Université Kasdi Merbah – Ouargla

Vice Rectorat de la Formation Supérieure, de la Formation Continue et des Diplômes

Fouille de données d'opinion des usagers de sites E-commerce

UKM Ouargla, Juin 2013.

UNIVERSITE KASDI MERBAH OUARGLA

Faculté des Sciences, de la Technologie et des Sciences de la Matière

Département des Mathématiques et d'Informatique

Mémoire

MASTER ACADEMIQUE

Domaine : Mathématiques et Informatique.

Filière : Informatique académique.

Spécialité : Informatique Industrielle.

Présenté par : ***Melle : Randa BENKHELIFA***

Melle : Saliha GAGUI.

Thème

Fouille de données d'opinion des usagers de sites E-commerce

Soutenu publiquement
le :../6/2013.

Devant le jury :

Mme.	Laallam F.Z.	Président	UKM Ouargla
M.	Herrouz A.	Encadreur/rapporteur	UKM Ouargla
M.	Mahdjoub M.B.	Examinateur	UKM Ouargla

Année Universitaire : 2012 /2013


بسم الله الرحمن الرحيم

Remerciements

En premier lieu,

Nous tenons à remercier ALLAH qui nous a aidé et nous adonné la patience

et le courage durant ces longues années d'études, et nous avoir

données la force à accomplir ce travail.

Nous remercions vivement Mr.Herrouz A.Hakim, notre encadreur pour son aide,

sa compréhension, ses conseils et critiques pertinentes.

Nous remercions également toutes les personnes qui nous ont aidés, de près ou de loin

pour la réalisation de ce travail en particulier

nos parents, et Melle Azzaoui Nadjet.

Nous tenons à remercier les membres du jury pour l'honneur qu'ils nous

ont fait en acceptant de juger notre travail.

Nous adressons aussi nos remerciements à tous nos enseignants qui ont veillé

sur notre formation

Table des matières

Liste des figures

Liste des tableaux

Résumé

L'objectif de ce travail est la réalisation d'un outil adapté à l'intérêt de la fouille d'opinion et l'Analyse des Sentiments, qui se sont développés en même temps avec la naissance du web 2.0 (les forums, les sites de discussion, les réseaux sociaux ou, plus généralement, dans les blogs) où les internautes expriment librement leur opinion.

Aujourd'hui, l'opinion Mining et le Sentiments Analysis font partie du même domaine de recherche. Ce domaine s'occupe de traitement d'opinion, de sentiment, et de la subjectivité dans le texte. Il cherche à classer le texte en repérant d'abord les phrases porteuses d'opinion (classement objectif/subjectif), puis d'attribuer une polarité (positive, négative ou neutre) à l'opinion. Cette classification aide à la prise de décision surtout dans le monde du business et de la politique.

Mots-clés : *Web, Fouille d'Opinion, Analyse de Sentiments, Fouille de Web, Opinion, subjectivité.*

Abstract

The objective of this work is the realization of a tool for the benefit of the Opinion Mining and Sentiments Analysis, which developed simultaneously with the birth of Web 2.0 (forums, websites discussion, social networks or, more generally, in blogs) where users express their opinions freely.

Today, opinion Mining and Sentiments Analysis are a part of the same research area. This domain deals with treatment of opinion, sentiment, and subjectivity in the text. It seeks to classify the text by first identifying carriers of opinion sentences (classification objective/ subjective), then assign a polarity (positive, negative or neutral) to the opinion. This classification helps to decision taken over everything in the world of business, and politics.

Keywords: *Web, Opinion Mining, Sentiments Analysis, Web Mining, Opinion, Subjectivity*

ملخص

الهدف من هذا العمل هو تحقيق *وتنفيذ برنامج* لفائدة تعدين الرأي وتحليل المشاعر،التي تطورت مع تطورالويب 2.0 (المنتديات،مواقع الدردشة،والشبكات الاجتماعية،أو بشكل أعم في بلوق) حيثما يمكن للمستخدمين التعبير عن آرائهم بحرية.

