République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Saâd DAHLAB, Blida

Faculté des sciences
Département d'informatique

Présenté par :

KHELIFA MOHAMED BENOMAR MOHAMED YACINE

En vue d'obtenir le diplôme de master
Domaine : Mathématique et informatique

Filière : Informatique
Spécialité : Informatique
Option : Ingénierie de logiciel

Sujet :

CONCEPTION ET IMPLEMENTATION D'UN SYSTEME DE CLASSIFICATION DES TWEETS

Encadré par :
Madame Madani Amina

Examiné par :

Mr CHERIF-ZAHAR S	Maitre-Assistant A	Président de jurys
Mme CHERFA I	Maitre-Assistant B	Examinatrice
Mme AMEUR K	Maitre-Assistant B	Examinatrice

2012/2013

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Saâd DAHLAB, Blida

Faculté des sciences
Département d'informatique

Présenté par :

KHELIFA MOHAMED

BENOMAR MOHAMED YACINE

En vue d'obtenir le diplôme de master
Domaine : Mathématique et informatique

Filière : Informatique
Spécialité : Informatique
Option : Ingénierie de logiciel

Sujet :

CONCEPTION ET IMPLEMENTATION D'UN SYSTEME DE CLASSIFICATION DES TWEETS

Encadré par :
Madame Madani Amina

2012/2013

Résumé

Les réseaux sociaux comme Twitter sont la dernière tendance dans le monde globalisé d'aujourd'hui. Twitter est utilisé par un large éventail d'utilisateurs et pour des différents besoins. La fouille et l'exploitation du contenu de twitter et spécifiquement les messages Tweets peut révéler des informations précieuses. Dans cette thèse, nous proposons une façon de classer automatiquement les Tweets, nous construisons une application à mettre en place et former une classification. En outre, module de détection est mis en œuvre. Après la construction de l'ensemble de données, une évaluation est réalisée pour mesurer la précision de la classification. Les résultats sont très prometteurs.

Abstarct :

Social networks like Twitter are the latest trend in today's globalized world. Twitter is used by a wide range of users and for different needs. The search and exploitation of content and specifically Tweets twitter messages can reveal valuable information. In this thesis, we propose a way to automatically classify Tweets, we build an application to build and train a classification. In addition, the detection module is implemented. After construction of the data set, an evaluation is performed to measure the accuracy of the classification. The results are very promising.

REMERCIMENTS

Tout d'abord on tient à remercier notre DIEU le tout miséricordieux de nous avoir permis d'achever ce travail.

Nous adressons nos vifs remerciements à Madame Madani, notre promotrice, pour ses conseils, sa collaboration et ses encouragements

Nous remercions notamment nos chers enseignants pour leurs efforts durant les 5 années d'étude.

Enfin, on remercie toute personne ayant contribuée de prés ou de loin à la progression de ce projet.

DEDICACES

Le premier mérite revient à mes chers parents pour leur grande patience, compréhension et soutien.

A toute ma famille

A tous mes amis

Ainsi qu'a tous ceux qui me sont chers

DEDICACES

TABLE DES MATIERES

Introduction générale 1
Motivation 1
Problématique 2
Objectif 2
Organisation du mémoire 2

CHAPITRE I TWITTER : GENERALITES

1. Introduction 5
2. Historique 5
3. Twitter 6
   - 3.1 Les tweets 7
   - 3.2 Avantage 7
   - 3.3 Followers 8
   - 3.4 Twitter Slang 8
   - 3.5 API de Twitter 9
   - 3.6 Statistiques 9
   - 3.7 Business Model 11
   - 3.8 Spam 12
4. L'usage de Twitter 13
   - 4.1 L'usage politique 13
   - 4.2 L'usage social 13

Conclusion 14

TABLE DES MATIERES

CHAPITRE II L'ETAT DE L'ART

1. Introduction..... 16
2. La fouille de texte (Text mining) 16

2.1 La fouille dans les tweets (Tweet mining) 16

3. La classification des données (Clustring Data) 16

3.1 La classification supervisée 17

- Méthodes 17

3.2 La classification non supervisée..... 17

- Méthodes 17

4. La classification des tweets..... 17

4.1 La Classification thématique..... 18

4.1.1 Extraction des tendances (topics) 18

4.1.2 La détection des maladies..... 20

4.1.3 La détection des catastrophes naturelles..... 20

4.2 L'analyse des sentiments..... 20

4.2.1 Les applications composites (ou mashup)..... 21

- TweetFeel..... 21

4.3 L'analyse des réseaux sociaux..... 22

4.3.1 La catégorisation des utilisateurs..... 22

4.3.2 Les amis..... 23

4.3.3 Anatomie des mises à jour de statut..... 23

4.3.4 La répartition géographique..... 24

4.3.5 Les systèmes de recommandation..... 24

4.3.6 E-learning..... 25

4.3.7 La retransmission des évènements en direct..... 25

4.3.8 L'acquisition de connaissances..... 26

Conclusion..... 26

TABLE DES MATIERES

CHAPITRE III CONCEPTION ET IMPLEMENTATION

1. Introduction..... 28
2. Présentation de la démarche utilisée..... 28

2.1 Le cycle de vie..... 28

3. Modèle en cascade..... 28

3.1 Expression des besoins..... 29

3.2 Conception..... 29

4. Expression des besoins..... 30

4.1 Identification des cas d'utilisation..... 30

- Les cas d'utilisations..... 30

4.2 Diagramme de cas d'utilisation..... 30

5. Conception..... 31

5.1 Diagramme de Classes..... 31

6. L'Architecture globale de l'application..... 33
7. L'approche de fouille dans les messages tweets..... 34

7.1 La Collection des données..... 36

7.1.1 Extraction des données à partir des documents XML..... 36

7.1.2 Prétraitement des données..... 36

- Elimination des mots vides (Stop words) 36
- Racinisation (stemming) 37
  - ➢ Détail de l'algorithme de Porter..... 37
  - ➢ Les étapes de l'algorithme Porter..... 38

7.1.3 Construction du dictionnaire..... 40

7.2 L'extraction des éléments de classification..... 40

7.2.1 Calcul du poids avec TF-IDF..... 40

➢ TF-IDF..... 40

7.2.2 Extraction des sujets..... 41

7.2.3 Construction des collections modèles..... 42

7.2.4 Construction des vecteurs / tweets..... 42

7.3 La classification (clustering)..... 42

TABLE DES MATIERES

8. Conception détaillée et implémentation..... 43
8.1 Technologie (Outil de développement) 43
8.1.1 Java..... 43
8.1.2 Eclipse..... 44
8.1.3 MySQL..... 44
9. Présentation de l'application « Whatup »..... 44
9.1 Interface d'importation du fichier XML..... 45
9.2 Calcul du poids..... 46
9.3 La classification..... 47

Conclusion..... 48

1. Introduction :

Les réseaux sociaux et les blogues prennent de plus en plus de place dans notre vie. Avec des millions d'utilisateurs, ils s'imposent comme outils de communication et d'échange d'information. Ces outils en général est surtout Twitter , qui est devenu une source constante de nouvelles (news) alternatives pour les utilisateurs d'Internet, et également un canal dans lequel les utilisateurs peuvent attirer et diriger l'attention des médias internationaux.

Twitter a pris une avance considérable sur les médias traditionnels par exemple lors du tremblement de terre 2011 de Washington DC aux états unis, les utilisateurs ont pratiquement posté des tweets au moment du séisme. Les utilisateurs de New York voyaient ces tweets avant même qu'ils sentent le séisme [1].

Grace à Twitter, nous serons tenus informés en temps réel sur n'importe quel sujet. Nous dépendons de membres de la famille, des amis proches, et ceux qui font partie de nos réseaux numériques à agir comme des journalistes, afin de nous alerter quand un évènement important s'est passé ou se passe, nous sommes toujours connectés alors nous sommes toujours au courant.

2. Motivation :

Les réseaux sociaux tels que Twitter sont la dernière tendance dans la mondialisation .**Twitter** est devenu un processus avec sa propre dynamique. Il est utilisé dans différents scénarios par un large éventail d'utilisateurs. Chacun d'eux a son propre comportement, son propre style d'écriture. L'exploitation de ces réseaux sociaux peut extraire des précieuses informations .Actuellement, il y a des nombreuses recherches dans ce domaine, avec des résultats prometteurs.

Par exemple :

Sitaram et al [2] ont utilisé Twitter pour prédire les recettes des films, ils ont atteints une précision de 97%. En outre, Twitter a été utilisé pour prédire les résultats des élections présidentielles américaines en 2008 [3] .les Tweeters ont favorisé Obama sur McCain et Obama a fini par gagné l'élection.

3. Problématique :

Nous vivons aujourd'hui dans un monde qui change chaque minute, un monde où les événements s' accélèrent de façon terrible, pour cette raison , le besoin de rester informé sur ce qui se passe à chaque instant a augmenté , et en raison du grand nombre d'informations qui circule dans les réseaux sociaux et la diversité de ces informations, l'utilisateur a besoin des moyens pour l'aider à atteindre les nouvelles appropriées pour lui au bon moment .

A travers ce travail, nous nous essayons de répondre à la question : Qu'est-ce qui se passe, et sur quel sujet ce concentre l'information sur twitter . Notre travail consiste à extraire et analyser les tendances ou les sujets émergents dans le réseau twitter afin d'enquêter sur ce qui ce passe en temps réel sur ce réseau.

4. Objectif :

L'objectif de notre travail est d'élaborer une approche capable d'extraire les tendances des messages twitter et de classifier ces messages en se basant sur ces tendances.Notre objectif et de fournir des bons résultats (avec une précision importante) en utilisent une collection de données appropriées à ce genre de travail.

5. Organisation du mémoire :

Dans le premier chapitre nous allons essayer de comprendre twitter en général , sa structure , son modèle économique , son rôle dans la vie politique est sociale , ensuite dans le deuxième chapitre nous proposons une étude approfondie sur l'ensemble de travaux réalisés dans le domaine du tweet mining . Dans le troisième chapitre nous proposons une nouvelle approche de tweet maining ainsi que l'implémentation de du système . Et dans le dernier chapitre nous détaillons les tests de l'approche.

1. Introduction :

En Mars 2006, un nouveau service de communication appelé Twitter fait ses débuts. Il commençait comme un petit projet dans une entreprise de podcasting à San Francisco, mais il ne fallut pas longtemps attendre avant que le projet de côté devient l'événement principal d'aujourd'hui. Un peu plus de cinq ans plus tard, Twitter est en plein essor.

En Octobre 2012, Twitter a annoncé que le nombre de tweets publiés par journée a atteint 58 millions [4]. , En outre, Twitter est maintenant disponible en 17 langues (et les gens utilisent plus de langues pour communiquer). [5]

Twitter est devenu un canal de communication clé lors des grands événements politiques et Catastrophes naturels, et les entreprises s'appuient désormais sur lui pour le marketing, les relations publiques et même pour le service destiné à la clientèle.

Dans ce chapitre, nous allons présenter twitter comme un réseau social avec son historique, son principe de fonctionnement, ses avantages sur les autres réseaux et ces déférents modes d'utilisation dans les domaines politiques et économiques.

2. Historique :

Il semblerait que la création de **Twitter** soit, moins glorieuse que la version officielle ne le dit. Officiellement, Deux anciens de Google, **Evan Williams** et **Biz Stone**, les deux fondateurs dans la société **Odeo** voulaient lancer une plateforme de podcasting, mais l'arrivée d'**iTunes** change la donne. Avec un autre employé d'**Odeo**, **Jack Dorsey**, ils lancent **Twitter** qui n'intéresse pas les investisseurs d'**Odeo**. **Williams** rachète généreusement leurs parts pour quelques millions de dollars. Aujourd'hui, cet étrange réseau social, cette plate-forme de mini-messages et de micro-blogging, est estimée à 5 milliards de dollars [6].

Mais d'autres sources affirment que **Twitter** a commencé comme une idée de **Jack Dorsey** 2006 . **Dorsey** avait initialement imaginé **Twitter** comme une plate-forme de communication basée sur le principe SMS (**Short Message Service**) ou des groupes d'amis pourront garder un œil sur ce que les autres faisaient en fonction de leurs mises à jour de statut.

Dorsey propose cette plate-forme basée sur SMS aux fondateurs de l'entreprise de podcasting **Odeo. Evan Williams** et **Biz Stone,** ces deux derniers donne à **Dorsey** le feu vert pour passer plus de temps sur le projet et de le développer d'avantage.

