الجمهورية الجزائرية الديمقراطية الشعبية

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

وزارة التعليم العالي والبحث العلمي

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

Université Dr. TaharMoulay SAIDA — جامعة د الطاهر مولاي سعيدة

Faculté : Technologie — كلية التكنولوجيا

Département : Informatique — قسم : الإعلام الآلي

MEMOIRE DE Master

Option : Sécurité Informatique et Cryptographie

THEME

Analyse de Contexte des Personnes

Sur

Les Réseaux Sociaux : *Analyse des sentiments sur Twitter*

Présenté par :

Djellouli Mohamed

TABTI Abdelkrim

Encadré par :

D. Mm Dib.S

Promotion : 2017 - 2018

Remerciement

Avant tout on tient à remercier le bon dieu qui nous à réunis et qui à offerts cette chance de poursuivre nos étude et grâce à lui nous avons pu accomplir ce modeste travail.

Nous tenons à exprimer nos sincères gratuitement et remerciement à nos professeurs qui nous a apportés de l'aide et qui ont contribué à l'élaboration de ce rapport de stage :

Tout d'abord :

Mm.Dib Soumia *nos encadreur pour avoir bien voulu accorder notre travail de recherche et dont ses conseils et ses orientations nous ont énormément aidés. Les mots nous manquent pour elle exprimer nos profonds remerciements.*

Nos enseignants pour l'aide et le temps qu'ils ont bien voulus nous consacrer durant cette formation.

Nous amis et spécialement « ……. » Qui nous ont énormément aidés.

En fin à tous ceux qui nous ont encouragés de près ou de loin à l'élaboration de ce travail de recherche.

Djellouli Mohammed.
TABTI Abdelkrim.

Table des Matière

Table des figures

Chapitre IV : Implémentation Résultats et discussion….....….....61

Liste des Tableaux

Résumé :

Les réseaux sociaux ne cessent de connaître un succès exponentiel auprès des internautes. Ils sont considérés comme des plateformes d'échanges et d'interactions entre un ensemble de personnes, de groupes ou des entités sociales.

Chaque jour on trouve plus 2,5 milliards de mise à jour Facebook, 400 millions de tweets. Une immense quantité de donnée, qui décrit le contexte de ces internautes (contexte professionnel tel que le poste de travail actuel et les expériences précédentes, ou bien un contexte général tel que les centres d'intérêts, le profil psychologique ou bien les personnes qu'ils l'entourent)

Plusieurs domaines sur le Web peuvent bénéficier de ce contexte des personnes, tel que la recommandation des produits qui peuvent intéresser une personne dans les sites de e- commerce en se basant sur le contexte de la personne, ou bien la génération de prospects dans les systèmes e-marketing, etc...

Ces informations sur le contexte des personnes sont éparpillées sur plusieurs postes sur les réseaux sociaux d'une part. D'autre part ces informations sont non-structure et difficile à extraire et exploiter. Pour cela, le présent projet a pour objectif de concevoir un système intelligent qui permet de récolter les informations sur une personne sur les réseaux sociaux, puis d'extraire et de structurer dans un graphe extensible le contexte professionnel et général d'une personne à des personnes dépressif ou non.

Le système doit être basé sur les algorithmes de machine Learning qui permet l'évolution continue et l'amélioration des résultats avec le temps.

Mots Clés : Réseau social, Méthodes d'analyse, Système d'information, Fouille des données, Modélisation des données, Graphe contextuel, Profilage, Graphe social, Graphe d'intérêt, Algorithmes de machine Learning, Big Data, Architecture Rest, Base de données graphe.

Abstract:

Social networks know an exponential success With of internet users. They are considered platforms of exchanges and interactions between a set of people, groups or social entities. Every day there are more than 2.5 billion Facebook updates, 400 million tweets. A huge amount of data that describes the context of these users (professional context such as the current workstation and previous experience, or a general context such as hobbies, the most interesting topic for a user).

Several areas on the Web can benefit from this context of people, such as the recommendation of products that may interest someone in e-commerce site, based on the context Of the person, or lead generation systems in e-marketing, etc.

This information on the context of people are scattered on several posts on social networks on the one hand. On the Other hand, this information is unstructured and difficult to

exploit and extract. The present project aims to design an intelligent system that allows harvesting information about a person in the social networks, then extract and structure in a scalable graph the context of a person for person depressive or not.

The system must be based on the machine learning algorithms which allow the continuous evolution and the improvement results with time.

Keywords : Social network, Analyze methods, Information System, Data mining, Data modeling, Contextual graph, Profiling, Interest graph, Social graph, Machin learning algorithms, Big Data, Rest Architecture, Graph database.

ملخص :

شبكات التواصل الاجتماعية تعرف نجاحا هائلا بين مستخدمي الانترنت.فهي تعتبر منصات التبادل و التفاعل بين الاشخاص او الجماعات او الكيانات الاجتماعية .كل يوم هناك اكثر من 2.5 مليار تحديثات الفيسبوك .400 مليون تويت.

هناك كمية هائلة من البيانات التي تصف سياق هؤلاء المستخدمين(بيانات مهنية مثل محطة العمل الحالية و الخبرات السابقة او سياق عام مثل الهوايات و الموضوعات الاكثر اثارة للاهتمام ,الحالة النفسية او الناس الذين يحيطون بالمستخدم)

يمكن أن العديد من المجالات على الويب ان تستفيد من السياق هؤلاء الأشخاص مثل توصية المنتجات التي قد تثير اهتمام شخص ما في موقع التجارةالاكترونية استنادا الى سياق شخص او اكتشاف زبائن محتملين في التسويق الالكتروني ....الخ

هذه المعلومات على سياق الاشخاص مبعثرة في العديد من المشاركات على الشبكات الاجتماعية من جهة ,من جهة اخرى هذه المعلومات غير مهيكلة و صعبة للاستغلال و الاستخراج.

و يهدف هذا المشروع الحالي الى تصميم نظام ذكي الذي يسمح بحصاد المعلومات عن شخص في الشبكات الاجتماعية ,ثم استخراج و هيكلتها في مخطط سياق الأشخاص الى شخص مضغوط أو شخص غير مضغوط

يجب ان يستند هذا النظام على الخوارزميات التعلم الالي التي تسمح بالتطور و التحسن المستمر للنتائج مع مرور الوقت.

الكلمات المفتاحية : التسويق الالكتروني , الخوارزميات التعلم الالي , السياق , شبكات التواصل الاجتماعي ,كلمات البحث شبكات الاجتماعية

Introduction Générale

Le développement important des réseaux techniques (Internet et la téléphonie mobile), depuis des années conduit à faciliter l'essor des réseaux sociaux. Aujourd'hui, ils sont considérés comme un espace d'échange de diverses informations et interactions entre différentes personnes ou organisations. Cette masse de données qui peuvent décrire le contexte de ces utilisateurs sont d'une grande utilité pour de nombreux domaines, que ce social, professionnel ou commercial. Ces données éparpillées et mal structurées sont difficiles à exploiter ou extraire, cependant plusieurs techniques sont mises à disposition pour analyser et structurer le contexte et les relations entre entités.

La fouille de données consiste à rechercher et extraire de l'information (utile et inconnue) de gros volumes de données stockées dans des bases ou des entrepôts de données.

Différentes approches, ont été utilisées pour la catégorisation et la classification de textes offrant ainsi, aux développeurs dans le domaine plusieurs issues, qui amène à poser une question très récurrente sur le choix du meilleur algorithme pour la classification automatique de textes.

Elle peut être :

- ✓ supervisée : les classes sont connues à priori, elles ont en général une sémantique associée
- ✓ non-supervisée (en anglais clustering) : les classes sont fondées sur la structure des objets, la sémantique associée aux classes est plus difficile à déterminer

L'analyse du contexte des personnes dans les réseaux sociaux est une approche à la fois graphique et analytique, cette approche basée sur des algorithmes machine Learning va permettre de récolter, extraire puis structurer dans un graphe contextuel les informations relatives d'une entité, puis améliorer ces résultats dans le temps.

La dépression peut être difficile à détecter de l'extérieur; tout le monde se sent parfois bleu ou triste. Mais la tristesse ou les humeurs sont une partie normale de la vie; la dépression est beaucoup plus que de la tristesse. Nous pouvons décrire la dépression comme «vivre dans un trou noir» ou avoir un sentiment de catastrophe imminente.

Cependant, beaucoup de gens éprouvent les premiers symptômes de la dépression au cours de leurs années de collège. Malheureusement, la dépression n'est pas souvent bien détectée et traitée comme conséquence les étudiants qui souffrent de dépression ne reçoivent pas l'aide dont ils ont besoin. Les étudiants utilisent des sites de réseautage social pour leurs activités personnelles et scolaires, de nombreuses études montrent que les sites de réseaux sociaux font partie intégrante de leur vie sociale et peuvent inclure des références affichées sur la dépression. Cependant, sur Twitter, les étudiants partagent leurs sentiments, leurs émotions, leurs pensées et leurs expériences personnelles avec d'autres partenaires.

Le présent sujet consiste en la conception et la réalisation d'un système qui permet d'extraire et structurer le contexte à partir des réseaux sociaux le contexte (dépressive, non dépressive), et représenter le résultat sous forme d'un graphe.

Ce rapport présente l'ensemble des étapes suivies pour créer une solution ; il contient 4 chapitres essentiels organisés comme suit :

- Chapitre1 : dans ce chapitre nous présentons les généralités sur l'analyse des réseaux sociaux, et puis nous aborderons à la présentation des réseaux sociaux les plus utilisés et analyserons ces réseaux sociaux

- Chapitre 2 : nous allons définir la notion de contexte d'une personne, puis faire une modélisation du processus de développement, et puis nous présenterons quelques techniques d'analyse de sentiment du contenu dans les micro-blogs

- Chapitre 3 : dans ce chapitre, nous discuterons les différentes étapes de notre approche proposée pour résoudre le problème de détection des personnes dépressives à travers une analyse décisionnelle des tweets. Ensuite, nous allons définir les outils utilisés pour la réalisation de la partie pratique de nos travaux.

- Chapitre 4 : dans celui-ci, nous verons la présentation de l'environnement matériel et logiciel nécessaire pour implémenter, puis nous exposerons quelques photos de l'interface de nos applications.

Nous clôturerons ce travail avec une conclusion générale dans laquelle nous présenterons quelques perspectives visant à enrichir ce travail.

Problématique :

Des masses de données concernant une personne sont de plus en plus importantes dans les réseaux sociaux. Il devient fatidique d'utiliser ces données pour aider l'utilisateur à accéder facilement à l'information qui correspond à ses besoins spécifiques. De là se pose les questions suivantes :

Comment récupérer ces données et analyser toute cette quantité de données ?

Est-il possible d'utiliser des données écrites en langage naturel et multi-langue ?

Comment traiter des données écrites en langage naturel ?

Comment récupérer les intérêts de la personne ?

- Obtenir le profil psychologique d'une personne ?
- Classer la personne à une personne dépressive ou non

Beaucoup de questions nous essaierons de répondre.

Objectif de l'étude :

L'objet de ce projet est de faire une conception puis la réalisation d'un système qui permet de répondre aux questions précédentes.

Système doit avoir accès aux données (Tweets, commentaires...etc.) du compte d'une personne (compte Twitter), cela dans le but de les traiter pour extraire ces sentiments.

Plus tard un de nos objectifs est d'intégrer notre système avec d'autres domaines.

Chapitre 1 :

Généralité sur l'analyse des réseaux sociaux

Dans ce chapitre nous présentons le contexte d'analyse des réseaux sociaux, pour cela nous commençons par définir le terme de réseau social ce qui va nous permettre par la suite d'introduire le concept d'analyse des réseaux sociaux ainsi que les méthodes d'analyse utilisées.

I. Introduction :

- Les réseaux sociaux ne cessent de connaître un succès exponentiel auprès des internautes, chaque jour on trouve plus 2,5 milliards de mise à jour Facebook, 400 millions de tweets. Cette immense quantité de donnée, qui décrit le contexte de ces utilisateurs, peut être bénéfique dans plusieurs domaines du Web, tel que le E-commerce ou bien le E-marketing.
- Ces information éparpillées et non structurées ce qui les rendent difficiles a extraire et à exploiter.

II. Réseau social :

1. Définition :

Le concept de « social » a été inventé en 1954 par un anthropologue du nom de. John A. Barnes Le principe de réseau se définit par deux éléments : les contacts et les liaisons entre les contacts. Plus nous avons de contacts plus notre réseau est important et donc plus nous sommes utiles (la notion d'utilité ici se résume à la capacité à transmettre des informations). **[1]**

D'autre part, les réseaux sociaux sont considérés comme des services web qui permettent aux individus de construire un profil public ou semi-public liée avec une liste d'autres profils qui nécessite une confirmation bidirectionnelle pour l'amitié, mais parfois unidirectionnels comme fans ou abonnés. **[2]**

Nous pouvons dire que le réseau social peut se voir comme étant une société virtuelle, ou l'utilisateur peut avoir une identité en créant son propre profil. Il permet également de tisser des liens avec d'autres membres, en publiant des messages, des articles, des Photographes,…etc Tout ceci est conservé et peut être analysé pour diverses raisons.

Après avoir introduit le concept de réseau social, nous présentons dans ce qui suit les types de réseaux sociaux existant actuellement sur internet, en mettant l'accent par la suite sur les réseaux les plus actifs,

2. Catégories des réseaux Sociaux dans l'internet :

Voici une liste de types de réseaux sociaux avec des exemples

- *Les réseaux sociaux généralistes pour discuter :* ils sont nombreux et diversifiés mais voici les principaux à retenir : Facebook, Myspace, Twitter.
- *Les Réseaux sociaux de partage :* ils permettent la mise en ligne et le partage de vidéo, de photos...telles que YouTube, Flicker.
- *Les réseaux sociaux professionnels :* ils permettent de présenter votre carrière professionnelle : LinkedIn, Viadeo.
- *Les réseaux de services :* ils proposent d'échanger des bonnes adresses, services... ex : ma-residence, voisineo...

2.1. Présentation des réseaux sociaux les plus utilisés :

Facebook :

Créé en 2006, Facebook est un réseau social permettant de partager des informations, des photos, des vidéos, des humeurs, de la musique avec des amis ou une communauté. Il est possible d'utiliser Facebook à titre privé ou professionnel.

Chaque publication sur le mur, elle s'affichera sur la page d'accueil des amis. Facebook propose aussi un service de messagerie électronique et de chat (discussion instantanée).

Figure I.1: logo Facebook

Twitter:

Site de microblogage permettant de publier de courts messages d'un maximum de 140 caractères appelés « twitte ». Dans les 140 caractères, il est possible de créer des liens vers d'autres sources et engager une discussion avec d'autres utilisateurs.

Chaque usager peut choisir de suivre les messages d'une ou plusieurs personnes ou organisations. Pour rendre la recherche plus facile, des mots peuvent être tagués de cette façon : #recherche, #enseignement...

Figure I.2: logo Twitter

YouTube:

YouTube est un site web d'hébergement de vidéos sur lequel les utilisateurs peuvent envoyer, visualiser et partager des séquences vidéo. Il a été créé en février 2005 par trois anciens employés de PayPal. Le service situé à San Bruno en Californie (États-Unis) emploie la technique Adobe Flash et/ou HTML 5 pour afficher toutes sortes de vidéos : des extraits de films, d'émissions de télé et des clips de musique, mais aussi des vidéos amateurs provenant de

blogs par exemple. En octobre 2006, Google a annoncé qu'après avoir conclu un accord, il deviendrait le propriétaire de l'entreprise en échange d'actions Google d'une valeur totale de 1,65 milliard de dollars américains. La transaction prit fin le 13 novembre 2006. En 2009, 350 millions de personnes visitent chaque mois ce site de partage de vidéos. Le 17 mai 2010, plus de 2 milliards de vidéos sont vues quotidiennement.

Figure I.3: Logo YouTube

III. Analyse des réseaux sociaux:

1. Définition :

L'analyse des réseaux sociaux (ARS) est avant tout une boîte à outils permettant de Visualiser et modéliser les relations sociales sous forme de graphes.

- **Les nœuds** représentent les individus, les organisations.
- **Les liens** sont les relations entre ces nœuds.

De ce fait, l'ARS repose sur des visualisations graphiques issues d'algorithmes permettant de calculer des degrés de force ou de densité entre les différents acteurs d'un réseau. Ainsi, l'analyse des réseaux sociaux est fondée sur une approche structurale des relations entre membres d'un milieu social organisé, elle s'attache à décrire les interdépendances entre acteurs. **[3]**

Nous pouvons dire donc que l'ARS est l'étude des entités sociales c'est-à-dire les personnes (acteurs) dans une organisation ou société et leurs interactions (relations).

Ces interactions peuvent être représentées par un graphe, ou chaque nœud représente un acteur et chaque lien est considéré comme une relation ou arc. Nous pouvons étudier après, les propriétés et le rôle de chaque arc ainsi que la position de chaque acteur social. Il s'agit en réalité d'une technique mathématique qui va nous permettre de comprendre les relations que les individus établissent entre eux grâce à l'étude de l'intensité de leurs interactions, à la fois dans le monde du travail et dans leur communauté sociale.

- ✓ Un sociogramme est un schéma qui illustre un réseau social. Il peut aussi être connu sous le nom de diagramme de réseau social ou graphe. Vous trouverez un exemple d'un sociogramme à la **Figure**. Les cercles représentent différents acteurs dans le réseau, et les lignes et les flèches représentent les liens entre eux.

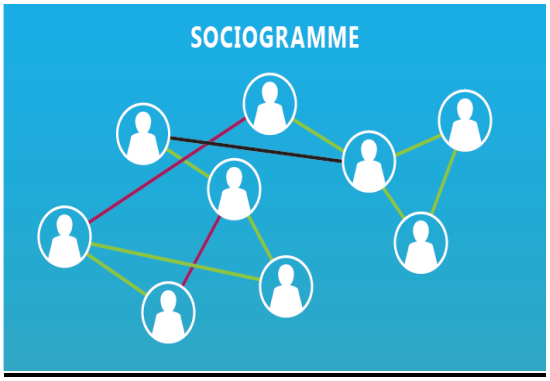

Figure I.4 : un exemple d'un sociogramme

ACTEURS : Les cercles représentent les individus ou les organismes du réseau.

LIENS : les liens représentent les relations entre les acteurs, les flèches illustrent la direction de la relation.

2. Développement historique :

L'ARS à près de 70 ans d'histoire. Elle peut être divisée en trois grandes périodes : les fondements de l'approche ; l'articulation de la méthode ; le développement actuel **[4]**. Les fondations de ces différents édifices ont été construites entre les années 1940 et les années 1960. Dans les années 1960 et 1970 se sont développées des recherches méthodologiques destinées à assurer la mise en œuvre rigoureuse. Dès les années 1980 à aujourd'hui, elles ont été amendées et perfectionnées, ce qui à mener à la découverte de nouvelles pistes. L'analyse des réseaux se fonde sur deux cadres **théorique** et **méthodologique**.**[5]**

- L'aspect théorique repose sur une conception large de la structure sociale et de nombreuses études empiriques démontrant que les comportements des individus sont liés aux structures dans lesquelles ils s'insèrent. La sociométrie a aussi contribué à l'essor de l'analyse des réseaux sociaux. L'apport empirique de la sociométrie est dû en partie à l'œuvre de Moreno, qui est perçu comme l'un des précurseurs de l'analyse de réseau et de la psychologie sociale. Enfin l'analyse de réseaux repose aussi sur l'apport des mathématiques aux sciences sociales : dans le développement théorique de l'analyse de réseaux, des chercheurs ont trouvé des utilités aux modèles mathématiques.

- L'aspect méthodologique se réfère quant à lui à l'usage fait par le chercheur des données de types quantitatifs et qualitatifs et de traitement analytique de ces données

En 1957, Elisabeth Bott publie son étude sur les systèmes de relations des familles. Elle émet l'hypothèse que Le degré de ségrégation des rôles entre mari et femme varie dans le même sens que la densité du réseau social de la famille ; c'est-à-dire que la séparation distincte dans la répartition des tâches domestiques selon le genre tend à être plus élevée dans un réseau social où les membres sont fortement liés entre eux.