اليوم، تعدين الرأي وتحليل المشاعر ينتمون إلى نفس مجال البحث. يهتم هذا المجال بمعالجة الرأي والمشاعر، والذاتية في النص التي تسعى لتصنيف النص من خلال استخراج الجمل الحاملة للآراء (تصنيف موضوعي / ذاتي)، ثم تعيين قطبية الرأي (سواء إيجابية أو سلبية أو محايدة). هذا التصنيف يساعد على اتخاذ القرار في كل من العالم التجاري والعالم السياسي.

كلمات البحث: *الويب، تعدين الرأي، تحليل المشاعر، تعدين الويب، رأي، الذاتية.*

Introduction générale

Introduction générale

Aujourd'hui, l'E-commerce est devenu une réalité économique et il est de plus en plus populaire. Le nombre de commentaires des internautes est en croissance constante. Les opinions sur le Web affectent nos choix et nos décisions. Il s'avère alors indispensable de traiter une quantité importante de critiques des clients afin de présenter à l'utilisateur l'information dont il a besoin dans la forme la plus appropriée [7].

"Qu'est-ce que les autres pensent ?" a toujours été un important élément d'information pour la plupart des gens au cours du processus de prise de décision [24]. De nombreux sites interactifs ont vu le jour avec la naissance du Web 2.0. Ils proposent à l'internaute de donner son avis sur des produits (livres, films, mobiles, ordinateurs portables) dans les groupes de discussions, les blogs, forums et autres sites spécialisés dans les critiques de produits (comme Amazone, Cnet, les réseaux sociaux, etc.).

La Fouille de Web (Web Mining) consiste à utiliser l'ensemble des techniques de la fouille de données afin d'analyser et de comprendre le comportement des internautes sur les sites Web permettent de valoriser le contenu des sites en améliorant l'organisation et les performances des sites. Le Web Content Mining est un des axes du Web Mining, il concerne l'analyse du contenu des pages Web. Le contenu d'une page peut être des textes ou du multimédia. Un texte est considéré comme une entité porteuse d'une information. Cette information peut être classée en deux catégories principales : faits et opinions.

La Fouille de données d'Opinion (Opinion Mining) et l'Analyse des Sentiments (Sentiments Analysis), font partie d'un domaine émergent. Ce dernier s'occupe de traitement de la subjectivité : opinions, avis, sentiments, émotions, évaluations, croyances ou jugements personnel. Ensuite, il attribue une polarité (positive, négative ou neutre) à cette opinion.
Ces données d'opinion revêtent aujourd'hui une importance stratégique et économique évidente car leur analyse permet de connaître les points forts et les points faibles des produits, d'estimer la perception du produit par les consommateurs, afin d'améliorer les profits. Il permet également au consommateur de donner son opinion, de l'aider à la prise de décision, en s'inspirant des sentiments et d'opinions d'autres clients sur un produit donné.

L'objectif de ce travail est de développer un nouveau système permettant d'analyser et classer les opinions des consommateurs sur tel ou tel produit dans des sites E-commerce.

Ce mémoire est organisé en cinq chapitres.

- Nous consacrerons un **premier chapitre** à présenter les caractéristiques du Web et en particulier le Web 2.0. La notion du Contenu Généré par les Utilisateurs retiendra tout particulièrement notre attention.
- Le **second chapitre** portera sur la présentation des concepts généraux relatifs aux Data Mining, Web Mining, Web Content Mining, et Text Mining.
- Le **troisième chapitre** fait le point sur le domaine de l'Opinion Mining et le Sentiments Analysis.
- Le **quatrième chapitre** est consacré à la présentation et à la comparaison des principaux travaux réalisés dans le domaine de la fouille du Web et spécialement sur le réseau social Twitter.
- Notre **cinquième chapitre**, quant à lui, proposera la réalisation de notre nouveau système qui permet, entre autres, de faire l'extraction des contenus des sites Web, le nettoyage et l'analyse des commentaires.

Enfin, en guise de conclusions, nous indiquerons quelques remarques sur ce travail en concluant par un bilan et nous exposerons les perspectives pour de futurs travaux.

