiques est l'API la plus polyvalente pour la collecte des données à partir de l'API Streaming, c'est celle qu'on va utiliser dans notre étude.

- **API Search** est le format public d'API proposé par Twitter qui permet d'effectuer des requêtes automatiques au sein des tweets de façon similaire à la requête pouvant être faite manuellement sur la version web ou mobile de Twitter. [10]
  Elle peut être utilisée pour réaliser de la veille ou de l'analyse de contenus au sein des tweets. Elle est cependant limitée en volume.[10]

3.5.Statistique :

Voici quelques chiffres sur Twitter à prendre en compte pour l'année 2016 :

La plate-forme

- 100 millions d'utilisateurs connectés chaque jour. [11]
- 320 millions d'utilisateurs actifs par mois. [11]
- 500 millions de tweets publiés par jour. [11]
- Twitter est le quatrième plus grand réseau social derrière Facebook, Google+ et Instagram. [11]
- C'est aussi le troisième plus grand réseau social générant le plus de trafic.[11]
- Si Twitter été un pays, il serait le 12ème plus grand du monde.[11]

Les tweets

- Le nombre moyen de tweets chez les hommes est de 567. [11]
- Le nombre moyen de tweets chez les femmes est de 610.[11]

- 63% des utilisateurs considèrent leur smartphone comme appareil principal pour publier des tweets. [11]
- 29% des utilisateurs utilisent un ordinateur de bureau ou portable comme support principal pour publier. [11]
- 550 millions de comptes n'ont jamais publié un seul tweet. [11]
- 44% des comptes n'ont aucun tweet. [11]

- Il y a 117 millions de twittos actifs qui publient des tweets chaque mois. [11]
- 13% des comptes publient chaque mois. [11]

Les marques et la publicité

- 58% des marques les plus influentes ont plus de 100 000 abonnés. [11]
- 47% des personnes qui suivent une marque sur Twitter sont plus enclins à visiter le site internet de la compagnie. [11]
- 86% des revenues de la publicité sur Twitter proviennent du mobile. [11]

Les images sur Twitter

- Les tweets intégrants des images obtiennent 18% de clics supplémentaires. [11]
- …89% de mentions « J'aime » supplémentaires. [11]
- …150% de retweets supplémentaires. [11]

Les utilisateurs

- En moyenne, les utilisateurs passent 170 minutes par mois sur Twitter.
- 78% des utilisateurs sont sur mobile. [11]
- 77% des comptes Twitter ne proviennent pas des Etat-Unis.
- 391 millions de comptes n'ont aucun abonné. [11]
- 81% des comptes ont moins de 50 abonnés. [11]
- 29% des 15/34 ans utilisent Twitter. [11]
- 26% des adolescents définissent Twitter comme étant leur réseau social préféré. [11]
- Il y a environ 20 millions de faux comptes Twitter (sans blague).[11]

- 9,89% des Américains utilisent Twitter au travail. [11]

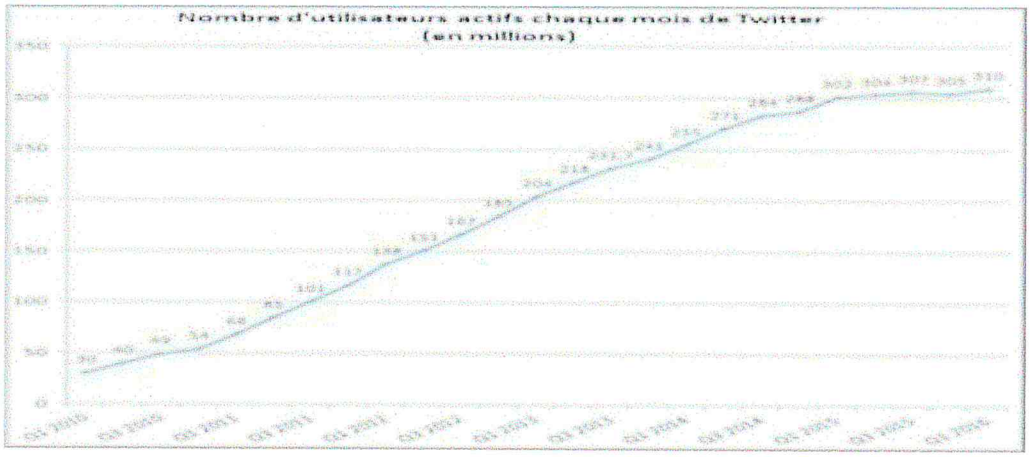

Figure 1-3 : Nombre d'utilisateurs actifs chaque mois de Twitter.[11]

4. Conclusion :

Dans ce chapitre nous avons présenté Twitter, nous avons vu en détail la structure des tweets qui constituent nos données à traiter dans cette étude ainsi que les APIs-Twitter permettant la collection de ces données.

Nous pouvons donc conclure que Twitter a certaines particularités qu'ils le rendent très utile pour la détection d'opinions des sujets émergents, deux aspects importants de ces particularités sont les suivantes :

- **Twitter** est utilisé par un large public pour s'exprimer sur de différents sujets ou des évènements qui se rapportent en temps réel. Il constitue donc, une source précieuse d'information.
- **L'API Twitter** permet un accès facile et efficace pour recueillir un très grand nombre de tweets. Le corpus recueilli peut être arbitrairement grand, contenant des mines d'information qui vont être exploitées par la suite.

Dans le chapitre suivant, nous allons présenter un état de l'art concernant la détection des opinions ou l'analyse des sentiments des tweets. Nous allons étudier en profondeur différentes approches réalisées dans ce domaine.

Chapitre 02 : état de l'art

1. Introduction :

Twitter est l'un des réseaux sociaux en ligne les plus populaires à ce jour, où les utilisateurs postent leurs opinions sous forme de texte court appelé «tweets». Ces tweets sont généralement limités à 140 caractères.

Twitter a environ 100 millions d'utilisateurs connectés chaque jour, environ 500 millions de Tweets sont envoyés chaque jour [11]. Selon Bollen, "un Tweet est un microscopique, instanciation temporellement authentique du sentiment ".Les tweets sont croustillantes et brèves, le sentiment public peut être facilement exploré. **[12]**

L'analyse des sentiments est connue comme l'exploitation de la fouille d'opinion qui se réfère à l'utilisation des ressources naturelles de traitement automatique du langage. Elle est utilisée pour déterminer l'attitude d'un auteur, par rapport à un sujet particulier ou la polarité contextuelle globale dans le texte. La croissance rapide des médias sociaux a suscité l'intérêt de l'analyse des sentiments. Le défi de la détection de la criminalité dans une zone géographique en analysant des tweets qui ont une relation avec un crime et en effectuant ensuite l'analyse des sentiments pour identifier les zones sujettes à la criminalité en temps réel. La plupart des études portant sur le patron du crime de détection utilisent des techniques d'exploration de données pour mieux comprendre les données historiques.
Dans ce chapitre, nous allons présenter une étude approfondie de la fouille des messages tweets et plus précisément **l'analyse des sentiments des tweets concernant les crimes.**
Nous présenterons la description des approches existantes dans ce domaine.

2. L'analyse des sentiments :

L'Analyse des sentiments est utilisée pour décrire l'analyse automatique de texte évaluatif et pour la recherche de valeur prédictive des jugements.

L'analyse des sentiments est une tâche de la fouille de données et la fouille de texte qui essaye de définir les sentiments présents dans un texte ou un ensemble de texte. Elle est utilisée pour analyser des grands volumes de texte. L'analyse de sentiments et la fouille d'opinion sont devenues des sujets de recherche à la mode. Les auteurs y décrivent les approches et les techniques mises en œuvre dans la recherche d'informations orientée vers la fouille d'opinion. Bien que l'analyse de sentiments et la fouille d'opinion soient souvent considérées ensemble dans les publications scientifiques, certains auteurs considèrent même les deux termes comme des synonymes. **[13]**

2.1.La Fouille de données (data mining) :

C'est un ensemble des techniques d'exploration de données permettant d'extraire des connaissances sous la forme de **modèles de description** afin de : **[14]**

- **Décrire** le comportement actuel des données.
- Et/ou **Prédire** le comportement futur des données.

2.2.La fouille de texte :

La fouille de textes (text mining) est l'héritière directe de la fouille de données (data-mining), née dans les années 90. Elle concerne les processus interactifs et itératifs de découverte de connaissances dans des grandes collections de documents. **[15]**

La Fouille de données et la fouille de textes possèdent en commun des méthodes et algorithmes tels que les algorithmes de recherche par niveaux de motifs vérifiant certaines propriétés, l'exemple le plus classique de propriété étant la fréquence. La fouille de textes présente ses propres spécificités : les documents peuvent être plus ou moins structurés, la phase de prétraitement joue un grand rôle. Il est courant et souvent intéressant de prendre en compte l'ordre des mots, et il est maintenant établi que des ressources provenant du traitement automatique des langues et/ou de la linguistique sont nécessaires pour apporter des résultats applicatifs avec une réelle plus-value. Beaucoup de méthodes de fouille de textes incluent des traitements d'analyse pré-lexicale (e.g., traitement des chiffres), d'analyse lexicale telle que l'élimination de « mots vides », morphologie, analyse syntaxique (e.g., détermination des

groupes nominaux), analyse sémantique incluant des apports linguistiques aussi bien que des particularités des textes étudiés. **[15]**

Les termes analyse de sentiments (sentiment analysis) et fouille d'opinion (opinion mining) sont souvent utilisés de manière interchangeable. Le vocable analyse de sentiments est préféré dans le domaine du traitement automatique des langues, tandis que le terme fouille d'opinion a été lui adopté par la communauté de la recherche d'informations. Bien que ces deux termes concernent des champs d'investigation très proches, qui pourraient même être considérés comme une seule et même entité.

2.3.La fouille des Tweets :

Le **tweets mining** est une technique permettant d'automatiser le traitement de gros volumes de contenus des messages Tweets pour extraire des connaissances comme les principales tendances et répertorier de manière statistique les différents sujets évoqués. Les techniques de tweet mining sont surtout utilisées pour des données déjà disponibles au format numérique.

3. La prédiction d'un crime à partir de Twitter :

La criminologie est l'étude pluridisciplinaire du phénomène criminel. Elle prend appui sur les sciences humaines (psychologie, sociologie, droit, etc.) pour connaître le délit, le délinquant, la victime, la criminalité et la réaction sociale face au crime.

Bien que le phénomène criminel soit connu de tous, notamment par l'entremise des médias, sa compréhension reste souvent anecdotique et fragmentaire. Le criminologue va au-delà du fait divers et aborde la question criminelle en toute rigueur, en privilégiant une approche critique. Il analyse d'abord le crime et ensuite le type d'intervention entourant celui-ci.

L'étude de criminologie est basée sur l'identification des caractéristiques criminelles, des trésors d'information cachés sur Twitter peuvent être extrêmement utiles dans la lutte contre la criminalité, certains délits ou agressions pouvant être détectés à l'avance à condition de les analyser correctement, donc l'analyse de tweets permet de prédire la criminalité.

3.1. Les différentes approches :

3.1.1. Première approche :

ASHOK BOLLA propose une approche pour détecter la criminalité basée sur le traitement de données à partir de Twitter, l'analyse géographique et l'analyse des sentiments sur les données extraite. Cette approche permet en premier de récupérer les dix premières villes dangereuses avec leur géo localisation. Les Tweets ont été captés dans une 50km autour de la géo localisation centrale de la ville. Il utilise une stratégie de recherche par mot clé pour identifier les tweets liés à la criminalité inclus «pistolet », « crime», «sinistre », « tuer». Chaque tweet a été analysé avant l'analyse des sentiments. Cette analyse permet de séparer les termes individuels dans un tweet en fonction des limites d'espace blanc, convertir le tweet en lettres minuscules et retirez tous les caractères non alphanumériques de tweets (par exemple, les signes et tirets). Il effectue une analyse géographique sur les tweets extraits dont l'objectif principal est d'identifier et de filtrer les tweets. Enfin L'analyse des sentiments est utilisée pour déterminer l'attitude de l'auteur par rapport soit à un sujet particulier ou à une polarité contextuelle globale du document.**[16]**

Cette analyse utilisée pour évaluer le sentiment qui extraits les tweets utilisant deux techniques de l'analyse des sentiments, la technique à savoir sur la base ANEW et le modèle de l'apprentissage en profondeur. **[16]**

L'ANEW (Affective Norms for English Words) est développé pour fournir un ensemble de notes émotionnelles normatives pour un grand nombre de mots dans la langue anglaise. Cela a été mis au point pour aider les chercheurs lors de l'étude des émotions ; il est souvent utilisé pour déterminer le sentiment d'un tweet. **[16]**

ANEW est une liste de mots anglais construite par Bradley et Lang . Il y a 1034 mots notés pour la valence , l' excitation et la domination . L'ANEW fournit un ensemble de notes émotionnelles normatives pour un grand nombre de mots dans la langue anglaise. Cet ensemble de verbes ont été notés en termes de plaisir, de l'excitation, et la domination dans le but de créer une norme pour une utilisation dans les études d'émotion et d'attention. **[16]**

Dans les techniques basées sur **ANEW**, un effort a été fait dans la cartographie de chaque terme **d'ANEW** à son équivalent dans le tweet. Cela a ensuite été appliqué pour améliorer la mise en correspondance. **[16]**

Le modèle d'apprentissage en profondeur est défini comme un programme d'analyse des sentiments multi-classe. Il permet de classer les tweets dans cinq classes (Très positive,

Positive, Neutre, négatif et Très négative), Les tweets qui ont été classés comme étant soit négative ou très négative ont été identifiés comme contribuant à l'intensité de la criminalité.il fonctionne réellement sur la structure de la phrase. Il est utilisé pour extraire le sentiment de chaque classe. **[16]**

3.1.2. Deuxième approche :

Wang et al ont proposé une approche de la modélisation de la criminalité .Dans leur approche les affaires criminelles antérieures sont utilisées comme données d'apprentissage supervisé dans le modèle prédictif .cette approche suit les étapes suivantes :

- recueillir un corpus de tweets. **[17]**
- extraire ensuite les événements du principal contenu textuel de chaque tweet en utilisant une technique de NLP connue sous le nom SRL **(Semantic Role Labeling)**.

La prédiction de la criminalité basée sur Twitter repose sur une compréhension sémantique de tweets, qui va au-delà de sac de mots et les représentations de sentiment. Une telle compréhension peut être dérivée d'un processus connu sous le nom (SRL), qui extrait les événements mentionnés dans les tweets, les entités impliquées dans les événements et les rôles des entités par rapport aux événements. **[17]**

- appliquer **l'allocation latente** de **Dirichlet (LDA)**, après le traitement des tweets avec les systèmes de SRL, ils ont de multiples événements $\mathbf{e_i}$ associés à un jour. En termes de modélisation du sujet, chaque jour **d** est associé à un document $\mathbf{do_{cd}}$ contient des "mots" $\{\mathbf{e_1}; \mathbf{e_2}; \ldots; \mathbf{e_{nd}}\}$, où **e** est la longueur de $\mathbf{do_{cd}}$. Ces mots décrivent ce qui est arrivé le jour **d**.