L'hypothèse de Bott demeure valide et non réfutée jusqu'à ce jour. Pour sa part, Stanley Milgram met en place, en 1967, un dispositif expérimental d'investigation qui fait encore

aujourd'hui référence dans les recherches sur le petit monde. Il a tenté de calculer le nombre de liens moyens qui séparent une personne de n'importe quelle autre personne sur terre. **[6]**

Aujourd'hui les sujets de recherches en analyse de réseaux sont multiples, la famille, les relations de travail, la camaraderie, etc. Cette approche est actuellement aussi utilisée à d'autres fins que celles de la recherche scientifique, par des conseillers en relations professionnelles ou encore à des fins commerciales, comme dans le cas du projet FOAF (Friend Of A Friend). **[7]**

Les réseaux sociaux prennent de plus en plus d'ampleur dans notre vie quotidienne, l'idée d'analyser ces derniers pour extraire de l'information offre sans doute un avantage important pour plusieurs domaines.

3. Utilisations de l'analyse des réseaux sociaux :

La grande masse d'informations conservées dans les réseaux sociaux peuvent êtres très utiles dans de nombreux domaine. Elles peuvent servir à surveiller la marque de son entreprise, en lui offrant des informations sur: l'opinons du public sur ses produit, avoir une idée sur l'état actuel du marché, les entreprises concurrentiels, ainsi acquérir une nouvelle stratégie de marketing car grâce à ses données elle pourra enfin établir une comparaison des produits existants dans le marché, assurer une meilleure gestion de gamme de produits, permettre un meilleur soutien à sa clientèle, pouvoir suivre les influenceurs **[3]**. Ces données peuvent aussi être utilisées lors des élections pour voir l'opinion du public.

Dans la suite de cette partie nous présentons les méthodes utilisées pour analyser un réseau social, puis établir une comparaison entre ses méthodes dont le but de faire apparaitre les inconvénients de chacune et enfin déterminer celle qui conviendra le mieux à analyser les réseaux sociaux d'aujourd'hui.

4. Méthodes d'analyse des réseaux sociaux :

Tous ces réseaux sociaux amassent de très nombreuses données les amis, les messages, les images, la fréquence d'utilisation, etc. Tous ces échanges et informations sont soigneusement enregistrés. Dès lors se pose le problème de l'exploitation de cette masse d'informations. Il faut tout d'abord modéliser le réseau sous forme mathématique. La structure de base est bien entendu le graphe : l'analyse des figures produites permet de tirer un grand nombre d'informations, et aussi de prédire en partie l'évolution future du réseau.

Des mises à jour sont effectuées, pour offrir des méthodes d'analyse flexibles qui tiennent compte de toutes les informations en termes de structure et de contenu. On distingue alors deux grandes familles : traditionnelle et fouille de données. **[8]**

4.1 Méthodes traditionnelles (classiques) :

Dans cette partie, nous détaillons les mesures utilisées dans l'analyse traditionnelle des réseaux sociaux. Les mesures locales sont les indices qui se focalisent sur les informations locales d'un acteur donné, par contre Les mesures globales apportent une information sur l'ensemble de la structure (réseau)**. [9]**

4.1.1 Les mesures locales :

Les mesures dites locales caractérisent un sommet ou un lien : on obtient autant de résultats qu'il y a de sommets ou de liens avec les mesures locales. Dans cette section, nous

allons introduire deux notions liées à l'analyse des hyperliens qui permet de mesurer le degré de pertinence d'un acteur dans sa communauté : la centralité et le prestige **[8]**

- **La centralité :**

Les acteurs importants sont ceux qui sont liés et impliqués avec les autres acteurs d'une façon extensive. Après avoir modélisé les contacts par des liens dans un réseau, nous pouvons définir un acteur central comme étant celui qui est impliqué dans plusieurs liens.

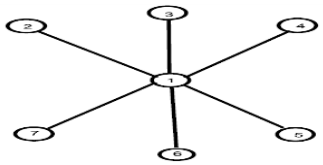

Figure I.5 : réseau sociaux centre **[8]**

Nous remarquons dans la **Figure** que l'acteur 1 est l'acteur le plus central parce qu'il/elle communique avec la majorité des autres acteurs, il figure sur tous les 6 plus cours chemins qui lient les 6 autres acteurs.

Il existe différents types de liens entre les nœuds (acteur), ce qui a permet de générer plusieurs types de centralités dont :

- **Degré de Centralité** : Soit n le nombre total de nœud (acteurs) dans un réseau (graphe). Le degré de centralité d'un acteur i noté $C_D(i)$ est le degré du nœud acteur (le nombre d'arêtes) noté d(i) normalisé par le degré maximal n-1.

$$C_D(i)= d(i) / (n-l)$$

- **Centralité de proximité :** la centralité est définie dans cette approche par la notion de proximité ou de distance. Si i un acteur central donc il peut interagir facilement avec les autres acteurs. Par conséquence, sa distance avec les autres doit être courte. Nous utilisons donc la distance la plus courte pour calculer cette mesure.

- **Le prestige :**

La notion de prestige est une autre manière de mesurer l'importance d'un nœud.

Un nœud prestigieux est un nœud à lequel un grand nombre d'autres nœud se lient- il reçoit un grand nombre de liens entrants. Nous distinguons les mesures suivantes :

- **Le degré de prestige :** avec dI(i) le degré entrant du nœud i, n le nombre nœuds dans le réseau, le degré de prestige est donné par la relation :

$$PD(i) = dl(i) / (n-1)$$

- **Le prestige de tri :** Les mesures proposées jusqu'au là sont fondées sur les liens entrants et sortants d'un acteur donné. La mesure du prestige du tri considère la réputation et I importance des acteurs choisissant l'acteur i. Ce type d'algorithme est utilisé pour trier les résultats de recherche comme Google (algorithme Rank Page).

4.1.2 Les mesures globales :

Dans le but d'avoir une idée sur la structure globale du réseau, certaines mesures globales ont été élaborées nous citons :

- La densité P du graphe G : cette mesure permet d'exprimer le degré connectivité au sein du graphe G représentant un réseau **[9]**. C'est le nombre de liens existant dans le graphe G, normalisé par le nombre maximal de lien dans le graphe.
- La distance géodésique entre deux nœuds est le plus court chemin entre les deux nœuds.
- La distance moyenne d'un graphe connecté est égale à la moyenne des distances géodésiques entre tous les pairs d'acteurs.
- Le diamètre d'un graphe connecté est l'excentricité maximale qui puisse exister entre deux de ses nœuds.

4.2 Fouille de données dans les réseaux sociaux :

Les réseaux sociaux amassent de très nombreuses données : les amis, les messages, les images, la fréquence d'utilisation...etc. Dès lors se pose le problème de l'exploitation de cette masse d'informations. La fouille de données s'avère être un outil incroyablement riche et puissant lorsqu'il est appliqué aux réseaux sociaux, puisque cette méthode nous permet de modéliser le réseau sous forme mathématique, l'analyser pour tirer le maximum d'information mais aussi prédire en partie l'évolution future du réseau.

4.2.1 Définition :

La fouille de données ou Datamining est l'ensemble des méthodes scientifiques destinées à l'exploration et l'analyse de grande quantité de données informatiques en vue de détecter des profils-type, des comportements récurrents, des règles, des liens, des tendances inconnues, des structures particulières restituant de façon concise l'essentiel de l'information utile pour l'aide à la décision. **[10]**

Appliquer cette technique dans notre cas, cette technique va nous permettre d'extraire depuis les réseaux sociaux ; les données relatives à un client puis les analyser et classer. Le datamining permet d'accomplir les quatre types d'analyse suivant : Classification, Estimation, Segmentation, Prévision. Ces types d'analyse se répartissent dans deux catégories descriptives et prédictives comme illustré dans le tableau suivant :

Techniques descriptives	**Techniques prédictives**
- Classification	- Estimation. - Segmentation. - Prévision.

Tableau I.1 : Type d'analyse du DataMining

Les techniques descriptives (ex : la classification) permettent de décrire, résumer, synthétiser et classer les ces techniques essaye de mettre en évidence des informations présentes mais cachées par le volume des données. Par contre, les techniques prédictives essayent d'extrapoler des nouvelles informations à partir des données présentées, ci-dessus :

- **La classification (technique descriptive) :**

Est une méthode qui permet de regrouper des objets (personne, intérêts...etc.) en groupes, ou familles de sorte que les objets d'un même groupe se ressemblent le plus possible, et ceux de groupes distincts diffèrent le plus possible.

Le nombre des groupes est parfois fixés, ils ne sont pas prédéfinis mais déterminés au cours de l'opération.

Dans le cas d'extraction de données à partir des réseaux sociaux, cette méthode nous permettra de regrouper les intérêts d'une personne par classe (personnage, produit, achat, consommation, activité...etc.). D'une autre façon nous pouvons dire que cette méthode descriptive permet de décrire de façon simple une réalité complexe en la résumant.

- **L'estimation (technique prédictive) :**

L'estimation consiste à définir le lien entre un ensemble de prédicteurs et une variable cible.

Ce lien est défini à partir de données complètes, c'est-à-dire dont les valeurs sont connues tant pour les prédicteurs que pour la variable cible. Ensuite, nous pouvons déduire une variable cible inconnue de la connaissance des prédicteurs. L'intérêt de cette technique est d'estimer à partir des caractéristiques d'un objet, la valeur d'un champ inconnu. L'estimation va être utile pour détecter la psychologie de la personne c'est-à-dire estimer le caractère de la personne à partir de ces activités et environnement.

- **La segmentation (technique prédictive) :**

Consiste à former des groupes (segments) homogènes à l'intérieur d'une population. Cette tâche est souvent effectuée avant les précédentes pour construire des groupes sur lesquels on applique des tâches de classification ou d'estimation. Dans l'analyse des réseaux sociaux, nous pouvons dire que cette technique va nous servir par exemple pour la détection de communautés.

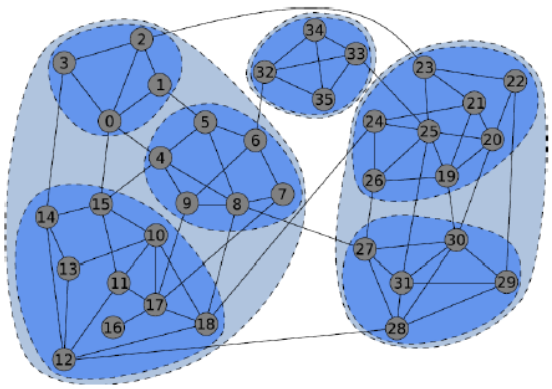

Figure I.6 : exemple de segmentation **[9]**

La prévision (technique prédictive) :

Consiste à estimer une valeur future d'un champ à partir des données réelles possédées. Les méthodes de classification et d'estimation peuvent être utilisées en prédiction. Dans notre cas, la prévision va nous servir de prévenir le comportement d'une personne dans le futur, les produits qui vont lui intéresser dans le futur. Cette technique est très utilisée dans le domaine d'intelligence artificielle, spécialement dans les algorithmes machine Learning.

5. Etude comparative des méthodes :

Dans notre cas « analyser le contexte d'une personne depuis les réseaux sociaux » , il est très important de trouver la méthode qui va nous permettre d'analyser cette grande masse de données c'est-à-dire : analyser des milliers de mise à jours sur Facebook, des milliers de tweets,...etc. Pour pouvoir par la suite extraire les informations dont on a besoin : les personnes qui fréquente le plus, les produits qu'il l'attire le plus, les activités qu'il exerce le plus souvent, puis enfin essayer de prévenir les besoins de cette personne. Toutes ces opérations doivent être dans un intervalle de temps très court

Nous présentons dans le **tableau** suivant les caractéristiques de chaque méthode, dans le but de sélectionner la méthode appropriée :

Méthodes Traditionnelles d'analyse de données	**Data Mining**
- Apparition 1960-1980. - Appliquées sur une centaine de données - Quelques dizaine de variable. - Les variables sont numériques. - Techniques simples.	- Apparition depuis 1990, - Appliquées sur plusieurs millions de données. - Quelques centaines de variables. Certaines variables sont non numériques. - Assemblage de techniques d'analyse. - Résultat plus exacte que celui obtenu par les techniques traditionnelles. - Analyse des données imparfaites, avec des erreu de saisie, valeurs manquantes.

	- Calcul rapide, parfois en temps réel. - Utilisation pour l'aide à la décision, Intelligence artificielle (Machine Learning). - Beaucoup de logiciels consacrés à la fouille de données.

Tableau I.2 : Comparaison entre les méthodes traditionnelles et Data Mining

5.1 Synthèse :

En analysant les deux méthodes d'analyse : traditionnelle et fouille des données, nous pouvons dire tout de suite que la méthode de fouille des réseaux sociaux est la plus adéquate à combler les besoins de notre système Cette dernière nous offre la technique de classification et de segmentation qui sera utile pour identifier les intérêts, la technique de descriptive utilisée dans les algorithmes du Machine Learning pour pouvoir anticiper les besoins de la personne.

Dans la suite de ce chapitre nous allons définir le concept de **Machine Learning**, puis nous présentons quelques outils d'analyse de réseaux sociaux.

VI. Machine Learning :

Définition :

Machine Learning est une branche d'intelligence artificielle qui permet à une machine d'analyser un système, de comprendre pas à pas son fonctionnement et comme résultat d'effectuer ou simuler des différentes taches de ce système. L'algorithme d'apprentissage a pour objectif d'apprendre le fonctionnement du système étudié de manière active. Il connaît les entrées possibles du système et compose des séquences qu'il soumet au système (requêtes) pour observer ses réponses (séquences autorisée/refusée, valeurs renvoyées, etc.). **[11]**

D'autre part, nous pouvons définir le terme Machine Learning comme un ensemble d'algorithmes qui permettent d'apprendre le fonctionnement d'un système en observant régulièrement les tâches qu'il réalise, puis prédire son comportement et ses décisions.

Figure I.7 : Machine Learning

1. Catégories de Machine Learning :

Il existe deux grandes catégories de prédictions en machine Learning : la régression et la classification. Pour la régression ce que nous souhaitons prédire est une valeur numérique continue (par exemple, le prix d'un appartement) alors que pour la classification, on cherchera à déterminer une valeur discrète et finie (par exemple, à quelle espèce appartient une fleur). Pour pouvoir faire ces prédictions, les algorithmes effectuent des calculs plus ou moins complexes sur des valeurs numériques **[12]**. Il est donc nécessaire de transformer es caractéristiques des éléments à analyser en un tableau numérique représentant ces caractéristiques. Afin d'obtenir de bonnes prédictions, il est nécessaire de rassembler toutes les caractéristiques importantes (pour le résultat qu'on souhaite établir) et de bien les modéliser. Par exemple, si on cherche à déterminer le prix de vente d'un appartement, il faudra probablement prendre en compte sa superficie, son emplacement, son entretien, s'il est meublé...etc.

Mais peut être aussi d'autres caractéristiques auxquelles on ne pense pas forcément au premier abord.

Il est donc essentiel de très bien connaître le domaine métier pour pouvoir modéliser correctement les éléments que l'on souhaite traiter. Une fois toutes ces caractéristiques transformées en valeurs numériques, on peut appliquer un algorithme de machine Learning à nos données pour pouvoir construire un modèle prédictif. Un algorithme très simple est par exemple de faire une régression linéaire sur les caractéristiques des éléments. La valeur à prédire sera donc une combinaison linéaire des caractéristiques. On peut aussi essayer des régressions logarithmique ou polynomiale mais il suffit simplement de créer de nouvelles fonctionnalités pour pouvoir se ramener à une simple régression linéaire.

2. Limites de Machine Learning :

Les algorithmes machines Learning ne sont pas non plus une boite noire magique qui permet de tout deviner et qui s'adapte à tout. Si l'on souhaite faire des prédictions de bonne qualité, il y a certaines choses à prendre en compte. Tout d'abord, il est nécessaire que les résultats que l'on souhaite prédire soient différentiables. Les algorithmes ne jouent que dans la qualité des prédictions, les 80% autres sont dus à la qualité des données.

Dans le monde réel, les choses sont bien plus complexes, il se peut y avoir des imprécisions sur les caractéristiques, des erreurs dans la classification des données ou beaucoup d'autres choses qui rendront les prédictions moins évidentes. Une autre limite du machine Learning est que les

algorithmes sont incapables d'extrapoler les données de manière fiable. Il est donc nécessaire de faire des prédictions uniquement sur le même domaine de données que celui utilisé pour l'apprentissage. Les algorithmes de machine Learning ne permettent donc pas d'apprendre de nouvelles choses mais seulement d'automatiser des choses connues ou de mettre en évidence des relations.

V. Conclusion :

Dans ce chapitre nous avons défini le concept d'analyse des réseaux sociaux ; son développement historique puis citer quelques domaines d'utilisation de cette analyse. Par la suite nous avons introduit les méthodes qui permettent l'analyse des réseaux sociaux, puis les comparer pour tirer celle qui possède les qualités idéales pour analyser des réseaux de très grande masse de taille de Twitter ou Facebook…etc.

Dans le chapitre suivant nous introduisons la notion de contexte d'une personne dans les réseaux sociaux, nous citons par la suite quelque outil d'analyse de contexte, puis proposer un processus de développement du contexte de la personne.

Chapitre 2 :

Contexte de la Personne dans les Réseaux sociaux

Nous avons défini la notion d'analyse des réseaux sociaux dans le chapitre précédent, cela dans le but d'introduire le concept d'analyse de contexte de la personne. Dans ce chapitre nous commençons par définir la notion de contexte d'une personne, puis faire une étude sur l'analyse de sentiment du contenu dans les micro-blogs.

I. Introduction :

Dans ce chapitre nous présentons le contexte des personnes dans les réseaux sociaux, pour cela nous commençons par définir le contexte d'une personne ce qui va nous permettre par la suite d'introduire la modélisation du processus de développement ainsi que les techniques d'analyse des sentiments dans le micro blog.

II. Le contexte d'une personne:

1. Définition:

Les différentes avancées dans le domaine de l'informatique ont créé le besoin de systèmes dépendants du contexte. L 'objectif de ces systèmes étant de fournir des informations qui dépendent de l'environnement de l'utilisateur afin d'améliorer son interaction avec les différents systèmes qu'ils utilisent quotidiennement.

La notion du contexte a été utilisée en linguistique et psychologie avant d'être adoptée en informatique, mais a aussi une origine lointaine et une longue histoire en philosophie. La mobilité a donné une dimension importante au contexte qui a touché à de nombreux champs d'application en informatique comme l'informatique ubiquitaire, l'intelligence artificielle, le traitement de la langue naturelle (informatique cognitive),...etc.

Nous pouvons définir le contexte comme étant toute information qui peut être utilisée pour caractériser la situation d'une entité. Une entité est une place, une personne, ou un objet qui est considéré pertinent à l'interaction entre un utilisateur avec les médias sociaux. Le contexte peut inclure la localisation, l'identité de la personne, les personnes qui l'entourent personne, les activités ou les produits qui l'intéressent.

III. Modélisation du processus de développement du contexte d'un utilisateur:

Il devient Crucial d'aider l'utilisateur à accéder facilement à l'information qui correspond à ses besoins spécifiques. Depuis plus d'une décennie, la conception du contexte d'une personne dans les systèmes d'information est devenue un enjeu majeur pour l'amélioration de la qualité des services rendus aux utilisateurs **[13]**. Le contexte des personnes construits à partir des réseaux sociaux sont alors utilisés dans divers systèmes tels que les systèmes de personnalisation, les systèmes adaptatifs, les systèmes de recommandation, les systèmes d'analyses comportementales...etc. Son application peut intéresser les moteurs de recherche. E-Commerce. E—Learning librairies digitales. La médecine, télécommunications, sécurité.....etc. L'usage du contexte d'une personne dans ces systèmes implique les étapes **:**

- Développement du contexte des utilisateurs qui nécessite la collecte de données sur les traces d'activités des utilisateurs et l'usage des techniques d'apprentissage automatique sur ces données (Fouille de données) surtout lorsque on aura à faire à des profils incomplet ou l'information sur l'utilisateur est insuffisante.