Chapitre

Les enjeux du Web 2.0

Chapitre 1 : Les enjeux du Web 2.0

La dernière décennie a vu l'émergence d'Internet qui a subi une énorme transformation. Après la prise de conscience du potentiel de cette technologie, les services se sont développés pour devenir de plus en plus nombreux, mais surtout de plus en plus interactifs.

1.1 Le Web : définition et caractéristiques

Internet, et plus spécialement le Web fait aujourd'hui partie intégrante de notre quotidien, et surtout pour les nouvelles générations. Le Web a amélioré et facilité notre vie, ce qui rend les processus plus efficaces. En effet, le Web facilite la recherche d'information, la communication à distance en temps réel, et il minimise les coûts d'envoi des messages qui étaient auparavant uniquement par correspondance.

Le Web est un réseau maillé dont l'entité de base est la page web et dont la structure repose sur la notion de liens [25]. Les pages Web sont un moyen facile de se promener sur Internet [47].

Le Web permet :

- ✓ de lire,
- ✓ d'utiliser les informations mises à disposition,
- ✓ de produire et diffuser ses propres informations : site internet, weblog, contribution à des sites collaboratifs,
- ✓ de faire du shopping,
- ✓ de voir les nouvelles,
- ✓ d'écouter le Coran, la musique, regarder les vidéos, les films, etc.,
- ✓ de jouer un rôle important en télémédecine,
- ✓ la formation à travers le net (télé-éducation),
- ✓ partage et échanges de données, des images, des connaissances, etc.

Le Web est un "système d'information multimédia" [25], une page web peut contenir :

- ✓ du Texte (html …),
- ✓ des images (Gif, JPEG, PNG, ...),
- ✓ de la vidéo/audio (mp3, mov, wav ...),
- ✓ des Animations (flash …),

- ✓ etc.

Les pages sont liées entre elles au moyen de liens hypertextes qui permettent de naviguer [25]:

- à l'intérieur même d'une page,
- vers une page différente située sur le même serveur (généralement sur le même site),
- vers une page qui se trouve sur n'importe quel autre serveur connecté.

C'est l'hypertextualité qui fait toute la force et l'intérêt du Web.

Le Web a permis la généralisation du document électronique en définissant un hypertexte en réseau. Cette conception du réseau informatique comme un immense document distribué sur plusieurs sites a été rendue possible par la normalisation portant sur trois éléments [42] :

a. une manière de nommer les ressources informatiques de façon à les rendre accessibles par divers outils de navigation : les URL (Uniform Resource Locator) ;
b. le protocole HTTP (HyperText Transfer Protocol), qui permet d'exploiter l'infrastructure technique du réseau à partir d'une généralisation du modèle client-serveur qui englobe tous les protocoles de transmission des données définis précédemment ;
c. le langage de balisage HTML (HyperText Markup Language), qui permet d'inscrire au sein même d'un document électronique des ancres permettant de naviguer facilement (par un simple clic de la souris dans le cas général) entre diverses ressources distribuées sur le réseau.

De nombreuses caractéristiques rendent le Web si intéressant :

a. C'est est un environnement hypermédia : L'usager n'est pas obligé de suivre un texte en séquence. Ce qui va lui permettre d'emprunter des itinéraires différents. Cet environnement permet donc une navigation non linéaire et un lecteur rapide.
b. Il est interactif : L'environnement hypermédia favorise l'interaction. Naviguer dans différents écrans d'un site peut interpeller l'internaute et le garder actif.

c. Il est archivé : les fichiers permanents des documents et les sessions interactives en ligne sont emmagasinés et disponibles à des fins de recherche ou utilisation éventuelle.
d. Il est dynamique : il est en constante évolution et les implications pour la mise en ligne des différents contenus sont énormes.
e. Il est ouvert : il est basé sur des protocoles et normes largement acceptés. Actuellement, pratiquement toutes les plates-formes informatiques supportent Internet. Ceci facilite pour l'utilisateur de mettre en ligne des contenus qui seront accessibles dans le monde entier sur toutes sortes de configurations.
f. Il est distribué : L'information est archivée dans des milliers de serveurs du réseau à travers le monde. Ceci veut dire que quelque soit la localisation de l'usager, il a accès à la même information.
g. Il est mondialement accessible.
h. Il est filtré : Les internautes sont cachés derrière des écrans et peuvent rester anonymes.
i. Enfin, le Web est –en partie asynchrone : Les événements ne se produisent pas tous en même temps ; il est donc nécessaire de considérer est environnement de manière asynchrone.