Au début le nom de la plateforme était **twttr**, Le développeur **Noah Glass** est crédité de venir avec le nom originale ainsi que son incarnation finale **Twitter**. [6]

Dorsey publie le premier message sur **Twitter** le 21 Mars 2006 à 21h50 : "*just setting up my twttr*". [6]

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 [7].

En septembre 2006, **Evan Williams** le PDG d'**Odeo** écrit une lettre à ses investisseurs :

"*Deux mois après son lancement, Twitter a moins de 5 000 utilisateurs enregistrés. Je vais continuer à investir dans Twitter, mais je ne peux pas dire que cela justifie le capital-risque mis dans Odeo, surtout qu'au départ il visait un marché complètement différent.*" [8]

Williams leur propose de leur racheter leurs parts, et finalement ils acceptent. **Williams** achète **Odeo** et **Twitter** avec. Le montant n'a jamais été rendu public, mais l'ensemble devait atteindre 5 millions de dollars, Williams aurait tout remboursé. Cinq ans après, **Odeo** vendu 5 millions de dollars vaut environ 1 000 fois plus : 5 milliards de dollars [8].

3. Twitter :

Twitter est un réseau social très populaire qui pose une seule question: **What are you doing?** (Qu'est-ce que tu fais?). La réponse est limitée à 140 caractères. Les Mises à jour peuvent être envoyées via un navigateur Web, SMS et E-mail et sont affichés sur le profil des utilisateurs.

Twitter repose sur le principe du **microblog** ou **microblogue** qui est un dérivé concis du **blog**, qui permet de publier un court article, plus court que dans les blogs classiques. Les articles pouvant être de type texte court, mais peuvent également contenir une image ou même une vidéo embarquée. Les flux d'agrégation sont plus

légers que dans les blogs traditionnels et peuvent contenir tout le message. La diffusion peut également être restreinte par l'éditeur à un cercle de personnes désirées.

Figure 1.1 : Screenshot de l'interface de Twitter

3.1 Les tweets :

Les tweets ou les gazouillis, sont des messages textuels brefs partagés et diffusés sur Twitter. Un tweet ne doit pas contenir plus de 140 caractères (soit une ou deux phrases). Chaque utilisateur dispose d'une page Twitter.Les tweets ne sont pas seulement affichés sur la page profil d'un utilisateur, mais ils peuvent être envoyés gratuitement aux followers par Internet, par messagerie instantanée, par SMS (Short Message Service), par flux RSS (Really Simple Syndication), par e-mail, ou d'autres plates-formes de réseaux sociaux, tels que Twitterrific ou Facebook.

3.2 Avantage :

Twitter est devenu un outil central qui propose quelques avantages :

1. Faire de la veille sur des sujets pointus.
2. Contacter des personnes intéressantes rapidement et sans être envahissant.
3. Interagir à distance de façon rapide et légère (moins intrusif que le mobile ou le SMS, moins lourd qu'Email).
4. Suivre la propagation d'une nouvelle information avant que tout autre média l'a encore captée

Figure 1. 2 : Screenshot d'un profil Twitter

3.3 Followers :

Twitter a mis en place un concept de followers ou adeptes en français (ceux qui suivent vos messages). Si un certain utilisateur met à jour son statut, tous les followers sont informés. Ceci est réalisé en cliquant sur le button **suivre** ou (Follow) sur une page **Twitter** .On peut suivre tous les autres utilisateurs à moins que cet utilisateur a mis son profil en mode privée. Dans ce cas, une demande d'approbation doit être envoyée en premier.

3.4 Twitter Slang :

En raison de la limite de 140 caractères, les utilisateurs ont développé des stratégies à fin de mettre autant d'informations que possible dans les messages. Cela comprend l'utilisation du caractère dièse (#) Pour marquer un message avec certains sujets. Par exemple, un message peut ressembler à ceci:

(Je suis en train de tester une nouvelle fonctionnalité Twitter. # Test # twitter #Blida, le message est maintenant tagués avec. « # Test # twitter #Blida», et d'autres personnes sont en mesure de rechercher ces balises à l'aide du site Twitter.

En outre, l'utilisation des services de raccourcissement d'URL est devenue très populaire. Ces services permettre de raccourcir une certaine URL afin qu'il puisse être

posté sur **Twitter**. Par exemple http://www.know-center.tugraz.at/forschung/ devient http://tinyurl.com/yh8lqhz .

3.5 API de Twitter :

Twitter fournit une application d'interface et de programmation « **Application Programming Interface** » (API) , pour accéder à ses données .En utilisant le protocole HTTP, **Twitter** propose une méthode pour chaque fonction qui peut être utilisée sur le site. Cela comprend les mises à jour de statut, les opérations de recherche et d'accéder au **timeline** (mur) des utilisateurs. L'utilisation de l'API est gratuite, elle nécessite seulement un compte **Twitter** actif. Elle est limitée à 20.000 requêtes par heure. L'API est utilisée par un large éventail d'applications.

3.6 Statistiques :

Dans le tableau ci-dessus, nous présentons quelques statistiques sur Twitter.

Description des Statistiques	**Montant**
Nombre total d'utilisateurs enregistrés actifs	**554, 750,000**
Nombre de nouveaux utilisateurs chaque jour	**135,000**
Nombre de visiteurs du site Twitter chaque mois	**190 million**
Nombre de tweets publié par jour	**58 million**
Nombre de requête qui interroge le moteur de recherche de Twitter chaque jour	**2.1 billion**
Pourcentage des utilisateurs qui utilisent leur téléphone pour tweeter	**43 %**
Pourcentage des tweets qui proviennent des applications tierce partie	**60%**
Nombre des employées de Twitter	**2,500**
Nombre d'utilisateurs actifs chaque mois	**115 million**
Nombre de jours qu'il faut pour attiendre 1 milliard de tweets	**5 jours**
Nombre de tweets qui se produisent chaque seconde	**9,100**

Tableau 1. 1 : Statiques générales sur Twitter [4].

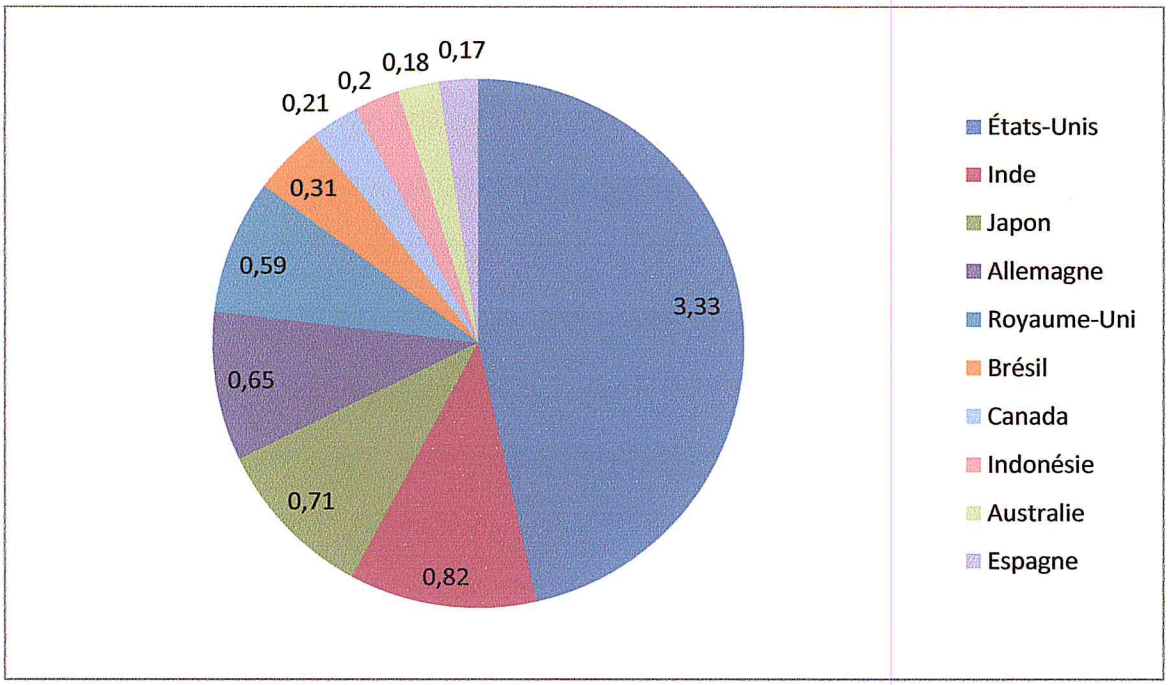

Figure 1.3: La répartition géographique des utilisateurs Twitter dans le monde [9]

Nous remarquons à travers cette figure que le principal trafic sur twitter se concentre aux états unis, mais en termes de pourcentage de la population utilisant Twitter, le Brésil arrive en tête de liste, un Brésilien sur quatre utilise Twitter, comparativement à environ un dixième de la population américaine. L'Indonésie arrive en deuxième position par rapport au pourcentage d'utilisateurs, suivie par les Pays-Bas, le Japon et le Venezuela.

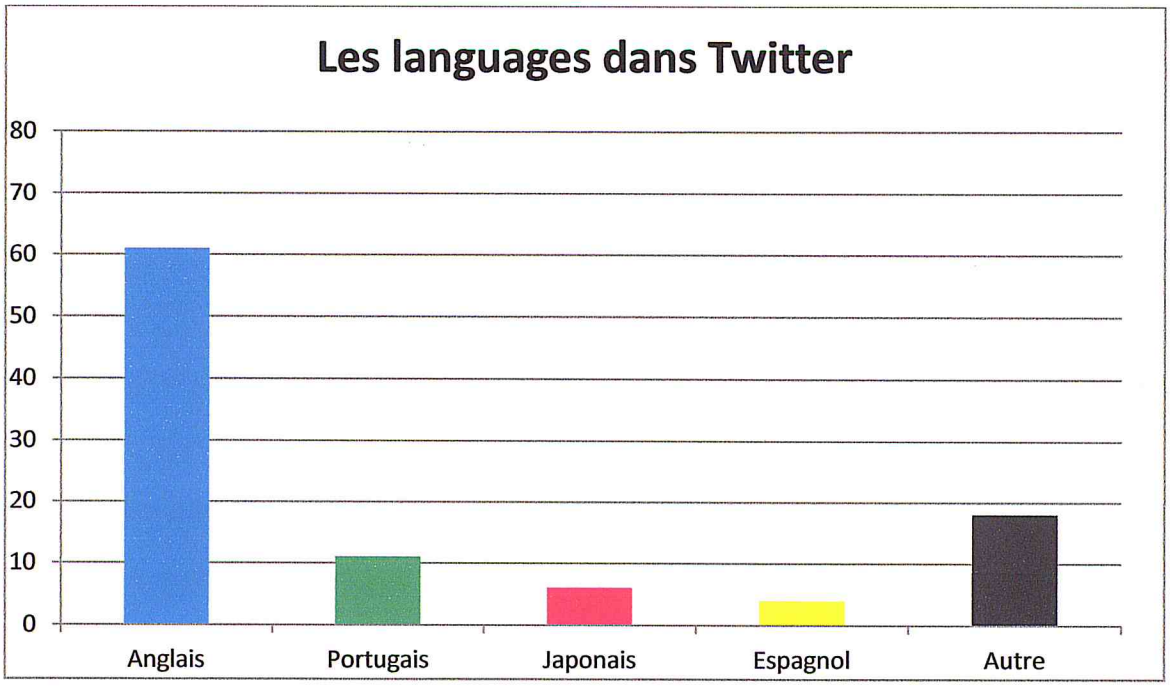

Figure 1. 4: Les languages populaires sur Twitter [9].

Cette figure montre que la langue anglaise reste la plus utilisée dans le réseau twitter suivie par le portugais et la langue japonaise.

3.7 Business Model :

Le modèle économique de **Twitter** est, depuis le lancement du service en 2006, un véritable mystère. Toutes les pistes ont été évoquées : **services premium payants, outils supplémentaires, publicité sur le site, publicité entre les tweets.**

Comme plusieurs autres réseaux sociaux populaires, **Twitter** se cherche un modèle économique valide qui génère effectivement des revenus. **Twitter** confirme sur le web:

> « ***Twitter has many appealing opportunities for generating revenue but we are holding of on implementation for now because we don't want to distract ourselves from the more important work at hand which is to create a compelling service and great user experience for millions of people around the world*** » [10].