Comme sujet modélisation de documents textuels réels, ils ont émis l'hypothèse que les événements de jour seraient latents de façon particulière. Ainsi, au lieu d'utiliser $\mathbf{do_{cd}}$ directement pour prédire les futurs incidents, ils ont extrait plus de sujets $\{\mathbf{t_1}; \mathbf{t_2}; \ldots; \mathbf{t_k}\}$ de $\mathbf{do_{cd}}$ utilisant **l'allocation Dirichlet latente** (LDA). **LDA** est un modèle de langage probabiliste qui peut être utilisé pour expliquer la façon dont une collection de documents est générée à partir d'un ensemble de sujets cachés (ou latents). **LDA** découvre les sujets en fonction de mots et réduit la dimensionnalité des documents mentir dans l'espace des sujets **k** dimensions. Étant donné le nombre de sujets **k**, **LDA** peut estimer la répartition thématique du document $\{T_{d,1}; T_{d,2}; \ldots; T_{d,k}\}$, où $T_{d,i}$ est la probabilité que le document **d** est lié au sujet **i**. Ils ont appliqué **LDA** pour dériver $\{T_{d,1}; T_{d,2}; \ldots; T_{d,k}\}$ pour les événements décrits dans les tweets le jour **d**. Intuitivement, cette analyse à propos de la relation entre les **k** événements majeurs (latentes) au jour **d** et les événements observables $\mathbf{e_i}$ qui étaient rapporté par les

agences de presse. Ceci réduit la dimension de do_{cd} . **[17]**

2. 1.3. Troisième approche:

Gerber propose une approche pour prédire où les crimes peuvent se produire. Ils ont développé un programme informatique, utilisé aujourd'hui par la ville de Chicago, visant à prédire les différents types de crimes. Les Tweets peuvent être utiles pour prédire 19 à 25 sortes de crimes, en particulier pour des infractions telles que le harcèlement, les vols, et certains types d'agression. Il commence a collecté de plus de 1,5 million de tweets publics marqués avec les coordonnées GPS dans les limites de la ville entre Janvier et Mars, 2013. Pendant ce temps, il ont rassemblé des informations sur tous les crimes documentés qui se sont produits au cours de la même période. Ensuite, ils ont créé un algorithme informatique qui séparait les tweets dans 1 km par 1 quartier de kilomètres, puis il analyse le contenu des tweets dans chaque quartier pour savoir ce que les gens ont tweetés, il utilise (KDE) **Kernel densité estimation**, ce dernier est un moyen non paramétrique pour estimer la fonction de densité de probabilité d'une variable aléatoire. Les densités de la criminalité pour les incidents de vol sont calculées par KDE. KDE correspond à une fonction de densité de probabilité spatiale à deux dimensions à un dossier de crime historique. Cette approche permet à l'analyste de visualiser rapidement et identifier les zones avec des concentrations historiquement élevées de criminalité. L'ajout de données Twitter améliore les performances de prédiction de la criminalité par rapport à une approche standard basée sur l'estimation de la densité du noyau. En décodant les formes d'expression utilisées dans les tweets, pour y repérer des indices indirects faisant allusion à différents types de criminalité. Le contenu a ensuite été regroupé dans des centaines des "sujets". Les choses deviennent un peu techniques à partir d'ici. En termes de base, le modèle de Gerber compare les sujets dans un quartier aux données sur la criminalité historique de ce même endroit dans la ville pour un mois donné. Le modèle forme des corrélations entre les sujets et les crimes, puis utilise ces corrélations pour prédire la criminalité dans le même quartier pour un mois ultérieur. **[18]**

3.1.4. Quatrième approche :

Chen et al ont discuté sur la prédiction de la criminalité utilisant les sentiments de Twitter et des données météorologiques. Ils ont appliqué l'analyse de data mining pour estimer la polarité des tweets, KDE pour déterminer la densité de la criminalité, et la méthode de régression logistique pour faire des prédictions sur les futurs incidents de vol. Ils ont utilisé

un dictionnaire de lexique de sentiment et «une fonction de polarité » pour estimer les valeurs de polarité de sentiment pour chaque tweet posté par les utilisateurs dans chaque quartier pendant 6 heures chaque journée. Dans la fonction de polarité, un cluster de contexte de mots est tiré d'autour d'un mot polarisé pour être utilisé comme variante shifters. Chaque mot dans ce contexte a été marqué dans quatre catégories différentes, y compris neutre, négateur, amplificateur, et de-amplificateur. Chaque mot polarisé dans chaque document a ensuite été calculé sur la base des scores de polarité. Afin de calculer la polarité, ils ont attribué des scores pour les mots en utilisant le dictionnaire de lexique. Des mots avec des connotations négatives ont été assignés des scores négatifs de polarité, et les mots avec des connotations positives ont été notés comme positifs. Après avoir évalué les scores de polarité des documents, ils ont ensuite utilisé amplificateurs ou amplificateurs DE (valeur par défaut est 8, d'amplifier le poids, et limité à -1 comme borne inférieure) pour trouvé la polarité de chaque document. Enfin, avec le pôle de contexte de mots sont cités, ils ont divisé par la racine carrée du nombre de mots dans chaque document, qui a produit le score de polarité pour chaque document. Ils ont évalué le modèle prédictif de la criminalité utilisant une analyse de sentiment et les facteurs météorologiques efficaces en comparant sa prédiction aux incidents de vol réels qui se sont produits à travers Chicago, entre le 25 Décembre 2013 et 30 Janvier 2014. La performance des modèles de prédiction de la criminalité logistique est mesurée par une courbe de surveillance. Ils concluent que s'il y a une tendance croissante du sentiment de polarité, un taux de criminalité plus élevé pourrait être prévu. Pendant ce temps, à une température plus élevée, il y a un risque plus élevé de crimes de vol, tandis que la faible humidité réduit le risque de crimes de vol. Il y a beaucoup de façons d'étendre ce travail. La façon la plus directe pour améliorer la précision de prédire est d'obtenir des données de prévisions météorologiques dans chaque période de temps de 6 heures. Pour l'instant, ils ont récupéré seulement des prévisions météorologiques quotidiennes à partir des ressources accessibles au public. De plus, une fois que nous avons accès à des données météorologiques spatialement différenciées pour chaque secteur spécial différent, le pouvoir prédictif du modèle peut améliorer. Enfin, ils ont utilisé la régression logistique qui permet d'utiliser les caractéristiques de Twitter et les facteurs météorologiques prévus pour prédire la probabilité d'incidents survenus dans une période spécifique. **[19]**

3.1.5. Cinquième approche :

ABBOUTE et al développent une application permettant d'utiliser le réseau social Twitter pour la prévention des suicides. Cette approche est basée sur les étapes suivantes :

- **Collecte des données :**

1. Définir un vocabulaire associé aux différentes thématiques critiques liées au suicide (e.g. dépression, peur, harcèlement, etc.). Du fait que la plupart des messages sont publiés en anglais sur Twitter, il était préférable de définir le vocabulaire dans cette même langue.
De plus, le vocabulaire doit être divisé en différentes catégories et sous-catégories afin de pouvoir identifier facilement le degré de menace provenant du tweet. Par exemple, dans le cas d'un harcèlement via un tweet, il est nécessaire d'identifier les destinataires des tweets (car ce sont eux qui risquent de passer à l'acte). Au contraire, pour les autres catégories du vocabulaire, ce sont les personnes qui ont publié le tweet qui doivent être identifiées.
Ils ont enregistré les différents termes au format CSV (Comma-separated values) ce qui permet de réutiliser ces données dans d'autres contextes **.[20]**
2. Récupération des tweets :
Collecter des tweets via l'API Twitter. Pour cela, ils utilisent la librairie Java Twitter4J.
Ils ont utilisé l'API Rest pour récupérer les tweets ayant un lien avec les différentes thématiques critiques collectées**.[20]**

- **Création de la base de données :**

Ils ont créé une base de données implémentée sous MySQL qui permet d'enregistrer les tweets suspects ainsi que les termes permettant d'affirmer si un tweet est effectivement suspect ou pas**.[20]**

- **Classification des tweets:**

ABBOUTE et al classer les Tweets en deux classes : classe pour les Tweets suspect risque de suicide et classe de Tweets suspect non risque de suicide.
Ces Tweets "classifiés manuellement" ont été chargés dans l'application et Weka les utilise pour apprendre un modèle prédictif. Par la suite,Weka sera capable de classifier les Tweets suspects selon les deux classes. En sortie, Weka fournit donc les Tweets suspects classifiés**.[20]**

4. Comparaison:

Dans cette partie nous allons comparer les travaux de l'analyse de sentiment sur les crimes à partir des tweets étudiés précédemment en utilisant un tableau comparatif et en se basant sur les critères de comparaison décrits ci-dessous :

- **Collecte de tweets :**
  1. **Zone géographique :** permettant de trouver pour une zone donnée des messages contenant certains éléments ou de visualiser les termes les plus utilisés sur cette zone.
  2. **Technique de recherche :** stratégie sur lequel les approches cherchent les tweets postés sur l'environnement Twitter.
- **La classification de données :**

C'est les méthodes d'analyse de données, leurs objectifs est d'obtenir une représentation schématique simple d'un tableau de données complexe à partir d'une typologie (segmentation), c'est à dire d'une partition des n individus dans des classes, définies par l'observation de p variables. Il existe deux types de classification : supervisée et non supervisée.

1. Classification supervisée :

L'objectif de la classification supervisée est principalement de définir des règles permettant de classer des objets dans des classes à partir de variables qualitatives ou quantitatives caractérisant ces objets.

Méthodes :

- Les k plus proches voisins
- Le classifieur Bayésien naïf
- Arbres de décision
- Réseaux de neurones
- SVM

2. Classification non supervisée :

Il s'agit pour un système de diviser un groupe hétérogène de données, en sous-groupes de manière à ce que les données considérées comme les plus similaires soient associées au sein d'un groupe homogène et qu'au contraire les données considérées comme différentes se retrouvent dans d'autres groupes distincts ; l'objectif étant de permettre une extraction de connaissance organisée à partir de ces données.

Méthodes :

- K-means.
- Les réseaux de neurones
- Les algorithmes génétiques

- **Temps réel :** ces travaux sont en temps réel ou non.
- **Facteurs d'entrées**: les facteurs d'entrées décrivent la source de l'information à analyser.

Les dimensions de couverture qu'on a choisies sont ''nombre de tweet'' et ''genre''

- **Langue** : quelle est la langue utilisée.
- **Algorithme** : le nom de l'algorithme de classification de chaque approche.

Synthèse :

Structure	Approche	Algorithme De classification	Collecte de tweets		Facteurs d'entrées		En Temps réel	Classification	langue
			Zone géographique	Technique de recherche	Nombre de tweet	Genre			
	Première	modèle d'apprentissage en profondeur	détecter la criminalité des tweets de la ville États-Unis	stratégie de recherche par mot clé	100,000 tweets	Sujet de criminalité historique	non	supervisée	anglais
	deuxième	LDA	Détecter la criminalité de tous les tweets postés sur Twitter	recherche par mot clé	140 million tweets	Sujet de criminalité Réelle	oui	supervisée	anglais

	Troisième	KDE	détecter la criminalité des tweets de la ville de Chicago	KDE (basé sur un dossier historique de crime avec leurs géo localisation)	1.5 million de tweets marqué par GPS	Sujet de criminalité Réelle	oui	Non	anglais
	Quatrième	KDE	Détecter la criminalité de tous les tweets postés sur Twitter	un vocabulaire de mot	1.5 million de tweets	Sujet qui concerne le vol	oui	Supervisée	anglais
	Cinquième	IBK	Prédire les personnes qui sont risque de suicide dans la France.	un vocabulaire de mot.	Un grand nombre de tweets.	Sujet qui concerne le suicide (prévention des suicides).	oui	supervisée	Français

Tableau 2-1 : comparaison des différentes approches

Les avantages des approches :

- Utilisation d'un modèle d'apprentissage en profondeur qui permet de classer les tweets automatiquement comme c'est le cas pour l'approche d'ASHOK BOLLA .[16]
- La compréhension sémantique des tweets pour la prédiction de la criminalité pour l'approche de Wang et al. [17]
- La classification automatique des tweets avec l'outil de classification weka pour les travaux de . [20]

Les inconvénients des approches :

- La récupération des dix premières villes dangereuses, pour l'approche d'ASHOK BOLLA [16], ce programme est limité on ne peut pas récupérer toutes les villes dangereuses dans un pays.
- La classification manuelle des tweets pour les travaux d'ABBOUTE et al. [20]

5. Conclusion :

Nous avons décrit des travaux qui nous intéressent à l'analyse des sentiments. Ceux-ci nous ont apporté des idées pour la détection des opinions sur les crimes à partir de Twitter.
Dans le chapitre suivant, nous allons proposer une nouvelle approche qui rentre dans le domaine du tweet mining pour le but de l'analyse des sentiments sur les crimes qui font l'actualité sur Twitter.

Chapitre 03 :Conception

Chapitre 03 : Conception

1. Introduction :

Ce chapitre présente une nouvelle approche de l'analyse des sentiments sur les crimes à partir de Twitter. Nous avons pris comme base les cinq travaux qu'on a étudiés dans le chapitre précédent. Nous nous sommes basés surtout sur l'approche de ABBOUTE et all [20] qui se traite un seul crime suicide, Celle-ci s'appuie sur l'extraction d'informations contenues dans la structure des tweets dans le but de les classifier et s'appuie aussi sur la classification avec outil Weka qui contient plusieurs classifieur sans les implémenter. Dans notre approche nous avons ajouté une étape de prétraitement on utilisant la bibliothèque Stanford NLP et appliquer la classification sur les tweets prétraités.

2. Présentation de la démarche utilisée :

Nous expliquons dans cette étape, les besoins de notre application afin de pouvoir passer à l'étape de conception et d'architecture. Nous présentons le cycle de vie que nous avons suivi pour la réalisation de ce projet. Nous illustrerons les solutions apportées par notre outil face aux problèmes posés, en se basant sur le langage UML (Unified Modeling Language) en utilisant le processus UP (Unified Process).UP est une méthode de prise en charge du cycle de vie d'un logiciel développé en orienté objet. Il représente les étapes du cycle de vie sous forme de diagrammes UML.

2.1 Le cycle de vie d'un logiciel :

Le cycle de vie d'un logiciel est la période située entre le début de la conception et l'arrêt de l'exploitation de ce logiciel. Il regroupe un ensemble d'activités et correspond à l'identification des états successifs d'une application ou d'un produit déterminé. Il est essentiellement dynamique, évolutif et presque toujours progressif. Il est envisagé à un instant donné et va comprendre les progrès technologiques et les contraintes organisationnelles. **[21]**
En ce qui concerne notre projet nous avons suivi le modèle en cascade.

2.2 Modèle en cascade :

Le modèle de cycle de vie en cascade a été mis au point dès 1966, puis formalisé aux alentours de 1970. Il définit des phases séquentielles (chaque étape doit être terminée avant que la suivante commence) qui ont pour but de réaliser un produit logiciel fini et testé **[22]**. La figure suivante représente le modèle en cascade :

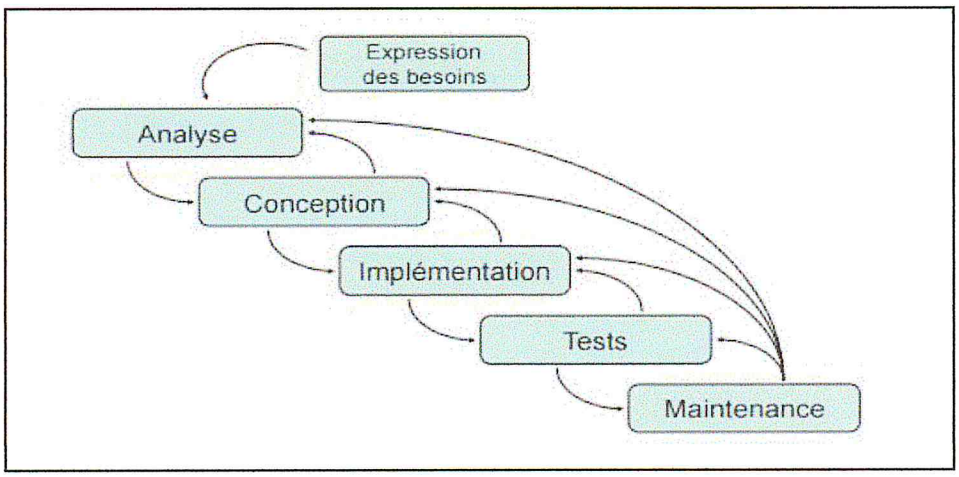

Figure 3-1 : le modèle en cascade.[23]

2.3. Description des phases de modèle:

2.3.1. Expression des besoins :

La spécification des besoins est une étape essentielle au début de processus de développement, elle consiste généralement à déterminer :

- La Description du problème à traiter afin d'identifier les besoins de l'utilisateur, de spécifier ce que doit faire le logiciel : informations manipulées, services rendus, interfaces, contraintes
- Et la mise en œuvre des principes : abstraction, séparation des problèmes, séparation des besoins fonctionnels.