1.2 Les enjeux du Web 2.0

Aujourd'hui, l'internaute n'est plus simplement spectateur, il peut s'il le souhaite devenir acteur. Et tout est mis en œuvre pour que le simple spectateur devienne acteur du Web afin, entre autres choses, de le fidéliser. Cette révolution a donnée naissance à ce que l'on nomme aujourd'hui le Web 2.0 appelée encore Web Social ou Web Participatif [13].

Bien que la définition du Web 2.0 ne soit pas encore parfaitement établie, un grand aspect caractérisant ce Web nouvelle génération peut être mis en avant : le rôle central de l'utilisateur.

Une caractéristique principale du Web 2.0 qui le distingue nettement du Web 1.0 est donc la prise de contrôle de l'information par les utilisateurs. N'importe quel internaute peut aujourd'hui apporter sa pierre à l'édifice. Il peut se faire une place sur la toile, collaborer, partager des informations, des outils, des fichiers multimédias, donner ses opinions, commenter, réagir, etc. et tout ceci sans connaissances spécifiques. En effet, quand auparavant

il fallait un minimum de savoir-faire en informatique et en programmation pour créer son espace sur Internet, aujourd'hui il suffit de savoir cliquer car de plus en plus d'outils sont mis à disposition de tout un chacun afin de faciliter toutes ces interactions. Parmi ces outils, nous pouvons citer les réseaux sociaux, les blogs, les wikis, les sites de partage de vidéos, de photos, de musiques, etc. [13].

De nombreux sites présents sur Internet aujourd'hui offrent la possibilité à tous leurs visiteurs de laisser, au minimum, une trace textuelle et ainsi s'exprimer publiquement. Tout ce contenu, qu'il soit textuel ou autre, est appelé Contenu Généré par les Utilisateurs ou UGC (pour User Generated Content). Il représente une quantité de données de plus en plus importante sur la toile et est composé, en très grande partie, de données textuelles. Ce nouvel espace d'expression représente une grosse quantité d'informations, notamment en termes d'avis et d'opinions, susceptibles d'être exploitées à des fins diverses. Les données textuelles, notamment, peuvent être analysées dans déférents buts. Par exemple, dans le domaine de la fouille d'opinion (Opinion Mining), les textes sont utilisés afin de permettre à des entreprises de connaître automatiquement l'image que les consommateurs ont d'eux, de même pour faire de la comparaison d'articles de vente, réaliser des sondages, détecter des rumeurs, etc.

Les textes rédigés par les internautes sont en général beaucoup plus subjectifs que les articles rédigés par des professionnels et donc beaucoup plus porteurs d'opinion. De plus, ils contiennent un éventail d'avis et de jugements souvent plus représentatifs du consommateur ordinaire, étant rédigés par un panel d'individus variés parmi lesquels on retrouve tout autant de profanes que de connaisseurs du sujet abordé. Par exemples, pour ce qui est du cas des films, il est possible de trouver des avis de spectateurs de tous âges et de tous horizons. Ce que n'offrent pas les critiques journalistiques qui sont généralement rédigés par des personnes du même milieu, c'est-à-dire dans ce cas, par des journalistes cinéphiles appartenant à la population active et ayant fait un certain nombre d'années d'études [13].