- Représentation du contexte des utilisateurs construits qui implique la structuration des données recueillies vu leur quantité et leur diversité (les activités les plus pratiquées, les personnes les plus fréquentées, les caractères psychologiques, les lieux les plus visités...etc.).

Nous pouvons dire que le processus de développement du contexte d'un utilisateur à partir des réseaux sociaux est similaire à tout processus d'extraction de connaissances à partir des données. Nous distinguons quatre grandes étapes dans ce processus : la collecte de données, la structuration des données, l'analyse de données et la représentation des données.

Figure II.1 : processus de développement du contexte d'une personne.

Nous avons vu que le domaine d'analyse d'un contexte dans les réseaux sociaux et très large alors nous avons limité nous travail à l'analyse des sentiments baser sur la théorie de Big Five.

1. Théorie des cinq grands facteurs de personnalité (Big Five) :

La technique « Big five » est l'une des méthodes qui sert à tirer le Comportement ou la Personnalité de l'utilisateur c'est à dire son profil psychologique. Elle consiste à analyser les cinq (5) grands traits de personnalité illustrés clatis **Figure** qui proviennent du regroupement de tous les traits de caractère d'un être humain. Voici les 5 items du Big five :

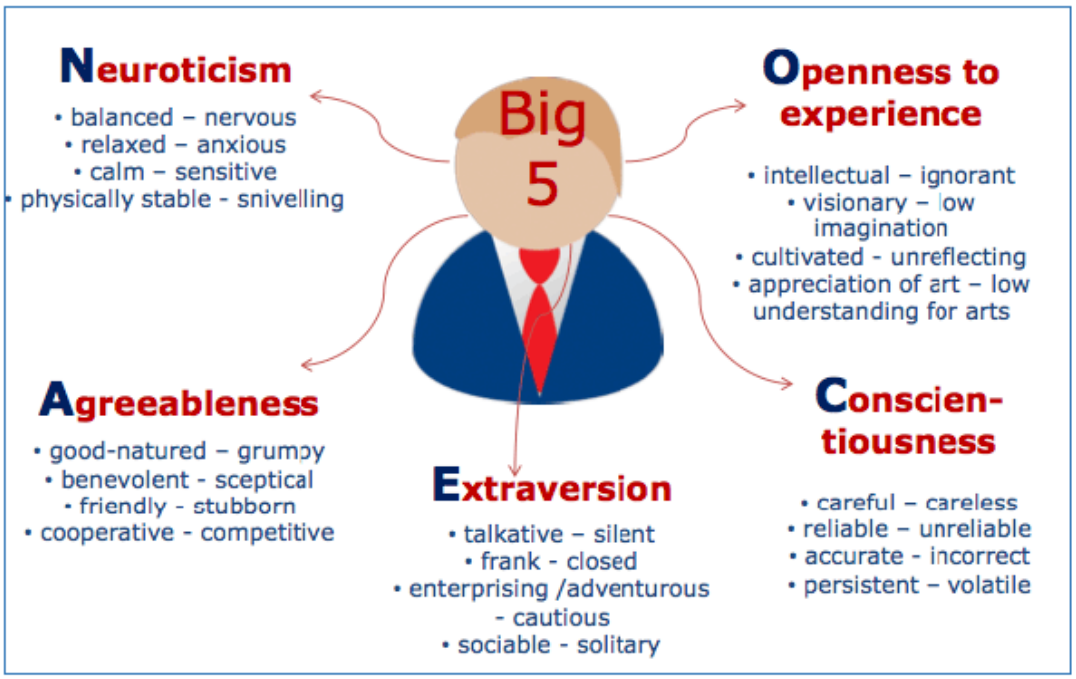

Figure II.2 : les 5 grands traits de personnalité.

- **Introverti ou extraverti :** une personne extravertie est sociable, sait s'affirmer, est énergique et active.
- **Degré d'agréabilité :** une personne à l'agréabilité élevée est altruiste, sensible, emplie de compassion. Modeste et aime la dimension collective dans l'action.
- **Degré de névrosisme :** une personne au névrosisme élevé a tendance à éprouver des émotions négatives telles que l'anxiété, la tension. les sautes d'humeur. La dépression, l'instabilité émotionnelle et à être très impulsive... Au contraire, avec un degré faible de névrosisme, on est stable émotionnellement, positif, bien dans sa peau.
- **Le degré d'ouverture à l'expérience :** s'il est élevé, la personne est curieuse. imaginative, originale, inventive, ingénieuse, à l'esprit. vif et des tendances artistiques.
- **Le caractère consciencieux :** quand une Personne a un tel caractère à un degré élevé, c'est une personne organisée. efficace, responsable, capable de résister aux impulsions du moment et de s'investir dans une tâche avec des gratifications relativement lointaines.

Aussi, avons-nous limité notre travail à l'analyse de **Degré de névrosisme** sur des Tweets si la personne est dépressive ou non.

IV. La dépression en psychologie :

1. Recherche sur la dépression en psychologie

La dépression est la quatrième plus grande maladie au monde et sera en deuxième place en 2020 selon les statistiques de l'Organisation mondiale de la santé **[14]** Le principal clinique symptôme de patients déprimés est l'humeur déprimée durable et le manque de Émotions positives. Ils préfèrent être seuls plutôt qu'avec les autres. De plus, la plupart des patients déprimés souffrent d'insomnie chronique.

La recherche de la dépression dans le réseau social en psychologie vient en deux types:

- l'un est de découvrir les disciplines d'une foule d'utilisateurs déprimés **[15-16]**
- L'autre est d'examiner minutieusement un cas particulier **[17].**

Littérature **[15]** observe linguistique marqueurs de la dépression à travers la collecte des messages par déprimé et non-déprimé les personnes du forum Internet. Il analyse le texte avec LIWC, un ordinateur outil de comptage de mots, et montre que les écrivains déprimés en ligne utilisent plus pronoms singuliers à la première personne, mais moins de pronoms au pluriel, plus négative les mots d'émotion mais les mots d'émotion moins positifs. La littérature **[17]** discute des relations entre les comportements SNS et les niveaux de dépression basés sur des événements. Il est établi par des outils de questionnaires et de statistiques, et révèle que Les résultats des publications initiales pourraient indiquer les niveaux dépressifs des micro-blogueurs. Aussi, la période de temps que les utilisateurs postent des micro-blogs est une considération comme la plupart des cas de dépression les patients souffrent d'insomnie chronique.

Ces recherches sur les caractéristiques de la dépression dans la perspective de la psychologie Voyez des connaissances de base fiables pour notre étude. Cependant, quand vient à l'information problèmes d'analyse, seuls quelques outils statistiques simples sont conçus pour eux, qui sans aucun doute limiter leurs recherches. Par conséquent, une technique d'exploration de données spécifique Pour détecter les utilisateurs déprimés est conçue dans cette étude en fonction de leurs résultats.

V. Techniques d'analyse de sentiment

Comme le symptôme cardinal de la dépression est des émotions négatives sévères et le manque d'émotions positives, l'analyse du sentiment est l'étape la plus importante dans la dépression détection. L'analyse des sentiments vise les opinions et le sentiment des utilisateurs la plupart des textes qu'ils ont publiés **[18]**. Récemment, de nombreux progrès ont été réalisés dans l'analyse des sentiments sur les données Twitter. Ces recherches comprennent deux aspects:

- Analyse indépendante du sujet, à savoir jugement de la polarité des tweets

Sans considérer si cela est pertinent pour un sujet. Les principales approches
Sont basés sur des hashtags, des smileys et quelques traits abstraits.

- Analyse dépendante du sujet, à savoir jugé de la polarité des tweets

Sur le sujet donné. Les sentiments des tweets sont positifs, négatifs dans, selon non seulement les caractéristiques abstraites mais aussi les fonctionnalités dépendantes de la cible, qui se réfèrent aux commentaires sur la cible lui-même et les choses connexes, qui sont définies comme des cibles étendues.

La recherche d'analyse de sentiment sur les textes est entrain de développement. Les Peu d'études ont été faites pour résoudre problèmes dans un domaine spécifique, bien que la stratégie d'analyse soit très différente pour différentes champs. Par exemple, les suer de la dépression ont tendance à penser au sujet de la «mort», donc ce genre de mots devrait faire l'objet d'une attention particulière lors de la construction du vocabulaire. Les micro-blogs sont souvent écrits dans un style familier, qui apporte également nouveaux défis lors de l'instauration des règles linguistiques dans la méthode proposée.

Le problème abordé dans ce sujet est l'analyse des sentiments dépendant de micro-blogs. Inspiré par le travail dans la littérature **[19]**, les caractéristiques abstraites et les fonctionnalités dépendantes de la cible sont prises en compte. Cette étude souligne la particularité de la dépression et le contenu de micro-blog, et le modèle entier est spécifiquement conçu en fonction d'eux. Comme le montre la **Fig.II.3**, une méthode d'analyse de sentiment est premièrement proposée d'utiliser le vocabulaire et les règles artificielles pour calculer l'inclination de chaque micro-blog de la Fig. (A). Le vocabulaire et les règles de l'homme dans Méthode d'analyse de sentiment sont construits sur la base des règles de syntaxe françaises et anglaise, la particularité de la dépression et des micro-blogs. Ensuite, comme indiqué dans Fig. (B), un modèle de détection de dépression est construit sur la base de la proposition méthode et caractéristiques des utilisateurs déprimés issus de la recherche psychologique. Enfin, la signification de chaque caractéristique est analysée et simplifiée. Modèle est proposé pour l'application dans Micro-blog.

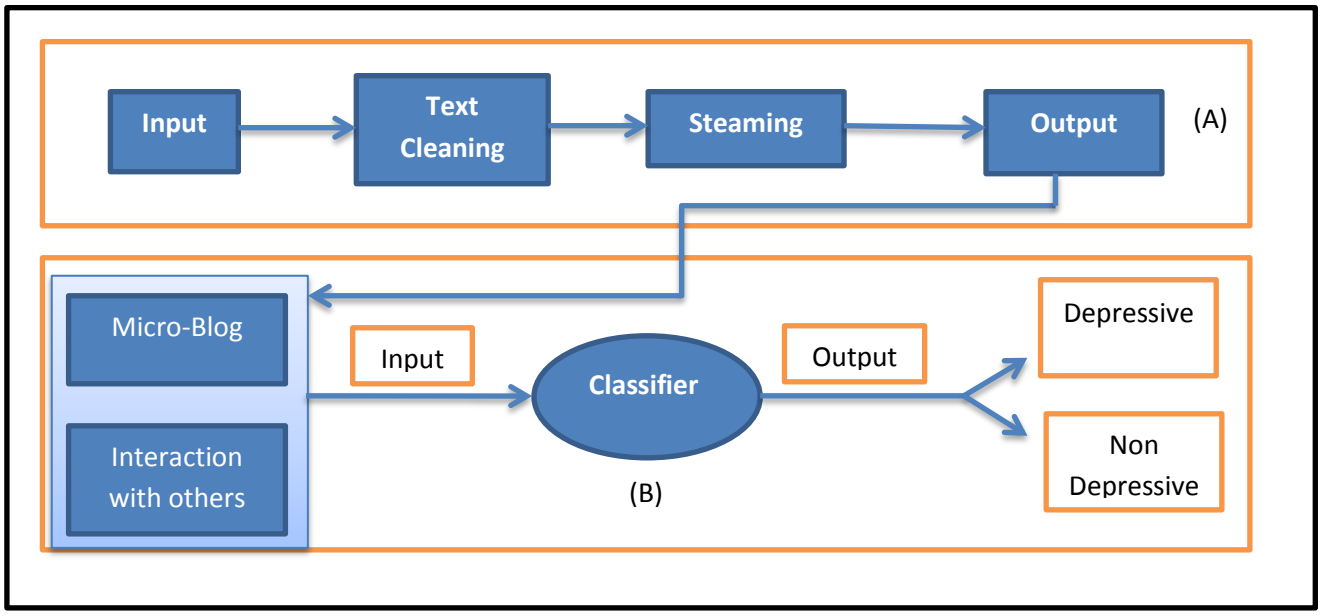

Figure II.3 : technique d'analyse des sentiments Proposer dans ce travail.

VI. Analyse de sentiment du contenu de micro-blog :

L'expression la plus directe de l'humeur dépressive est le contenu du micro-blog des utilisateurs, donc la méthode d'analyse des sentiments dans cette section aide à comprendre la

polarité de chaque morceau de micro-blog, qui souligne l'inclinaison de la dépression reflétée par le contenu. Un vocabulaire est construit sur la base de **HowNet** , et les modèles de structures de phrases et les règles de calcul sont dérivées selon Règles de syntaxe française ou anglaise. Comme décrit ci-dessus, la particularité de la dépression et les micro-blogs portent une attention particulière à l'ensemble du processus.

1. Construction du vocabulaire :

La particularité la plus essentielle de la dépression et des micro-blogs est l'utilisation des mots. Un vocabulaire pour la détection de la dépression est construit sur la base de **HowNet**, un vocabulaire complet de mots français et anglais, comme indiqué dans le tableau.

2 Item		Num.	Example
Emotion Words	Positive	4566	pretty, love, like, happy, good
	Negative	4370	ugly, sad, depressed, unhappy, bad
Degree Modifiers		219	most(2), over(1.75), very(1.5), more(1), -ish(0.75),insufficient(0.5)

Figure II.4: Words in HowNet vocabulary **[20]**

HowNet contient la plupart des mots d'émotion populaires et des modificateurs de degré. Les poids des modificateurs de degré sont quantifiés en six niveaux en fonction de leurs intensités. **HowNet** est conçu pour l'analyse du sentiment général. Pour le faire pour le calcul de l'inclinaison en dépression, plusieurs ajustements sont effectués comme suit:

1. *Les mots d'émotion, les cyberspeaks, les particules modales et les mots négatifs sont ajoutés:*

1. Les utilisateurs déprimés ont tendance à utiliser plus de mots d'émotion, en particulier les émotions négatives. mots d'instruction dont certains ne sont même que pour eux. Par exemple, « bye » est un mot neutre pour les gens normaux, mais c'est un négatif typique pour déprimé utilisateurs. Donc, ces Mots d'émotion typiques pour la dépression sont ajoutés.

2. Considérant que les cyberspeaks sont répandus sur Internet, ils sont un rôle important dans les micro-blogs. Par conséquent, ces mots sont également ajoutés, par exemple, « smilence », qui signifie" souriez silencieusement ", dans le vocabulaire.
3. Comme les micro-blogs sont souvent écrits dans un style familier, les particules modales souvent se produisent dans des micro-blogs pour exprimer des sentiments directement, tels que « ha-ha » et « a-ha », donc ces particules modales sont aussi ajoutées au vocabulaire.

2. *La partie du discours de chaque mot est reconnue:*

Les règles de calcul proposées sont dérivées des règles de syntaxe française et anglaise, sont définis par les parties du discours. Ainsi, la partie du discours de chaque mot est reconnue niés et également importé en tant qu'attribut dans le vocabulaire.

2. Règles Construction linguistiques :

Le sens d'une phrase ne peut pas être décidé uniquement par les mots qu'elle utilise, mais aussi par l'ordre des mots, nommé la structure de la phrase. Par exemple (un peu malheureux) " (très malheureux)", les deux phrases partager le même mot, mais ont une étendue différente évidente de la façon dont heureux c'est le cas, donc la structure de la phrase devrait être prise en compte dans le processus de calcul de la polarité. La structure des phrases pourrait être décrite comme règles linguistiques, ce qui reflète la complexité de la langue dans un aspect. Dans cette section, les règles linguistiques basées sur le vocabulaire proposé sont construites en prenant en compte le style familier du micro-blog.

VII. Sentiment et données Twitter :

1. Ensemble de données :

Twitter contient ensemble des données et des tweets qui nous vont récupérer automatiquement avec l'API Twitter. Ces tweets ont été annotés automatiquement dépressive et non dépressive. Ceux qui contenaient les deux, n'ont pas été gardés. L'ensemble d'entraînement est annoté en deux classes (dépressive et non dépressive) alors que l'ensemble de test est annoté à la main sur deux différentes classes (dépressive, non dépressive). Pour nos expériences, nous n'utilisons que les classes dépressive et non dépressive de l'ensemble de test. Chaque ligne du fichier contient un seul tweet contenant au maximum 140 caractères et peut contenir plusieurs phrases (selon la longueur). Parce que les tweets ont été collectés Directement sur l'API twitter, ils peuvent donc contenir des adresses HTML, des hashtags # et des noms d'utilisateurs (précédés d'un @). Finalement la structure de chaque ligne est la suivante :

1. la polarité du tweet (e.g, 1 = dépressive 0 = non dépressive)
2. l'id du tweet (e.g, 4510)
3. la date du tweet (e.g Sat May 16 23 :58 :44 UTC 2009)
4. le nom de l'utilisateur qui a posté le tweet (e.g, robotickilldozr)
5. Le text du tweet (e.g. "I must think about positive.").

2. Caractéristiques des tweets :

Nous présentons rapidement dans cette sous-partie les principales caractéristiques d'un tweet et de Twitter.

2.1. Longueur :

La longueur maximale d'un message posté sur Twitter est de 140 caractères. D'après Go et al. **[21]**, la longueur moyenne des tweets est de 14 mots ou 78 caractères sur ce corpus. Cette longueur est très Courte contrairement à celles utilisées dans d'autres corpus pour la classification de sentiments (comme les critiques de films).

2.2. Disponibilité des données et modèles du langage :

Les sujets abordés sur Twitter sont très divers et l'API twitter permet de récolter des millions de messages. En effet, le nombre de tweets postés chaque jour est immense. Les utilisateurs peuvent poster des messages depuis n'importe quel lieu et avec différents appareils. Il est à noter qu'un tweet peut contenir des fautes d'orthographe liées à l'utilisation de Smartphones et à la limitation de caractères. De plus, le registre de langue utilisé peut être familier.

IIX. Conclusion :

Dans ce chapitre nous avons défini le contexte de la personne dans les réseaux sociaux Par la suite nous avons introduit la modélisation et le processus du développement du contexte d'une personne, puis nous avons précisé nous travaille sur l'analyse de sentiment et on 'a défini les technique d'analyse des sentiments sure les tweets et leur caractéristiques.

Dans le chapitre suivant nous introduisons l'implémentation, résultats et discussion, nous citons la description de l'approche, puis le prétraitement et les différents algorithmes de classification supervise et compare avec discussion sur les différents résultats.

Chapitre 3 :

Implémentation Résultat et discussion

Dans ce chapitre nous présentons les approches, et puis évaluer les résultats et puis on ’a fait une phase de conception et implémentation et nous avons discuté sur les résultats et les comparer.

I. Introduction :

Dans les dernières années, nous sommes dans un monde numérique où l'information est disponible en grande quantité et sous diverses formes. 80% de cette masse d'informations est sous format textuelles. Cependant, toutes ces informations n'auront aucune utilité si on ne peut pas accéder aux connaissances qu'elle porte. C'est pour cette raison, nous avons besoin d'outils spécifiques d'accès à l'information, afin de réduire l'intervention humaine.

L'analyse contextuelle des réseaux sociaux est un domaine qui a attiré beaucoup de chercheurs, ce qui a donné naissance à de nombreux travaux. Dans notre travail nous sommes basées sur un principe fondé sur le corpus (Corpus-based Approach) qui consiste à attribuer des données à un classificateur pour l'apprentissage d'une façon supervisée, ce dernier génère un modèle qui est utilisé pour la partie test.

Dans notre travail nous avons utilisé le twitter comme réseau social qui est un site de réseautage permet aux utilisateurs d'écrire de courts articles, appelés «tweets». Le contenu de ce chapitre permet de discuter les différentes étapes de notre approche proposée pour résoudre le problème de détection des personnes dépressive à travers une analyse décisionnelle des tweets. Ensuite nous allons définir les outils utilisé pour la réalisation de la partie pratique de nos travaux avec une présentation générale des résultats obtenus en discutant les différents comparaisons appliquer entre les différents techniques utilisées et proposée durant notre travail.