Cette étape ne préoccupe pas des solutions, mais des questions : elle identifie le «quoi faire?»Et identifie les entités de l'environnement du système. **[23]**

2.3.2. Analyse :

Cette étape consiste à :

- Répondre au « Que fait le système ? », Modélisation du domaine d'application
- Analyse de l'existant et des contraintes de réalisation
- Abstraction et séparation des problèmes, séparation en unités cohérentes Afin de déterminer un dossier d'analyse (Modèle du domaine, Modèle de l'existant (éventuellement) , modèle conceptuel) et un plan de validation, dossier de tests d'intégration. **[23]**

2.3.3. Conception :

C'est la phase la plus importante du processus de développement d'un logiciel. Elle permet de :

- répondre au « comment réaliser le système ».
- Décomposition modulaire, définition de chaque constituant du logiciel : informations traitées, traitements effectués, résultats fournis, contraintes à respecter Pour déterminer un dossier de conception et un plan de test global et par module. Dans ce projet en utilisant le diagramme de classe d'UML. **[23]**

2.3.4. Implémentation :

L'objectif de cette phase est la réalisation des programmes dans un (des) langage(s) de programmation et tests selon les plans définis lors de la conception pour construire un (les) dossier (s) de programmation et le(s) code(s) source(s). **[23]**

4. Notre approche :

Notre approche est constituée de trois grandes phases dont chacune est composée de différentes étapes.

Phase 1 : l'extraction des données :

- La collection des données à partir de Twitter.
- Stockage des données collectées.
- Construction du vocabulaire.
- Prétraitement des données.

Phase 2 : Apprentissage :

- Calcul du poids de thématiques avec TF-IDF.
- Construction de la matrice de poids.

Phase 3 : Classification :

- La classification des tweets en utilisant Weka.

4.1. Schéma global de notre approche :

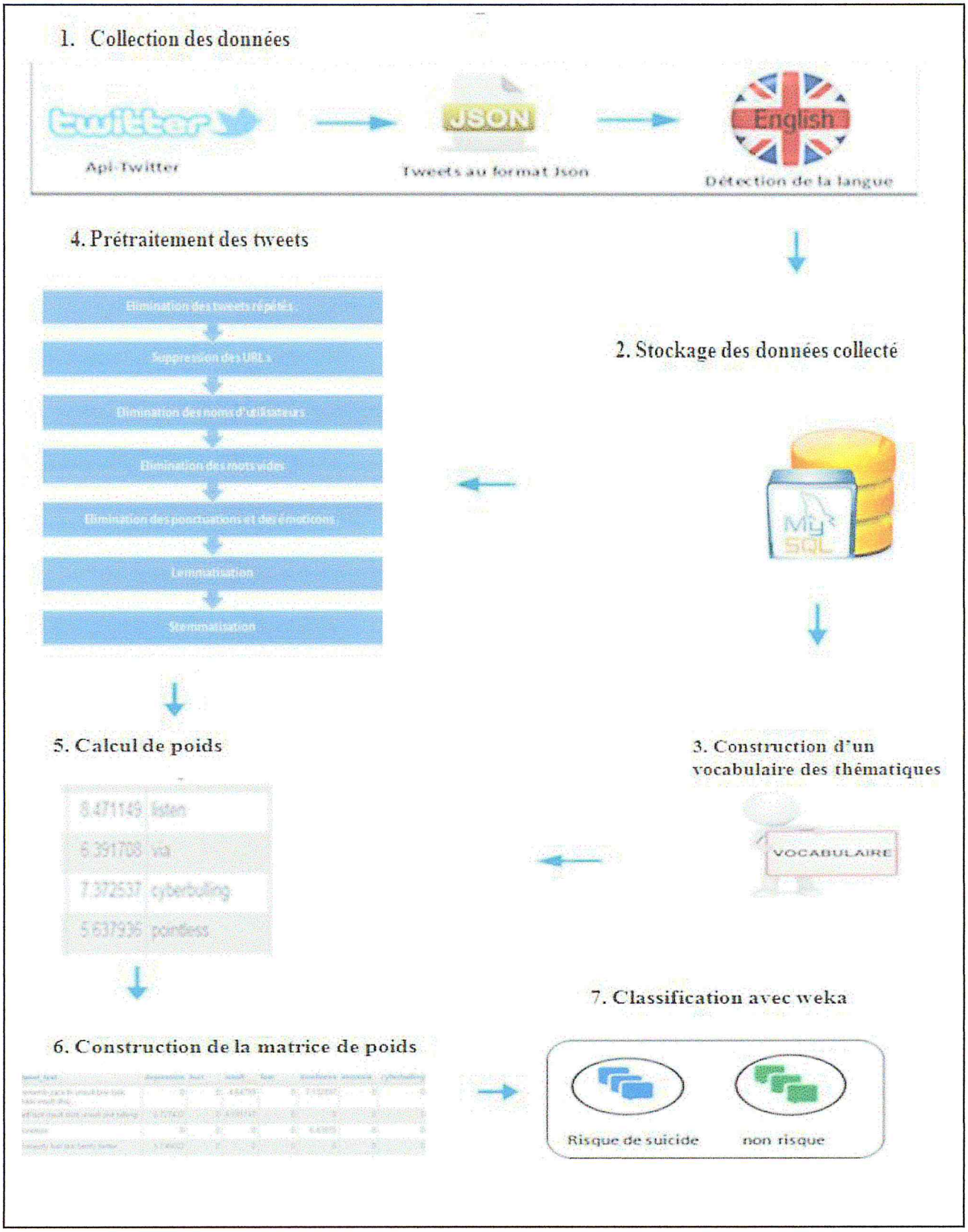

Figure 3-3: Schéma global de notre approche

4.2. Description des étapes de notre approche :

4.2.1 Collection des données :

Cette phase est la tâche la plus importante dans le processus d'analyse des données. Elle consiste à récupérer l'ensemble des données à traiter dans notre approche.

- **L'accès à l'API** :

Comme décrit dans le chapitre 01 concernant l'API Twitter, il y a plusieurs méthodes pour collecter les tweets. Ici nous avons utilisé le ''flux public'' (Public streams) où les données retournées par l'api de Twitter sont en format JavaScript Object Notation (JSON). **JSON** (JavaScript Object Notation) est un format d'échange de données en texte lisible**[24]**. Il est utilisé pour représenter des structures de données et des objets simples dans un code qui repose sur un navigateur Web.

Les API Twitter ne sont accessibles que via des requêtes authentifiées. Twitter utilise l'authentification ouverte Oauth (Open authentification) et chaque demande doit être signée avec des informations d'identification valable d'un utilisateur donné. L'accès à l'API Twitter est également limité à un certain nombre de demandes dans un laps de temps réel appelé la limite de vitesse (rate limit). Une fenêtre de limite de vitesse est utilisée pour renouveler le quota d'appels de l'API autorisé périodiquement. Nous avons limité le nombre de tweets retournés à 400 tweets.

Actuellement, à l'aide des APIs Twitter nous pouvons récupérer facilement des tweets, qui répondent à un ensemble de paramètres (e.g. mots-clés). Ces APIs renvoient des données bien structurées pour faciliter leur analyse et l'accès à l'information désirée. Nous avons utilisé l'API streaming qui permet de récupérer les tweets en temps réel.

On commence par créer une application dans Twitter en utilisant le site de développeur **dev.twitter.com**, les applications sont connues en tant que consommateur. Après la création de l'application, **l'API Twitter** fournit 4 codes : (**consumer key**) et (**consumer secret**) pour l'authentification, (**access token**) et (**access secret**) pour la vérification de l'authentification. Ce protocole fournit une alternative plus sure au niveau de la sécurité des mots de passe. Ces 4 codes par la suite sont utilisés via une bibliothèque spéciale pour l'authentification et la récupération des données. La recherche est effectuée par **mot clé** avec la langue anglais et dans un pays choisi. Le mot clé utilisé c'est « suicide » de plus, nous avons collecté des phrases entières qui sont souvent utilisées par des gens qui ont l'intention de se suicider comme par exemple, "I want to die" ou "My family would be better off without me". "hate my life".

- **choix de pays :**

Le tableau 3-1 représente une liste des taux de suicide par pays selon les données de l'Organisation mondiale de la Santé et d' autres sources, dans lesquelles le rang d'un pays est déterminé par le total de ses décès de taux officiellement enregistrés comme des suicides dans la dernière année disponible. Les statistiques OMS sont basées sur les rapports officiels de chaque pays respectif.

Les deux sexes rang	Pays	Les deux sexes	Mâle rang	Mâle	Femelle rang	Femelle
1	Guyana (plus d' info)	44.2	1	70,8	1	22.1
2	Corée du Sud (plus d'info)	28,9	5	41,7	5	18,0
3	Sri Lanka (plus d' info)	28,8	3	46,4	7	12.8
4	Lituanie (plus d' info)	28.2	2	51,0	29	8.4
5	Suriname	27,8	4	44,5	11	11.9
6	Mozambique (plus d' info)	27.4	8	34.2	2	21.1
7	Tanzanie	24.9	13	31,6	4	18.3
	Népal (plus d' info)	24.9	17	30.1	3	20,0
9	Kazakhstan (plus d' info)	23,8	6	40,6	21	9.3
dix	Burundi	23.1	9	34.1	9	12.5

Tableau 3-1 : Suicides par 100.000 personnes par an (normalisés selon âge) [25]

Dans notre approche nous allons choisir le pays avec le taux le plus élevé dans le monde Qui représente « GUYANA ».

4.2.2 Stockage des tweets :

Chaque tweet récupéré est stocké directement dans la base de données MYSQL. les mots clés de la collection sont choisis par l'utilisateur, la langue et le nom de la ville ou la géo localisation de la ville sont choisis par défaut en tant que mot clé de la recherche.

4.2.3. Construction d'un vocabulaire des thématiques :

Dans cette étape, nous identifions un vocabulaire associé aux thématiques critiques liées au suicide (Depression, Hurt, Insult, Fear, Loneliness, anorexia, cyberbulling).

Du fait que la plupart des messages soient publiés en anglais sur Twitter, donc nous avons préféré de définir le vocabulaire dans cette même langue. Pour chaque thématique, nous constituons un vocabulaire avec une liste des termes enregistrés dans un document.

4.2.4. Prétraitement des données :

Le contenu textuel des tweets récupérés contient des données non structurées, des incohérences typographiques, des mots vides de sens, des ponctuations …etc, donc ces données nécessitent toutefois un nettoyage et une normalisation. Cette étape est appelée le prétraitement des tweets et elle est très essentielle, elle est constituée des étapes suivantes :

- **Elimination des tweets répétés :**

Dans la phase d'apprentissage, il faut avoir une seule occurrence pour chaque tweet, pour ne pas affecter les probabilités d'occurrences des mots, qui est la caractéristique la Plus importante dans la phase d'apprentissage.

- **Suppression des URL :**

Les liens du web dans un tweet, certainement, ne contiennent pas des sentiments, donc il est préférable de les supprimer.

- **Elimination des noms d'utilisateurs :**

Les noms d'utilisateur dans les tweets sont toujours précédés par un \@", ils ne fournissent Pas une information sur le sentiment.

- **Elimination des mots vides (Stop Words) :**

Les **mots vides** (ou stop Words, en anglais) sont des mots très communs et utilisés dans pratiquement tous les textes. Leur présence peut dégrader la performance de l'algorithme de classification en termes de coût et en termes de précision de la classification. En anglais, des mots vides évidents pourraient être:

« are », « after », « and », « such », « before », « because », « between», «she »,…un mot vide est un mot non significatif figurant dans un texte, Ce mot apparaît avec une fréquence semblable dans chacun des textes de la collection n'est pas discriminant, ne permet pas de distinguer les textes les uns par rapport aux autres. Il existe différentes collections des mots vides de la langue anglaise sur le web, nous avons utilisé une liste qui contient les mots vides puis nous allons faire le filtrage.

- **Elimination des ponctuations et des émoticons :**

Cette étape consiste à filtrer toutes les ponctuations et les émoticons qui se trouvent dans les tweets puisqu'ils peuvent porter plusieurs sentiments, mais leur forte présence peut affecter négativement la précision de l'algorithme.

- **Lemmatisation :**

La lemmatisation est la réduction d'un mot à sa forme la plus simple, par exemple transformer tous les verbes conjugués à l'infinitif, tous les mots pluriel au singulier et le féminin au masculin. Ce traitement permet de normaliser les mots qui dérivent de la même racine et de les traiter comme s'ils étaient le même afin de leur attribuer un poids unique. [26]

- **Stemmatisation (stemming) :**

La stemmatisation est le processus d'élimination de suffixes des mots afin d'obtenir leur racine commune. Cela permet de générer la forme de base (souvent tronquée) appelée le stem (Racine en français). Par exemple :{computer, computing, computation} devient « comput »**[27].** Un des algorithmes les plus populaires de stemmatisation est celui de porter.

- **Détail de l'algorithme de porter :**

Soient v représente une voyelle (y est considéré comme une voyelle s'il est précédé par une consonne), c représente une consonne, V représente une suite de voyelles et C représente une suite de consonnes. Donc un mot en anglais peut être de l'une des 4 formes suivantes:

- CVCV…… C
- CVCV…… V
- VCVC…… C
- VCVC…… V

Ce qui peut se représenter par :

[C]V CV C … [V] Ou **[C](V C)^{m}[V]**

Où **m** est appelée la mesure d'un mot.

m = 0: tree, by.
m = 1: trouble, oats, trees, ivy.
m = 2 : troubles, private, oaten, orrery.

Les règles de dé suffixation sont exprimées sous la forme $(condition)S_1 \rightarrow S_2$ ce qui signifie que si un mot se termine par S_1 et que le préfixe satisfait la condition alors le suffixe S_1 est remplacé par S_2

- $*e$: le préfixe se termine par la lettre e
- $*v*$: le préfixe contient une voyelle
- $*d$: le préfixe se termine par une consonne doublée
- $*o$: le préfixe se termine par cvc où le second c n'est ni w, ni x, ni y.

Il est possible d'utiliser des opérateurs booléens: et, ou, [28]

- **Les étapes de l'algorithme de porter :**

Étape		Règles	Exemples
1	A	□ SSES → SS □ IES → I □ SS → SS □ S →	caresses→caress ponies→poni caress→caress cats → cat
	B	□ (m>0) EED → EE □ (*v*) ED → □ (*v*) ING →	feed → feed, agreed → agree plastered → plaster, bled → bled motoring → motor, sing → sing
	C	□ (*v*) Y → I	happy → happi, sky → sky

2	□ (m>0) ATIONAL → ATE □ (m>0) TIONAL → TION □ (m>0) ENCI → ENCE □ (m>0) ANCI → ANCE □ …	relational → relate conditional → condition, rational → rational valenci → valence hesitansi → hesitance …
3	□ (m>0) ICATE → IC □ (m>0) ATIVE → □ (m>0) ALIZE → AL □ (m>0) ICITI → IC □ …	triplicate → triplic formative → form formalize → formal electriciti → electric …
4	□ (m>1) AL → □ (m>1) ANCE → □ (m>1) ENCE → □ (m>1) ER → □ …	revival → reviv allowance → allow inference → infer airliner → airlin …
5	□ (m>1) E → □ (m=1 and not *o) E → □ (m>1 and *d and *L) → lettre non doublée	probate → probat, rate → rate cease → ceas controll → control, roll → roll

Tableau 3-2: Les étapes de l'algorithme de porter . **[28]**

4.2.5. Calcul des poids :

Dans cette étape nous avons utilisé la mesure TF-IDF (term frequency, inverse document frequency) pour calculer le poids des termes qui constitue notre vocabulaire dans chaque tweets.

Dans un ensemble de document donné, les termes ont des différentes importances dans un certain document.