Généralement, le Contenu Généré par les Utilisateurs est donc composé de données textuelles qui sont porteuses d'opinions et de sentiments. L'accès au contenu sémantique de ces données, préalable à la connaissance des opinions qu'elles véhiculent, représente un enjeu pour de nombreux acteurs:

- le consommateur, c'est-à-dire chacun de nous, qui veut s'informer avant toute décision qu'elle soit d'achat ou autre;

- les fournisseurs de biens et de services qui cherchent à se positionner les uns par rapport aux autres dans un univers hautement compétitif et face à une demande de plus en plus complexe à identifier;
- les chercheurs : économistes, sociologues,… ou simplement les responsables publics qui cherchent à comprendre le comportement individuel ou collectif pour anticiper, réguler ou ajuster les rapports entre les différents agents socioéconomiques.

C'est dans ce contexte que s'introduit le domaine de la fouille d'opinion [7].

Chapitre

La fouille du Web

Chapitre 2 : La fouille du Web

2.1 La fouille de données (Data Mining (DM))

Compte tenu de la grande taille des bases de données actuelles, les données brutes sont généralement de faible qualité. Elles peuvent être incomplètes (valeurs manquantes ou agrégées), bruitées (valeurs erronées ou aberrantes) ou incohérentes (divergence entre attributs). L'application d'algorithmes du DM sur de telles données complexifie l'apprentissage et nuit à la performance ainsi qu'à la fiabilité du modèle [46].

Aujourd'hui, le DM est un domaine très en vogue et qui «s'intéresse à la découverte d'informations utiles et nouvelles dans une quantité importante de données» [10].

2.1.1 Définitions

Le DM est un processus inductif, itératif et intéractif de découverte dans les Bases de Données larges, de modèles de données valides, nouveaux, utiles et compréhensibles [51].

- ✓ **Itératif** : nécessite plusieurs passes.
- ✓ **Intéractif** : l'utilisateur est dans la boucle du processus.
- ✓ **Valides** : valables dans le futur.
- ✓ **Nouveaux**: non prévisibles.
- ✓ **Utiles** : permettent à l'utilisateur de prendre des décisions.
- ✓ **Compréhensibles**: présentation simple.

L'extraction de connaissances à partir de bases de données (ECBD) : est un processus non trivial qui construit un modèle valide, nouveau, potentiellement utile et au final compréhensible, à partir de données [18].