II. Description de l'approche :

Notre approche est constituée de 3 modules principaux comme le montre la figure suivante :

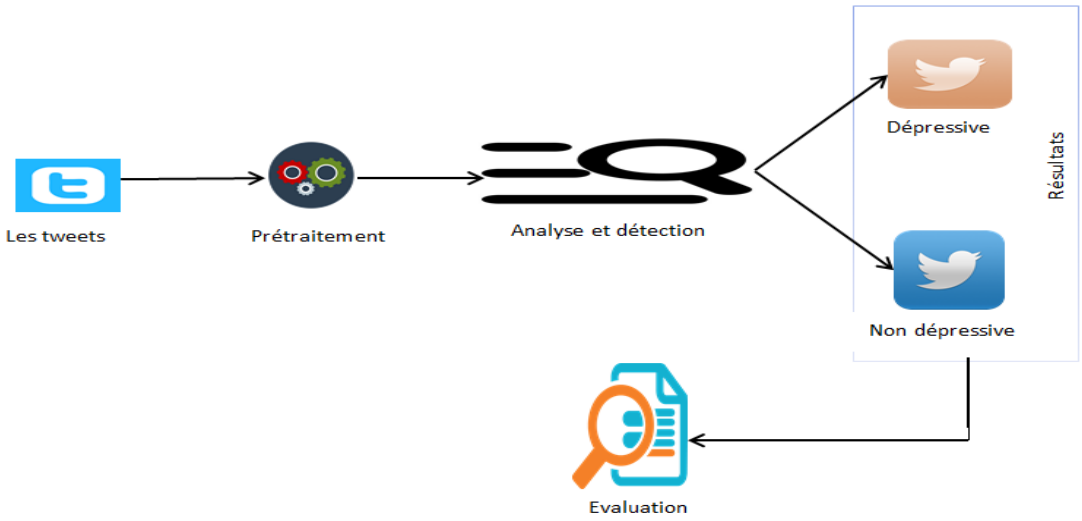

Figure III .1: architecture générale de notre approche

1. Tweets2011 corpus (les tweets):

Pour nos expérimentations nous avons utilisé le corpus Tweets2011 qui a été utilisé dans la compétition célèbre de recherche d'information appelée TREC 2011. C'est corpus spécialisé construit selon les mots clés. Les auteurs de ce corpus ont utilisés l'API Twitter4J pour extraire 649 tweets où ils ont utilisé les mots clés (politique, cinéma, sport, music, guerre, science). Après dans TREC 2012 ces tweets ont été classé en deux classes (tweet dépressive, tweet non dépressive) **[22]**. Le tableau suivant résume la classification des tweets.

Catégorie	Dépressive	Non dépressive
Cinéma	85	62
Politique	49	33
Guerre	64	13
Sport	33	58
Music	119	56
Science	19	58
Total	369	280

Tableau III .1 : statistique générale du dataset Tweets2011.

2. Prétraitement des données textuel:

Dans le cas des données textuel, la première étape est la reconnaissance des termes, des ponctuations, des fins de paragraphes et des phrases. Nous devons aussi unifier l'écriture des lettres en minuscule afin de facilité la mise en correspondance.

Figure III .2: les étapes de prétraitement des données tweets.

La figure précédente regroupe les différentes étapes nécessaires à la vectorisation des tweets.

2.1. Nettoyage :

Nous éliminons tous les caractères non alphabétiques comme les chiffres et les caractères spéciaux.

2.2. Représentation de texte :

Cette étape assure la transformation des textes vers une liste de termes. Nous avons implémenté et intégrer dans l'EBIRI différentes techniques de représentation comme:

- **La représentation Sac de mots :**

Cette technique permet d'isoler les ponctuations et de découper les séquences de caractères liés, en fonction de la présence ou l'absence des caractères de séparation (de type espace, tabulation, ou retour à la ligne). Cette méthode permet de rendre les textes dans une liste de mots appelée sac, par exemple la phrase «Je suis étudiant à l'université » deviendra une liste des mots {je, suis, étudiant, à, université}.

- **La représentation N-grammes caractères :**

Cette technique est directement liée à un paramètre N où une fenêtre de N cases sera construite (par exemple N=2 alors une fenêtre de deux cases). Cette fenêtre se déplace dans tout le texte du début à la fin, caractère par caractère et chaque N-grammes capturées sera enregistrée dans une liste.

- **La représentation par racinisation (stemming):**

Cette technique de représentation est basée sur le regroupement des mots ayant la même racine (stem). Pour l'extraction des stems nous avons appliqués l'algorithme de porter qu'est simple basée sur des règles de remplacement des chaines de caractères pour supprimer les suffixes les plus utilisée et les signes pluriel, par exemple le mot « troubling » deviendra « trouble » ou « relation» deviendra « relate ». Cette représentation réduit de 30% la taille moyenne d'un document.

2.3. Le codage:

Cette étape permet de calculer l'importance de chaque attribut (composant) dans chaque document en utilisant différentes méthodes de pondération comme:

- **La pondération fréquentiel brute (PFB):**

La PFB permet de calculer le nombre d'occurrences du chaque terme $\boldsymbol{t_j}$ dans chaque document$\boldsymbol{d_i}$. En d'autre terme, un document sera transformer en un vecteur dont les composantes vont correspondre au nombre de fois où le terme $\boldsymbol{t_j}$ apparaît dans le document $\boldsymbol{d_i}$.

Pondération fréquentielle brute (t_j, d_i) = le nombre d'occurrences de terme T_i dans le document D_i

- **La pondération TF*IDF:**

Terme frequency*inversed documents frequency (IDF) est basé sur la loi de Zipf. Elle permet de calculer le poids de chaque terme en multipliant un facteur concernant l'importance du terme T dans le texte avec un autre qui concerne l'importance de ce terme dans tout l'ensemble de données.

La fréquence inversée des documents(T) $= \log \frac{|D|}{DF(T)}$.

- DF (T) représente le nombre de documents qui comprennent le mot T.
- D: Le nombre de documents dans l'ensemble de données.

3. Analyse des tweet :

Pour cette étape nous avons testé trois algorithmes classiques supervisée qui vont être détaillé par la suite :

3.1. Algorithme K Plus Proches Voisins (KPPV)

Le KPPV appelé en anglais K nearest neighbor (KNN) est un algorithme de classification supervisée simple et naïve. L'objectif c'est de classé chaque nouvel exemple (de la base de teste) sur la base de leur distance avec les exemples de la base d'apprentissage. Il nécessite la présence des paramètres comme : Base d'apprentissage, La valeur du K et Une mesure de distance.

Pour prédire la classe d'un nouvel exemple « X » l'algorithme calcule la distance de X avec chaque exemple de la base d'apprentissage afin de trouver les « K » plus proches voisin de X. Enfin la classe majoritaire parmi les K classes sera attribuée à X.

Pseudo code algorithme K plus proches voisins

i : le numéro d'exemple de la base d'apprentissage.

C : la classe de l'exemple numéro i de la base d'apprentissage.

Entrée : -choix d'une mesure de distance

- La valeur du paramètre K.
- Base d'apprentissage D.
- X : le nouvel exemple que l'on veut connaitre sa classe.

Début

pour chaque *((i, c) ∈ D)* ***faire***

Calculer la distance *dist(x, i)*

fin

pour chaque *{i ∈ kplus proches voisins (x)}* ***faire***

Compter le nombre d'occurrence de chaque classe.

fin

Attribuer à *x la classe la plus fréquente (classe majoritaire).*

fin

3.2. Algorithme naïve bayes :

Naive Bayes est l'une des méthodes de classification supervisée les plus courantes peut être utilisé pour effectuer une classification de texte. Tous les textes de tweet seront transformés en vecteurs de terme pour être traités par l'algorithme. Habituellement, le terme vecteur est généré à partir d'un vocabulaire unique, généré à partir de l'ensemble de données d'apprentissage, et il n'y a pas de mots en double dans le vocabulaire. La taille du terme vecteur est la taille du vocabulaire.

Le classificateur Naive Bayesian est basé sur le théorème de Bayes avec les hypothèses d'indépendance entre les prédicteurs. Un modèle bayésien naïf est facile à construire, sans estimation compliquée des paramètres itératifs, ce qui le rend particulièrement utile pour de très grands ensembles de données. En dépit de sa simplicité, le classificateur Naive Bayesien fait souvent étonnamment bien et est largement utilisé parce qu'il surpasse souvent les méthodes de classification plus sophistiquées **[23]**.

Le théorème de Bayes fournit un moyen de calculer la probabilité postérieure, P (c | x), à partir de P (c), P (x) et P (x | c). Les classificateurs Naive Bayes supposent que l'effet de la valeur d'un terme (x) sur une classe donnée (c) est indépendant des valeurs d'autres termes. Cette hypothèse est appelée indépendance conditionnelle de classe.

$$P(C|X)\frac{P(X|C)P(C)}{P(X)}$$

P(C|X)=P(x1|c)*P(x2|c)……P(Xn|c)P(c).

- P (c | x) est la probabilité a posteriori de la classe (cible) donnée avec prédicteur (attribut).
- P (c) est la probabilité a priori de la classe.
- P (x | c) est la vraisemblance qui est la probabilité de l'attribut de donné la classe c.
- P (x) est la probabilité préalable du prédicteur.

3.3. Algorithme arbre de décision :

L'arbre de décision se situe dans le cadre de l'apprentissage supervisé, à partir d'une base d'apprentissage, on construit l'arbre (un modèle prédictif) dont chaque chemin depuis la racine jusqu'à une feuille correspond à une règle de classement.

Le but est de construire un modèle à partir d'un ensemble d'exemples associés aux classes pour trouver une description pour chaque classe à partir des propriétés communes entre les exemples. Une fois ce modèle construit, on peut extraire un ensemble de règles qui vont être utilisés pour classer de nouveaux textes dont la classe est inconnue **[23]**.

3.3. 1. Construction de l'arbre de décision :

Le principe de construire un arbre de décision x consiste à diviser les exemples de la base d'apprentissage récursivement en se basant sur l'idée de Top-Down Induction On commence par construire la racine de l'arbre en continuant jusqu'à la feuille par des teste défini à l'aide des attribues c'est à dire jusqu'à obtenir des sous ensemble d'exemple ne contenant que des exemples appartenant tous à une même classe.

C4.5 construit l'arbre de décision récursivement comme et à chaque itération il calcule l'entropie E(S) (gain ratio) de chaque attribue de la base d'apprentissage. L'attribut ayant le plus grand gain ratio sera choisis comme racine pour produire des sous-ensembles .l'algorithme continuera d'une façon récursive pour chaque sous ensemble d'attributs.Lorsque tous les éléments dans un sous ensemble appartient à la même classe , ce sous ensemble ne sera plus parcouru et ce nœud dans l'arbre de décision devient un nœud terminal étiqueté avec une étiquette de la même classe que la classe où tous ses éléments appartiennent. L'algorithme c4.5 se termine lorsque tous les sous-ensembles seront classés.

- **Entropie :** ID3 algorithme utilise l'entropie pour calculer l'homogénéité de l'échantillon. Si l'échantillon « S » est complètement homogène l'entropie est nulle et si l'échantillon est un partage égal des voix qu'il a une entropie d'un seul. E(X) comme la quantité d'information apportée par la réalisation de l'événement « x ». Donc :

$$\mathbf{E(X) = -P(xi)logP(xi)}$$

- P(xi) : probabilité de l'attribut xi.

- **Gain d'information :** Le gain d'information est basée sur la diminution de l'entropie après un ensemble de données est divisé en un attribut. Construire un arbre de décision est tout attribut de trouver qui retourne le plus grand gain d'information (les branches les plus homogènes).

$$\mathbf{Gain(T,X)=Entropy(T)-Entropy(T,X)}$$

- Entropy(T) : entropie de la classe.
- Entropy(T,X) : entropie de l'attribut x par rapport à la classe.

- **Gain ratio :** Le c4.5 est basé sur le principe du gain ratio pour choisir l'attribut racine

$$\mathbf{Gainratio(T,X)=\frac{Gain(T,X)}{Splitinfo(x)}}$$

$$\textbf{Splitinfo(x)}=-\sum_{V\in X} \boldsymbol{P(V)log_2P(V)}$$

- **V:** valeur de l'attribut x.

4. Evaluation :

Les mesures d'évaluation utilisées pour évaluer nos algorithmes sont différentes et chaque mesure a un objectif d'utilisation comme le montre les parties suivants.

4.1. Matrice de confusion :

Matrice de contingence		Jugement de expert	
		Vrais	Faux
Jugement de l'algorithme	Vrais	VP_i	FP_i
	Faux	FN_i	VN_i
Vrais positive (VP)	Le nombre d'instances attribués à une catégorie convenablement (dépressive).		
Vrais Négative (VN)	Le nombre d'instances correctement attribués à la classe non dépressive. (Qui n'ont pas à être attribués à une catégorie, et ne l'ont pas été)		
Faux positive (FP)	Le nombre d'instances non dépressive et qui ont été attribués à la classe dépressive. (instances attribués à des mauvaises catégories)		
Faux négative (FN)	Le nombre d'instances réellement dépressive et qui ont été attribué à la classe non dépressive. (Qui auraient dû être attribués à une catégorie mais qui ne l'ont pas été)		

Tableau III .2: matrice de confusion

4.2. Les tweets correctement analysés :

Le pourcentage de personnes bien classées (comme dépressive et non dépressive).

4.3. Les tweets incorrectement analysés :

Le pourcentage de personnes mal classées (comme dépressive et non dépressive).

4.4. Kappa static (K):

L'évaluation de l'étendue de l'accord entre 2 ou plusieurs évaluateurs est courante en sciences sociales, comportementales et médicales. Les deux évaluateurs sont l'algorithme et la classe réelle de l'exemple. La cohérence entre les deux évaluateurs est lit dans la matrice de confusion. La valeur de K est toujours comprise entre -1 et 1.

- K=1 si les algorithmes et le jugement de l'expert sont les mêmes.
- K=-1 si les algorithmes et le jugement de l'expert sont complètement différent.

$$K=\frac{P_0-P_C}{1-P_C}$$

- P_0 : nombre des personnes bien classé

$$P_C = \frac{\sum_i A_i * R_i}{total^2}$$

- Total : nombre des instances total dans le dataset.
- A_i : somme des éléments de la ligne i de la matrice de confusion.
- R_i : somme des colonnes de la ligne i de la matrice de confusion.

4.5. TP rate : Le taux du true positifs.

$$\textbf{TP} = \frac{\textbf{Nombre de vrais positifs}}{\textbf{nombre d'exemplesde cette classe}} = \frac{\textbf{Nombre de vrais positifs}}{\textbf{Nombre de vrais positifs+nombre de vrais négatifs}}$$

C'est le rapport entre le nombre de bien classé et le nombre total des personnes qui devrais être bien classé

4.6. FP rate : Le taux des faux positifs

$$\textbf{FP} = \frac{\textbf{Nombre de faux positifs}}{\textbf{nombre es personnes n'étantpas de cette classe}} = \frac{\textbf{Nombre de faux positifs}}{\textbf{Nombre de faux positifs+nombre de vrais négatifs}}$$

4.7. Précision (P):

La précision permet de mesurer la capacité d'un algorithme à retourner seulement les personnes dépressive. Elle représente le rapport entre le nombre des personnes correctement classé par l'algorithme dans la classe dépressive par rapport au nombre des personnes total classées par l'algorithme dans la classe dépressive.

$$P = \frac{VP_i}{VP_i + FP_i}$$

4.8. Rappel (R) :

Le rappel mesure la capacité de notre système à retourner les instances bien classées. Elle représente le rapport entre le nombre d'instances correctement classés par notre système dans la classe dépressive par rapport au nombre total des documents réellement dans la classe c.

$$R = \frac{VP_i}{VP_i + FN_i}$$

4.9. f-mesure :

Permet de regrouper en un seul nombre la performance du l'algorithme. Elle est basée sur les résultats du rappel et précision.

$$\mathbf{F} = \frac{2 * R * P}{R + P}$$

III. Phase d'implémentation et Conception :

1. Conception

1.1 Le langage de modélisation :

UML (en anglais Unified Modeling Language ou « langage de modélisation unifié ») est un langage de modélisation graphique à base de pictogrammes. Il est apparu dans le monde du génie logiciel, dans le cadre de la « conception orientée objet ». Couramment utilisé dans les projets logiciels, il peut être appliqué à toutes sortes de systèmes ne se limitant pas au domaine informatique. En effet, l'UML nous permet une meilleure conception du côté de l'application avec ses notions d'objets et de classes, et nous donne une décomposition claire et simple afin de dégager les entités et les classes nécessaires.

1.2 Analyse :

L'analyse est la phase qui répond à la question « que faut-il faire ? », elle a pour but de se doter d'une vision claire et rigoureuse du problème posé et du système à réaliser en déterminant

Ses éléments et leurs interactions. Nous allons commencer par une analyse des besoins.

1.2.1 Analyse des besoins :

L'analyse des besoins(ou d'application) nous permet d'effectuer la modélisation des fonctionnalités de l'application les plus importantes.

Les uses cases permettent de structurer les besoins de l'utilisateur et les objectifs correspondants d'un système. Ils centrent l'expression des exigences du système sur l'utilisateur : il part du principe que les objectifs du système sont tous motivés.

Cette phase propose une réalisation de l'analyse et des cas d'utilisation en prenant en compte toutes leurs exigences.

1.3 Diagramme de séquence système (DSS):

Nous allons dans cette étape présenter graphiquement sur des diagrammes de séquence UML.

Pour chaque étape, le DSS montre non seulement le User externes qui interagissent avec le système, mais également les événements système déclenchés par le User.

Nous allons présenter le DSS correspondant à chaque cas d'utilisation développée.