(**Twitter** offre de nombreuses possibilités pour générer des revenus, mais nous tenons juste sur la mise en œuvre pour l'instant parce que nous ne voulons pas nous distraire de travail le plus important qui est de créer un service convaincant et une grande expérience pour des millions de personnes à travers le monde).

Le financement du service **Twitter** s'appuie fortement sur les investisseurs et a ainsi généré un financement total de 155 millions de dollars. Selon le Financial Times, les investisseurs valorise actuellement le site avec 1 milliard de dollars [11].

En Avril 2010, **Twitter** fait une nouvelle tentative de trouver un modèle économique par la mise en place des **"Promoted accounts"**, des comptes de clients qui sont mis en avant sur le réseau. Selon leur blog, ils veulent afficher la publicité sur la page de résultat après une recherche sur le site. Si le retour est positif, ils prévoient également d'afficher de la publicité dans les tweets. Cette approche semble être similaire aux **(Tweets sponsorisés).**

Dernièrement **Twitter** vient d'acquérir **AdGrok**, une société spécialisée dans la publicité en ligne qui pourrait permettre à **Twitter** de booster ses recettes publicitaires, qui se sont élevées, en 2010, à 45 millions de dollars (contre plus de 1,3 milliards de dollars pour Facebook) [12].

3.8Spam :

Comme chaque plate-forme de communication populaire, **Twitter** est confronté aux problèmes des Spams. Sur le Blog officiel de la société, les opérateurs de **Twitter** définissent un spam comme « une variété de différèrent comportements qui vont de l'insidieuse a l'ennuyeux ». Cela inclut le « following »agressif (suivre un utilisateur twitter de manière excessive) ou bien le cas contraire, le « unfollowing », et aussi des liens vers des sites de phishing.

Twitter lutte dur pour éviter les spam. Par exemple, chaque utilisateur dispose d'un bouton « REPORT FOR SPAM» dédié à indiquer un spam. Selon les opérateurs, ils ont réussi à faire baisser le niveau des spam à 2%.

4. L'usage de Twitter :

Twitter est utlisé dans plusieurs domaines et pour plusieurs raisons, pour capter l'intention sur un évènement ou un produit ou même une idée politique.

4.1 L'usage politique :

En 2009, Au cours des élections présidentiels iranienne , Twitter a joué un rôle très important. Alors que les journaux et les blogs ont été fortement censurés par le gouvernement, les utilisateurs de Twitter ont publié des nouvelles de la rue en temps réel en utilisant les hashtags **#Iran** ou **# iranelection**. Les médias partout dans le monde ont affichées les derniers messages de Twitter. Après un certain temps, le gouvernement iranien a tenté de réprimer ces messages. Les utilisateurs ont réagi et demandé à tous les utilisateurs de Twitter de modifier leur emplacement pour **Téhéran, Iran**, ce qui rend la tâche de localiser ces personnes impossible pour le ministère du Renseignement Iranien.

Le Département d'Etat américain était au courant de l'importance de cet outil de communication et demandait à la direction de **Twitter** de reporter une mise à niveau régulière qui a eu lieu 3 jours après l'élection.[13]

Le 15 Juin 2009. **Twitter** a accepté et a effectué la mise à niveau à 14h00 heure américaine, qui est 01h30 heure de Téhéran. Le chef de l'opposition, **MirHossein Moussavi** a même twitté son arrestation:« *Cher peuple iranien, Moussavi ne vous a pas laissé seul, il a été placé en résidence par le ministère du Renseignement*». [14]

4.2 L'usage social :

Twitter a été utilisé pour recueillir des fonds pour les victimes du séisme en Haïti [15], Peu de temps après le séisme, la Croix-Rouge américaine a envoyé un Tweet pour informer les donateurs. Cela montre que **Twitter** joue un rôle important dans la société de nos jours et peut être utilisé pour autre chose que des mises à jour de statut ordinaire.

Conclusion :

Depuis son lancement en 2006, **Twitter** est un outil social qui trouve un vrai succès dans l'univers brutal du web 2.0. Avec plus de 140 millions d'utilisateurs et 340 millions de tweets quotidiens. **Twitter** a grignoté une place confortable dans l'ombre de Facebook, rendant des services insoupçonnables à son lancement.

Twitter regroupe maintenant tous les utilisateurs directement dans son écosystème, il va être intéressant de suivre quelle direction les dirigeants de **Twitter** vont finalement privilégier dans le futur :

1. **Twitter hyper social:** un site social cherchant à concurrencer Facebook sur son propre terrain : la vie de tous les jours et les échanges du quotidien pour le grand public avec plus de façons simples de regrouper ses amis.
2. **Twitter nouveau média :** un site de diffusion et d'analyse de l'information mondiale orienté vers le data-mining et la visualisation des données, pour concurrencer la recherche de **Google.**

Dans les deux cas **Twitter** semble avoir une place à prendre. Mais les modèles économiques seraient probablement très différents.

Dans ce chapitre on a essayé de bien présenter twitter ainsi que son rôle dans la vie, dans le chapitre suivant nous proposons une étude de l'état de l'art concernant la fouille des messages tweets.

Chapitre II :

L'ÉTAT DE L'ART

1. Introduction :

Dans ce chapitre nous allons présenter une étude approfondie sur l'état de l'art dans le fouille des messages tweets, La plus part des recherches dans le tweet mining se focalise sur la classification des utilisateurs de ce réseau à travers leurs publications ou la détection des sentiments concernant un évènement ou un produit.

2. La fouille de texte (Text mining) :

La fouille de textes ou l'extraction de connaissances à partir d'un texte est une spécialisation de la fouille de données. C'est un ensemble de traitements informatiques consistant à extraire des connaissances selon un critère de nouveauté ou de similarité dans des textes produits par des humains pour des humains. Dans la pratique, cela revient à mettre en algorithmes un modèle simplifié des théories linguistiques dans des systèmes informatiques d'apprentissage et de statistiques.

2.1 La fouille dans les tweets (Tweet mining) :

Le tweet mining est une technique permettant d'automatiser le traitement de gros volumes de contenus des massages tweets pour en 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.
Beaucoup de recherches dans le tweet mining se focalise sur la classification des tweets.

3. La classification des données (Clustring Data) :

La classification est une méthode d'analyse de données, L'objectif le plus simple d'une classification est de répartir l'échantillon en groupes d'observations homogènes, chaque groupe étant bien différencié des autres [16].Il existe deux type de classification : Supervisée et non supervisée.

3.1 La 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
- SVM

3.2 La 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.

4. La classification des tweets :

La classification est le processus d'attribution des tweets, sous une forme ou une autre, à des groupes ou classes à partir d'un ensemble prédéfini, ce processus fait partie de la fouille dans les tweets, La classification qu'elle soit supervisée ou non supervisée, permet d'analyser les tweets, de les classifier dans une, plusieurs ou aucune catégorie afin d'extraire des connaissances afin d'établir automatiquement qui participe à la production de l'information dans twitter ou de la conversation autour des événements ou des produits afin d'améliorer la consommation de contenu de l'événement pour aider à exposer

les parties prenantes de cet événement et leurs intérêts variés, et même aider à orienter la couverture d'un événement par les médias.

Contrairement à la classification des autres types de documents semi-structurés tels que XML, qui est basée sur la structure plus que le contenu des documents, la classification des tweets exploite moins l'information structurelle pourtant elle est très riche. La classification est basée beaucoup plus sur le contenu des tweets. La combinaison du contenu et de l'information structurelle peut aider à extraire des informations utiles et significatives.

La date et l'heure des tweets est l'information structurelle la plus exploitée en vue de faire l'analyse des tweets et l'extraction des thèmes en temps réel. Nous pensons que l'analyse des tweets en temps réel peut aider à comprendre ce qui se passe actuellement, que pensent les gens, partout dans le monde.

Voici la présentation des différentes recherches dans le tweet mining selon le but de chaque recherche :

- Classification thématique : Extraction des tendances la détection des maladies et des catastrophes naturelles.
- l'analyse des sentiments
- l'analyse du réseau social.

4.1 La Classification thématique:

4.1.1 Extraction des tendances (topics)

Une tendance est un sujet qui est affiché fortement sur Twitter (par exemple, la catastrophe d'Haïti en Janvier 2010).

Cheong et al [17] ont étudié l'anatomie des sujets émergents ou les tendances sur Twitter . Ils ont divisé les sujets tendances en 3 catégories :

- Sujets à long terme : sujets à long terme se produisent rarement mais reste une longue période de temps dans la les (timeline public),
- Sujets à moyen terme : tandis que les sujets à moyen terme se produit plus fréquemment mais ils sont limités a quelques jours
- Sujets à court terme : les Sujets à court terme sont des sujets très discutés, et se réfèrent souvent à l'actualité.

Les résultats classent les utilisateurs en 3 grands groupes:

Personal ou (personnel),(**Aggregator** ou Agrégateur) et **Marketing** , Les résultats montrent que la plupart des utilisateurs qui parlent de leur vie personnelle contribuent à émerger des nouveaux sujets émergents ou tendances, et que les spams (Marketing) sont généralement fondées sur les tendance afin de susciter l'attention.

Naaman et al, [18] ont classés les tweets en fonction de leur contenu. Ils ont créé 9 catégories et attribué manuellement les 10 derniers tweets des 350 utilisateurs sélectionnés au hasard. Les catégories sont les suivantes:

- IS («**Information sharing** ou Partage de l'information) .
- SP (**Self Promotion** ou la promotion du soi).
- OC (**Opinions/Complaints** ou Avis et plaintes).
- RT (**Statements and Random Thoughts** ou déclaration et pensées aléatoires).
- ME (**It's all about me** ou à propos de moi).
- QF (**Questions to followers** ou Questions aux fans).
- PM (**Presence maintenance** ou maintenance de présence).
- AM (**Anecdote – me** ou Anecdote - moi).
- AO (**Anecdote – other** ou Anecdote - autre).

En utilisant l'apprentissage automatique pour classer les utilisateurs de Twitter dans des catégories, chaque utilisateur est représenté en tant que vecteur des caractéristiques pour capturer les différences entre les catégories. Les résultats montrent que plus de 40% des Tweets sont classés comme ME (par exemple :(j'ai faim)), suivie de RT, OC avec environ avec 20% chacun. En d'autres termes, cela signifie que 20% des Tweets ont un caractère de nouvelles, tandis que 80% peut être caractérisée comme communication entre les utilisateurs.

4.1.2 La détection des maladies :

Culotta et al, [19] proposent la classification afin de prédire des maladies en analysant seulement le contenu textuel des tweets. Ils utilisent un classifieur de documents pour analyser le contenu des messages de Twitter. Le classifieur de documents est basé sur un sac de mots. La classification permet de prédire si un tweet rapporte un symptôme ILI (Influenza-Like Illnesses) afin de calculer des taux de grippe dans une population et les comparer avec les statistiques du CDC (The U.S. Centers for Disease Control and Prevention). La prédiction est basée sur la fréquence des messages qui contiennent certains mots-clés.

4.1.3 La détection des catastrophes naturelles :

Sakaki et al, [20] proposent une méthode pour contrôler les tweets et détecter la cible d'un événement en temps réel. La méthode détecte rapidement les tremblements de terre au Japon via les tweets et envoie des e-mails aux utilisateurs enregistrés avec une notification très rapide, et peut-être avant qu'un tremblement de terre arrive à un certain emplacement. (Sakaki et al., 2010) considèrent les tweets comme autant des capteurs produisant des informations sensorielles. Pour cela ils conçoivent un classifieur de tweets en utilisent un vecteur SVM (support vector machine) (Joachims, 1998) basé sur les composants d'un tweet tels que les mots-clés, le nombre de mots et leur contexte. Par la suite, ils produisent un modèle probabiliste spatio-temporel qui peut trouver le centre et la trajectoire de l'emplacement de l'événement.

4.2 L'analyse des sentiments :

La détection des sentiments (aussi connu comme l'analyse des sentiments, ou «Opinion maining » est l'approche de détecter le sentiment (ou des sentiments) de l'auteur du message en ce qui concerne un sujet particulier. Ceci est particulièrement intéressant pour les entreprises qui souhaitent savoir le degré d'appréciation de leurs produits chez les utilisateurs de twitter ou

d'autres réseaux sociaux. Par exemple, si le mot iPad est affecté avec plus de positifs ou négatifs sentiments. La même chose s'applique pour les films, les chansons, les voitures, les destinations de vacances, les partis politiques et ainsi de suite.