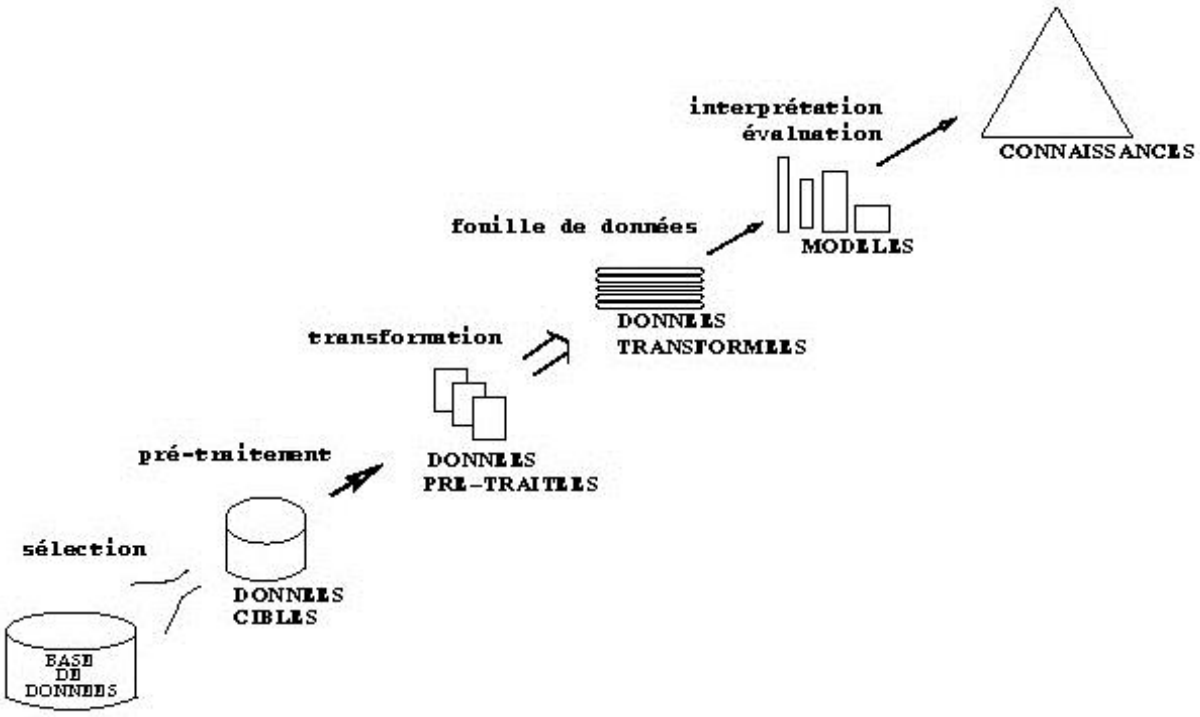

Figure 2.1 . Schéma global de l'ECBD [18].

1. **Analyse du problème d'application [**12]
   - ✓ Exposer le problème.
   - ✓ Définir les objectifs tangibles et quantifiables.
   - ✓ Formuler le problème à résoudre et connaître la typologie du problème (structuration).
   - ✓ Définir la manière dont la solution sera déployée (spécification de la solution).
2. **Sélection des données** [36] [14]
   - ✓ Effectuer l'inventaire de toutes les données existantes ;
   - ✓ Evaluer la qualité des données, détecter leurs insuffisances ;
   - ✓ Visualiser, analyser les distributions et les regroupements;
   - ✓ Créer un ensemble de données à étudier ;
3. **Prétraitement des données [29] [36]**

   Cette étape a pour objectif d'assurer la fiabilité des données créées durant l'étape précédente incluant les deux opérations suivantes :
   - ✓ Nettoyage : suppression du bruit et des valeurs manquantes (aberrantes) ;
   - ✓ Définition des stratégies pour traiter les données manquantes.

4. **Transformation des données [**14**]**

   Cette étape prépare les données pour le DM en transformant les données de l'étape précédente. Cette transformation concerne essentiellement deux aspects :

   - ✓ Réduction de dimension et d'extraction de données ;
   - ✓ Discrétisation des attributs numériques et transformations fonctionnelles.

5. **Fouille de données [**8**]**

   C'est le cœur de l'ECBD et qui consiste à identifier les motifs qui structurent les données, ou produit des modèles explicatifs ou prédictif des données. Cette phase fait appel à un lot de méthodes issues de différentes techniques.

6. **Evaluation et interprétation des résultats [51]**

   - ✓ Analyser la connaissance (intérêt) ;
   - ✓ Vérifier sa validité (sur le reste de la base de données) ;
   - ✓ Réitérer le processus si nécessaire.

La connaissance qui en est ainsi extraite est alors stockée dans la base de connaissances [18].

7. **Déploiement de la solution [51]**

   - ✓ La mettre à la disposition des décideurs ;
   - ✓ L'échanger avec d'autres applications (système expert, …) ;
   - ✓ etc.

2.1.2 Techniques du DM

Les outils du DM utilisent les mêmes fondements théoriques que les techniques statistiques traditionnelles. Mais, ils représentent une remarquable évolution par rapport à ces dernières **[**12].