Ci-dessous le DSS correspondant au cas d'utilisation « Générale» :

Diagramme de Séquence :

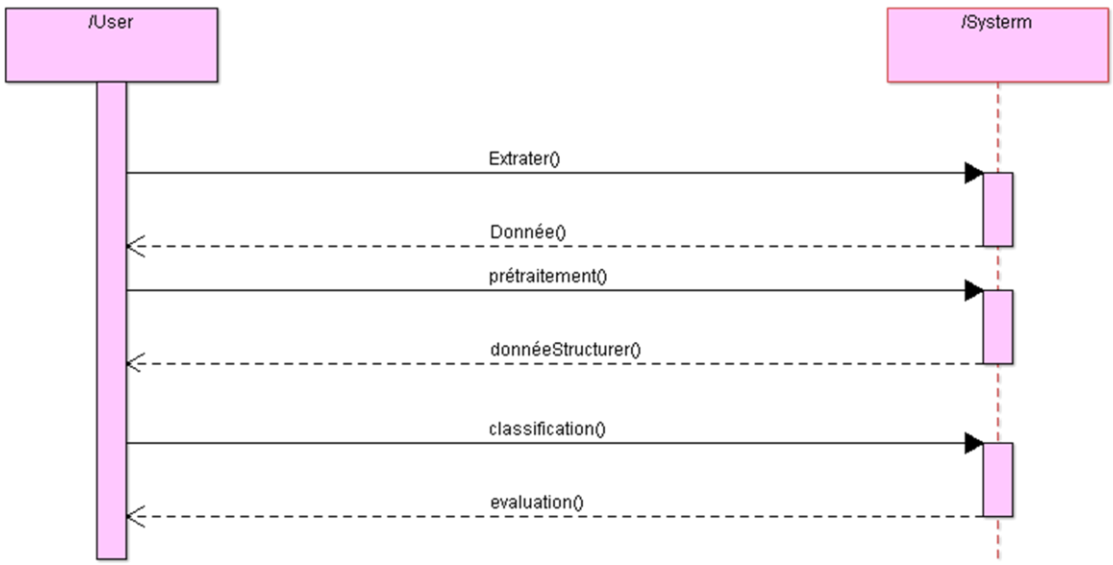

Figure III.3 : Diagramme de Séquence Système «Générale »

Ci-dessous le diagramme de séquence système « Extraction » :

Diagramme de Séquence :

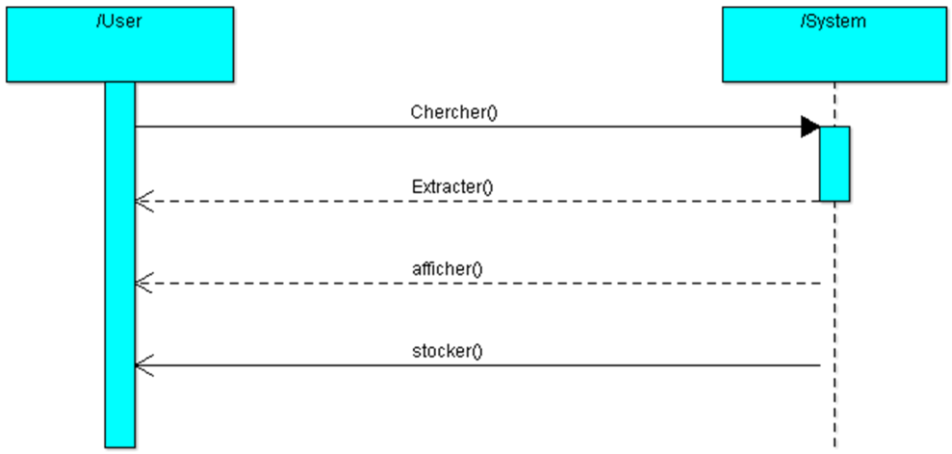

Figure III.4 : Diagramme de Séquence Système « Extraction»

Ci-dessous le diagramme de séquence système « Prétraitement»

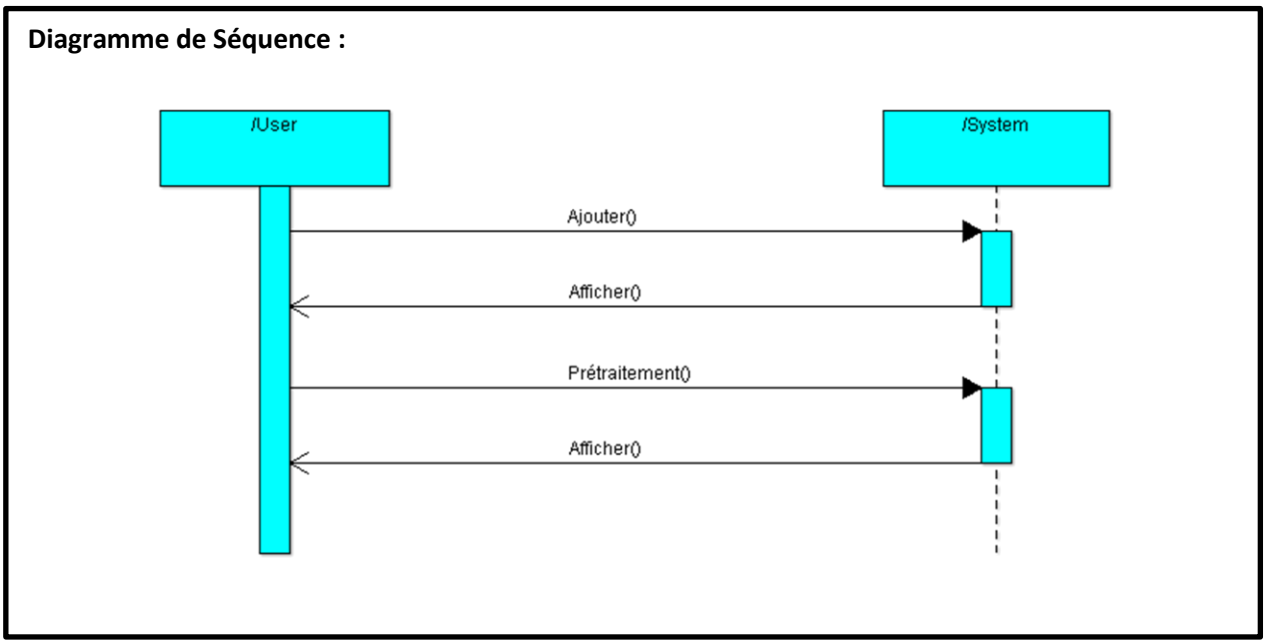

Figure III.5 : Diagramme de Séquence Système « prétraitement»

Ci-dessous le diagramme de séquence système « Classification » :

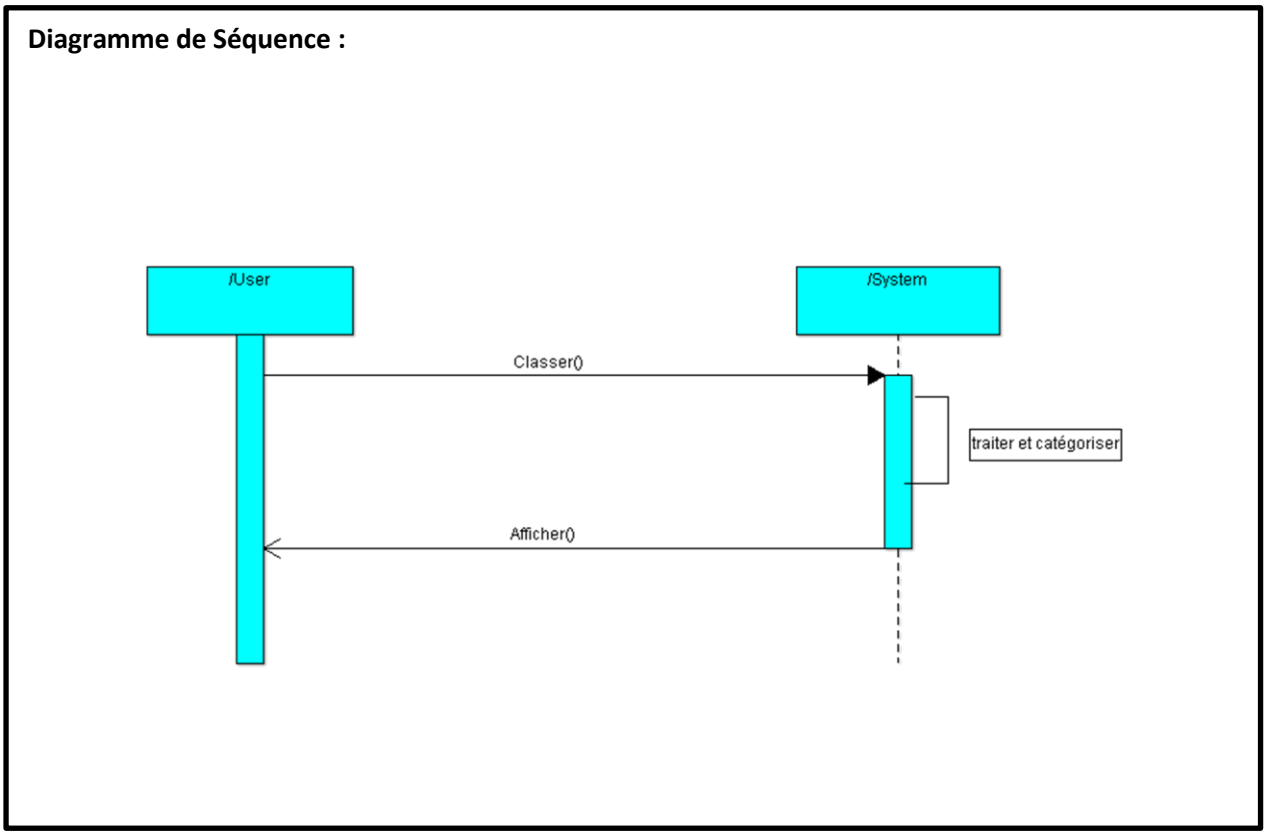

Figure III.6: Diagramme de Séquence Système « Classification»

1.4 Diagramme de classe de conception :

Les diagrammes de classes de conception représentent bien la structure statique du code, par le biais des attributs et des relations entre classes, mais ils contiennent également les opérations (aussi appelées méthodes) qui décrivent les responsabilités dynamiques des classes logicielles. L'attribution des bonnes responsabilités aux bonnes classes est l'un des problèmes les plus délicats de la conception orientée objet. Pour chaque service ou fonction, il faut décider quelle est la classe qui va le contenir.

Nous devons ainsi répartir tout le comportement du système entre les classes de conception, et décrire les collaborations induites.

Diagramme de Class :

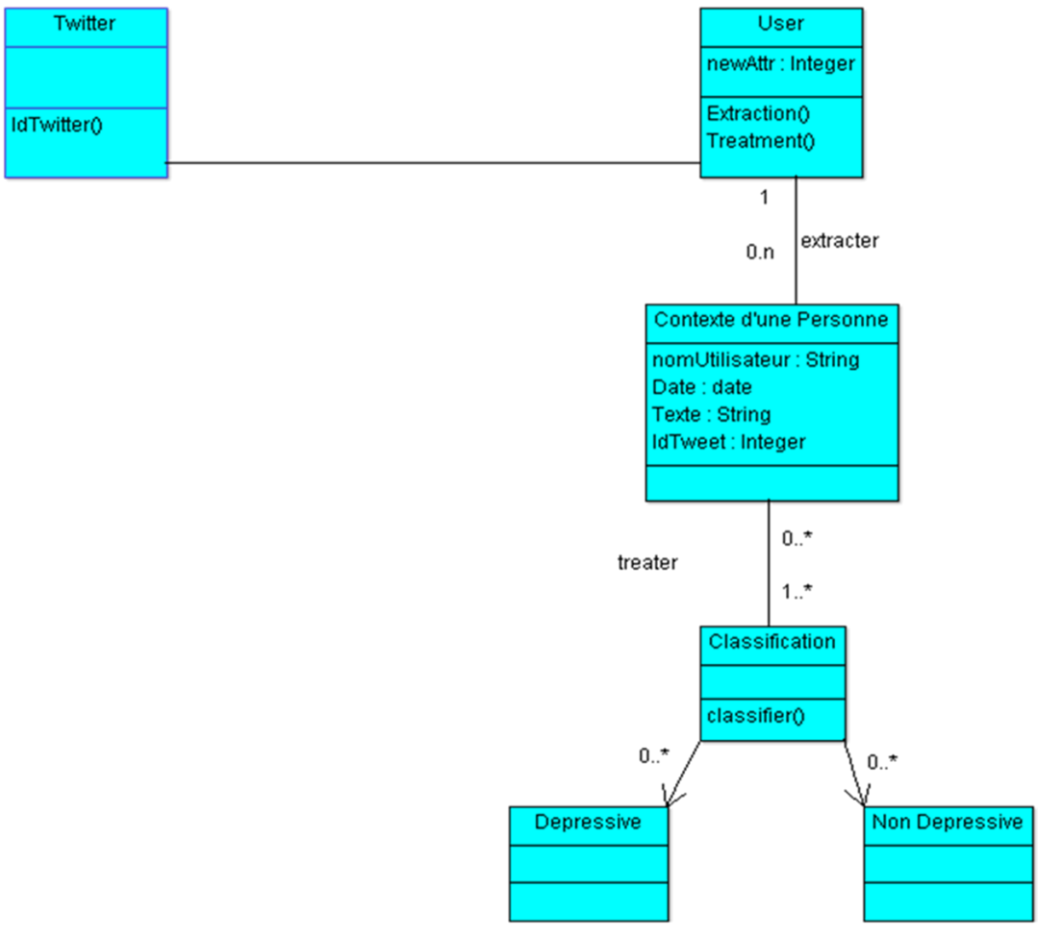

Figure III.7 : Diagramme de Class

2. Implémentation

Cette phase est importante pour décrire les étapes de mise en œuvre déférentes pour notre système et la structure de notre application. Pour la réalisation de notre travail nous avons utilisé java comme langage de programmation et eclips version ADT comme IDE avec quelques API comme jtatoo, jsplite, jfreechart. La figure suivante présente les classes qui composent l'application:

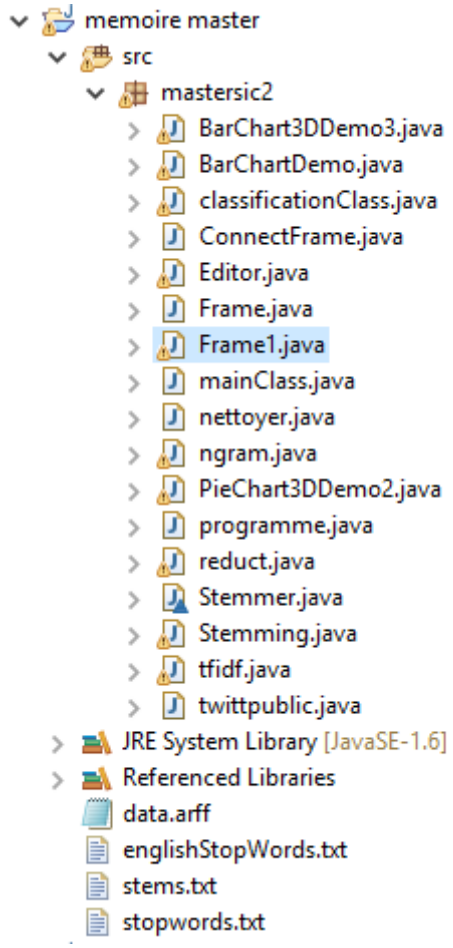

Figure III .8 : les classes de notre approche

Le package mastersic2 est composé de trois parties :

- Les classes (stemmer, n-gram, tfidf, stemming, nettoyer) ont un rôle qui est le prétraitement des données pour la vectorisations des tweets.
- Les classes (classificationclasse, programme) ont un role c'est la classification des tweet.
- Les classes (barchart3D et barchart) ont un role de visualisation des résultats.

IV. Résultat et discussion :

Dans cette partie nous allons à chaque fois tester les différentes techniques de représentations afin de fixer les paramètres idéals pour le problème d'analyse des tweets. Nous allons diviser cette partie en 3 comparaisons :

- Comparaison en terme de représentation (sac de mots, stemming, n-gram caractères).
- Comparaison entre les algorithmes (naive bayes, arbre de décision c4.5 et KPPV).
- Comparaison avec les algorithmes bioinspirée (intégrés dans l'outil de l'EBIRI).

Les tableaux suivant regroupent les meilleurs résultats obtenus après différent teste les cases colorés en bleu signifier les meilleurs résultats et les cases colorés en rouge signifiés les mauvais résultats obtenus pour chaque algorithme avec la variation des techniques de représentation et le codage tf*idf.

1. Le Tableau III .3 et les figures Figure III .9 et III .10 regroupent les meilleurs résultats obtenus par l'algorithme Kplus proches voisins (K=1 et distance cosinus) avec la variation des techniques de représentation et le codage tf*idf.

		Evaluation mesures							
		Précision	**Rappel**	**f-mesure**	**TS(%)**	**TE(%)**	**Kappa statique**	**Matrice de confusion**	
Techniques de représentation	**Sac de mots**	0. 724	0. 699	0.7	67.79%	32.21%	0.354	258	98
								111	182
	Stemming	0. 786	0.617	0.6913	68.72%	31.28%	0.386	228	62
								141	218
	2-gramme caractères	0.819	0.7235	0.769	75.19%	24.81%	0.5038	267	59
								102	221
	3-gramme caractères	0.86	0.764	0.811	79.81	20.19	0.596	282	44
								87	236
	4-gramme caractères	0.918	0.791	0.854	84.12%	15.86%	0.688	292	26
								77	254
	5-grammes caractères	0.844	0.764	0.802	78.58	21.42	0.56	282	52
								87	228

Tableau III .3: les résultats d'analyse utilisant l'algorithme Kplus proche voisins et la variation des techniques de représentation (K=1 et distance cosinus).

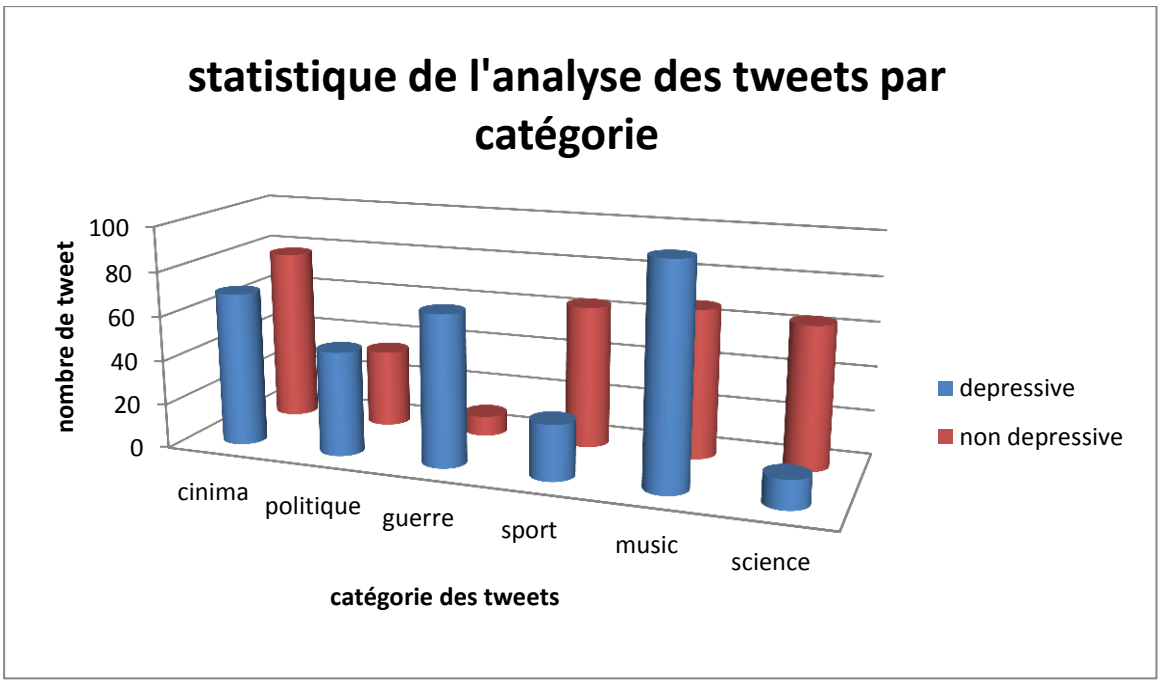

Figure III .9 : Nombre des tweets dépressive et non dépressive classer par catégorie obtenu après l'analyse de l'algorithme K plus proches voisins (K=1 et distance cosinus)

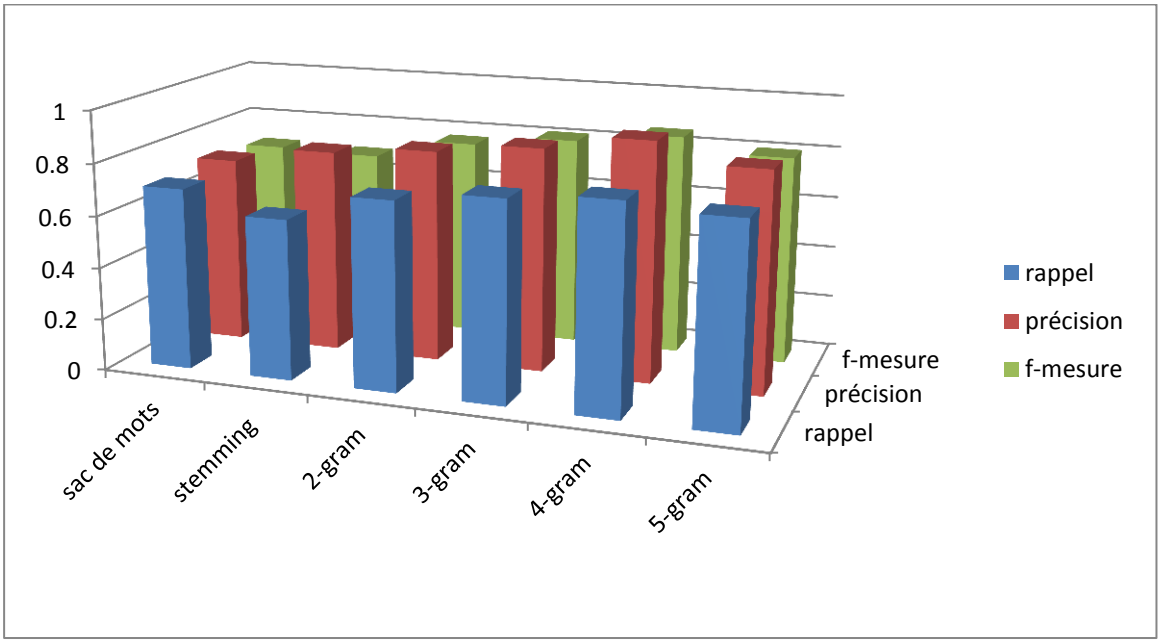

Figure III .10: Comparaison entre les techniques de représentation en utilisant l'algorithme Kplus proches voisins (K=1 et distance cosinus).