Différentes approches existent quand il s'agit d'essayer de déterminer le sentiment, allant de l'analyse lexicographique à des techniques d'apprentissage automatique utilisant SVM (Les machines à vecteurs de support).

Pang et al [21] ont évalué la performance des techniques d'apprentissage de machine sur la détection de sentiment. En utilisent un ensemble de données (tweets) de critiques des films, ils atteignent une précision de 80% en utilisant les Les machines à vecteurs de support.

Kamps et al [22] proposent des mesures qui déterminent l'orientation sémantique des adjectifs en utilisent la base des mots wordnet[1], Les Machines à vecteurs de support (SVM) et l'application de sémantique différentiel Osgood, ces mesures peuvent être utilisées pour mesurer les opinions, sentiments .L'evaluation à l'égard des jugements humains montre l'efficacité de ces mesures .

4.2.1 Les applications composites (ou mashup) :

Une application composite est une application qui combine du contenu ou du service provenant de plusieurs applications plus ou moins hétérogènes afin de proposer un nouveau service, par exemple l'application Wikimapia[2] (une fusion de Wikipedia et Google maps) .

Le meme principe s'applique sur twitter car son API est ouverte au public et les données peuvent être facilement consultées.

Par exemple :

TweetFeel :

TweetFeel est une application de recherche des sentiments en temps réel dans les Tweets, l'utilisateur saisie le nom de la personne, un produit ou un

1 : Base de données lexicale

2 :http://www.wikimapia.org/

événement etc…., ensuite l'application recherche les Tweets actuelles concernant les termes saisies . Il applique ensuite un algorithme essayer pour déterminer si la phrase de chaque tweet est positive ou négative, les mots comme «bon», «mauvais» ,«aime», «déteste» etc…. sont prise en compte pour déterminer si quelqu'un est positive ou négative envers notre recherche.

4.3 L'analyse des réseaux sociaux :

4.3.1 La catégorisation des utilisateurs :

Chaque utilisateur peut être représenté par 3 entités : le nombre de mises à jour de statut, le nombre d'amis et le nombre d'adeptes(Followers).

Sur la base de ces entités, Krishnamurthy et al [23] ont classé les utilisateurs de Twitter en 3 catégories:

- **broadcasters** ou (diffuseurs) : Les diffuseurs sont des utilisateurs ayant un nombre élevé d'adeptes. Cela inclut les utilisateurs tels que : Les chaînes télévisées, journaux ou les utilisateurs célèbres
- **Acquaintances** ou (connaissances) : le deuxième groupe (acquaintances) possèdent un nombre équilibrée d'amis et de followers (ils ont tendance à présenter réciprocité dans leurs relations) ,
- **miscreants** ou (mécréants) :Le dernier groupe , les mécréants, sont des followers d' un grand nombre d' entités twitter , ce nombre dépasse me nombre de leur adeptes . Les spammeurs entrent dans ce type.

Java et al, [24] Ont appliqué l'algorithme HITS[1] qui était initialement développé pour le classement des pages dans le WWW[2], afin de localiser dans Twitter :

- **les centres** ou (hubs) : un (hub) est un utilisateur twitter qui a beaucoup d'adeptes et beaucoup d'amis.
- **les autorités** ou (authorities) : Une autorité, d'un autre côté, c'est quelqu'un qui a beaucoup d'adeptes, mais moins d'amis.

1 : un algorithme qui permet de mesurer l'autorité d'une page Web par rapport à d'autres

2 :Les système hypertexte World Wide Web

Quelqu'un avec presque aucun amis et adeptes n'est ni un hub ni une autorité. Sur la base de cette classification, ils ont séparés l'intention des utilisateurs en 3 parties:

- **Information sharing** (le partage de l'information).
- **Information seeking** (recherche d'information).
- **FriendWise-relations** ou les relations d'amitié.

4.3.2 Les amis :

Chaque utilisateur de Twitter peut suivre d'autres utilisateurs, devenant ainsi (follower) d'autre.

Huberman et al, [25] ont enquêté sur la question avec combien de ces followers un utilisateur est en vraie communication. Il propose un ami en vraie communication c'est un ami dont l'utilisateur a envoyé à moins deux messages directes, Les résultats montrent que les gens communiquent directement avec seulement 13% de leurs tes adeptes, Cela révèle deux réseaux différents dans Twitter: Le premier est constitué des relations entre les utilisateurs et leurs adeptes, et le second des relations entre les amis réels.

4.3.3 Anatomie des mises à jour de statut :

Krishnamurthy et al [23] ont analysé les mises à jour des statuts. 60% des mises à jour étaient envoyées sur le site de Twitter, tandis que l'autre 40% ont été envoyés avec les mobiles et applications personnalisées, concernant le temps, ils découvrent aussi que le montant augmente dans les premières heures du matin, et baisse pendant les heures de nuit, d'un autre coté les utilisateurs qui ont plus d'adeptes affichent plus de mises à jour de statut que les utilisateurs avec moins d'adeptes.

Ringel et al [26] étudiés quels types de questions sont posées dans un réseau social comme Twitter. Ils ont mené une enquête auprès de 624 personnes , plus de 50% ont répondu qu'ils ont utilisé leurs messages pour obtenir des

informations utiles en posant des questions explicites , Selon l'enquête, les participants préféraient les réseaux sociaux plutôt que les moteurs de recherche, car ils ont confiance dans les réponses de leurs amis.

Ils croyaient que les moteurs de recherche tout simplement

ne sont pas en mesure de répondre à leur question, 17% des questions sont des faits (en demandant une réponse objective), tandis que plus de 50% sont questions opiniâtre ou recommandation. En ce qui concerne le sujet, la majorité a demande des conseils techniques (29%), suivi d'un spectacle (17%) et les questions de la famille (12%).

4.3.4 La répartition géographique :

Tseng, et al [27] ont effectué une analyse détaillée de Twitter en 2007,étant l'un des premières recherches dans ce sujet. Ils ont effectué une analyse géographique détaillée. Les résultats montrent que Twitter est surtout utilisé dans les États-Unis (en particulier la côte est), en Europe et en Asie (principalement Japon). Twitter est populaire le plus dans les villes de Tokyo, New York et San Francisco.

4.3.5 Les systèmes de recommandation :

Phelan et al [28] ont construit un système de recommandation appelé **Buzzer** qui propose les infos d'actualité aux utilisateurs en forme de flux RSS, en fonction de leurs préférences, twitter est utilisé pour classer ces news, le système construit un vecteur de termes des flux RSS et des tweets de l'utilisateur, ensuite utilise TF-IDF pour localiser les flux RSS qui correspondent le plus aux Tweets. Les Tweets peuvent être soit Tweets publics (timeline), ou tweets uniquement des amis. L'évaluation a révélé que l'utilisation du « timeline » produit les meilleurs résultats. 67% des utilisateurs préféraient les résultats produits par le (timeline) , tandis que 22% ont indiqué une préférence pour les résultats obtenus en se basant sur les amis , (11% n'ont pas préférer une stratégie).

4.3.6 E-learning :

Ebner et al [29] ont étudié si le principe du micro-blogging comme twitter peut être utile pour l'e-learning, en utilisant les mobiles , Ils ont créé un groupe sur la plateforme de micro-blogging , et demaner aux utilisateurs de joindre ce groupe et de discuter sur le theme e-learning, 23 utilisateurs ont participé , Le principal but de cette enquête est que les utilisateurs utilisent principalement les plates-formes de micro-blogging afin de connecter les uns avec les autres et de partager des nouvelles . Ceci s'applique en particulier sur la communauté scientifique, par exemple pour transmettre en direct des conférences ou d'informer la communauté de nouveaux papiers. Ils voient les plates-formes de micro-blogging comme une bonne alternative aux systèmes classiques de E-learning comme Blackboard [1] ou Moodle [2] vue que moins d'engagement est prévu dans ces platformes.

4.3.7 La retransmission des évènements en direct :

Kennedy et al [30] ont utilisé Twitter pour découvrir la sémantique et la structure des événements médiatiques en direct, Ils ont construit un système appelé **Statler** capable de réticuler un événement médiatique avec des annotations communautaires, La figure suivante présente une capture d'écran du système avec une analyse du débat présidentiel américain en 2008,la vidéo est segmentée en différentes parties, et chaque partie est annoté avec plusieurs balises. Cette liste de tags est généré en puisant dans la timeline Twitter publique tout le média est diffusé.

1 : Un système d'apprentissage virtuel

2 : une plateforme d'apprentissage en ligne

www. moodle.org

Figure 2.1 : Screenshot de l'interface Statler

4.3.8 L'acquisition de connaissances :

Weng et al [31] ont développé une approche pour localiser les utilisateurs importants dans Twitter en calculant un score qu'ils appellent (TwitterRank) (inspirée par PageRank de Google). Ce score prend le terme d'études de réseaux sociaux (homophilie), Dans ce contexte ce terme décrit le phénomène des gens qui sont intéressés par les mêmes sujets sont plus susceptibles de créer des liens les uns avec les autres que d'autres personnes, De cette façon, il est possible d'identifier les utilisateurs qui sont intéressés aux mêmes sujets, Les sujet sont distillés en utilisant une machine apprentissage non supervisée (Latent Dirichlet Allocation). En utilisant cette approche, ils sont capables de générer une liste des meilleurs sujets avec le&a plus d'utilisateurs

Conclusion :

Bien que les tweets soient très échangés sur le web, nous avons constaté qu'il y a peu de travaux qui s'intéressent à la fouille des tweets. Le problème majeur dans ce domaine consiste à déterminer, les informations à extraire à partir des tweets pour servir dans différents domaines. Dans le chapitre suivant nous proposons une nouvelle approche qui rentre dans le domaine du tweet mining pour le but d'extraire les sujets émergents à partir des tweets.

Chapitre III :

CONCEPTION ET IMPLEMENTAION

1. Introduction :

Dans le chapitre précédant, nous avons réalisé une analyse approfondie de L'état de l'art du domaine de fouille des tweets. En prenant comme base cette analyse, nous allons proposer dans ce chapitre la conception générale de notre système et une nouvelle approche de tweet mining. Celle-ci prend en considération le contenu et la structure spéciale des messages tweets dans le but de les classifier par sujet.
Ensuite nous allons définir les outils de développement choisis pour l'implémentation de notre système. Ensuite, nous présentons notre application, les tests et les résultats de l'évaluation.

2. Présentation de la démarche utilisée :

Nous expliquons dans cette étape du chapitre, 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 formes de diagrammes UML.

2.1 Le cycle de vie :

Le cycle de vie d'un logiciel est un ensemble séquentiel de phases, dont le nom et le nombre sont déterminés en fonction des besoins du projet, permettant généralement le développement d'un service ou d'un produit, en ce qui concerne notre projet nous avons suivi le modèle en cascade. [32]

3. Modèle en cascade :

Ce modèle est constitué d'une suite d'étapes qui ont pour but de réaliser un produit logiciel fini et testé. Le résultat de chaque étape est testé et on ne passe à l'étape suivante que lorsque l'étape actuelle est satisfaisante. Ce modèle est un cycle de vie linéaire, séquentiel, défini dans les années 70.

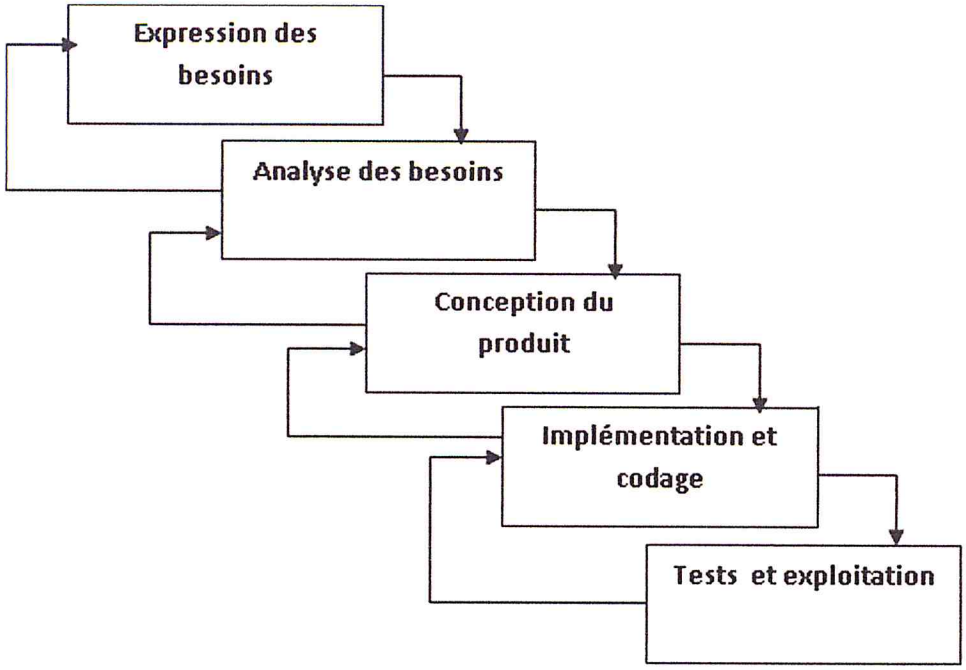

Figure 3.1 : le modèle en cascade.