TF-IDF calcule le poids de chaque terme dans un document, en prenant tout document exposé en compte. Plus un mot apparaît dans un document, et moins il apparait dans les autres documents de l'ensemble, plus son poids sera élevé. Pour attribuer un poids à un terme dans un document, la formule suivante est utilisée :

$$\mathrm{tfidf}_{i,j} = \mathrm{tf}_{i,j} \cdot \mathrm{idf}_i \quad \textbf{(3.1)}$$

$$\mathrm{Tf} = \frac{\text{Nbr de réptition}}{\text{nbr de terme de document}} \quad \textbf{(3.2)}$$

$$\mathrm{idf}_i = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|} \quad \textbf{(3.3)}$$

Où

$|D|$: Nombre total de documents dans le corpus

$|\{d_j : t_i \in d_j\}|$: Nombre de documents où le terme t_i apparaît c'est-à-dire $n_{i,j} \neq 0$

4.2.6. Construction de la matrice de poids :

Cette étape consiste à construire une matrice de poids pour tous les tweets, tel que les colonnes représentent les différentes thématiques qui constituent le vocabulaire, les lignes représentent l'ensemble de tweets collectés et chaque case détermine une valeur calculée par la mesure TF_IDF. Le tableau [3-3] montre un exemple de la matrice de poids pour deux tweets.

tweet_text	depression	hurt	insult	fear	loneliness	anorexia	cyberbulling
want die	3.407203	0	0	0	0	0	3.69891
life total mess pointless	4.458748	7.826044	0	0	4.992831	0	0

Tableau 3-3 : matrice de poid des tweets

Par exemple, le tweet « want die » contient deux termes, le premier terme 'want' appartient au thématique Depression avec une valeur TF_IDF égal a 3.407203.le deuxième terme 'die' appartient au thématique cyberbulling avec une valeur TF_IDF égal a 3.69891 et la valeur 0 signifie qu'il n'existe aucun terme dans le tweet qui appartient a ces thématiques.

4.2.7. Classification des tweets :

Cette étape consiste à effectuer une classification via l'outil Weka, après le chargement d'un fichier sous format ARFF(voir la **Figure3-5**), qui se compose d'une liste d'instance qui représente les différentes valeurs d'attributs (depression,hurt,insult,fear,loneliness,anorexia,cyberbulling,classe).

```
% 1. Title: twittersuicide Database
%
% 2. Sources:

%      (a) Date: Jun, 2016
%
@RELATION classification

@ATTRIBUTE  depression        NUMERIC
@ATTRIBUTE  hurt              NUMERIC
@ATTRIBUTE  insult            NUMERIC
@ATTRIBUTE  fear              NUMERIC
@ATTRIBUTE  loneliness        NUMERIC
@ATTRIBUTE  anorexia          NUMERIC
@ATTRIBUTE  cyberbulling      NUMERIC
@ATTRIBUTE classe  {R,N}
```

Figure 3-4: Le début d'un fichier ARFF

Dans un premier temps, nous avons effectué ce que nous appelons une "classification manuelle". En effet, nous avons sélectionné des tweets pour lesquels il y a une forte probabilité que leurs auteurs passent à l'acte (tweets risque de suicide). Nous avons fait de même pour les tweets sans risque. Après, nous avons effectué une classification de 5 algorithmes de l'API Weka (Naïve Bayes, SMO, IBK, JRIP, J48) dans outil de classification weka.

A partir des valeurs de mesures d'exactitude (TP Rate, FP Rate, Precision, Recall, F-Measure) nous avons choisis le classifieur qui a une valeur de précision la plus élevée (voir le **Tableau 3-4**).

Algorithme	Naïve Bayes	SMO	IBK	JRIP	J48
Précision de classification	67%	70%	91%	79%	80%

Tableau 3-4 : comparaison entre les différents algorithmes de classification

Le tableau ci-dessus montre que l'algorithme IBK possède le pourcentage de précision le plus élevé.

Alors, nous avons choisi l'algorithme IBK qui nous permet d'avoir un meilleur résultat lors de l'application de ce dernier du modèle d'apprentissage sur notre jeu de données.

- **Présentation de l'algorithme IBK :**

IBKreprésente l'implantation du classificateur les plus proches k voisins qui utilise la métrique distance. Par défaut, il utilise juste le voisin le plus proche (k=1); le nombre k peut être spécifié manuellement ou déterminé automatiquement en utilisant la validation croisée «leave-one-out ». Il normalise les attributs par défaut et peut aussi pondérer les distances. **[29]**

- **K plus proche voisin (KNN où k nearest neighbor en anglais) :** c'est une approche statistique de classification très connue. Il a été prouvé que c'est une des méthodes les plus performantes après des tests réalisés sur des corpus de données .Le principe de l'algorithme kNN est le suivant : étant donné un texte à classer, l'algorithme cherche les k voisins les plus proches parmi les documents utilisés au cours de la phase d'apprentissage, les catégories de ces k voisins les plus proches serviront à donner des poids aux catégories candidates de classification. C'est le degré de similarité entre le document test et le document voisin qui est utilisé comme poids de la catégorie de ce dernier, si plusieurs voisins partagent la même catégorie alors le poids attribué à cette catégorie est égal à la somme des degrés de similarité entre le document test et chacun des voisins appartenant à cette catégorie. Par cette méthode on peut obtenir une liste des poids attribués à chaque catégorie, le document test est classé dans une catégorie si le poids attribué à celle-ci est supérieur à un seuil fixé à l'avance. **[29]**

5. Conclusion :

Dans ce chapitre nous avons présenté une nouvelle approche de l'analyse des tweets sur le crime de suicide.

Cette approche est basée sur les étapes suivantes:

Collecter les tweets qui portent un mot clé suicide et une géo localisation précise (Guyana), et stocker ces derniers dans une base de données. Ensuite l'étape de prétraitement qui permet de rendre le tweet clair et lisible après l'élimination des mots vides, les émoticônes, ponctuations, tweets répétés, la stemmatisation et lemmatisation. Enfin classifier les tweets prétraités.

Chapitre 04 : Test et Implémentation

1. Introduction :

Dans ce chapitre, nous allons présenter la partie pratique qui constitue la mise en œuvre d'une plateforme pour notre approche d'analyse des sentiments **à** partir de Twitter concernant un crime donné. Nous commençons par les outils utilisés, puis nous donnons une présentation de l'application et enfin nous présentons l'évaluation de notre nouvelle approche.

2. Technologie (Outils de développement) :

2.1 Java :

Pour la réalisation de notre application, nous avons utilisé le langage de programmation JAVA.

Java est un langage de programmation récent développé par Sun Microsystems en 1995. Java fait partie de la « grande famille » des langages orientés objet. Il répond donc aux trois principes fondamentaux de l'approche orientée objet (POO) : l'encapsulation, le polymorphisme et l'héritage. **[31]**

Nous avons développé en JAVA pour les raisons suivantes :

- **Distribué :**

Java possède une importante bibliothèque de routines permettant de gérer les protocoles TCP/IP tels que HTTP et FTP. Les applications Java peuvent charger et accéder à des données sur Internet via des URL avec la même facilité qu'elles accèdent à un fichier local sur le système ce qui nous permettent dans notre recherche d'accéder facilement aux API de Twitter à travers le protocole HTTP.

- **Fiabilité :**

Java a été conçue pour que les programmes qui l'utilisent soient fiables sous différents aspects. Sa conception encourage le programmeur à traquer préventivement les éventuels problèmes.

- **Sécurité :**

Java a été conçue pour être exploité dans des environnements serveurs et distribués. Dans ce but, la sécurité n'a pas été négligée. Java permet la construction de systèmes inaltérables et sans virus.

- **Interprété :**

L'interpréteur Java peut exécuter les bytes code directement sur n'importe quelle machine sur laquelle il a été porté.

2.2. Netbeans :

Netbeans est un environnement de développement intégré (EDI), placé en Open Source par Sun. En plus de Java, Netbeans permet également de supporter différents autres langages, comme C, C++, JavaScript, PHP, HTML … Il comprend toutes les caractéristiques d'un IDE moderne (éditeur en couleur, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages Web). Conçu en Java, Netbeans est disponible sous Windows, Linux, Solaris, Mac OS X ou sous une version indépendante des systèmes d'exploitation (requérant une machine virtuelle Java). Un environnement Java Développement Kit (JDK) est requis pour les développements en Java. **[32]**

L'IDE Netbeans repose sur un noyau robuste, la plateforme Netbeans, que vous pouvez également utiliser pour développer vos propres applications Java, et un système de plugins performant, qui permet d'avoir un IDE modulable.

Enfin, cet IDE possède un débogueur de grande qualité ainsi qu'une interface graphique améliorée.

2.3. WampServer :

WampServer est une plate-forme de développement Web sous Windows. Il permet de développer des applications Web dynamiques à l'aide du serveur Apache2, du langage de scripts PHP et d'une base de données MySQL. Il possède également PHPMyAdmin pour gérer plus facilement les bases de données. **[33]**

2.3.1. PHPMyAdmin :

C'est une interface d'administration pour le SGBD MySQL. Il est écrit en langage PHP et s'appuie sur le serveur HTTP. **[34]**

Il permet d'administrer les éléments suivants :

- les bases de données
- les tables et leurs champs (ajout, suppression, définition du type)
- les index, les clés primaires et étrangères
- les utilisateurs de la base et leurs permissions
- exporter les données dans divers formats (CSV, XML, PDF, Open Document, Word, Excel et LaTeX)

2.4. MYSQL :

MySQL est un système de gestion de bases de données relationnelles. Le SQL dans "MySQL" signifie "Structured Query Language" : le langage standard pour les traitements de bases de données. MySQL est Open Source. Open Source (Standard Ouvert) signifie qu'il est possible à chacun d'utiliser et de modifier le logiciel. Tout le monde peut le télécharger sur Internet et l'utiliser sans payer aucun droit. Toute personne en ayant la volonté peut étudier et modifier le code source pour l'adapter à ses besoins propres. Toutefois, si vous devez intégrer MySQL dans une application commerciale, vous devez vous procurer une licence auprès de MySQL AB. **[34]**

2.5.Weka :

Weka (Waikato Environment for Knowledge Analysis) est un environnement de fouille de données développé par le groupe de recherche "machine Learning" du département d'informatique de l'université de Waikato en Nouvelle-Zélande. Il est utilisé dans le domaine de la recherche, de l'éducation et de l'industrie. Il est écrit dans le langage Java et testé sur plusieurs plateformes telles que Linux et Windows. Cet environnement est un logiciel "open source" et disponible sur le site du groupe de recherche "machine Learning" du département d'informatique de l'université de Waikato. **[35]**

Weka est une collection d'algorithmes d'apprentissage dont le but est de réaliser des tâches de fouille de données. Les algorithmes peuvent être appliqués directement à un ensemble de données ou appelés via un programme Java. Weka contient les outils pour le prétraitement de données, la classification, la régression, le groupement (clustering), les règles d'association et la visualisation. En effet, Weka permet d'effectuer un prétraitement sur un ensemble de données, d'appliquer un algorithme d'apprentissage, et d'analyser les résultats et les performances d'un classificateur. Il est aussi bien adapté pour intégrer de nouveaux algorithmes d'apprentissage. **[35]**

2.6.Bibliothèques trières :

2.6.1. Twitter 4j :

Twitter4j est une librairie Java permettant d'intégrer facilement l'API Twitter dans toute application Java. La librairie propose différentes classes et méthodes permettant de manipuler les méthodes qu'offre l'API Twitter **[36]**. Pour utiliser cette librairie, il suffit de télécharger un fichier au format "jar" et de l'ajouter au classpath de l'application JAVA. La JavaDoc de la librairie permet une prise en main rapide et facile de cette librairie.

2.6.2. Stanford CoreNLP :

La bibliothèque Stanford CoreNLP est un outil qui permet de parser, tokonizer, lemmatiser et d'extraire les entités nommés d'un texte. Il a été créé à l'université de Stanford **[37].** Il fournit un ensemble d'outils d'analyse du langage naturel. Elle peut donner les formes de base de mots, leurs parties du discours (POS Tags), et même la reconnaissance des entités nommées telles que les organisations, le lieu, etc. Cette bibliothèque peut même faire la lemmatisation (que pour la langue anglaise), et elle peut marquer la structure des phrases et les dépendances de mots, etc.

Nous avons utilisé cette bibliothèque pour faire la lemmatisation.

3. Présentation de l'application:

Dans notre travail nous avons collecté dans une période de un moins 500 tweets qui portent le mot clé suicide et des phrases entières qui sont souvent utilisées par des gens qui ont l'intention de se suicider comme par exemple, "I want to die", "My family would be better off without me", "I hate my life"et "kill my self" de la zone géographique Guyana.
Après le stockage des données collecté dans une base de données, nous avons construit un vocabulaire de sept thématiques liées aux suicides où chaque thématique contient un ensemble de termes. La **(Figure 4-1)** représente la thématique (Depression) avec la liste de ces termes.

```
Depression
Ability
Abnormal
Abuse
Adolescents
Affect
Agency
Aid
Alarm
Alienation
All ages
Alone
Anger
Anguish
Antidepressant
Anxiety
Anxious
Attempt
Attention
```

Figure 4-1 : Liste des termes de la thématique Depression

Par la suite, nous avons nettoyé les tweets stockés en utilisant des API et plusieurs algorithmes qui permettent d'éliminer des ponctuations, nom d'utilisateur, les émoticônes, et les mots vides, et puis nous avons calculé pour chaque tweet le poids TF_IDF des termes appartenant au vocabulaire pour construire la matrice de poids. Ensuite nous avons classifié les tweets en deux classes (risque et non risque de suicide) en utilisant l'outil de classification Weka.
Toutes ces étapes sont présentées dans une interface simple comportant la description de l'application (voir Figure **4-2**).
Le menu se compose de deux boutons. Le premier « home » affiche le sous-menu qui permet de récupérer et stocker les tweets dans la base de données« Streaming_Tweets », le

Sous-menu qui permet de prétraiter les tweets « Parsing_Tweets », le sous-menu qui permet de construire le modèle d'apprentissage « Preparing Data Set », et le sous-menu qui permet d'afficher les différentes statistiques de classification « Statistics ».

Le deuxième bouton « Help» pour aider les utilisateurs à comprendre les étapes du système.

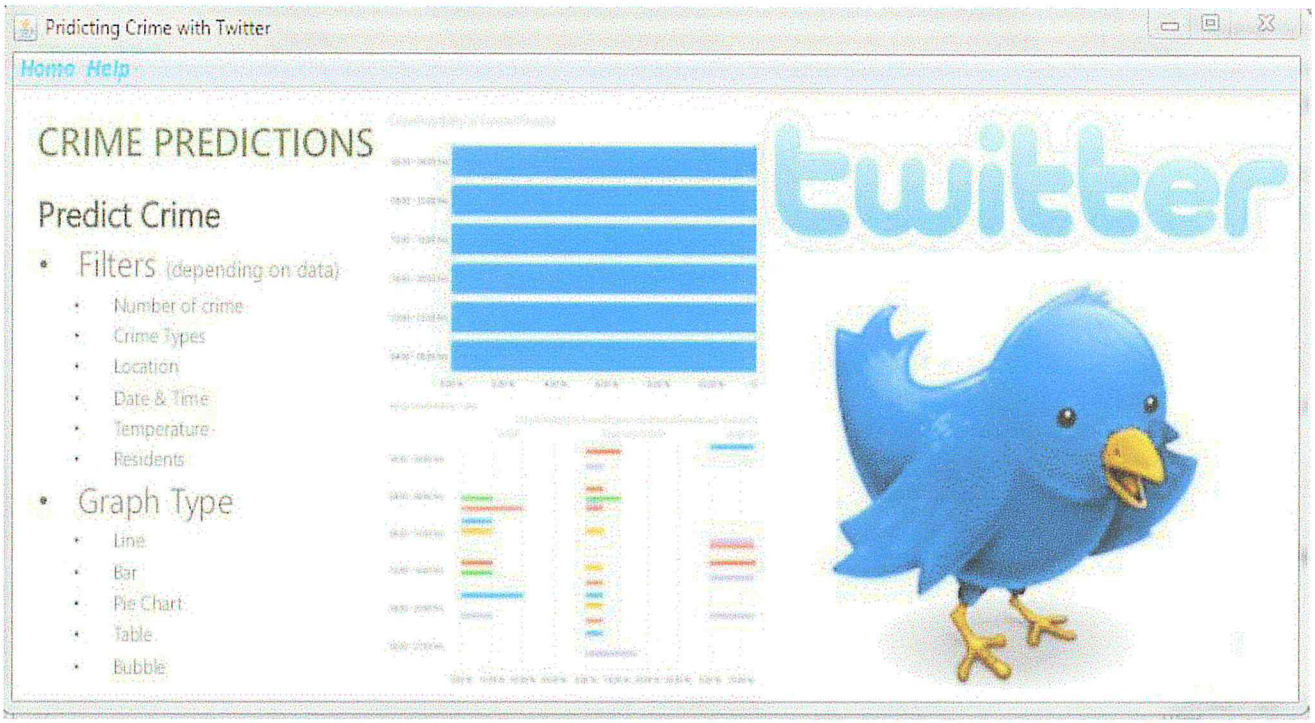

Figure 4-2 : Interface d'accueil.