2. Le Tableau III .4 et les figures Figure III .11 et III .12 regroupent les meilleurs résultats obtenus par l'algorithme naive bayes avec la variation des techniques de représentation et le codage tf*idf.

		Mesure d'évaluation							
		Précision	**Rappel**	**f-mesur**	**TS(%)**	**TE(%)**	**Kappa statique**	**Matrice de confusion**	
Technique de représentation	**Sac de mots**	0.74	0.59	0.654	65.48%	34.52%	0.313	221	76
								148	204
	Stemming	0.64	0.56	0.6	57.62%	42.37%	0.156	207	113
								162	167
	2-gramme caractére	0.7222	0.577	0.645	63.32%	36.67%	0.274	213	82
								156	198
	3-gramme caractére	0.781	0.715	0.745	72.41%	27.59%	0.45	264	74
								105	206
	4-gramme caractéres	0.7217	0.674	0.699	66.71	33.29	0.347	249	96
								120	184
	5-grammes caractéres	0.708	0.672	0.689	65.63	34.37	0.32	248	102
								121	178

Tableau III .4: les résultats d'analyse utilisant l'algorithme naïve bayes et la variation des techniques de représentation.

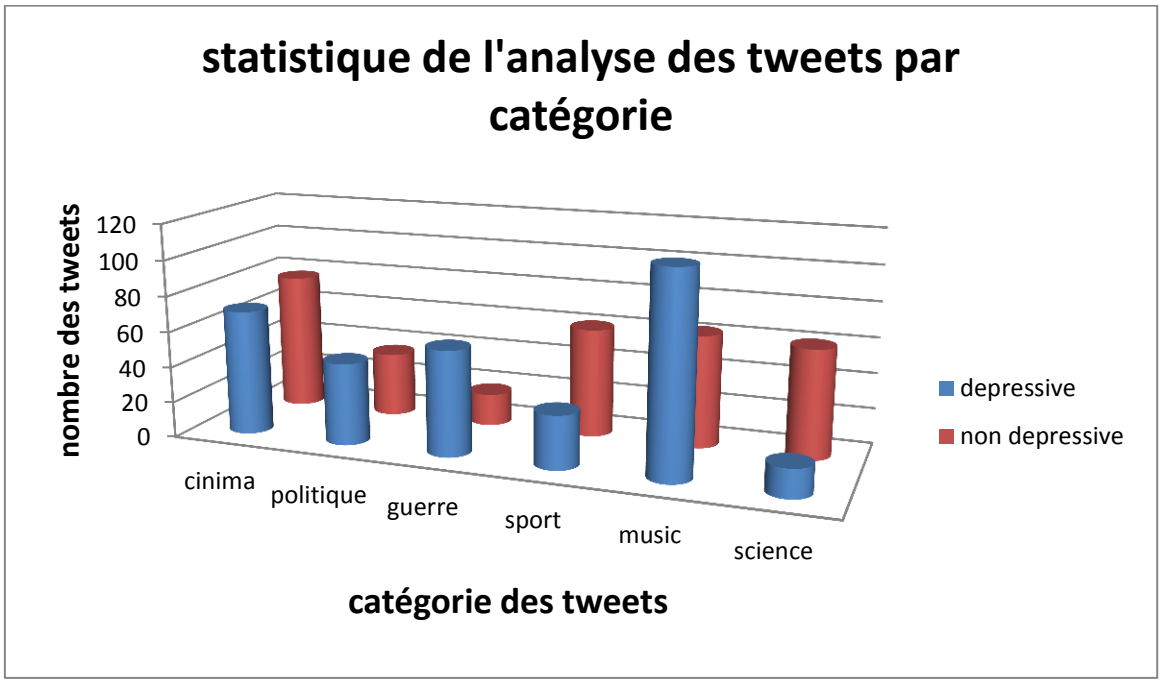

Figure III .11: Nombre des tweets dépressive et non dépressive classer par catégorie obtenu après l'analyse de l'algorithme naïve bayes

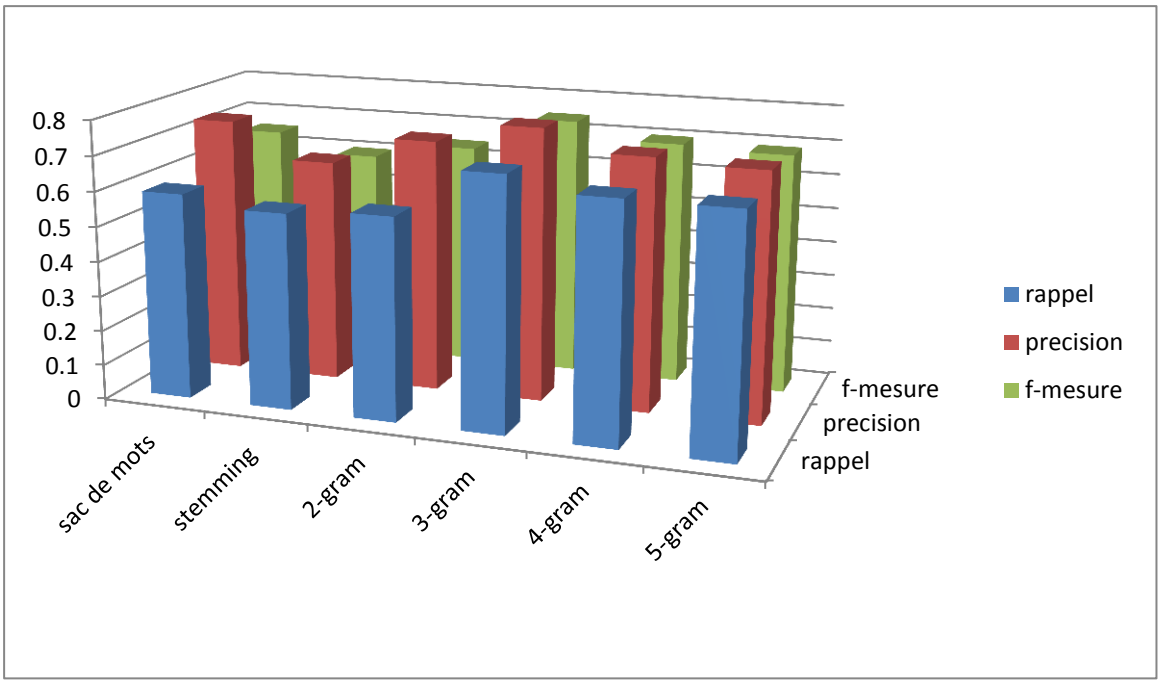

Figure III .12: Comparaison entre les techniques de représentation en utilisant l'algorithme naive bayes en termes de rappel, précision et f-mesure

3. Le Tableau III .5 et les figures Figure III .13 et III .14 regroupent les meilleurs résultats obtenus par l'algorithme C4.5 avec la variation des techniques de représentation et le codage tf*idf.

		Mesure d'évaluation							
		Précision	**Rappel**	**f-mesure**	**TS(%)**	**TE(%)**	**Kappa statique**	**Matrice de confusion**	
Technique de représentation	**Sac de mots**	**0.52**	**0.509**	**0.51**	**46.22%**	**53.78%**	**- 0.05**	**188**	**168**
								181	**112**
	Stemming	**0.526**	**0.463**	**0.49**	**45.76%**	**54.24%**	**-0.06**	**171**	**154**
								198	**126**
	2-gramme caractères	**0.608**	**0.51**	**0.558**	**53.62%**	**46.48%**	**0.09**	**190**	**122**
								179	**158**
	3-gramme caractères	**0.6**	**0.5257**	**0.557**	**53.15%**	**47.85%**	**0.083**	**194**	**129**
								175	**151**
	4-gramme caractères	**0.6**	**0.536**	**0.569**	**53.31%**	**47.69%**	**0.06**	**198**	**132**
								171	**148**
	5-grammes caractères	**0.613**	**0.533**	**0.57**	**54.39%**	**45.61%**	**0.098**	**197**	**124**
								172	**156**

Tableau III .5: les résultats d'analyse utilisant l'algorithme c4.5 et la variation des techniques de représentation.

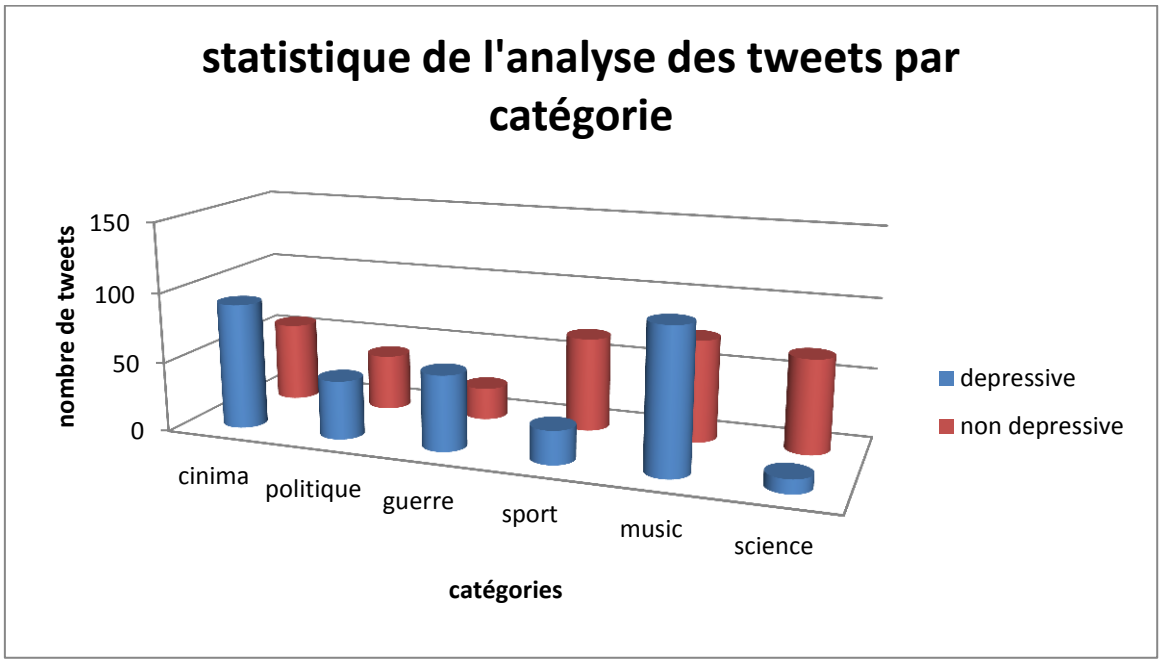

Figure III .13: Nombre des tweets dépressive et non dépressive classer par catégorie obtenu après l'analyse de l'algorithme arbre de décision c4.5.

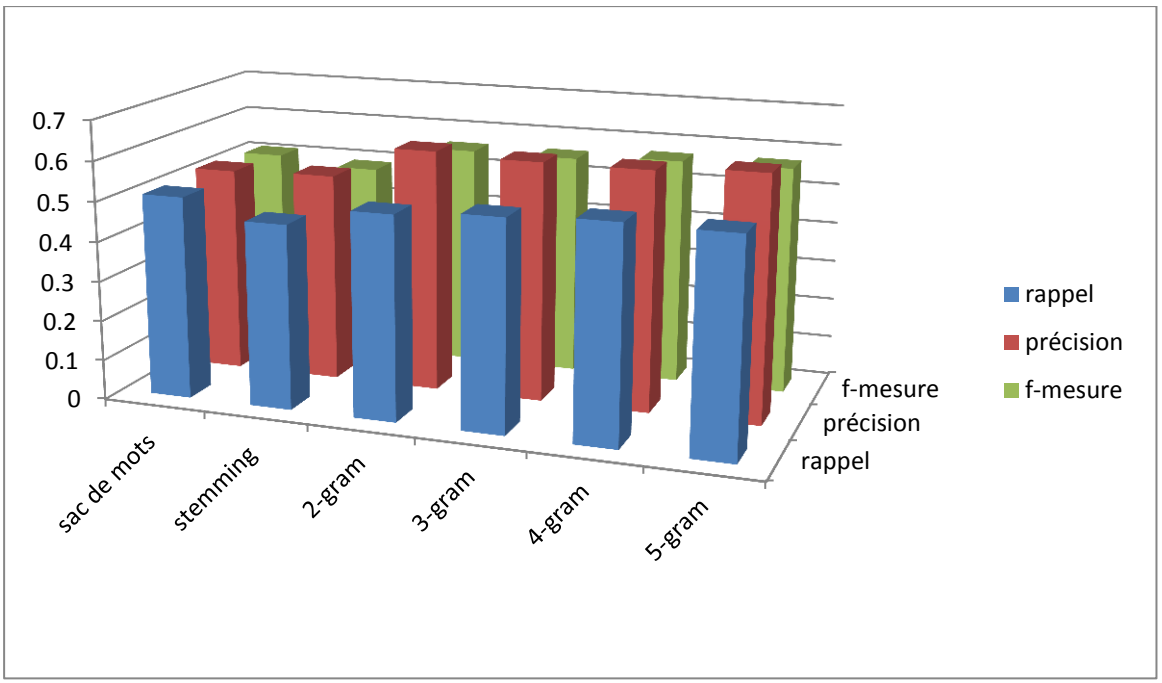

Figure III .14: Comparaison entre les techniques de représentation en utilisant l'algorithme arbre de décision c4.5 en termes de rappel, précision et f-mesure.

1. Discussions des résultats:

- **En terme de représentation :** Les résultats précédents montrent clairement que la technique n-gramme caractères donne les meilleurs résultats par rapport aux techniques (stemming, sac de mot) parce que les utilisateurs du twitter peuvent écrire leurs tweets en utilisant différents langue ce qui a provoquer des difficultés au technique sac de mots et stemming a bien représenté ces tweets.
- En termes de l'algorithme d'analyse : le K plus proches voisins donne meilleur résultats avec la représentation 4-gram caractère et codage tf*idf.

V. Comparaison externe :

Cette partie permet de comparer les meilleurs résultats obtenus par les algorithmes du datamining classique avec les résultats des algorithmes bio-inspirées comme l'algorithme des cafards **[24]** et machine cœur poumons **[25]** intégrés dans la boite à outil EBIRI **[26]** comme le montre le tableau suivant.

		Mesure d'évaluation							
		Précision	**Rappel**	**f-mesure**	**TS(%)**	**TE(%)**	**Kappa statique**	**Matrice de contingence**	
Les algorithmes	**Naive bayes**	0.781	0.715	0.745	72.41%	27.59%	0.45	264	74
								105	206
	Arbre de décisio	0.608	0.51	0.558	53.62%	46.48%	0.09	190	122
								179	158
	KPPV	0.918	0.791	0.854	84.12%	15.86%	0.688	292	26
								77	254
	Algorithme des cafards	**0.92**	**0.74**	**0.82**	**82.43**	**17.57**	**0.65**	**276**	**21**
								93	**259**
	Machine cœur poumons	**0.88**	**0.82**	**0.848**	**84.28**	**15.72**	**0.68**	**306**	**39**
								63	**241**

Tableau III .6: comparaison entre les algorithmes d'apprentissage supervisés classiques et les algorithmes bio-inspirés.

VI. Conclusion :

Dans notre travail nous avons sortie avec plusieurs décisions qui peuvent être utilisée par d'autres chercheurs et étudiants :

1. Les algorithmes supervisés classique donnent des meilleurs résultats comparés aux algorithmes bio-inspiré.
2. La technique n-gramme donne toujours les meilleurs résultats dans le domaine d'analyse des tweets.
3. L'algorithme k plus proches voisins donne des meilleurs résultats comparé aux autres algorithmes parce qu'il est basé sur un principe simple mais il nécessite beaucoup du temps.
4. Il y'a plus de possibilité qu'une personne soit dépressive si son message parle de la music et du sport par contre il y'a moins de possibilité qu'une personne soit non dépressive si son message parle du science ou du cinéma.

.

Chapitre 4 :

Réalisation

Dans ce chapitre nous avons discuté sur les outils et langage utilisé. Puis Nous avons présenté quelques interfaces de notre application.

I. Introduction :

La réalisation vient couronner le travail de l'étude préalable et de l'étude conceptuelle. Elle présente la dernière étape et elle est très importante puisque grâce à elle le projet informatique va exister réellement, sa réussite est conditionnée par une multitude de choix essentiellement d'ordre technique concernant l'exécution de ce qui a été conçu et proposé comme solution afin de répondre aux besoins des utilisateurs et de remédier aux insuffisances perçues.

Ce chapitre est donc consacré à la présentation de l'environnement matériel et logiciel nécessaire pour implémenter cette structure tout en veillant à garantir les spécificités de sécurité et les services de gestion énoncés au cahier des charges.

II. Outils et langage utilisé :

1. Spécification technique :

Une étape intéressante de ce projet était la mise en place de l'environnement matériel et logiciel nécessaire pour la conception, le développement et le test de l'application. Dans ce qui suit, nous présenterons l'environnement logiciel et matériel exploité dans notre projet.

1.1 Configuration matérielle :

- Ordinateur portable : HP
- Système d'exploitation : Windows 7
- Processeur : Core I3
- Mémoire : 4 G RAM
- Disque dur : 512 Gb

1.2 Configuration logicielle :

1.2.1 Eclipse IDE :

Eclipse IDE est un environnement de développement libre permettant de créer des programmes dans de nombreux langages de programmation (Java, C++, PHP…). C'est l'outil que nous allons utiliser pour programmer.

Figure IV.1 : Logo Eclipse

1.2.2 JRE :

L'environnement d'exécution Java (abr. *JRE* pour *Java Runtime Environment*), parfois nommé simplement « Java », est une famille de logiciels qui permet l'exécution des programmes écrits en langage de programmation Java, sur différentes plateformes informatiques.
Il est distribué gratuitement par Oracle Corporation, sous forme de différentes versions destinées aux systèmes d'exploitation Windows, Mac OS X et Linux, toutes conformes aux Java Specification Requests (JSR).

Figure IV.2 : logo JRE

1.2.4 Java 8 (langage de programmation) :

Java 8 est la dernière version de Java et offre de nouvelles fonctionnalités, des performances accrues et des corrections de bug pour améliorer l'efficacité de développement et d'exécution des programmes Java. La nouvelle version de Java est d'abord mise à disposition des développeurs afin qu'ils disposent du temps adéquat pour effectuer les opérations de test et de certification ; les utilisateurs finals pourront ensuite la télécharger sur le site Web java.com.

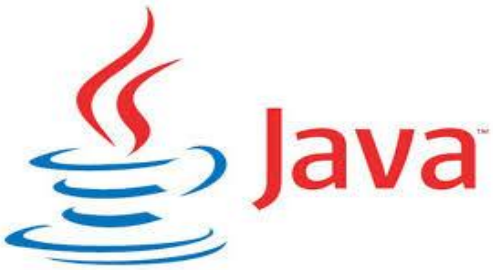

Figure IV .3 : logo java

1.2.5 C'est quoi WEKA ? :

Weka (Waikato Environnement for KnowledgeAnalysis) apprentissage automatique établi pour permettre de manipuler et d'analyser des fichiers de données, c'est un ensemble d'algorithmes qui peuvent soit être appliqués directement soit appeler à partir d'un code Java.