3.1. Expression des besoins :

La spécification des besoins est une étape essentielle au début de processus de développement, elle consiste généralement à déterminer précisément les besoins des utilisateurs du système afin d'éviter de développer un logiciel non adéquat.

Cette étape ne préoccupe pas des solutions mais des questions : elle identifie le « quoi faire ?»Et identifie les entités de l'environnement du système. Pour modéliser ces besoins on utilise le diagramme des cas d'utilisation d'UML. [33]

3.2. Conception :

C'est la phase la plus importante du processus de développement d'un logiciel. Elle s'intéresse d'abord au «comment ? », à savoir la solution du problème énoncé.

La conception a pour but de décomposer le logiciel en module, de préciser les interfaces et les fonctions de chaque module. A l'issue de cette étape, on obtient une description de l'architecture du logiciel et un ensemble de spécifications de ces divers composants en utilisant le diagramme de classe d'UML. [33]

4. Expression des besoins :

Cette phase consiste à définir les besoins fonctionnels de notre futur système, nous allons parler des fonctionnalités que peut offrir ce dernier, par la suite nous allons les modéliser en utilisant le diagramme des cas d'utilisation d'UML. [32]

4.1 Identification des cas d'utilisation :

- Les cas d'utilisations :

Un cas d'utilisation est une unité cohérente représentant une fonctionnalité visible de l'extérieur. Il réalise un service de bout en bout, avec un déclenchement, un déroulement et une fin, pour l'acteur qui l'initie. Un cas d'utilisation modélise donc un service rendu par le système, sans imposer le mode de réalisation de ce service.

Nous allons présenter notre diagramme de cas d'utilisation global, suivi par chaque cas d'utilisation détaillé.

4.2 Diagramme de cas d'utilisation :

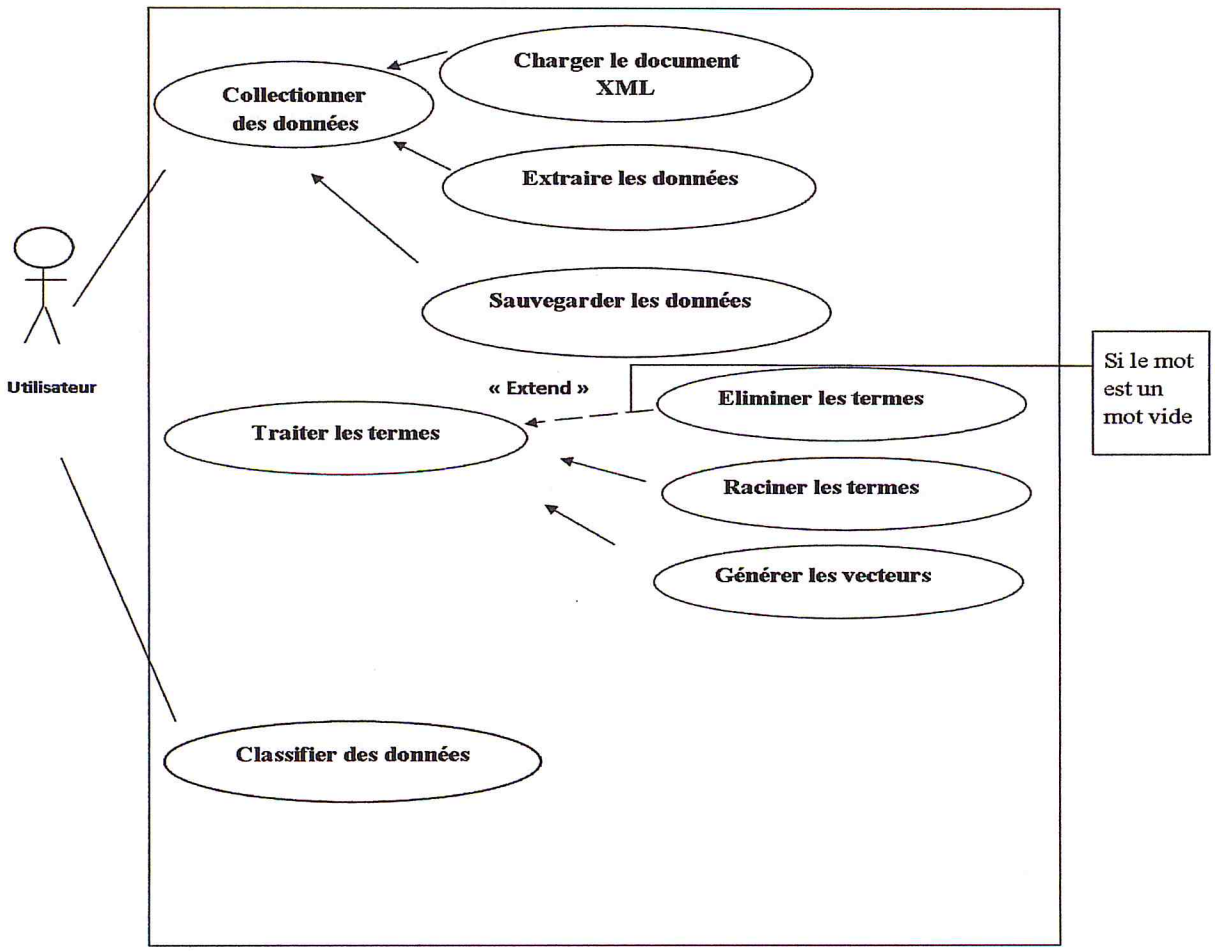

Figure 3.1 : Diagramme de cas d'utilisation global.

5. Conception :

5.1 Diagramme de Classes :

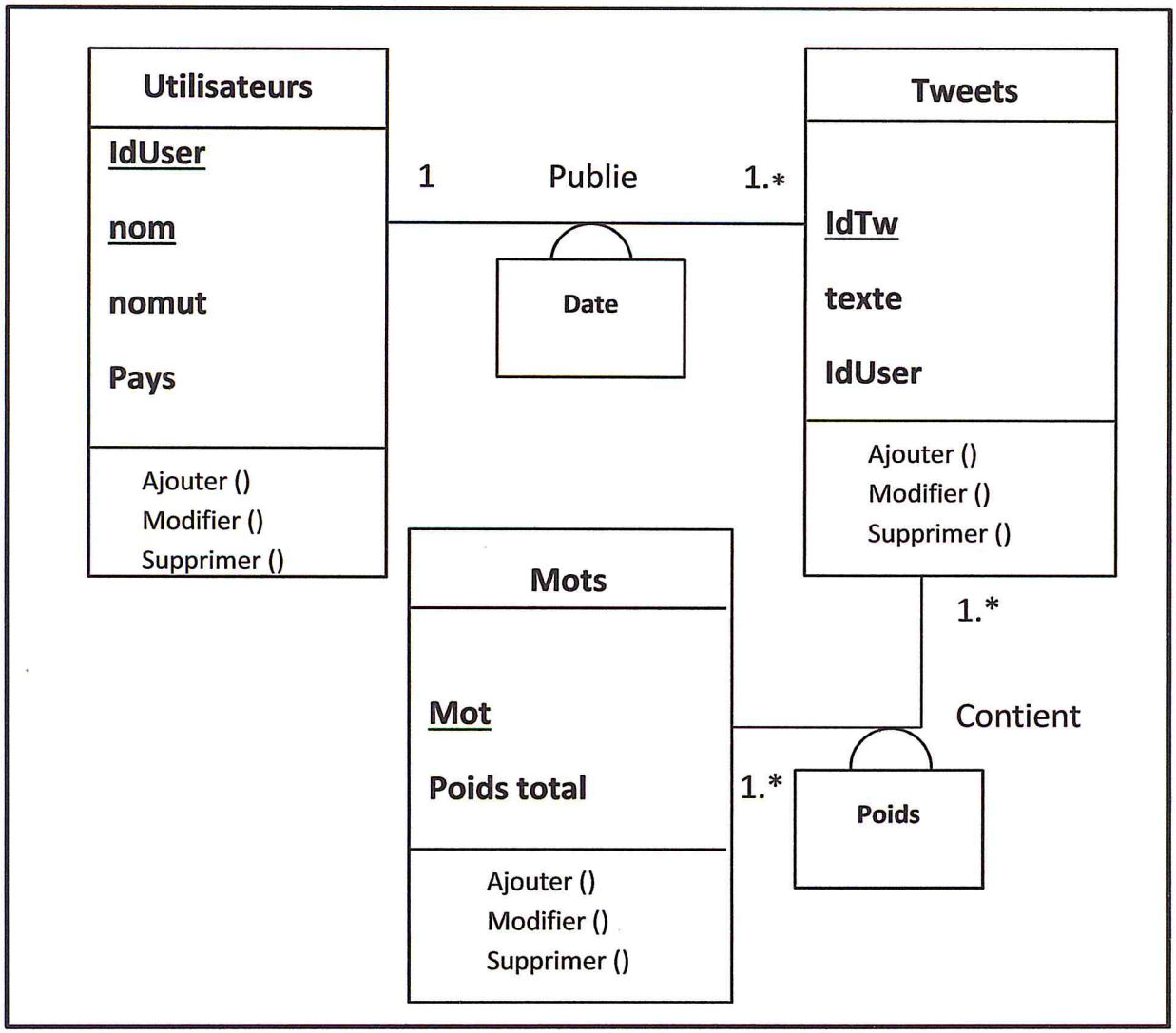

Figure 3.2 : Diagramme de classe

- Chaque tweet est identifié par son IdTw et représenté par son contenu.
- L'utilisateur est identifié par son IdUser et représenté son nom, son nom de profil sur twitter, et son pays.
- Un mot est identifié par son contenu (le mot lui-même) est représenté par, son poids total dans la collection.
- Chaque utilisateur publie un ou plusieurs tweets , un tweet publie dans une date donnée appartient à un et plusieurs utilisateurs.

- Un tweet contient un ou plusieurs mots, un mot appartient à un ou plusieurs tweets , chaque mot est représenté par son poids de son tweet.