3.1. Collection des tweets :

L'onglet « Streaming_Tweets » permet de collectionner un ensemble de tweets en temps réel, il contient un boutons : « load » pour récupérer et charger les tweets collectés dans une base de données (voir Figure **4-3**).

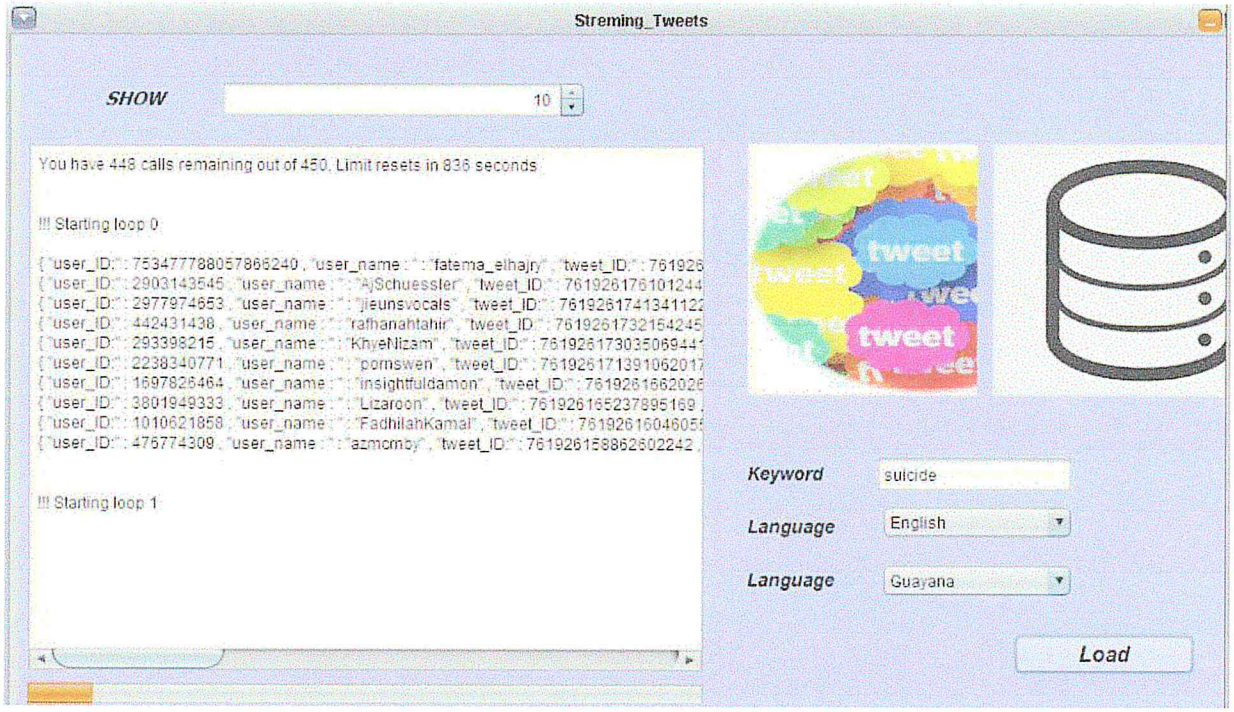

Figure 4-3 :L'onglet Streaming_Tweets.

3.2. Prétraitement :

L'onglet « Parsing_Tweets » permet de nettoyer un ensemble de tweets, il contient deux boutons : « tweet before clearing» pour récupérer que le contenu des tweets à partir de la base de données et « tweet after clearing » pour afficher les tweets après l'étape de prétraitement (voir Figure 4-4).

Figure 4-4 :L'onglet Parsing_Tweets.

3.3. Classification :

L'onglet « Preparing Data Set » permet de préparer une base d'apprentissage pour notre jeu de donnée (voir Figure **4-5**), il contient trois boutons : « calculating TF-IDF » permet de calculer le poids de chaque mot dans un tweet (voir Figure **4-5-.1**), «Generating ARFF file »permet de générer le fichier arff qui représente notre modèle d'apprentissage (voir Figure **4-5-2**), « WEKA classification » permet de visualiser le résultat de classification sous forme d'un arbre (voir Figure **4.5.3**).

La classification avec l'algorithme (**IBK**) sous Weka (voir la figure **4.5.4**).

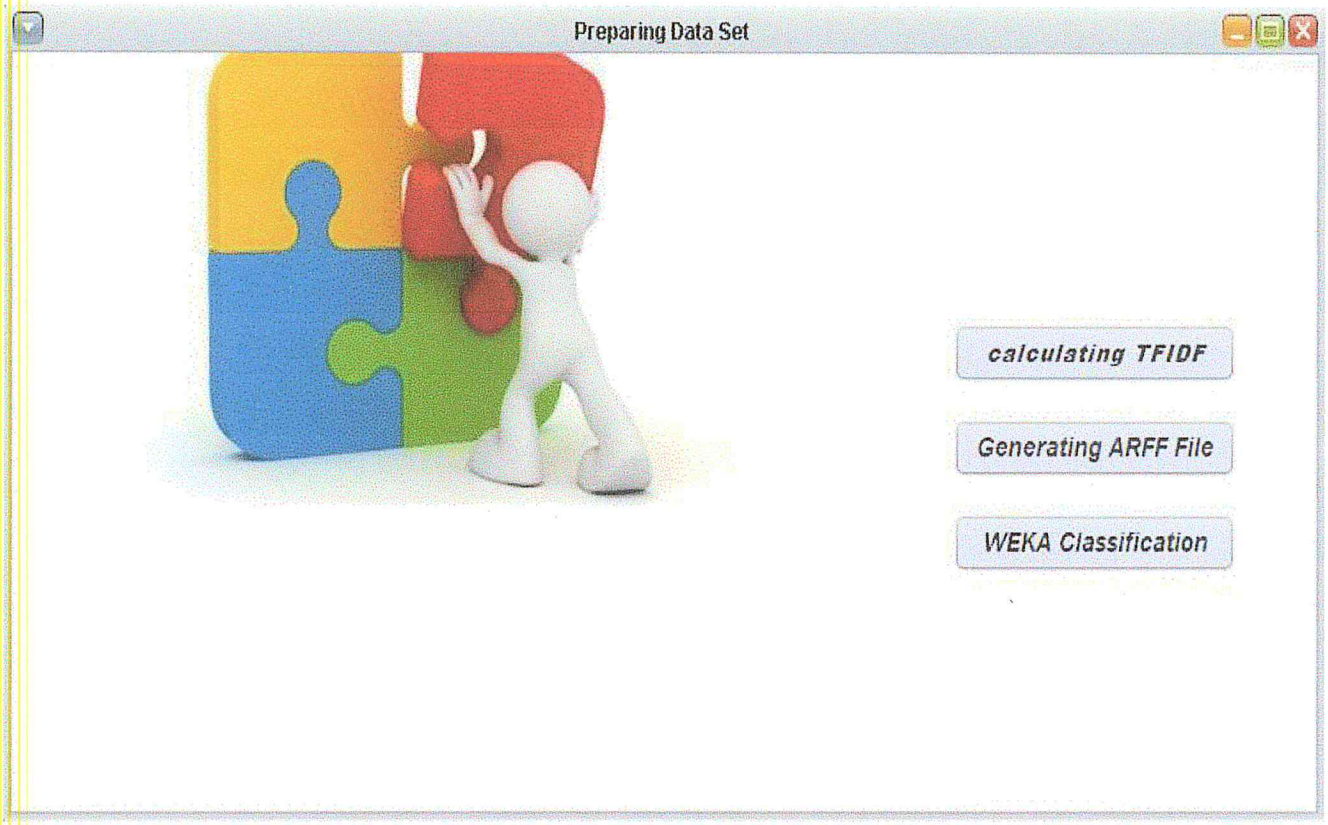

Figure 4-5:L'onglet Preparing Data Set.