Weka a été développé à l'université de Waikato en Nouvelle-Zélande, et a été mis en œuvre dans sa forme moderne en 1997 et le nom est synonyme de Waikato Environnement pour l'analyse des connaissances.

En dehors de l'université du Weka, est un oiseau d'une nature curieuse trouvée que sur les îles de la Nouvelle-Zélande. Le système est écrit en Java et distribué sous les termes de la GNU Général Public licence. Il fonctionne sur presque toute plate-forme et a été testé sous Linux, Windows et Macintosh systèmes d'exploitation.

Il fournit une interface uniforme à de nombreux algorithmes d'apprentissage différents, ainsi que des méthodes pour prétraitement et pour évaluer le résultat sur tout ensemble de données d'apprentissage.

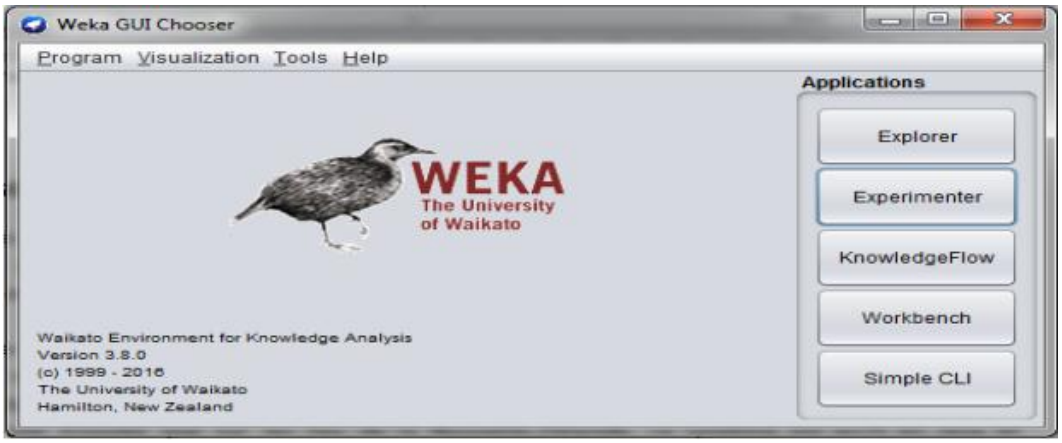

Figure IV.4 : Écran de démarrage WEKA.

1.2.6 La campagne TREC :

Le mot TREC signifie «Text Retrieval Conference» et désigne l'ensemble des conférences organisées par le NIST «National Institute of Standard and Technology» sur la recherche d'information. **[28]**

Cette compagne a organisé entre 2006 et 2008 une tâche de recherche d'information dans la blogosphère, qui comportait une tâche de recherche d'opinion, elle avait deux objectifs. La recherche de billets de blogs pertinents (baseline adhoc blog post retrieval task), une séparation des posts de blogs en objectif/subjectif (ceux qui expriment, ou non, une opinion sur une cible donnée). La recherche d'opinion dans les blogs (polarised opinion finding blog post retrieval task) et une séparation des posts en opinion positive/négative, avec un classement dans l'ordre de positivité/négativité décroissante Cette tâche a été introduite pour la première fois dans TREC 2007**. [27]**

La campagne d'évaluation TREC propose chaque année quelques tâches : La recherche de blog dont le principal intérêt porte sur un sujet (blog finding distillation task). Cette tâche a été introduite pour la première fois dans TREC 2007 et traitée aussi dans TREC 2008. **[29]**

La recherche de blog dont le principal intérêt porte sur un sujet, et tenant compte de certaines facettes (Faceted blog distillation task). Trois facettes ont été spécifiées pour TREC 2009 .

La première porte sur l'opinion (la valeur de cette facette est égale à «opinionated» ou bien à «factual »).

La deuxième facette porte sur le caractère personnel ou officiel des documents recherchés (la valeur de la facette est «personale» ou «official »).

La troisième facette est égale à «in depth» si l'analyse sur le topic est importante autrement elle est égale à «shallow». Cette tâche a été aussi proposée dans TREC2010.**[28]**

Une catégorisation des nouvelles : international, national, politique, sport, technologie, business, science) a été proposée pour TREC 2010 **[28]**. Plusieurs universités et laboratoires de recherche participent à ces conférences TREC etutilisent les collections de test (une collection de test comporte trois parties, un ensemble de documents, un ensemble de topics et les jugements de pertinence «relevance judgements») proposées par TREC pour évaluer leurs résultats.

Dix-sept groupes ont participé à TREC2006, vingt-quatre à TREC2007 (vingt pour la fouille d'opinions, onze pour la polarité et neuf pour le blog distillation), vingt groupes dans TREC2008 (vingt pour la baseline, dix-neuf pour la fouille d'opinions, seize pour la polarité et douze pour le blog distillation).

Pour l'évaluation de l'efficacité des systèmes de fouille d'opinions TREC utilise la technique du pooling, cette dernière consiste à choisir un certain nombre de runs soumis par les participants, (Le «run» est le résultat d'une exécution d'un système de recherche exécutant une tâche sur une collection de test), ensuite à choisir parmi ces runs les n (généralement n = 100) premiers documents en éliminant les doublons.

Le run soumis sera évalué, cette évaluation est faite par le logiciel TREC_EVAL à travers les mesures d'évaluation plus utilisées sont le rappel, la précision (qui est le ratio entre le nombre de documents pertinents retrouvés et le nombre total de documents retrouvés), la MAP la R-Précision. **[28]**

Figure IV.5 : représentation de TRE**C**

IV. Réalisation du projet:

Nous avons présenté dans cette partie quelques interfaces de notre application.

- **Interface de l'application :**

Dans la premier interface de l'application, l'utilisateur peut faire l'extraction Api des tweets et faire un prétraitement utilisons les boutons voir la **figure IV.6.**

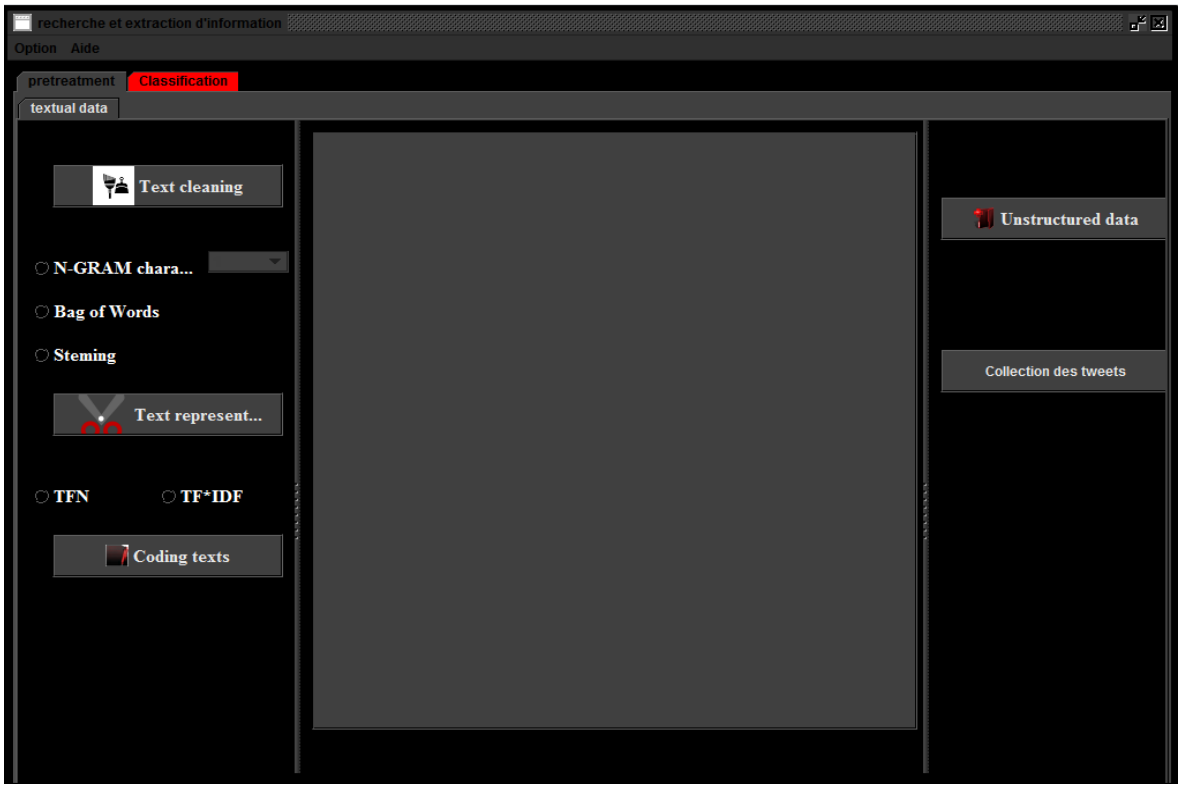

Figure IV.6 : Interface 1 de l'application

❖ **L'importation des données non structuré :**

En cliquant sur le bouton « unstructured data » un menu qui s'affiche qui nous permet de choisir le dataset qu'on va traiter.

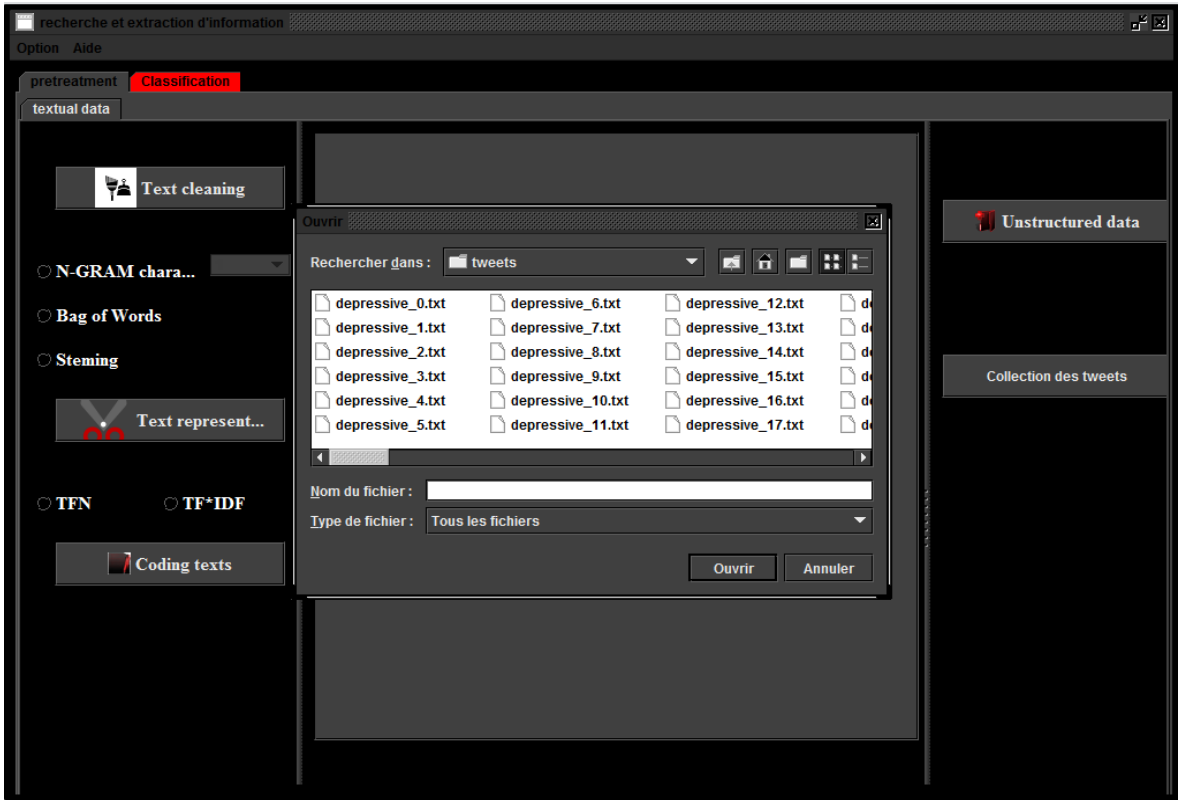

Figure IV.7 : Import Dataset « non structuré »

En cliquant sur le bouton « Ouvrir », le dataset sera affiché. Apres en cliquant sur le bouton « Text Cleaning » pour supprimer les caractères spéciaux les mots vide.

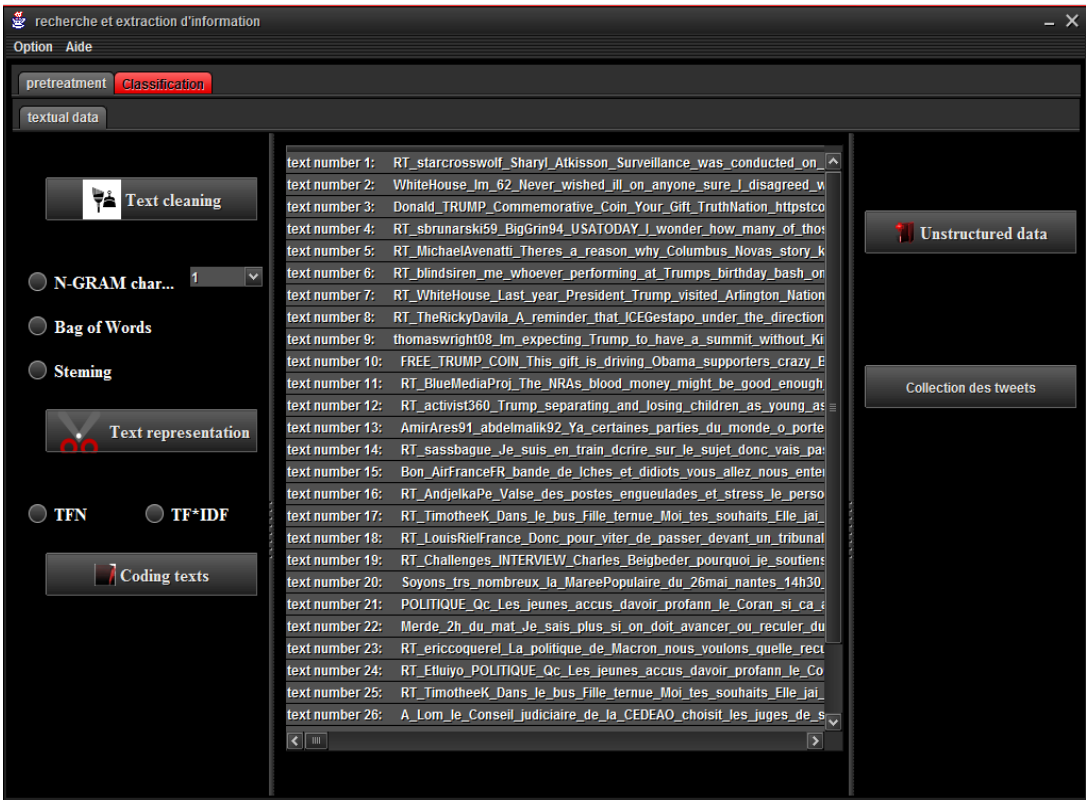

Figure IV.8: Interface 1 « Text cleaning »

En cliquant sur le bouton « Text representation » pour avoir une pondération binaire. Et nous avons le choix « N-gramme char ou sac de mot ».

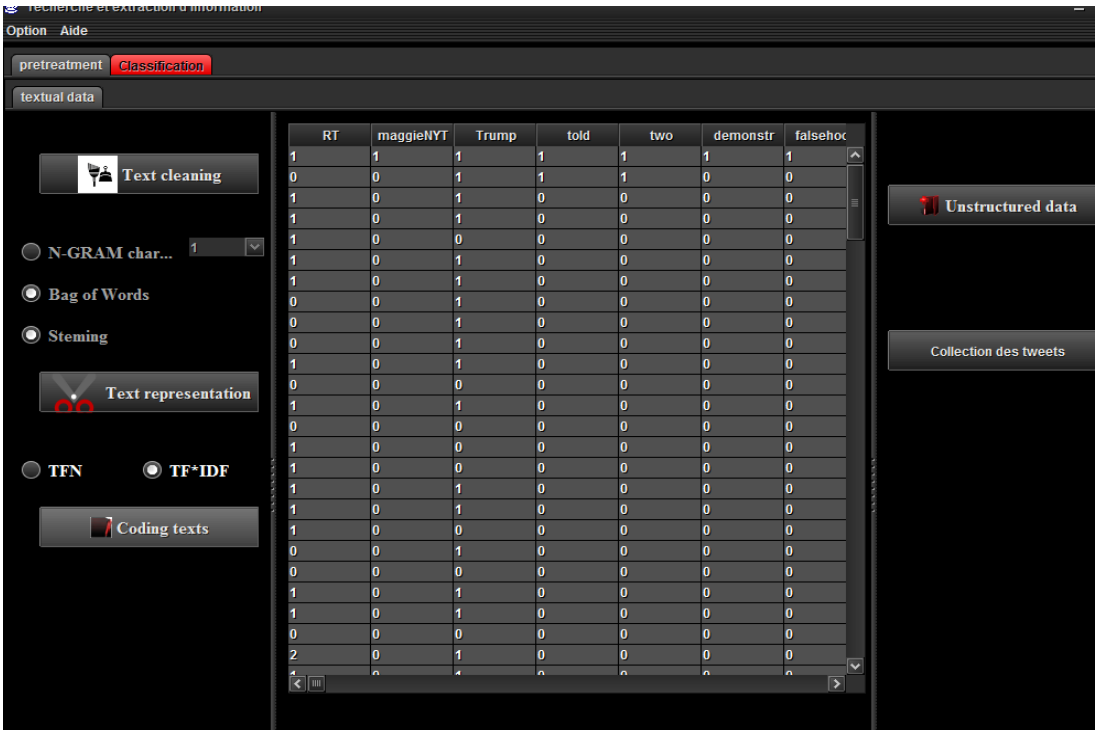

Figure IV.9 : Interface 1 « Représentation De Texte »

Pour coding le Texte on le choix de pondération « TF*IDF ou TFn ».

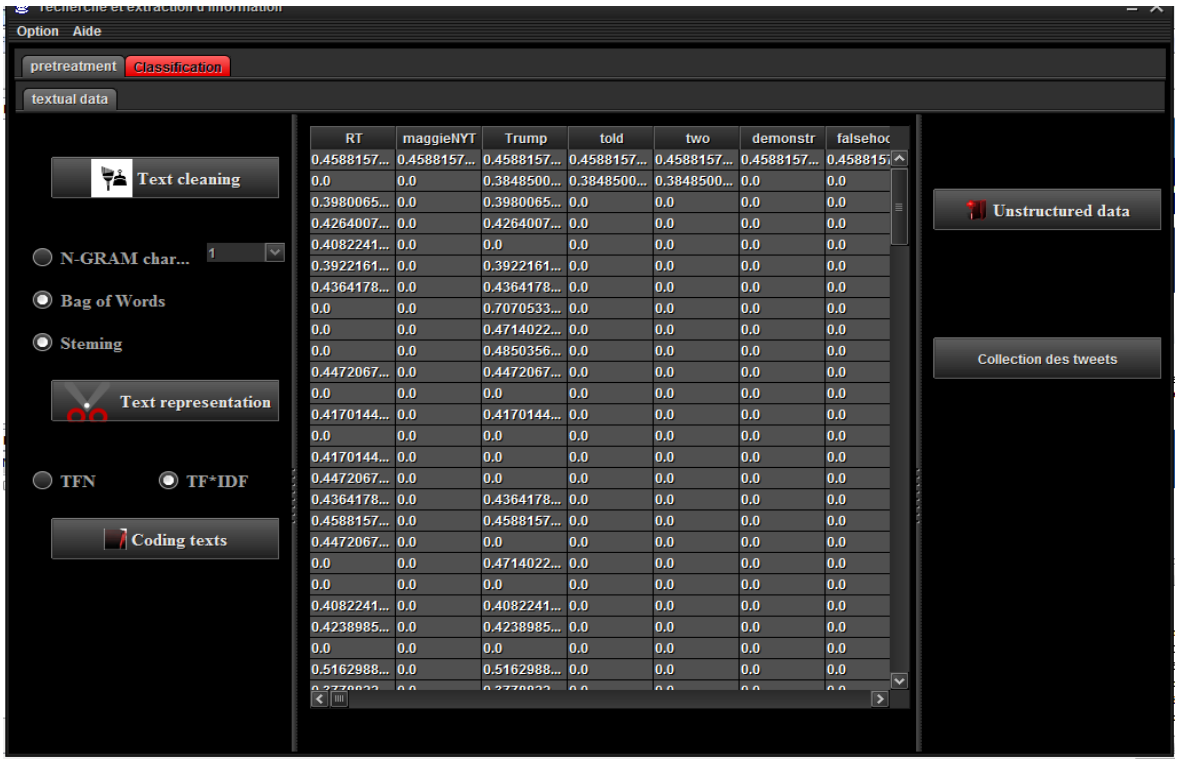

Figure IV.10: Interface 1 « Coding Text»

❖ Interface 2 « Classification » :

Dans la deuxième interface nous avons le choix de l'algorithme pour classifier les données structurer puis l'application nous donnons des résultats on peut les comparer.