Entité	Description	Type
Utilisateurs		
IdUser	Identifiant de l'utilisateur twitter	Numérique
nom	Nom de l'utilisateur	alphabétique
nomut	Nom de l'utilisateur twitter	alphanumérique
Pays	Pays de l'utilisateur twitter	alphabétique
Tweets		
IdTw	Code identifiant de tweet	Numérique
texte	Contenu textuel du tweet	alphabétique
IdUser	Identifiant de l'utilisateur twitter	Numérique
Mots		
IdMot	Code identifiant du mot	Numérique
Mot	Le mot lui-même	alphabétique
Poids total	Poids total du mot dans la collection	Numérique
Publie		
Date	La date de publication de tweet	Date
Contient		
Poids	Poids du mot dans son tweet	Numérique

Tableau 3.1 : description du diagramme de classe

6. L'Architecture globale de l'application:

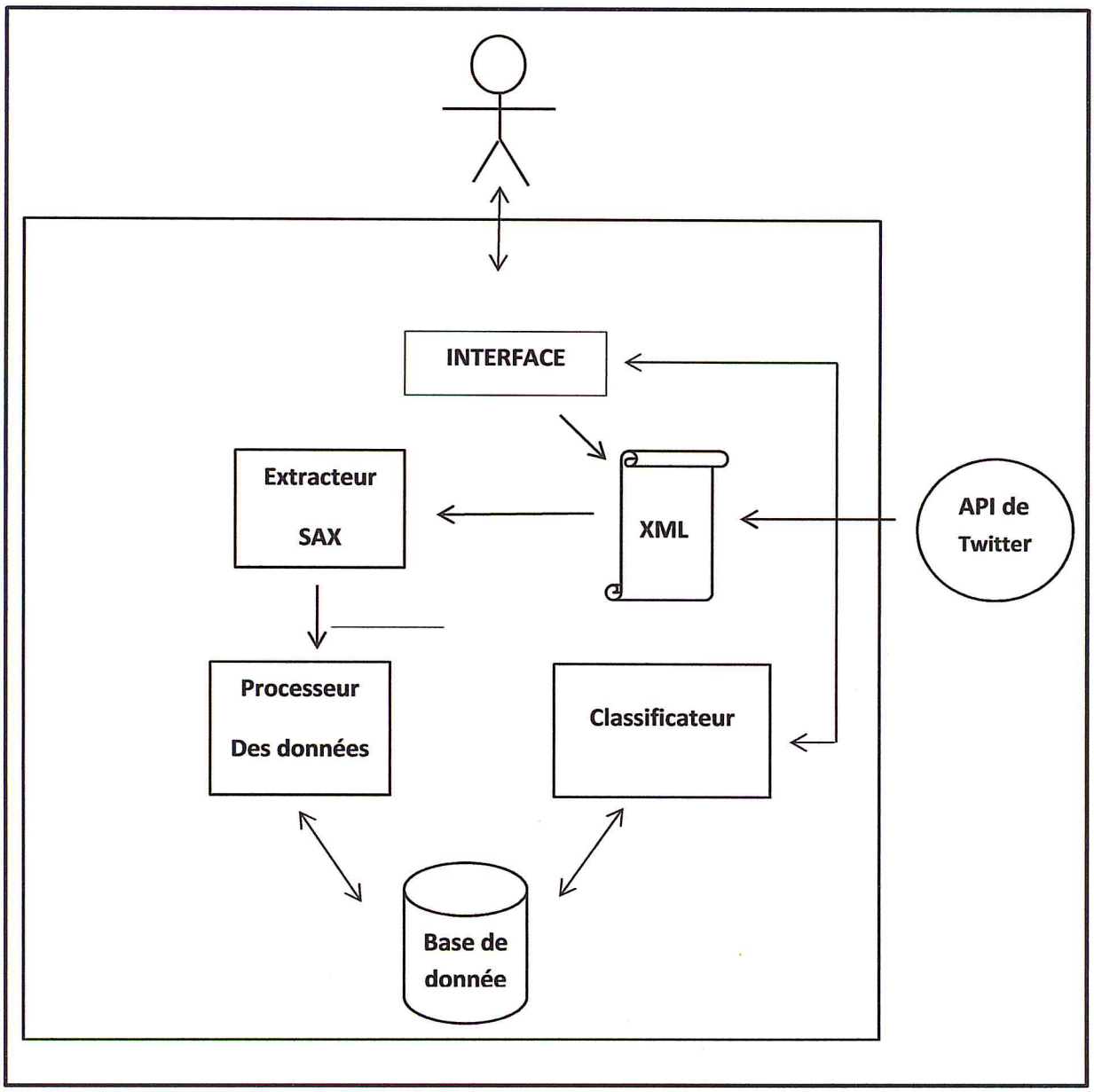

Figure 3.3: Architecture globale de l'application

La figure 3 .3 visualise comment l'application fonctionne en interne. L'utilisateur communique avec l'ouverture du fichier XML de l'origine de API Twitter , ce ficher est transmet au programme SAX afin d'extraire les données , en suite l'ensemble des données subit des traitements avant de se charger dans la base de données .une fois les paramètres de classification trouvées , le classificateur termine le système.

7. L'approche de fouille dans les messages tweets :

Notre approche consiste à extraire les messages tweets des documents Xml afin d'effectuer un Clustering (Classification) de leurs contenu. Les tweets sont représentés par des ensembles de vecteurs. La méthode TF-IDF est utilisée a fin de générer les collections modèles des déférentes sujets trouvés. Ces collections modèles des sujets sont ensuite utilisés pour construire les vecteurs qui représentent les tweets dans l'étape de classification.

Donc globalement, l'approche est constituée de trois grandes phases dont chacune est composée de différentes étapes. La figure 3.3 présente ces trois phases qui sont :

- La Collection des données.
- L'extraction des éléments de classification.
- Le clustering.

Les trois phases se résument comme suit :

Phase 1 : La Collection des données :

1. Extraction des données à partir des documents XML.
2. Prétraitement des données.
3. Construction du dictionnaire.

Phase 2 : L'extraction des éléments de classification :

4. Calcul du poids avec TF-IDF
5. Extraction des sujets.
6. Construction des collections modèles.
7. Construction des vecteurs / tweets.

Phase 3 : Le clustering :

8. Application de l'algorithme de clustering K-means

Dans ce qui suit, nous allons détailler les différentes phases et étapes Constituant notre approche :

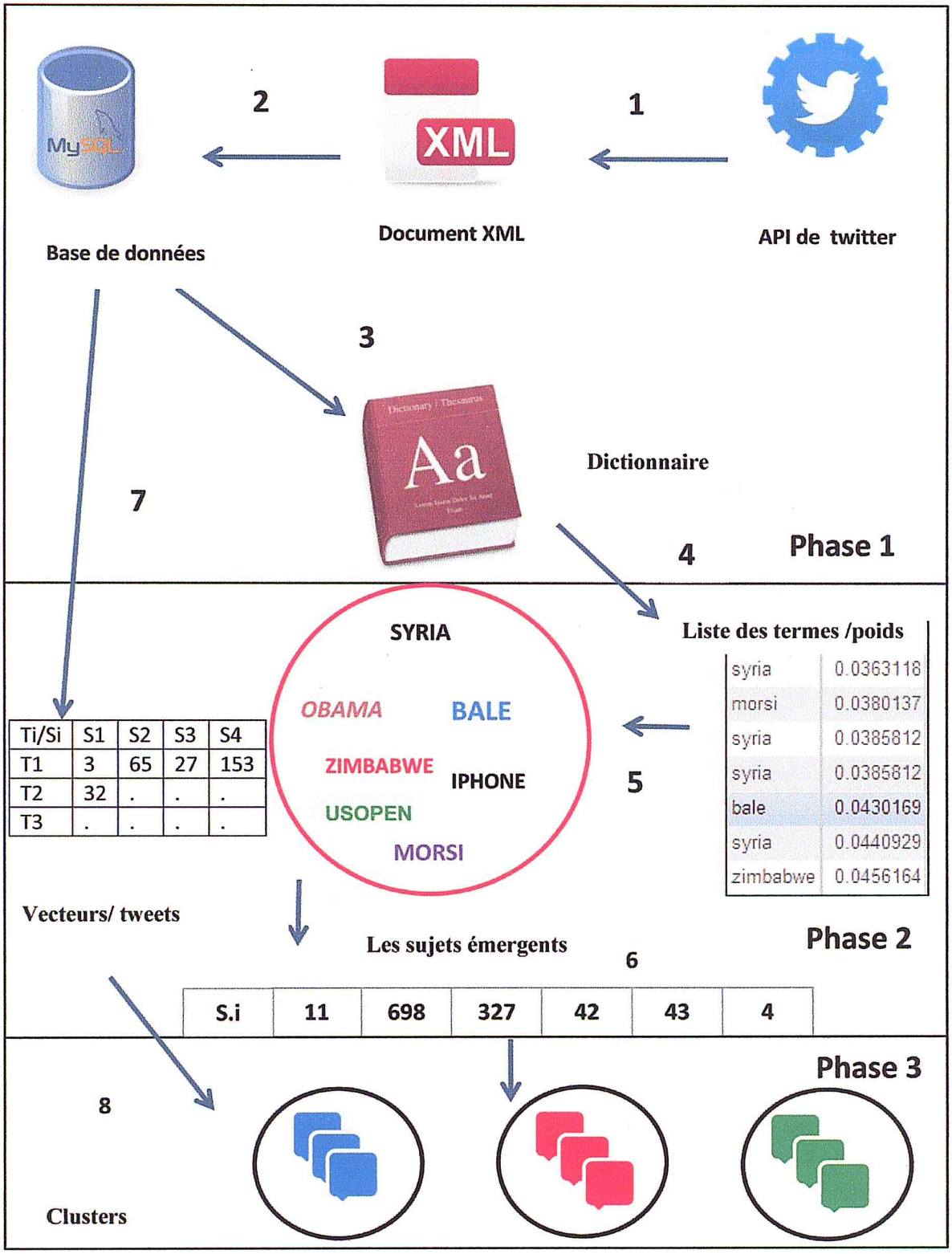

Figure 3.4 : Schéma global de l'approche

7.1 La Collection des données :

Cette phase consiste à collectionner l'ensemble des données à traiter dans notre approche. Ces données sont le résultat des requêtes HTTP envoyées à API de twitter .Ces requêtes permet de récupérer les tweets récents d'un utilisateur donnée en forme XML.

7.1.1 Extraction des données à partir des documents XML :

Cette étape consiste à extraire l'ensemble des données importantes pour notre approche, les documents XML comporte plusieurs informations sur un tweet, les données à extraire sont : le contenu de tweet (texte), date de publication, localisation (Pays). Ces données sont stockées dans la base de données pour les exploiter ensuite. Afin d'extraire ces données, nous utilisons la méthode SAX (Simple API XML), c'est donc une API qui permet de lire des flux (donc des fichiers) XML.

SAX se charge de lire le flux XML et à chaque fois qu'il rencontre un élément particulier, il appelle une méthode correspondante. Par exemple à chaque nouvelle balise XML qu'il rencontre il va appeler la méthode "startElement"[34].

StartElement	Appelé lorsque le parser rencontre une nouvelle balise XML	les balises d'ouverture <maBalise>
endElement	Appelé lorsque le parser rencontre une balise fermante XML	les balises de fermeture </maBalise>

Tableau 3.2 : Les méthodes SAX

7.1.2 Prétraitement des données :

Cette étape consiste à préparer l'ensemble des tweets pour la classification à fin de gagner le temps et l'espace de stockage de notre système.

, le prétraitement se compose de toutes les tâches de traitement automatique de langue qui ont lieu avant la transformation des tweets en vecteurs, ces tâches sont :

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

Un mot vide est un mot non significatif. 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 des déffenrets collection des mots vides de la langue anglaise sur le web, nous avons choisi la collection (Full-Text Stopwords) Cette liste est chargée directement du SGBD (Système de Gestion de Base de Données) MySQl[35] .

- **Racinisation (stemming) :**

La racinisation ou désuffixation (stemming) est un procédé de transformation des mots en leur radical ou racine .La racine d'un mot correspond à la partie du mot restante une fois que l'on a supprimé son préfixe et son suffixe, à savoir son radical, la racine ne correspond généralement pas à un mot réel. Par exemple, le mot « chercher » a pour radical « cherch » qui ne correspond pas à un mot réel. Par contre dans l'exemple de « frontal », le radical est « front ». Pour établir le stemming sur notre collection de tweets nous avons choisi l'algorithme PORTER. L'algorithme PORTER se compose d'une cinquantaine de règles de désuffixation classées en sept phases successives (traitement des pluriels et verbes à la troisième personne du singulier, traitement du passé et du progressif,...). Les mots à analyser passent par tous les stades et, dans le cas où plusieurs règles pourraient leur être appliquées, c'est toujours celle comprenant le suffixe le plus long qui est choisie. La désuffixation est accompagnée, dans la même étape, de règles de recodage. Ainsi, par exemple, "troubling" deviendra "troubl" par enlèvement du suffixe marqueur du progressif -ing et sera ensuite transformé en "trouble" par application de la règle "bl" devient "ble". Cet algorithme comprend aussi cinq règles de contexte, qui indiquent les conditions dans lesquelles un suffixe devra être supprimé. La terminaison en -ing, par exemple, ne sera enlevée que si le radical comporte au moins une voyelle. De cette manière, "troubling" deviendra "troubl", nous l'avons vu, alors que "sing" restera "sing".

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

Soit v représente une voyelle (y est considéré comme une voyelle s'il est précédé par une consonne), c représente une consonne; et soit V représente une suite de voyelles, C représente une suite de consonnes.