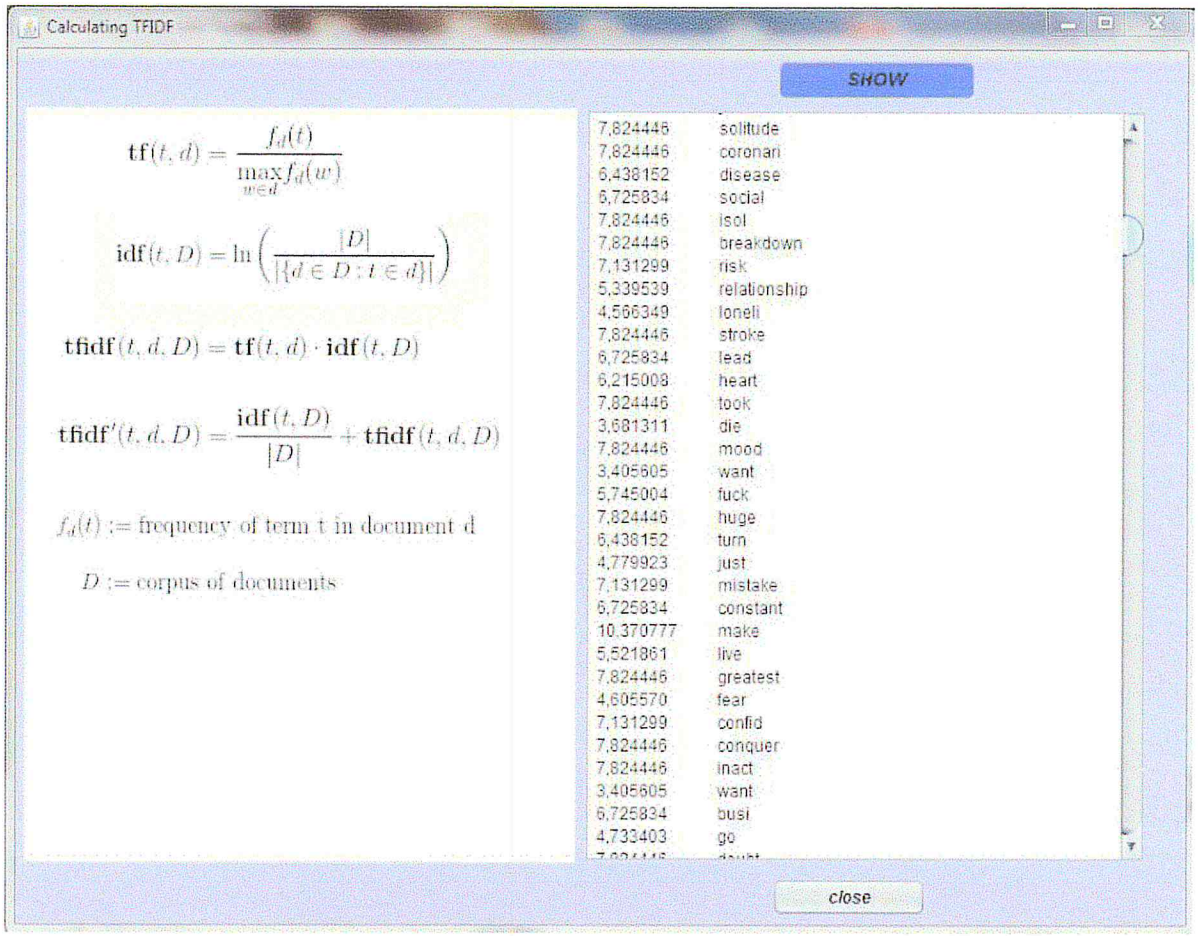

Figure 4-5-1: Calculer le poids des mots

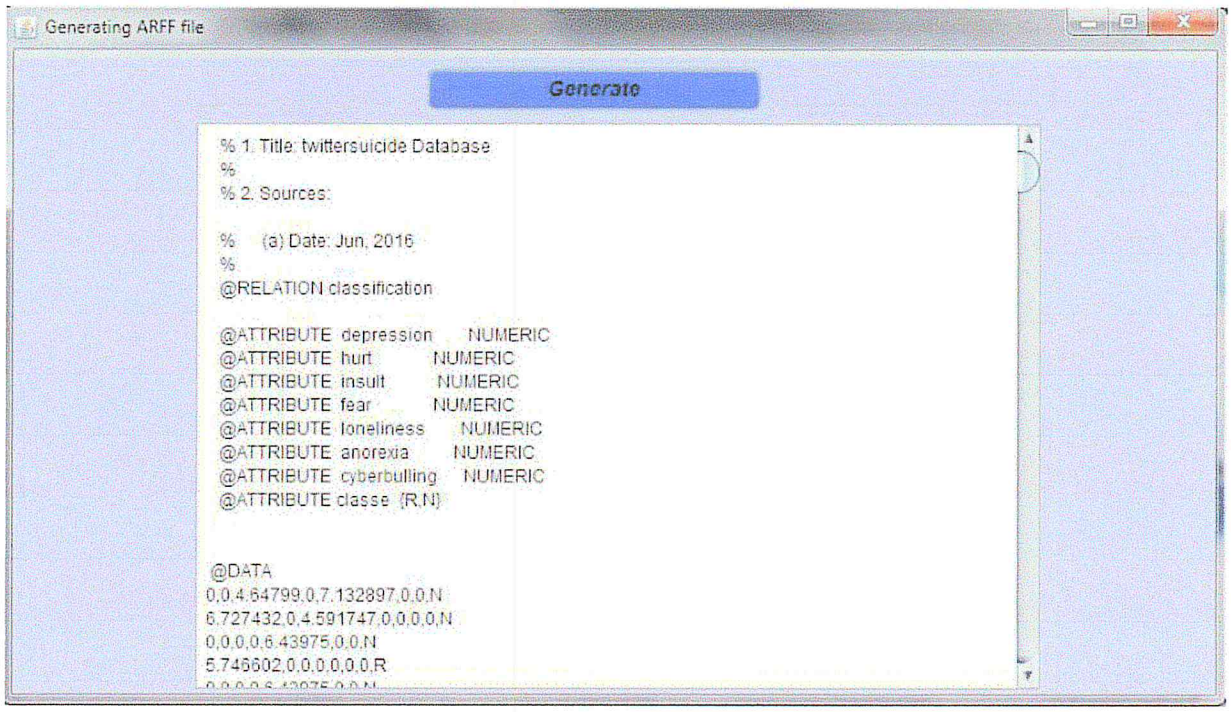

Figure 4-5-2 : Générer le fichier ARFF

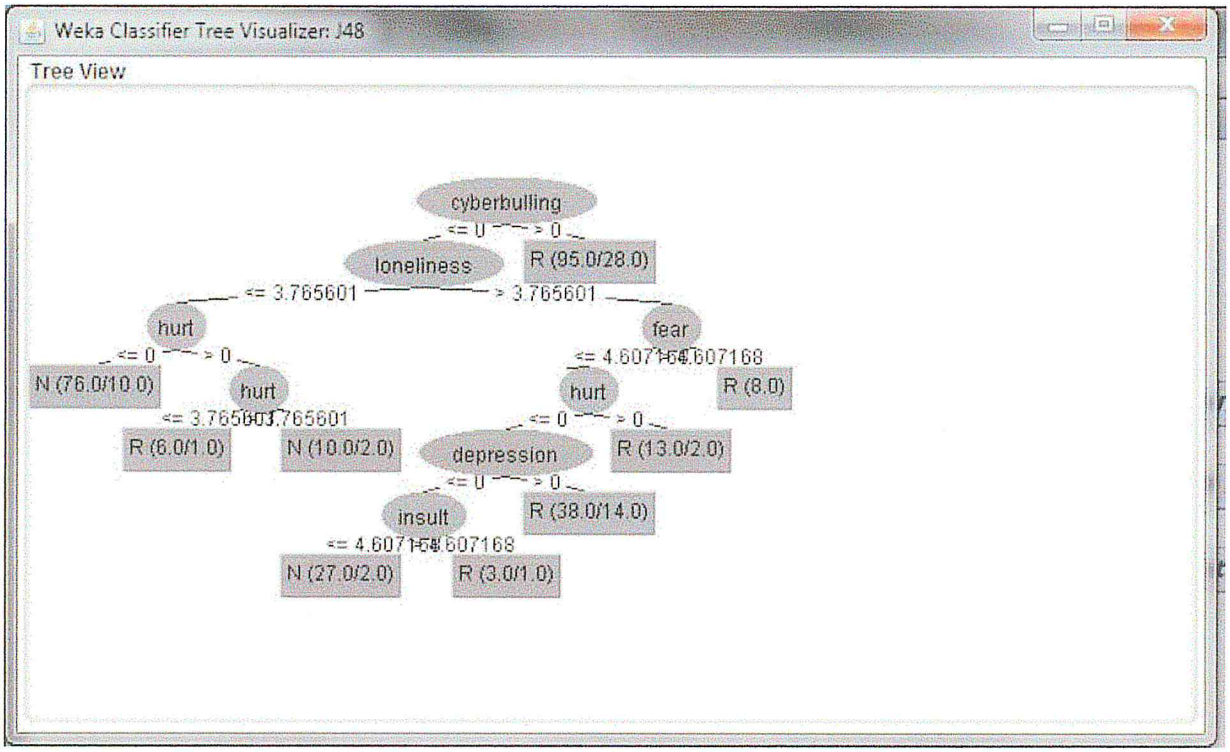

Figure 4-5-3 : l'arbre de décision après la classification avec J48

3.3.1. Phase de test :

Dans la phase de test, la classification a été appliquée sur le fichier ARFF généré dans outils de classification weka.

3.3.2. Qu'est ce qu'un bon classifieur :

On à 4 cas :

Vrai positif (true positive): ex: positif classé positif.

Vrai négatif (true negatif): ex: négatif classé positif.

Faux négatif (false negatif): ex: positif classé négatif.

Faux positif (false positive): ex: négatif classé positif.

Evaluation de résultats :

Après le chargement de fichier ARFF dans outil de classification Weka, avec le classifieur IBK nous avons trouvé les résultats illustré dans la (figure 4-6).

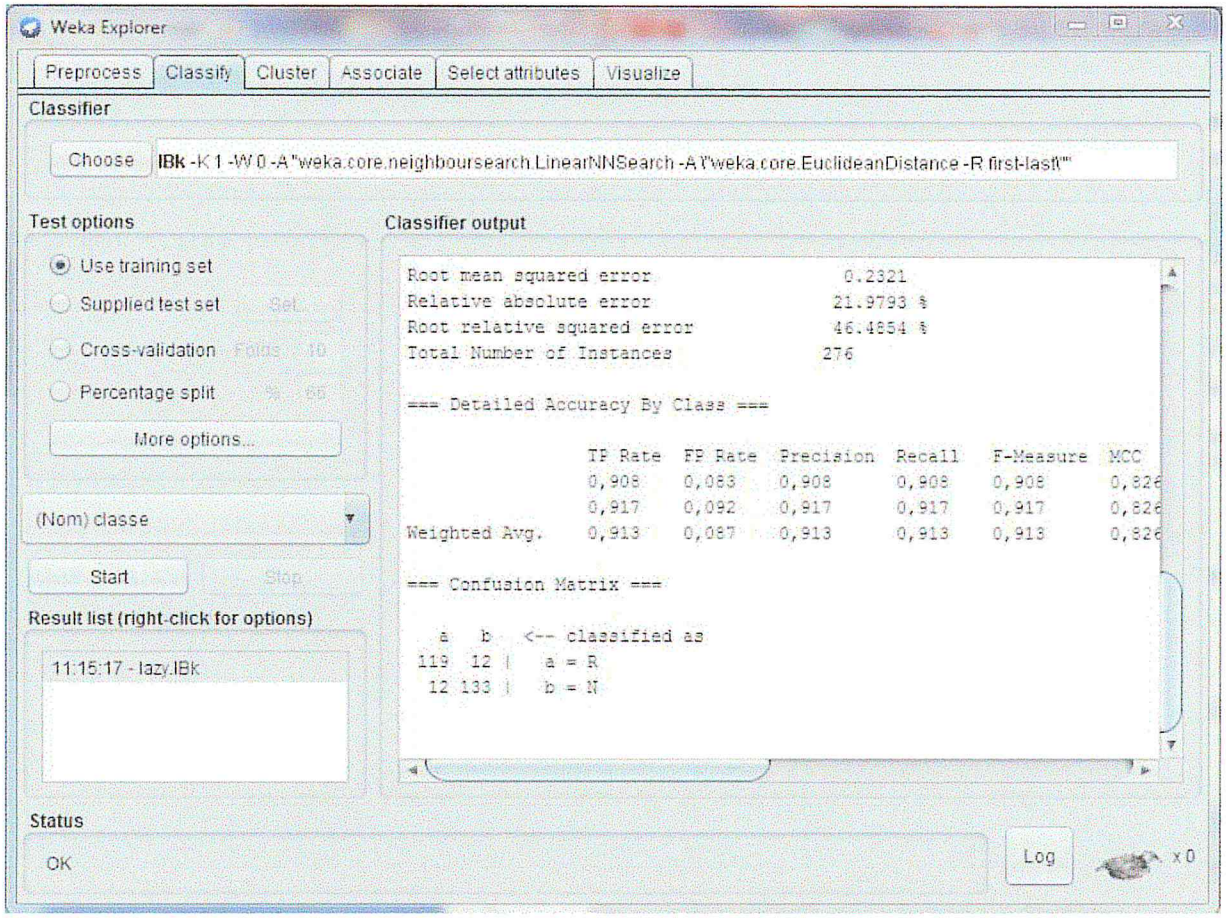

Figure 4-6 : classification avec l'algorithme IBK

- **Correctly Classified Instances**

Le nombre d'exemples bien classés, en valeur absolue, puis en pourcentage du nombre total d'exemples.

- **Incorrectly Classified Instances**

Sous le même format, le nombre d'exemples mal classés.

- **Kappa statistic :**

Le coefficient Kappa est censé mesuré le degré de concordance de deux ou de plusieurs juges. Dans Weka, on est toujours dans le cas de deux juges. On mesure la différence entre l'accord constaté entre les deux juges, et l'accord qui existerait si les juges classaient les exemples au hasard.

Dans Weka, le jugement, c'est la classe d'un exemple, et les deux juges sont le classifieur et la classe réelle de l'exemple.

L'accord/ désaccord entre les deux juges se lit directement dans la matrice de confusion : c'est une mesure dont la valeur est d'autant plus grande que la matrice est diagonale.

Le coefficient Kappa se calcule de la façon suivante :

$$\kappa = \frac{P_o - P_e}{1 - P_e} \qquad (4.1)$$

Avec Po : La proportion de l''echantillon sur laquelle les deux juges sont d'accord (i.e. la diagonale principale de la matrice de confusion).

Et :

$$P_e = \frac{\sum_i p_i p_{.i}}{n^2}$$

Ou

– pi : somme des éléments de la ligne i

– $p_{.i}$: somme des éléments de la colonne i

– n : taille de l''echantillon

Sur l'exemple de la (**Figure 4-6**), dont la matrice de confusion était :

	a	**b**	**Total :**
	119	12	131
	12	133	145
Total :	131	125	276

Tableau 4-1 : matrice de confusion

On a :

$$P0 = \frac{119 + 133}{276} = \frac{252}{276} = 0.91$$

$$P_e = \frac{(131*131)+(145*145)}{(276*276)} = 0.5$$

Et donc

$$K = \frac{091-05}{1-0.5} = 0.8256$$

Le coefficient Kappa prend ses valeurs entre -1 et 1.

– Il est maximal quand les deux jugements sont les mêmes : tous les exemples sont sur la diagonale, et

P0 = 1

– Il vaut 0 lorsque les deux jugements sont indépendants (P0 = Pe)

– Il vaut -1 lorsque les juges sont en total désaccord

Certains auteurs ont proposé une échelle de degré d'accord selon la valeur du coefficient voir la **(Figure 4-7)** :

Accord	Kappa
Excellent	>0.81
Bon	0.80-0.61
Modéré	0.4-0.41
Médiocre	0.4-0.21
Mauvais	0.20-0.0
Très mauvais	<0

Figure 4-7 : coefficient de Kappa

- **Mean absolute error :**

Erreur absolue en moyenne : pour chaque exemple, on calcule la différence entre la probabilité (calculée par le classifieur) pour un exemple d'appartenir `à sa véritable classe, et sa probabilité initiale d'appartenir à la classe qui lui a été fixée dans l'ensemble d'exemples (en général, cette probabilité vaut 1). On divise ensuite la somme de ces erreurs par le nombre d'instances dans l'ensemble d'exemples.

Plus formellement :

– Soient p1, p2, …, pn les probabilités calculées par le classifieur pour chaque exemple d'appartenir à sa vraie classe.

– Soient a1, a2, …, an les probabilités à priori pour chaque exemple d'appartenir `a la classe qui leur a été fixée par définition (en g´général, les a_i valent toujours 1, mais on peut imaginer qu'on soit un peu moins catégorique, et que la classe attribuée ne le soit qu'avec une certaine confiance).

– Alors on calcule :

$$\text{Mean Absolute Error} = \frac{|p_1 - a_1| + |p_2 - a_2| + \cdots + |p_n - a_n|}{n} \qquad \textbf{(4.2)}$$

Dans le cas ou le classifieur est un prédicateur, c'est-à-dire qu'il retourne une valeur réelle au lieu d'une classe discrète, c'est la différence entre la valeur calculée et la valeur attendue qui sont utilisées pour p_i et a_i ; ca peut par exemple être le cas pour les réseaux de neurones.

- **Root mean-squared error :**

Cette mesure d'erreur concerne principalement les prédicateurs Racine carrée de l'erreur quadratique moyenne : avec les mêmes notations que ci-dessus, elle correspond à :

$$\sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}} \qquad (4.3)$$

L'erreur quadratique avantage les solutions ou il y a beaucoup de petits écarts, par rapport `a celles qui sont exactes presque partout, mais qui font de grosses erreurs en un petit nombre de points. Le fait de prendre la racine carrée permet de manipuler des quantités qui ont la même dimension que les valeurs à prévoir.

- **Relative absolute error :**

Cette mesure d'erreur concerne principalement les prédicateurs Erreur absolue relative .
On compare l'erreur absolue avec l'erreur absolue d'un prédicateur très simple, qui retournerait toujours la valeur moyenne des a_i, soit

$$\bar{a} = \frac{1}{n}\sum_i a_i \ :$$

$$\frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{|a_1 - \bar{a}| + \cdots |a_n - \bar{a}|} \qquad (4.4)$$

- **Root relative squared error :**

Cette mesure d'erreur concerne principalement les prédicateurs Racine carrée de l'erreur quadratique relative : rapport entre l'erreur quadratique et ce que serait l'erreur quadratique d'un prédicateur qui retournerait toujours la valeur moyenne des a_i :

$$\sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \cdots (a_n - \bar{a})^2}} \qquad (4.5)$$

- **Les mesures d'exactitude:**

Ces valeurs se trouvent dans la partie "Detailed Accuracy By Class". Pour chaque classe, Weka fournit cinq mesures.

TP Rate : Rapport des vrais positifs. Il correspond à :

$$\frac{\text{Nombre de vrais positifs}}{\text{Nombre de vrais positifs + Nombre de faux négatif}} = \frac{\text{Nombre de vrais positifs}}{\text{Nombre d'exemples de cette classe}}$$

C'est donc le rapport entre le nombre de bien classé et le nombre total d'éléments qui devraient être bien classés.

FP Rate : Rapport des faux positifs. Il correspond, symétriquement à la définition précédente, à :

$$\frac{\text{Nombre de faux positifs}}{\text{Nombre de faux positifs + Nombre de vrais négatifs}} = \frac{\text{Nombre de faux positifs}}{\text{Nombre d'exemples n'étant pas de cette classe}}$$

La donnée des taux TP Rate et FP Rate permet de reconstruire la matrice de confusion pour une classe donnée.

Symétriquement, la matrice de confusion permet de calculer TP Rate et FP Rate. Prenons l'exemple de la (**Figure 4-6**):

– 119 exemples de classe R (risque de suicide) sont bien classés, mais les 12 dernier exemples sont mal classés: donc TP Rate = $\frac{119}{131} = 0{,}908$

– 133exemples de classe N (non risque) sont bien class´es, avec 12 exemples sont mal classés : TP Rate = $\frac{133}{145} = 0{,}917$

– 12 exemples sont classés à tort parmi les exemples de R (risque de suicide), mais 145 exemples, qui ne sont pas des exemples R (risque de suicide), n'ont pas été reconnus comme tels : FP Rate = $\frac{12}{145} = 0.083$

– 12 exemples sont classés à tort parmi les exemples de N (non risque), mais 131 exemples, qui ne sont pas des exemples N (non risque), n'ont pas été reconnus comme tels :

FP Rate = $\frac{12}{131} = 0.092$

Precision : C'est le rapport entre le nombre de vrais positifs et la somme des vrais positifs et des faux positifs. Le résultat suivant représente la Precision de la classe R (risque de suicide) :

$$\textbf{Precision} = \frac{119}{119+12} = 0.908$$

Recall : C'est le rapport entre le nombre de vrais positifs et la somme des vrais positifs et des faux négatifs. Le résultat suivant représente le rappel de la classe R (risque de suicide) :

$$\textbf{Recall} = \frac{119}{119+12} = 0.908$$

F-Measure : Cette quantité permet de regrouper en un seul nombre les performances du classifieur (pour une classe donnée) pour ce qui concerne le Recall et la Precision.

$$\text{F-Measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \qquad \textbf{(4.6)}$$

Ou F-Measure de la classe R :

$$\text{F-Measure} = \frac{2*0{,}908*0{,}908}{(0{,}908+0{,}908)} = 0.908$$

D'après les résultats obtenus (Figure 4-6) l'algorithme de classification IBK est arrivé à une bonne précision de 91%, même cas pour le Recall et le F-mesure par rapport aux autres algorithmes de classification.

3.4 Statistique :

L'onglet « Statistics » permet de présenter des statistiques par rapport aux résultats de la classification sous Weka (voir Figure 4-8), il contient deux boutons avec deux types de statistiques sont affichées :

« Pie graduent» Pourcentage des tweets suspects à risque par rapport aux tweets suspects sans risque (voir Figure 4-8-1), tel que 131 tweets (47.5%) représente les tweets suspects à risque de suicide et 145 tweets (52.5%) les tweets suspects sans risque de suicide.

« ROC Weka » statistique de chaque instance en Weka par rapport aux d'autres instances (Voir Figure 4-8-2).

Statistique de chaque classe (risque et non risque) en colonne (voir la figure 4-8-3) sous Weka.

« personne suicide » représente les identifiant des personne qui ont la probabilité de suicidé, illustré dans la (Figure 4-8-4).