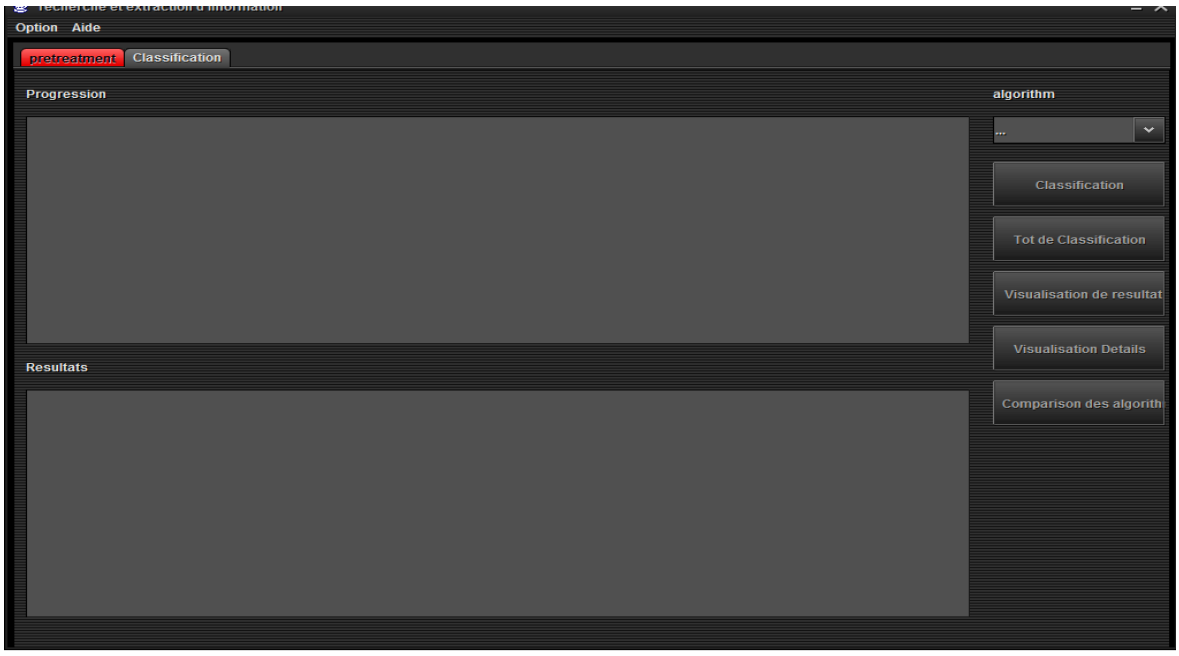

Figure IV.11 : Interface 2 « Classification »

En choisie l'algorithme puis on cliquant sur le bouton « Classification » pour faire les mesures. Puis nous avons voir les résultats afficher dans la **Figure IV.12.**

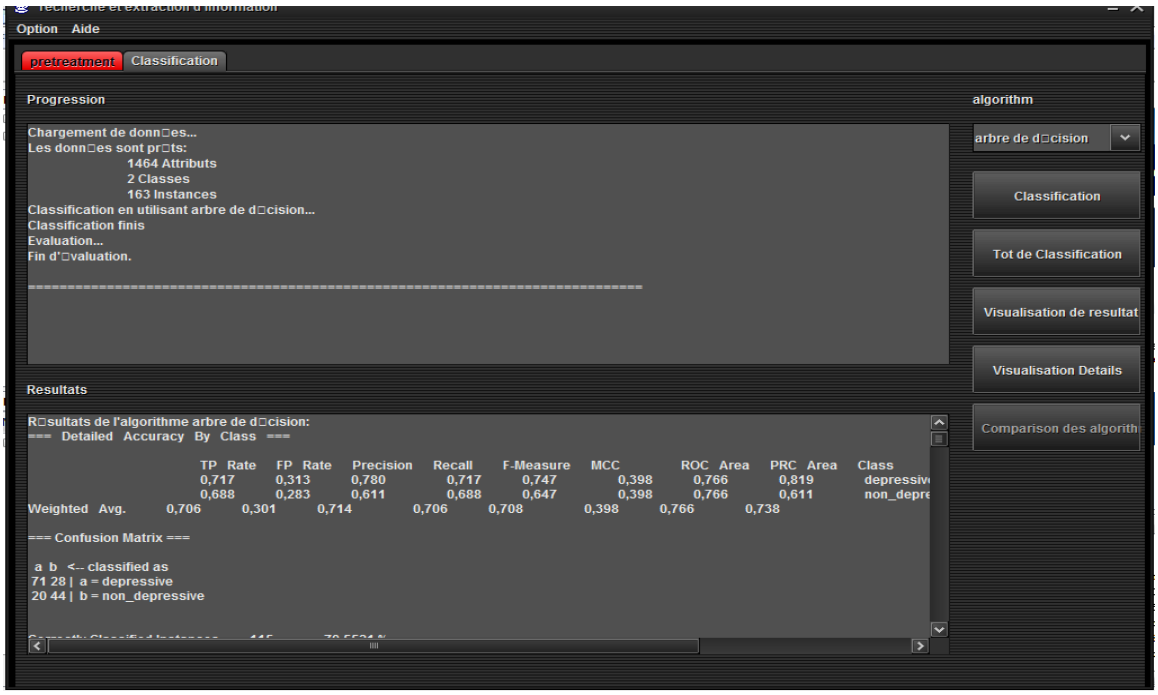

Figure IV.12 : Interface 2 « Résultat de Mesures »

❖ **Tôt de Classification :**

Apres avoir un résultat nous Application peut présenter les mesures sur un graphe comme la **Figure IV.13**.

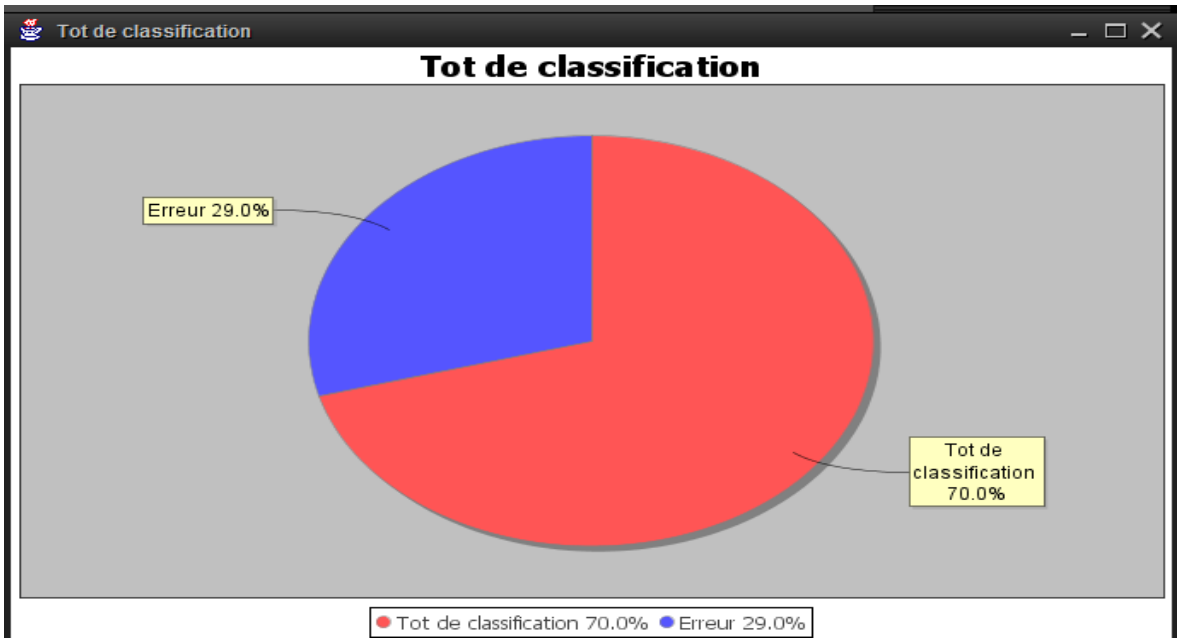

Figure IV.13 : « Tôt de classification »

Nous Application permettre de comparer les résultats de trois Algorithme et représenter dans un graphe comme **la Figure IV.14**.

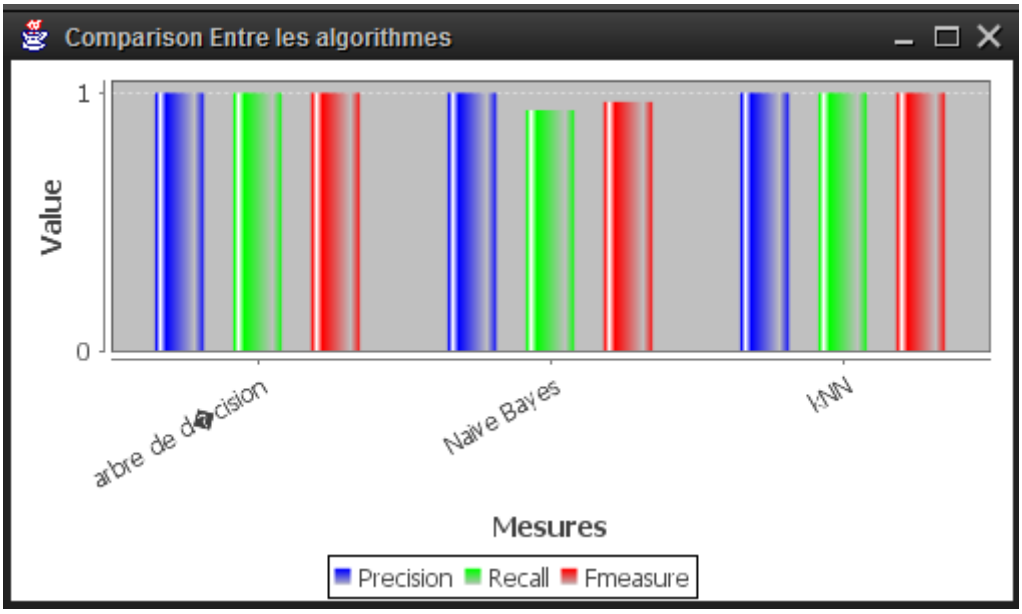

Figure IV.14 : « Comparaison Entre les Algorithmes »

Conclusion Générale

L'objectif de ce projet de fin d'étude est de concevoir un système intelligent qui permet de récolter les informations sur une personne depuis les réseaux sociaux, puis d'extraire, structuré et présente dans un graphe le contexte professionnel et général d'une personne à des personnes dépressives ou non.

Nous avons présenté le contexte d'analyse des réseaux sociaux, pour cela nous commençons par définir le terme de réseau social ce qui va nous permettre par la suite d'introduire le concept d'analyse des réseaux sociaux ainsi que les méthodes d'analyse utilisées.

Pour cela, dans le but d'introduire le concept d'analyse de contexte de la personne nous commençons par définir la notion de contexte d'une personne, puis faire une étude sur l'analyse de sentiments du contenu dans les micro-blogs.

Nous avons besoin d'outils spécifiques pour l'accès à l'information cette masse d'informations est sous format textuel et en grande quantité.

L'analyse contextuelle des réseaux sociaux est un domaine qui a attiré beaucoup de chercheurs, ce qui a donné naissance à de nombreux travaux. Dans notre travail nous nous sommes basés sur un principe fondé sur le corpus (Corpus-based Approach) qui consiste à attribuer des données à un classificateur pour l'apprentissage d'une façon supervisée, ce dernier génère un modèle qui est utilisé pour la partie test.

Dans notre travail nous avons utilisé le twitter comme réseau social qui est un site de réseautage permettant aux utilisateurs d'écrire de courts articles, appelés «tweets». Nous permettant de discuter les différentes étapes de notre approche proposée pour résoudre le problème de détection des personnes dépressives à travers une analyse décisionnelle des tweets. Ensuite, nous allons définir les outils utilisés pour la réalisation de la partie pratique de nos travaux avec une présentation générale des résultats obtenus en discutant les différentes comparaisons appliquées entre les différentes techniques utilisées et proposées durant notre travail.

En fin, à la prochain Nous souhaitons que ce projet peut être amélioré en essayant de couvrir plus des domaines qui s'intéresse une personne dans réseaux sociaux et intégrer les techniques de machine Learning pour améliorer les résultats.

Perceptive :

Nous souhaitons que ce projet peut être amélioré en essayant de couvrir plus des domaines qui s'intéresse sur contexte d'une personne par exemple les images qui veut être publie dans réseaux sociaux et intégrer les techniques de machine Learning pour améliorer les résultats.

Bibliographie

List d'abréviation

Chapitre1 : Généralité sur l'analyse des réseaux sociaux

HTML : hyper text markup langage

A**RS :** analyse réseaux sociaux

FOAF : Friend Of A Friend

PD : Degré prestige

CD : Degré Centralité

Chapitre 2 : Contexte de la personne dans les réseaux sociaux

LIWC : linguistic inquiry and word count

SNS : social network site

Chapitre 3 : implémentation résultats et discussion

TREC : Text Retrieval Conference

PFB : pondération fréquentiel brute

TF : Terme frequency

IDF : inversed documents frequency

DF : documents frequency

KPPV : K Plus Proches Voisins

KNN : K nearest neighbor

VP: Vrais positive

VN : Vrais Négative

FP : Faux positive

FN : Faux négative

Chapitre 4 : Réalisation

JRE : Java Runtime Environment

JSR : Java Specification Requests

Weka : Waikato Environnement for Knowledge Analysis

GNU : Général Public licence

NIST: National Institute of Standard and Technology

Références

[1] Romain Rissoan. Les réseaux sociaux Facebook. Twitter. Linkedin. Viadeo,Google+ Comprendre et maitriser ces nouveaux outils de communication ,ENI. 2011

[2] D. Boyd and E. Nicole. "Social Network Sites Definition, History," .Journal Of Computer—Mediated communication. 2008

[3] Mercanti-Onérin Maria. Analyse des réseaux sociaux et communautés en ligne : **Quelles applications et marketing. Centre (le recherche D\ISP, DRM** (CNRS UMR 7088), Université Paris Dauphine. FRANCE. 2010.

[4] Pierre Mercklé. « La « découverte » des réseaux sociaux » À propos de John A. Barnes et d'une expérience de traduction collaborative ouverte en sciences sociales, Réseaux, no 182, p.187-208. DOl:10.3917/res.182.0189.2013

[5] Eve Michael. « Deux traditions d'analyse des réseanx sociaux », Réseaux no 115. L 183-212. 2002

[6] http://fr.wikipedia.org/wiki/Analyse_des_r%C3%A9seaux_sociaux.

[7] Michel Bertrand. Claire Lenmercier. Sandro Guzzi-Heeb. Introduction : où en est l'analyse de réseaux en histoire ? , Vol 21. 2011

[8] Maria Malek. Laris-Eisti. Introduction à l'analyse des réseaux sociaux. 2009

[9] Erick Stattner. Introduction à l'Analyse des Réseaux Sociaux. Laboratoire LAMIA Université des Antilles et de la Guyane. Guadeloupe, FRANCE.2012

[10] Matthieu. Gaëtan. Benoît. Stépliane. Fouille de réseaux sociaux en ligne.
Synthèse bibliographique sur le Data Mining Sociale. 2012

[11] Orkhan ,lafarov et Subhan Gasimov. Machine Learning. rapport de projet dans le cadre d'un master 2 informatique, Université de Franche-Comté,FRANCE. 2012

[12] Christophe Thovex. Réseaux de Compétences : de l'Analyse des Réseaux Sociaux à l'Analyse Prédictive de Connaissances. Artificiel Intelligence.Université de Nantes. FRANCE. 2012

[13] Dietidonné Tcliuente, André Peiiiiiou. Marie-Francoise Canut. Nadine BaI)tiste-.Jessel, Florence Sedes. Niudélisation du processus de développement des profils utilisateurs dans les systèmes d'nformation.Université de Toulouse Institut de Recherche en Informatique de Toulouse.Système dinformation Généralisés. FRANCE. 2012

[14] World Health Organization, http://www.who.int/en/

[15] Ramirez-Esparza, N., Chung, C.K., Kacewicz, E., Pennebaker, J.W.: The Psychology of Word Use in Depression Forums in English and in Spanish: Testing Two Text Analytic Approaches. In: Proceedings of the International Conference on Weblogs and Social Media, pp. 102{108. Menlo Park, CA:AAAI Press (2008)

[16] Moreno, M., Jelenchick, L., Egan, K., Cox, E., Young, H.,Gannon, K.,etal.: Feeling Bad on Facebook: Depression Disclosures by College Students on Social Networking Site. Depression and Anxiety. 28, 447{455 (2011)

[17] Ji, Y.: Social Displacement, Homophily and Depression Levels: The Case of Zoufan on a Chinese Social Network Site. Cyberpsychology, Behavior, and Social Networking. For Peer Review (2012)

[18] Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Now Publisher Inc (2008).

[19] Jiang, L., Yu, M., Zhou, M., Liu, X., Zhao, T.: Target-dependent Twitter Sentiment Classifcation. In: Proceeding of 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 151{160 (2011)

[20] Dong, Z., Dong, Q.: HowNet)a Hybrid Language and Knowledge Resource.In: Proceedings of International Conference on Natural Language Processing and Knowledge Engineering, pp. 820{824. Los Alamitos, Ca: IEEE Press (2003)

[21] Go, Alec and Bhayani, Richa and Huang, Lei, 2009.

[22] McCreadie, R., Soboroff, I., Lin, J., Macdonald, C., Ounis, I., & McCullough, D. (2012, August). On building a reusable Twitter corpus. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* (pp. 1113-1114). ACM.

[23] Bickel, P. J., & Levina, E. (2004). Some theory for Fisher's linear discriminant function,naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, *10*(6), 989-1010.

[24] Bouarara, H. A., Hamou, R. M., & Amine, A. (2015). Novel Bio-Inspired Technique of Artificial Social Cockroaches (ASC). *International Journal of Organizational and Collective Intelligence (IJOCI)*, *5*(2), 47-79.

[25] Bouarara, H. A., Hamou, R. M., & Amine, A. (2015). A Novel Bio-Inspired Approach for Multilingual Spam Filtering. *International Journal of Intelligent Information Technologies (IJIIT)*, *11*(3), 45-87.

[26] Bouarara, H. A., & Hamou, R. M. (2017). *Environnement Bio-Inspirée pour la Recherche d'Information (EBIRI): les innovations issues de la nature*. Éditions universitaires européennes

[27] Cyril Grouin, Martine Hurault-Plantet Patrick Paroubek, Jean-Baptiste Berthelin, 2007,DEFT'07 : *une campagne d'évaluation en fouille d'opinion*

[28] FAIZABelbachir, juin2010, Recherche de l'Université Paul SabatierToulouse,*Expérimentation de fonctions pour la détection d'opinions dans les blogs*

[29]Ounis I., Macdonald C. etSoboroff I, Overview of TREC-2008 Blog Track. *Dans TREC:Proceedings of the TextRetrievalConference*, ,2008.