Alors 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] , Où :[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 désuffixation 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

- $*c$: le préfixe se termine par la lettre c
- $*v*$: 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, non

- **<u>Les étapes de l'algorithme Porter :</u>**

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 …
Étape 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 3.3: les déférentes Étapes et Racines obtenues de l'algorithme Porter[36]

7.1.3 Construction du dictionnaire :

Cette étape consiste à collectionner l'ensemble des termes traités dans les étapes précédentes dans un dictionnaire globale ou chaque mot possède un identifiant unique, ces identifiants sont ensuite utilisés dans la représentation vectorielle des tweets.

7.2 L'extraction des éléments de classification :

Le but de phase est de générer les éléments de classification, les collections de tendances ou chaque collection représente un sujet émergent et les vecteurs / tweets à fin d'utiliser ces vecteurs dans la phase de classification.

7.2.1 Calcul du poids avec TF-IDF (Term Frequency-Inverse Document Frequency) :

Cette étape consiste à calculer le poids de chaque mot ou terme dans son tweet en utilisant la méthode TF-IDF, ensuite calculer le poids total de ce terme dans l'ensemble des tweets a fin de différencier ces termes dans leurs tweets et dans toute la collection. Ce poids permet d'évaluer l'importance d'un terme contenu dans un tweet, relativement à une collection. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document. Il varie également en fonction de la fréquence du mot dans la collection.

- TF-IDF (Term Frequency-Inverse Document Frequency) :

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 :

$$\text{tfidf}_{i,j} = \text{tf}_{i,j} \cdot \text{idf}_i$$

$$\text{Tf} = \frac{\text{Nbr de réptition}}{\text{nbr de terme de le document}}$$

$$\text{idf}_i = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

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 $n_{i,j} \neq 0$

Le score est le plus élevé lorsque le terme apparaît souvent dans un petit sous-ensemble des documents, et sera plus bas lorsqu'il apparait plusieurs fois dans d'autres documents. TF-IDF est largement utilisé pour comparer la similarité entre les documents, fournissant une liste triée des documents les plus pertinents.

Les Tweets sont des messages courts (limité à 140 caractères), alors la fréquence (TF) d'un terme est généralement 1. Cela signifie que la fréquence inverse (IDF) prend plus d'importance, mais la pondération globale perd une puissance due à l'absence de richesses locale. Pour éviter ce problème, on a décidé d'augmenter le poids des termes qui sont précédés par le dièse #, par exemple : #Obama.

7.2.2 Extraction des sujets :

Cette étape consiste à utiliser le calcul de poids pour déterminer les sujets émergents, nous organisons les termes par ordre décroissant du poids, les termes avec les poids élevés ont plus de chance d'être un sujet car les sujets sont des termes qui reviennent souvent dans les tweets mais plus d'une fois dans le même tweet. Pour notre approche nous considérons les sujets comme 10% des termes avec les poids élevés qui compose la collection. Si on n'arrive pas à classifier tous les tweets à la fin de l'approche nous pouvons ajouter sujet par sujet jusqu'à la classification de toute la collection.

7.2.3 Construction des collections modèles:

Cette étape consiste à représenter chaque sujet par une collection qui regroupe l'ensemble des termes qui compose les tweets de ce sujet, par exemple pour le sujet (SYRIA) .L'ensemble de termes est : BACHAR, ONU, WAR, REFUGEES.....etc. Ces vecteurs sont ensuite utilisés comme référence dans la phase suivante.

7.2.4 Construction des vecteurs / tweets :

Cette étape consiste à représenter chaque tweet par un vecteur. La taille du vecteur est déterminer par le nombre du sujet ou tendances, chaque case représente le nombre total des termes d'un tweet donné présent dans une collection d'un sujet donné, par exemple :

Si la deuxième case d'un vecteur est 7 alors le tweet concerné contient 7 termes de la collection du deuxième sujet émergent.

7.3 La classification (clustering) :

La dernière phase consiste à classifier l'ensemble des vecteurs, chaque vecteur représente un tweet, la taille du vecteur est déterminé par le nombre de sujets émergents trouvées, chaque valeur représente le nombre de terme d'un sujet dans le tweet, par exemple un vecteur avec les valeurs : 6, 5, 2, 0, 1, ect.....Il existe 6 terme dans le tweet concerné qui appartient à la collection du premier sujet émergent, et 5 terme du deuxième sujet ainsi de suite. Ensuite on procède à la classification en utilisent l'algorithme de classification K means.

L'algorithme k-means est utilisé avec *k* un paramètre défini arbitrairement en entrée qui indique le nombre de clusters à construire. La distance euclidienne est Utilisée pour mesurer la distance entre paires de documents où *n* est le nombre des chemins. Elle est calculée par la formule suivante :

Les étapes de *k-means* sont les suivantes [37] :

1. Choisir aléatoirement *k* documents qui formeront l'ensemble des centroïdes Initiaux représentant les *k* clusters à construire.

2. Assigner chaque document au cluster dont le centroïde le plus proche selon la distance *d* (si un minimum de *d* est trouvé entre deux objets).

3. Si aucun document ne change de cluster d'une itération à l'autre alors arrêt et sortir les clusters. Sinon, mettre à jour les centroïdes des clusters en fonction des objets qui leur sont associés.

4. Aller à 2.

8. Conception détaillée et implémentation :

8.1 Technologie (Outil de développement) :

8.1.1 Java :

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

Java est un langage de programmation et une plate-forme informatique créée par Sun Microsystems en 1995. Il s'agit de la technologie sous-jacente qui permet l'exécution de programmes dernier cri, notamment des utilitaires, des jeux et des applications professionnelles. Java est utilisé sur plus de 850 millions d'ordinateurs de bureau et un milliard de périphériques dans le monde, dont des périphériques mobiles et des systèmes de diffusion télévisuelle [38].

Nous avons décidé de développer 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 sur Internet via des URL avec la même facilité qu'elles accèdent à un fichier local sur le système ce qui nous permettent ce qui nous permettent dans notre recherche d'accéder facilement aux API de Twitter des tweets à travers le protocole HTTP.
- Fiabilité :
  Java a été conçu 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 conçu pour être exploité dans des environnements serveur 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 bytecode directement sur n'importe quelle machine sur laquelle il a été porté.

8.1.2 Eclipse :

Eclipse est un éditeur, décliné et organisé en un ensemble de sous-projets de développements logiciels, de la Fondation Eclipse visant à développer un environnement de production de logiciels libres qui soit extensible, universel et polyvalent, en s'appuyant principalement sur Java.
Son objectif est de produire et fournir des outils pour la réalisation de logiciels, englobant les activités de programmation (notamment environnement de développement intégré et frameworks) mais aussi d'ATL recouvrant modélisation, conception, gestion de configuration [39].

8.1.3 MySQL :

MySQL est un système de gestion de base de données (SGBD). Il est distribué sous une double licence GPL. Il fait partie des logiciels de gestion de base de données les plus utilisés au monde, autant par le grand public (applications web principalement) que par des professionnels, en concurrence avec Oracle, Informix et Microsoft SQL Server [40].

9. Présentation de l'application « Whatup » :

Figure 3.5. Interface d'accueil de WhatUp.

Après l'exécution de l'application Whatup, une interface simple est affichée, Le menu se compose de trois boutons « File» ouvre le menu indique à l'utilisateur qu'il doit importer le fichier contenant l'ensemble de données à analyser, « Edit» éditer un fichier XML avant de l'analyser et « ?» Le bouton HELP pour aider les utilisateurs à comprendre les étapes du système.

9.1 Interface d'importation du fichier XML :

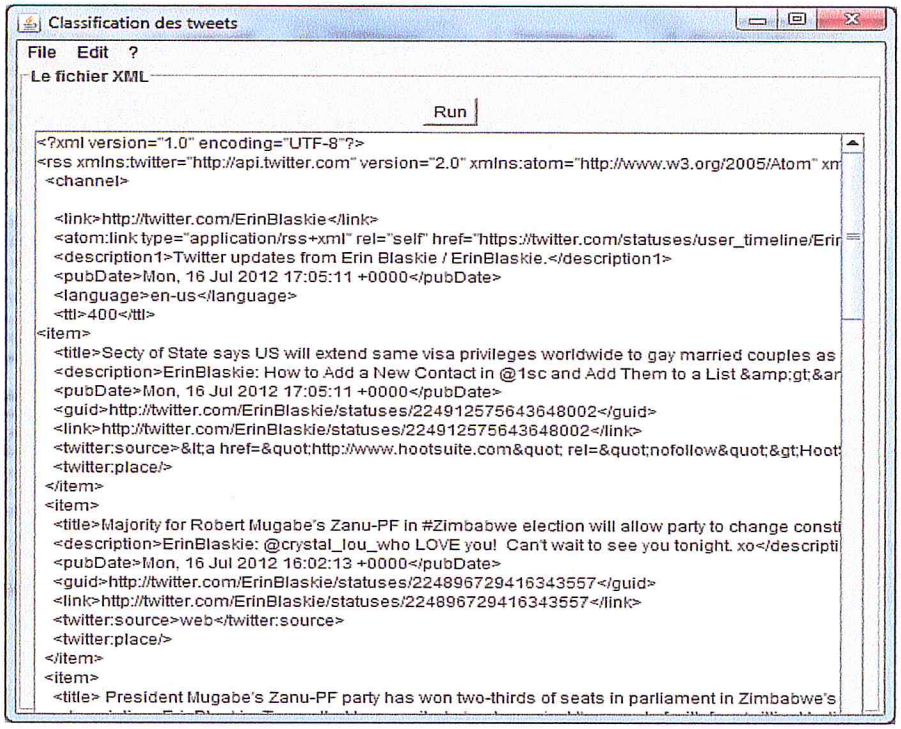

Figure 3.6. Interface de l'importation de fichier

A ce niveau, l'utilisateur visualise le type de fichier XML (1), le button « RUN » permet de lancer l'opération d'extraction et traitement du contenu de ce ficher a d'ajouter ces données à la collection des données déjà présente.

9.2 Calcul du poids :

Une fois le traitement de données on procède à l'étape du calcul du poids de chaque termes dans la collection en utilisent la méthode de calcul TF-IDF pour permettre de distinguer les termes les plus lourd ont plus de chance d'être un sujet émergent ou une tendance sur twitter,

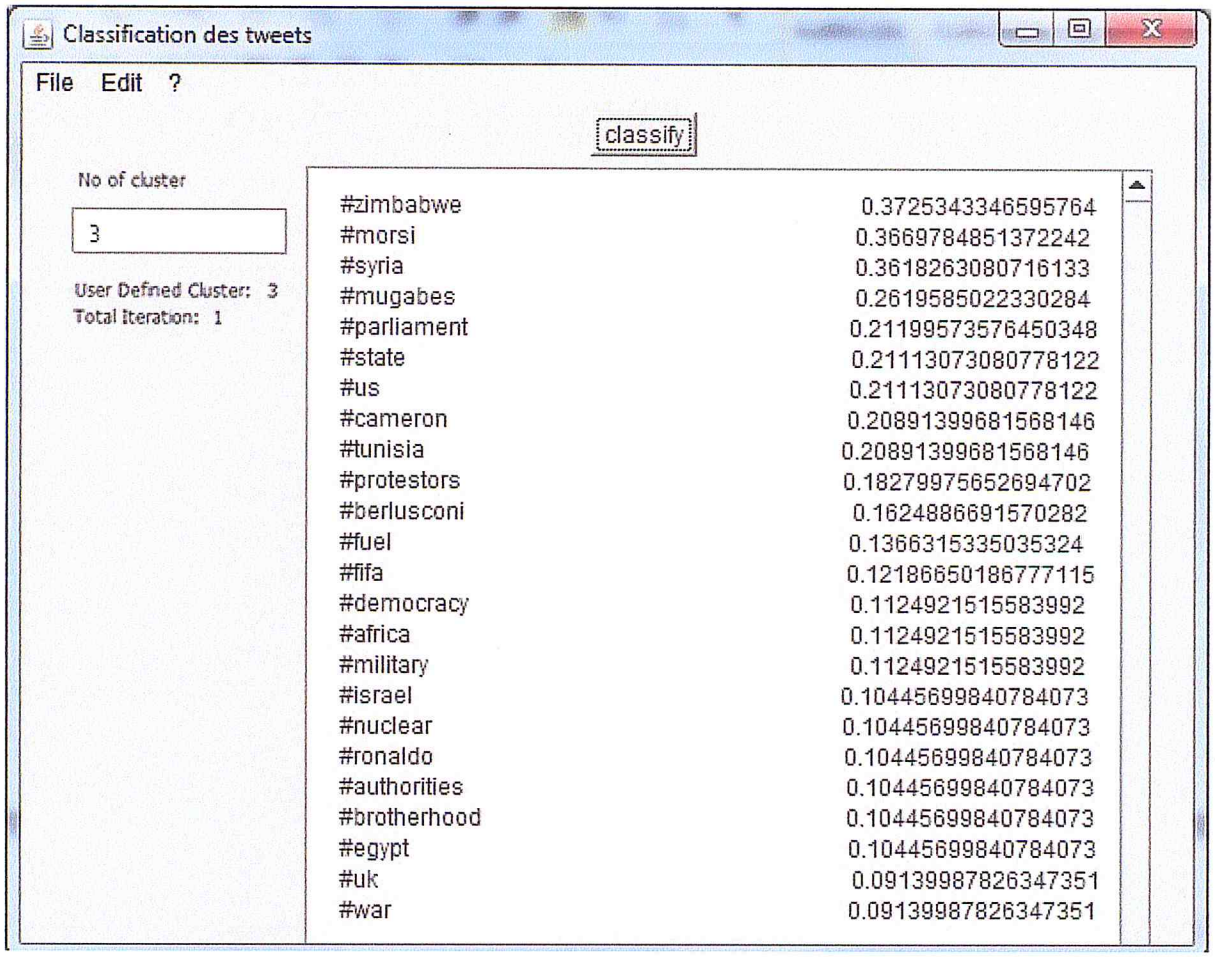

Figure 3.7 : Interface des termes les plus lourds

9.3 La classification :

Voici les résultat du classification sur la collection de tweets choisie, en utilsent l'algorithme de K means avec une itération, avec la distance euclidienne pour mesurer les distances entre les tweets . Chaque centre initial représente un sujet émergent dans le cas de cette collection 3 centres initiaux.