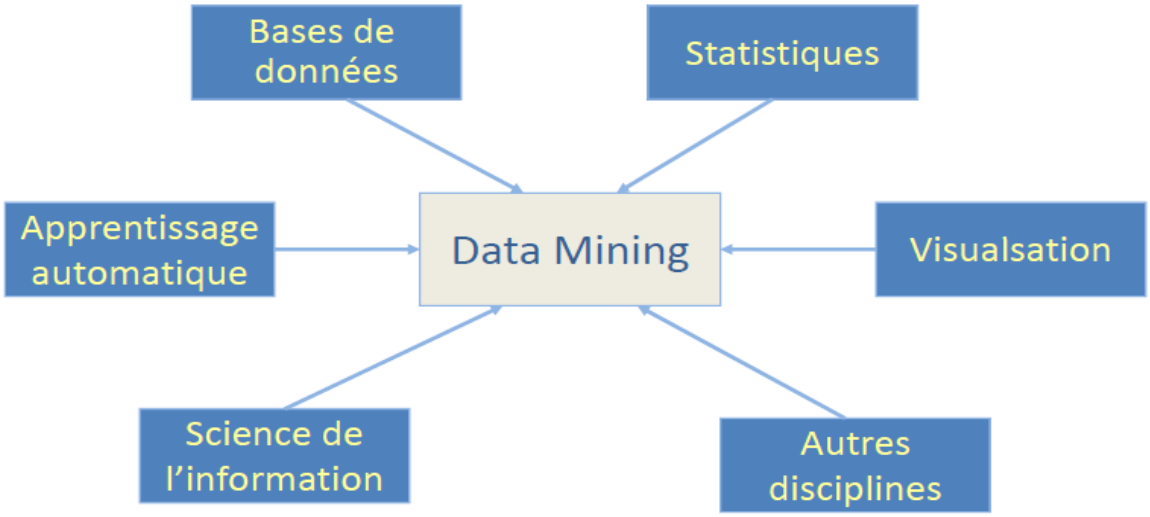

Figure 2.2. Le DM : union de disciplines variées.

2.1.3 Les tâches du DM

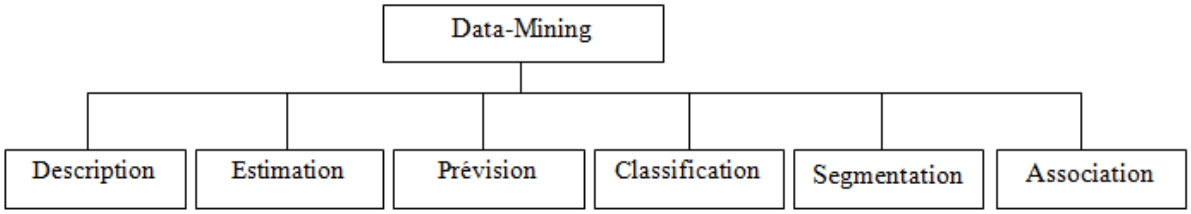

Figure 2.3. Les tâches du DM.

- **Description :** Les descriptions des modèles et des tendances servent à expliquer ou vérifier un fait.
- **Estimation :** l'estimation consiste à compléter une valeur manquante dans un champ particulier [1].
- **Prévision :** liée à la classification, cette tâche vise à prédire une ou plusieurs caractéristiques inconnues à partir d'un ensemble de caractéristiques connues.
- **Classification** : la capacité de classer des objets et des événements comme membres de classes prédéfinies.
- **Segmentation** (clustering): la segmentation a pour but de découvrir dans les données des groupes, non identifiés à l'avance, ayant les mêmes caractéristiques [14].

- **Association :** L'association a pour but de grouper par similitude ; elle consiste à déterminer quels attributs vont ensemble.

Le DM sert à donner un sens aux données, en extraire les relations masquées et non triviales, bref à exploiter la base de donnée. Le Web peut être considéré comme une gigantesque base de données.

2.2 La Fouille du Web (Web Mining)

Le Web Mining (WM) fut développé à la fin des années 1990. Il consiste à utiliser l'ensemble des techniques de la fouille de données afin de développer des outils permettant l'extraction d'informations pertinentes à partir de données du Web (documents, traces d'interactions, structure des liens, etc.) [9].

2.2.1 L'objectif du WM

1. L'amélioration et la valorisation des sites Web : L'analyse et la compréhension du comportement des internautes sur les sites Web en permettant de valoriser le contenu des sites tout en améliorant l'organisation et les performances des sites.

2. La personnalisation : Les techniques du DM appliquées aux données collectées sur le Web permettent d'extraire des informations intéressantes relatives à l'utilisation du site par les internautes. L'analyse de ces informations permet de personnaliser le contenu proposé aux internautes en tenant compte de leurs préférences et de leur profil.