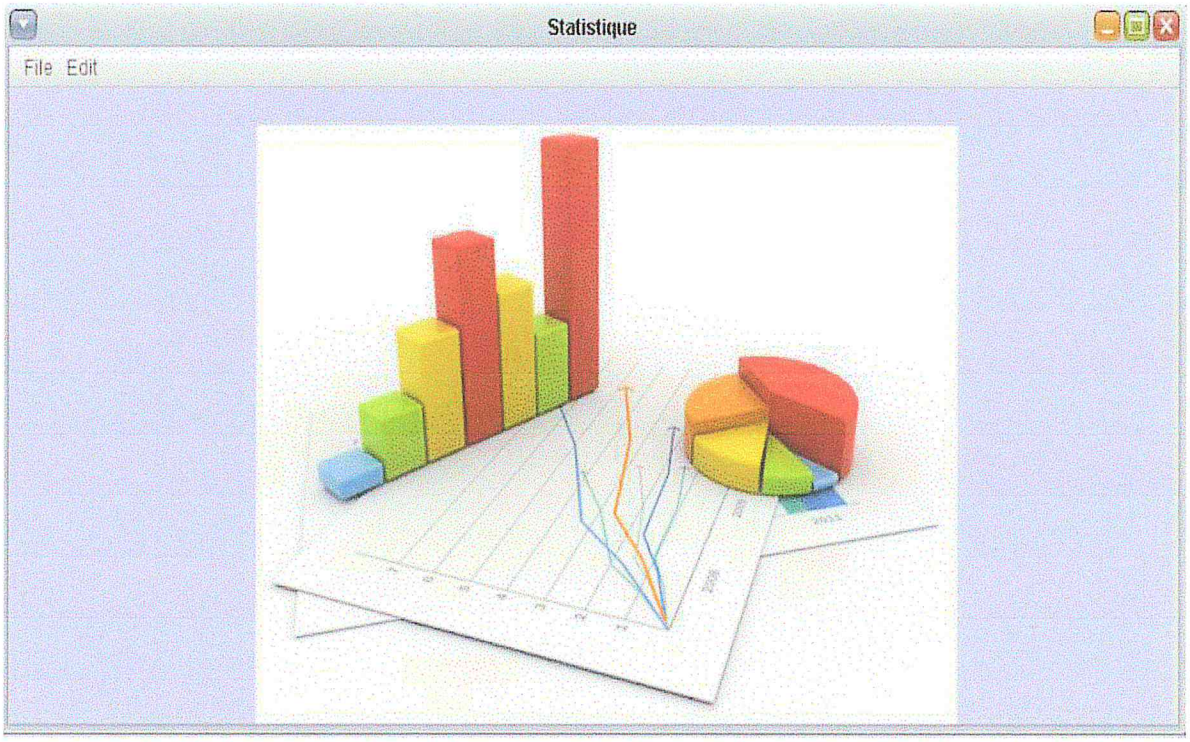

Figure 4-8 :L'onglet statistique

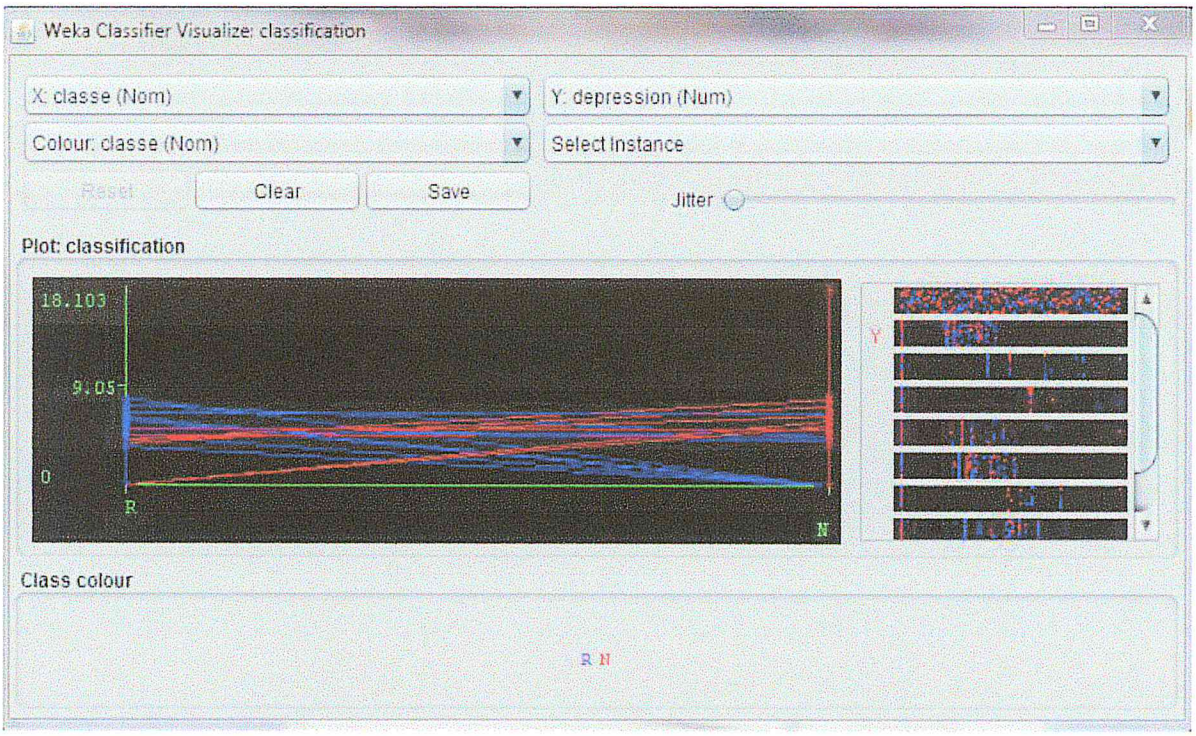

Figure 4-8-1 : statistique sous Weka

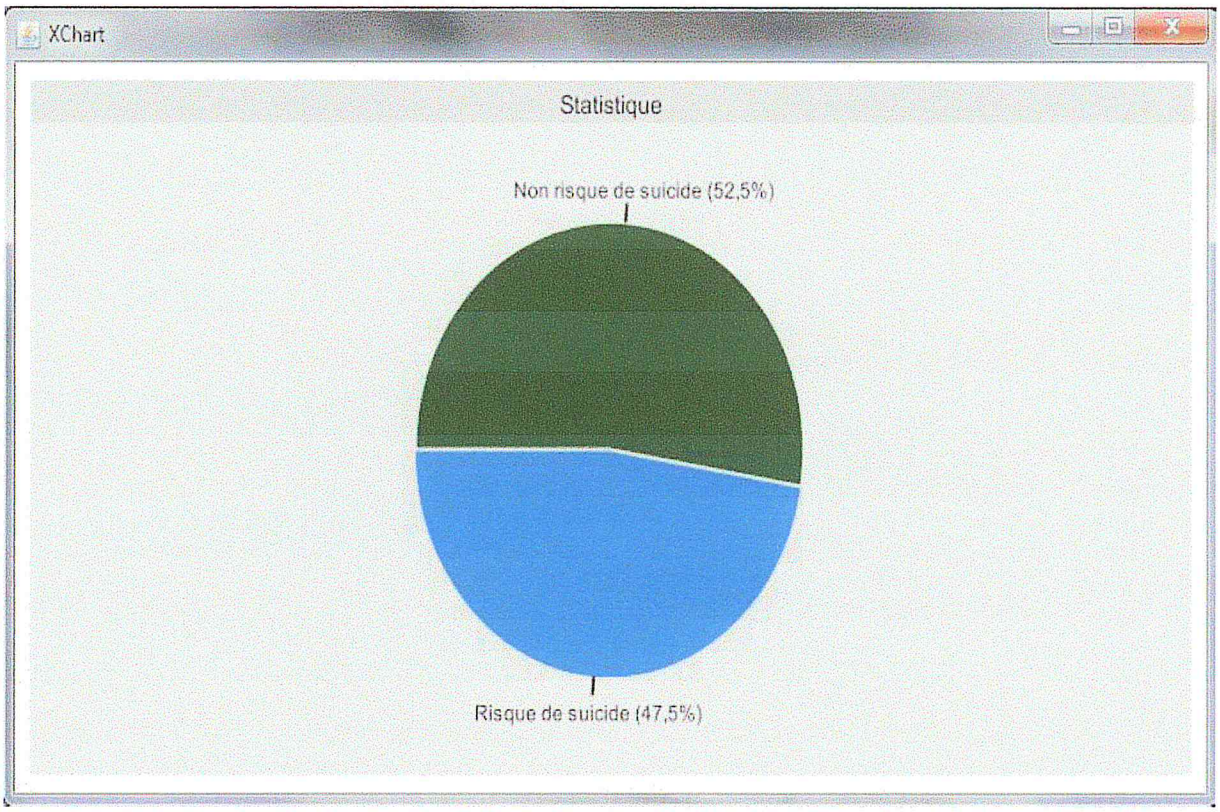

Figure 4-8-2 : statistique

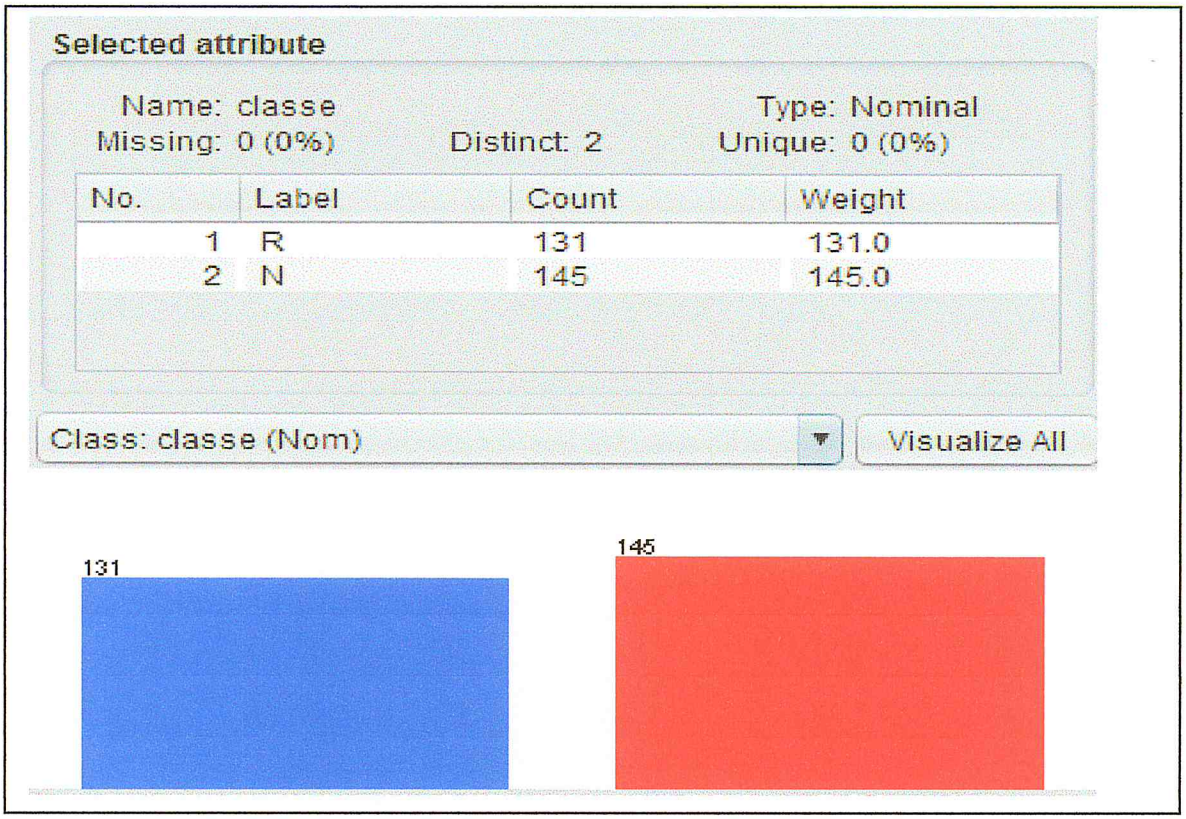

Figure 4-8-3 : statistique des classes en weka

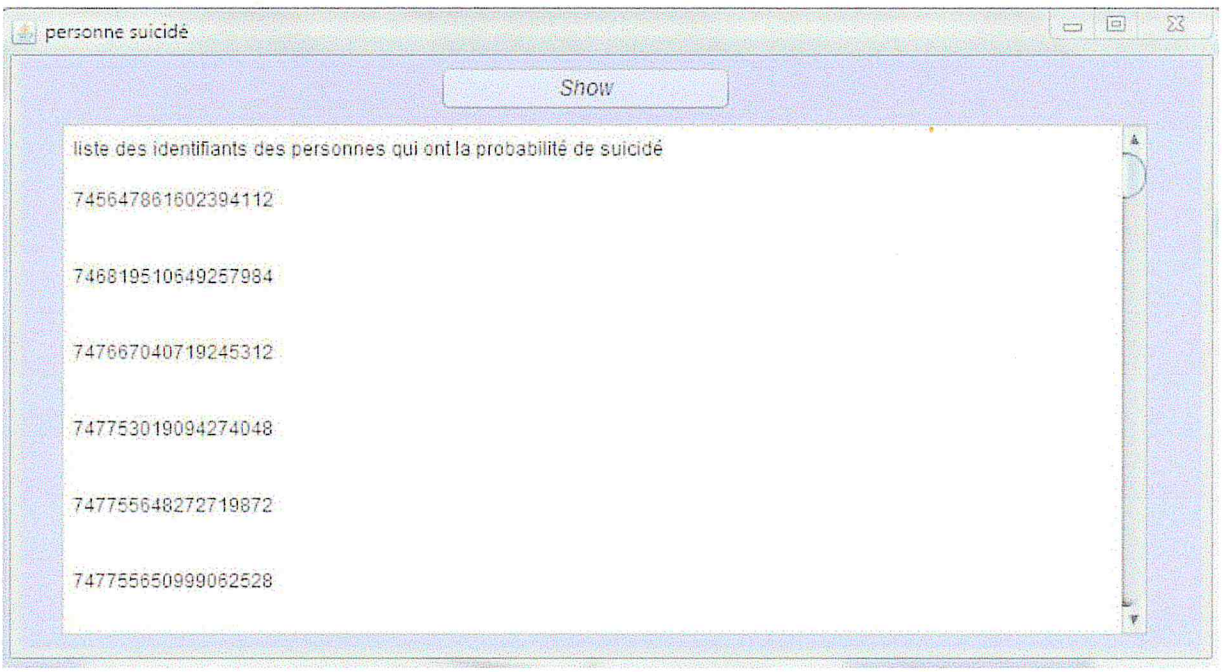

Figure 4-8-4 : liste de personnes suicidées.

4. Conclusion :

Dans ce chapitre, nous avons présenté en détail notre approche implémentée et l'interface de notre application. Nous avons commencé par la définition des outils qu'on a utilisé et après nous avons donné une présentation de notre application.

Conclusion générale:

Notre travail de mémoire ce déroule dans le domaine d'analyse et de classification des sentiments des tweets collecté.

Dans ce projet, nous avons développé un outil permettant potentiellement d'utiliser le réseau social Twitter comme force préventive dans la lutte contre le suicide et de proposer un système d'analyse et de classification des tweets concernant les thématiques de suicide (dépression, peur, harcèlement ,tuer ….etc), en se basant sur l'algorithme probabiliste IBK et la méthode de pondération TF-IDF pour différencier les termes.

Notre contribution dans les objectifs du travail est :

- D'extraire les tweets liés aux thématiques de suicide. Le problème major dans notre travail était de trouver les bonnes collections de tweets afin de répondre aux besoins de notre approche.
- De définir un vocabulaire associé à différentes thématiques critiques liées au suicide. Ce vocabulaire a été stocké dans une base de données relationnelle MySQL.
- De faire différents traitements nécessaires pour classifier les tweets suspects.
- De classer les tweets suspects en deux classes : classe de tweets suspects risque de suicide et classe de tweets suspects non risque de suicide .

Après avoir vu les résultats de test, nous pouvons conclure que la classification des sentiments avec les méthodes d'apprentissage supervisé implémenté dans les nouvelles technologies comme weka sont des méthodes essentielles et profondes dans le domaine d'analyse des sentiments.