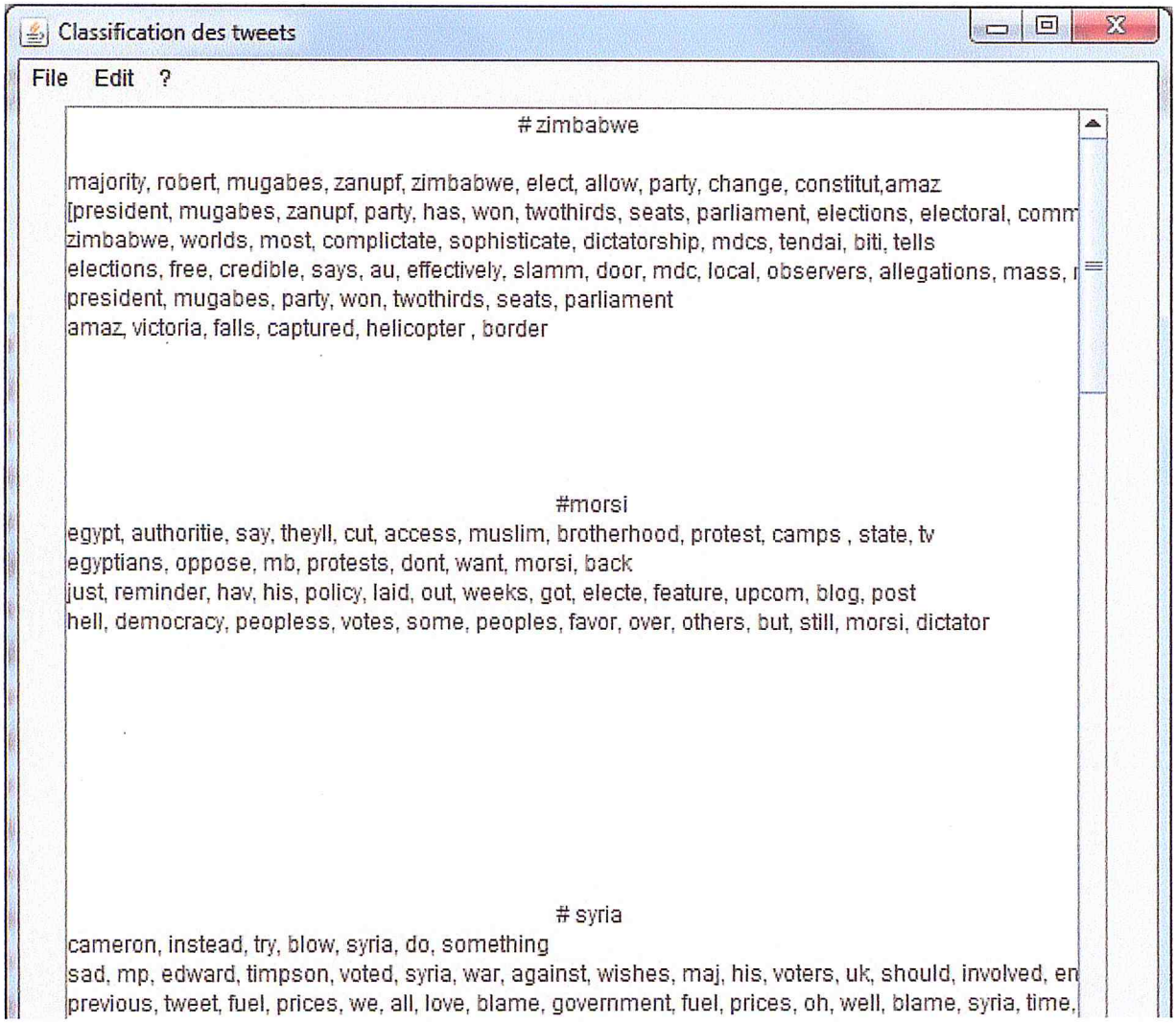

Figure 3.8 : Screenshot des clusters

Conclusion :

Dans ce chapitre nous avons proposé les différentes phases constituant notre nouvelle approche de fouille dans les tweets et plus précisément la en prenant en considération le contenu textuel. Notre approche nécessite une phase de validation sur des corpus de tweets. Pour ce faire. Dans le chapitre suivant, nous allons présenter les tests de cette approche.

Conclusion :

Dans cette étude, nous avons proposé un système d'analyse et de classification des messages tweets, en se basant sur l'algorithme K-means et la méthode de pondération TF-IDF différencier les termes.

Notre contribution dans les objectifs du travail est que par sa capacité à extraire les sujets émergents d'une collection tweets. Le problème majeur dans notre travail était de trouver les bonnes collections de tweets afin de répondre aux besoins de notre approche. Apres la recherche des collections le défi était de trouver une représentation numérique aux tweets , une représentation qui prend en considération la particularité du contenu des messages tweets et surtout l'absence du poids local .

Il reste un énorme travail a accomplir dans le domaine du tweet mining, le plus difficile serait de suivre le rythme infernal imposé par les utilisateurs de twitter, qui font introduire chaque jour des nouveaux termes et phrases et de nouvelle manière d'exprimer.

Bibliographie

[1] Ford, R. (2011) . Earthquake: Twitter Users Learned of Tremors Seconds Before Feeling Them . The Hollywood reporter [en ligne], (page consultée le 17/02/2013)
URL : http://www.hollywoodreporter.com/news/earthquake-twitter-users-learned-tremors-226481

[2] Asur , S et Huberman Bernardo A(2010). Predicting the future with social media[en ligne] . (page consultée le 12/01/2013).
URL : http://arxiv.org/pdf/1003.5699.pdf

[3] N.A. Diakopoulos et D.A. Shamma. (2010). Characterizing debate performance via aggregated twitter sentiment. (page consultée le 02/02/2013).
URL : http://www.research.yahoo.com/pub/3090

[4] Statisticbrain .Twitter Statistics : Statistic Verification.Twitter Research. Huffington Post. [en ligne] .(page consultée le 05/12/2012).

URL : http://www.statisticbrain.com/twitter-statistics/

[5] O'Reilly, T et Milstein, S. (2011). The Twitter Book. Sebastopol :O'Reilly Media . 248 pages.

[6] MacArthur, A . (2011)The Real History of Twitter . [en ligne], (page consultée le 17/03/2013)
URL : http://twitter.about.com/od/Twitter-Basics/a/The-Real-History-Of-Twitter-In-Brief.htm

[7] Petit Larousse illustré.(2012). les «twitteurs» et les «twitteuses».Paris : Edition Larousse. 1910 pages

[8] (2011).La véritable histoire de Twitter . [en ligne], (page consultée le 09/02/2013)

URL : http://www.atlantico.fr/pepites/twitter-williams-trahison-investisseurs-start-78046.html

[9] Gervai , A . (2011).Twitter Statistics – Updated stats for 2011 ., [en ligne], (page consultée le 08/02/2013)

URL : http://www.marketinggum.com/twitter-statistics-2011-updated-stats/

[10] Timothy , N.(2009). Does Twitter have the ability to generate revenue? [en ligne], (page consultée le 01/02/2013)

URL : http://www.examiner.com/article/does-twitter-have-the-ability-to-generate-revenue

[11] Bradshaw, T et Gelles, D .(2009) Twitter now worth $7.14m per character. Financial Times. [en ligne], (page consultée le 29/01/2013)

URL : http://www.ft.com/cms/s/0/641057ae-a933-11de-9b7f-00144feabdc0.html.

[12] ROSTAGNAT ,M .(2011) TWITTER PEAUFINE SON MODÈLE ÉCONOMIQUE. [en ligne], (page consultée le10/02/2013)

URL : http://www.ozap.com/actu/twitter-peaufine-modele-economique/425580

[13] Lev Grossman ,(2009). Iran protests: Twitter, the medium ofthe movement. Website . Time Magazine. [en ligne], (page consultée le 20/01/2013)

http://www.time.com/time/world/article/0,8599,1905125,00.html.

[14] Statut de . MirHossein Mousavi .Date de publication : 13 Juin 2009
URL : https://twitter.com/mousavi1388/status/2159159988

[15] Pepitone ,J. (2010). Twitter mobilizes haiti aid efforts . [en ligne], (page consultée le 13/12/2012)

[16] ALGORITHMES DE CLASSIFICATION, Maurice ROUX

Professeur émérite Université Paul Cézanne Marseille, France. Éditions Masson, Paris, en 1985 .Version électronique consultée le : 16/05/2013

URL : http://www.imep-cnrs.com/docu/mroux/algoclas.pdf

[17] Cheong,M .et Lee ,V. (2009)Integrating web-based intelligence retrievaland decision-making from the twitter trends knowledge base. In

SWSM '09: Proceeding of the 2nd ACM workshop on Social web search

and mining, pages 1/8, New York, Etat unis,. ACM.

[18] Naaman , M. Boase , (2010) J et Chih-Hui, L. Is it really about me?:message content in social awareness streams. In CSCW 10: Proceedingsof the 2010 ACM conference on Computer supported cooperative work,

pages 189/192, New York, NY, USA, 2010. ACM.

[19] Culotta , A. (2010)Towards detecting influenza epidemics by analyzing twitter messages. In KDD Workshop on Social Media Analytics.

[20] SAKAKI T., OKAZAKI M., MATSUON Y., Earthquake shakes Twitter users: real-time event detection by social sensors, Proceedings of the 19th international conference on World wide web (WWW), p.851–860, New York, NY, USA, 2010.

[21] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment classi_cation using machine learning techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79{86, 2002.

[22] Jaap Kamps, , Jaap Kamps, and Maarten Marx. Words with attitude.

In In 1st International WordNet Conference, pages 332{341, 2002.Language & Inference Technology Group, University of Amsterdam

[23] Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. A few chirps about twitter. In WOSP '08: Proceedings of the _rst workshop on Online social networks, pages 19{24, New York, NY, USA, 2008. ACM.

[24] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pages 56/65, New York, NY, USA, 2007. ACM.

[25] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. CoRR, abs/0812.1045, 2008.

[26] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. What do people ask their social networks, and why?: a survey study of status message q&a behavior. In CHI '10: Proceedings of the 28th international conference on Human factors in computing systems, pages 1739{1748, New York, NY, USA, 2010. ACM.

[27] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pages 56{65, New York, NY, USA, 2007. ACM.

URL : http://www.java.com/fr

[39] Site officiel d'Eclipse, «Définition d'Eclipse» [En ligne] (Page Consulter le : 18/02/2013).

URL : http://www.eclipse.org/

[39] [38] Site officiel de MySQL, «Définition de MySQL» [En ligne] (Page Consulter le : 19/04/2013). http://www.mysql.com/.