Perspective :

Plusieurs perspectives pour ce travail telles que :

- Collecter plus d'informations (Tweets) d'apprentissage pour améliorer la classification automatique dans Weka ;
- Ajout de statistiques supplémentaires : Pourcentage de tweets suspects à risque et sans risque par thématique ;
- Ordonner les tweets suspects par probabilités décroissantes du risque dans l'interface afin de faciliter l'analyse des tweets par des experts ;
- Application en ligne.

Annexe A : Lexique de Twitter.

- **Twitto** : est un utilisateur de Twitter. [6]

Figure A-1 : Capture d'écran de la page profile de l'utilisateur de Twitter [6]

- **Tweets** « gazouillis » : sont les messages postés sur Twitter. Ils sont limités à 140 caractères.[6]

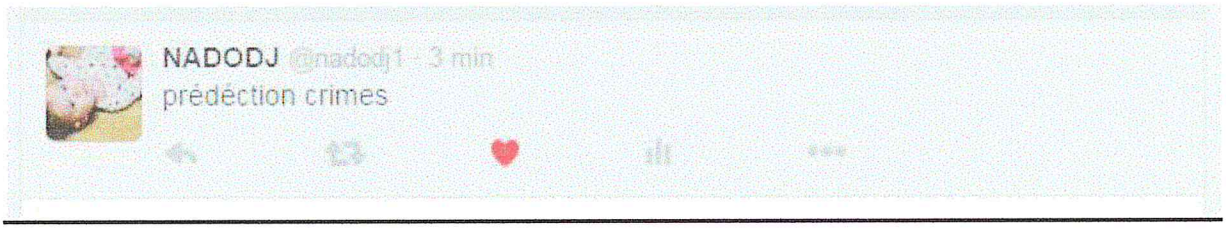

Figure A-2 : Capture d'écran d'un exemple de tweet[1]

- **J'aime** : Cliquer sur J'aime est un moyen simple de montrer que vous appréciez un Tweet. Vous pouvez par ailleurs utiliser cette fonctionnalité pour facilement retrouver ce Tweet plus tard. Cliquez sur l'icône en forme de cœur pour aimer un Tweet ; l'auteur verra ainsi que vous l'appréciez.[6]
- **Following / Abonnements** : correspondent aux nombre des comptes Twitter que vous suivez. Pour connaitre le nombre d'abonnements, allez sur votre page d'accueil Twitter le nombre se trouve dans la colonne de droite tout en haut. Et pour voir tous vos following (personnes que vous suivez) cliquez sur le nombre ou « Abonnements».[6]

[1] https://twitter.com/?lang=fr

Figure A-3 : Capture d'écran d'un exemple d'abonnement [6]

➢ **Followers / Abonnés** : c'est le nombre de comptes Twitter qui suit cette personne.Tout comme pour les abonnements, le nombre se situe sur la page d'accueil dans la colonne de droite et vous pouvez voir qui vous suit en cliquant sur le nombre ou «Abonnés ».[6]

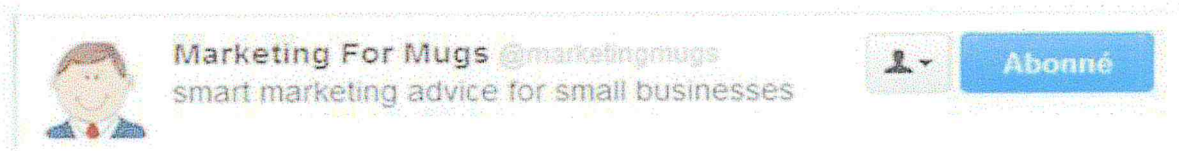

Figure A-4 : Capture d'écran d'un exemple d'un abonné. [6]

➢ **@Réponses** : si vous souhaitez répondre à un tweet, vous pouvez envoyer un tweet en débutant par le nom du compte précédé par un "@". Si nous prenons par exemple le tweet "@Antoine Bonjour !", vous allez ici envoyer le message "Bonjour" au compte d'Antoine, celui-ci verra votre réponse dans l'onglet "Réponses" de son profil. A noter que votre réponse est visible par tout le monde, du moins ceux qui vous suivent et qui suivent également le destinataire de votre message, et apparaît dans votre historique de tweets.[6]

➢ **Timeline** : Il s'agit du flux d'actualités de Twitter. La timeline générale présente l'ensemble des tweets postés par vos abonnements, et votre timeline personnelle affiche les différents tweets que vous avez mis en ligne. La timeline affiche les messages par ordre antéchronologique, c'est-à-dire du plus récent au plus ancien.[6]

Figure A-5 : Capture d'écran d'un exemple de Timeline. [6]

➢ **Les mentions (@) :** Un nom précédé d'arobase « @ » est un lien vers le compte Twitter de l'utilisateur de ce nom (qui permet de voir tous ses tweets, sauf s'ils sont protégés). Chaque utilisateur peut consulter les mentions qu'il a reçues dans l'onglet « @ Connect ». Si un tweet débute par une mention, seuls les followers suivant le compte mentionné verront le tweet dans leur fil d'actualité (par exemple @Eve rédige un tweet en commençant par @Bob, donc parmi les followers de @Eve, seuls ceux qui suivent également @Bob liront le tweet depuis leur fil d'actualité). [6]

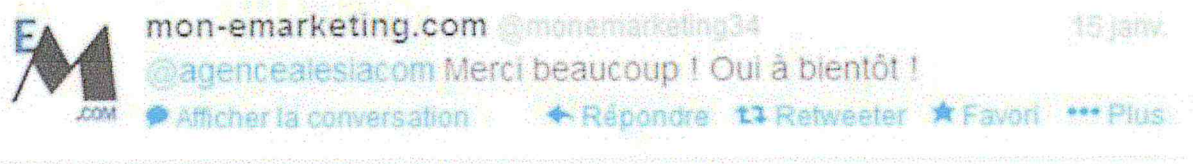

Figure A-6 : Capture d'écran d'un exemple de mention. [6]

➢ **RT (retweeter) :** Action qui consiste à rediffuser le message d'un autre utilisateur à vos abonnés. Un retweet (également désigné par l'abréviation RT) est donc un message rediffusé.

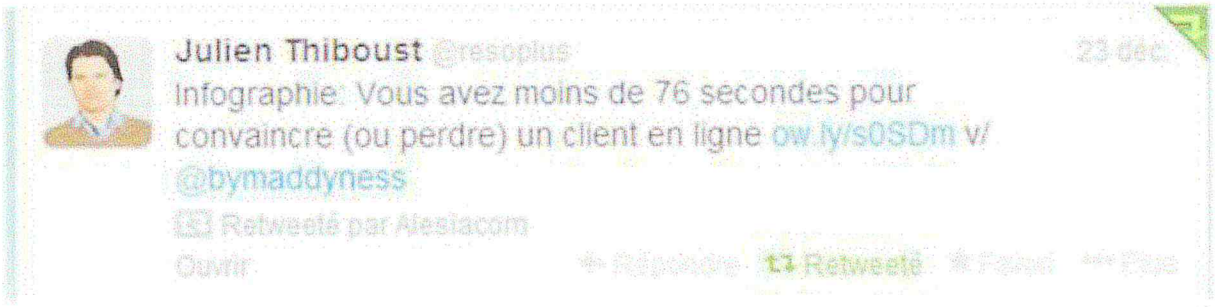

Figure A-7 : Capture d'écran d'un exemple de RT [6]

- **Message Privé (MP)** : (se dit « DM », pour « Direct Message » en anglais). Cette fonction permet d'envoyer un message privé à un utilisateur. Les MP sont eux-aussi limités à 140 caractères mais ils n'apparaissent pas dans les timeline : ils arrivent sur une messagerie interne à Twitter. On ne peut envoyer un MP à une personne que lorsqu'on la suit sur Twitter, et elle ne peut nous répondre que si elle nous suit également.[6]
- **Hashtag(#) :** Le « # » suivi d'un mot (sans espace et éviter les accents et autres caractères spéciaux) fonctionne un peu comme un mot clé ou un tag. Il permet de définir de manière générale le sujet principal du tweet. Lors d'un événement, il permet de suivre toutes les conversations sur Twitter relatives à cet événement. Ce qui est intéressant avec les hashtags, ils permettent de découvrir de nouvelles personnes qui parlent ou s'intéressent aux mêmes sujets que vous. [6]

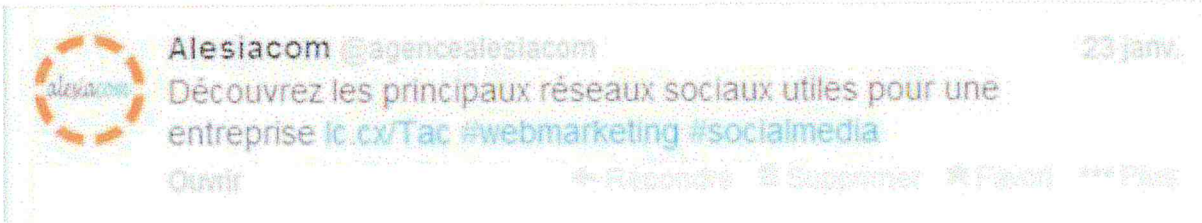

Figure A-8: Capture d'écran d'un exemple d'un Hashtag. [6]

- **Tendances :** Les tendances désignent en quelque sorte les sujets à la mode sur Twitter. Elles sont personnalisées en fonction de votre localisation et de vos abonnements.[6]

Figure A-9 : Capture d'écran d'un exemple des tendances [6]

Bibliographie :

[1] Unblog.fr. (2015). Qu'est-ce qu'un blog ?, [enligne].
URL:http://unblog.fr/2006/05/01/quest-ce-quun-blog/

[2]Futura-Sciences. Microblogging, [enligne].
URL :http://fr.cdn.v5.futura-sciences.com/builds/pdf/glossaire/10000-10999/10257-mic

[3] Robert, Péraro. (page consultée le février 2012). Twitter, c'est quoi exactement ?, [enligne].
URL: http://www.plateforme-echange.org/IMG/pdf/e-change_twitter_bat_26112012.pdf

[4] Arthur C. Clarke. (Page consultée le 2016).Twitter : Historique,présentation, chiffres-clés - Numerama, [enligne].URL : http://www.numerama.com/startup/twitter

[5] Maxime, Guernion. Histoire de Twitter, [enligne].URL : http://oseox.fr/twitter/histoire-twitter.html

[6] agencealesiacom. (Page consultée le 30/06/2016). 10 définitions pour maîtriser le vocabulaire de Twitter, [enligne]. URL: http://www.alesiacom.com/blog/maitriser-le-vocabulaire-de-twitter

[7] Twitter, Inc. (page consultée le 2016). Les différents types de Tweets et leur lieu d'apparition,[enligne].URL : https://support.twitter.com/articles/228517?lang=fr

[8]admin. (page consulté le 04/03/2015). Tutoriel – Utiliser l'API Twitter pour collecter des tweets avec Talend (et sans coder !),
[enligne].URL:http://www.erwanlenagard.com/general/tutoriel-utiliser-lapi-twitter-pour-collecter-d

[9] B, Bathelot. (page consulté 21/07/2015). Définition : API Twitter, [enligne].
URL: http://www.definitions-marketing.com/definition/api-twitter/

[10] B, Bathelot. (Page consulté 22/07/2015). Définition : Search API de Twitter, [enligne].URL : http://www.definitions-marketing.com/definition/search-api-de-twitter/

[11] Hoareau. (Page consultée le 06/01/2016). (Infographie) Des statistiques à savoir sur Twitter pour 2016, [en ligne].URL : http://boulevardduweb.com/stats-sur-twitter/

[12] Bollen, Johan, Huina Mao, and Alberto Pepe. "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena." *ICWSM* 11 (2011): 450-453.

[13] Cynthia Van Hee. (2013). L'analyse des sentiments appliquée sur des tweets politiques: une étude de corpus, [enligne].
URL: http://www.scriptiebank.be/sites/default/files/webform/scriptie/Masterproef_Cynt
[14] Jamal Atif. (2015). Analyse et Fouille de Données,[enligne].
URL: http://www.lamsade.dauphine.fr/~atif/lib/exe/fetch.php?media=teaching:coursafd
[15] université de Paris 3 - Sorbonne Nouvelle. Introduction à la fouille de textes.
URL: http://www.lattice.cnrs.fr/sites/itellier/poly_fouille_textes/fouille-textes.pdf
[16] Bolla, Raja Ashok. "Crime pattern detection using online social media." (2014).
[17] Wang, Xiaofeng, Donald E. Brown, and Matthew S. Gerber. "Spatio-temporal modeling of criminal incidents using geographic, demographic, and Twitter-derived information." In Intelligence and Security Informatics (ISI), 2012 IEEE International Conference on, pp. 36-41. IEEE, 2012.
[18] Gerber, Matthew S. "Predicting crime using Twitter and kernel density estimation." Decision Support Systems 61 (2014): 115-125.
[19] Chen, Xinyu, Youngwoon Cho, and Suk Young Jang. "Crime prediction using Twitter sentiment and weather." In Systems and Information Engineering Design Symposium (SIEDS), 2015, pp. 63-68. IEEE, 2015.
[20] BRINGAY, Sandra, Jérôme AZÉ, Pascal PONCELET, Amayas ABBOUTE, Yasser BOUDJERIOU, and Gilles ENTRINGER. "Prévention des suicides via Twitter."
[21] bordeaux, I. (2004).conduite de projet, [enligne].
URL: http://dept-info.labri.fr/~counilh/systeme-d-information/SI_0202.pdf
[22] Jean-François, PILLOU.(2015). Cycle de vie d'un logiciel, [enligne].
URL: http://www.commentcamarche.net/contents/473-cycle-de-vie-d-un-logiciel
[23]G, picard. (2009). Conduite et gestion de projets informatiques : une introduction,[enligne].
URL: http://www.univ-tebessa.dz/fichiers/master/master_1367.pdf
[24] Margaret, Rouse. (2015).que signifie JSON (Java Script Object notation), [enligne].
URL: http://www.lemagit.fr/definition/JSON-JavaScript-Object-Notation
[25] World Health Organization. "World Health rankings Live Longer Live Better." (2014).
[26] Ameni, Bouaziz. (2014).Catégorisation automatique de news à l'aide de technique d'apprentissage supervisé, [enligne].URL : http://www.zone-project.org/wp content/uploads/2013/03/ZONE-project_SVM_Ameni_Bouaziz.pdf
[27] Torres-Moreno, Juan-Manuel. Résume automatique de documents. Lavoisier, 2011.
[28] Porter, Martin F. "An algorithm for suffix stripping." *Program* 14, no. 3 (1980): 130-137.

[29] Yazid, Houria. "Les algorithmes d'apprentissage automatique offerts par l'environnement Weka." Université du Québec (2006).

[30] SENOUSSI Hafida.(2015). Sélection de Données pour l'Apprentissage des Réseaux de Neurones, Arbres de Décision et les k-Plus Proches Voisins : Application en Diagnostic de Pannes,[enligne].
URL : http://dspace.univ-usto.dz/bitstream/123456789/224/2/Memoire_Senoussi%20H

[31] Ismael Ngoma Serge Ntoto. (2016).mon java,[enligne].URL :https://www.editions-ue.com/catalog/details//store/fr/book/978-3-639-50367-8/m

[32] Pierre LACHEVRE et all. (2011). Site d'aide à la compréhension de l'anglais (SACA), [enligne]. URL :http://matis.univ-lehavre.fr/Projets/2010-2011/groupeD.pdf

[33] Guigouz, McPeter.PHPMyAdmin,[enligne].URL : phpmyadmin - Documentation Ubuntu Francophone.html

[34] atelier freelance. Introduction à MySQL, [enligne].URL: http://www.prosygma.com/telechargement/mysql_tutorial.pdf

[35] Yazid, H. "Document d'utilisation: Environnement Weka."

[36] japanese, korean.TWITTER4J, [enligne]. URL: http://twitter4j.org/en/

[37] http://nlp.stanford.edu/software/corenlp.shtml