2.2.2 Les axes de développement du WM

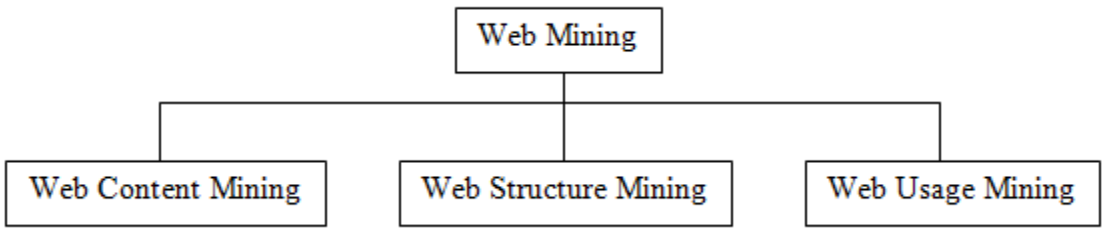

Figure 2.4. Les axes de développement du WM.

Il existe trois axes d'application dans le domaine du WM :

- **Web Structure Mining** : qui s'intéresse à l'analyse de la structure des sites Web.

- **Web Usage Mining** : qui analyse le comportement des utilisateurs des sites Web.
- **Web Content Mining** : qui concerne l'analyse du contenu des pages Web.

2.2.2.1 Web Structure Mining (WSM)

Le WSM consiste à analyser la structure des liens entre les pages ou les sites Web, qui constituent une source riche d'information, afin d'améliorer leur ergonomie par la suppression ou l'ajout de nouveaux liens entre les pages [2]. Les recherches consacrées à cette branche du WM sont inspirées des travaux sur l'étude des réseaux sociaux.

L'analyse de la structure du Web utilise plusieurs algorithmes, dont les plus célèbres sont PagRank et HITS.

Cette catégorie peut être divisée en deux types (lien hypertexte et structure de document)

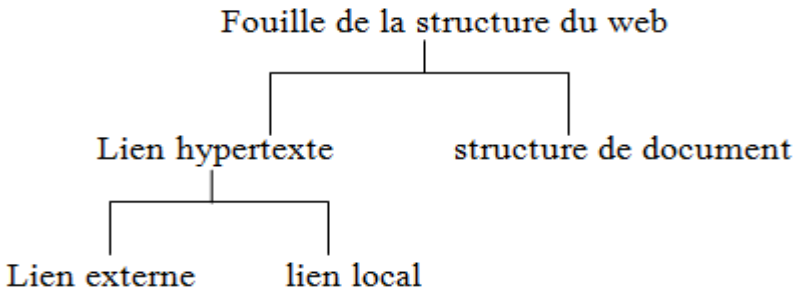

Figure 2.5.fouille de la structure de Web.

a. **Hypertextes** : un lien hypertexte est une unité structurelle qui relie une page Web vers un autre emplacement, que ce soit au sein de la même page Web (lien local) ou à une page Web différente (lien externe).
b. **La structure du document** : le contenu d'une page Web peuvent être organisée sous forme d'arborescence structurée, basée sur des balises HTML et XML.

L'analyse des liens entre les pages Web (Web Structure Mining) est utilisée pour:

a. la commande des documents correspondants à une requête de l'utilisateur.
b. décider quelles pages ajouter à une collection.
c. la catégorisation de la page.
d. pour trouver des pages liées.
e. pour trouver des sites Web dupliqués.

L'analyse des liens entre les pages Web permet :

- ✓ De déterminer combien de pages consultent les internautes en moyenne.
- ✓ D'adapter l'arborescence du site pour que les pages les plus recherchées soient dans les premières pages du site.
- ✓ D'améliorer l'ergonomie du site par création de nouveaux liens.

2.2.2.2 Web Usage Mining (WUM)

Le WUM ou Web Log Mining est le processus d'extraction d'informations utiles stockées dans les logs des serveurs Web (l'historique des transactions des utilisateurs) ou bien les informations données par les intervenants du Web. Ceci revient à analyser le comportement de l'utilisateur à travers un ensemble de clics qui modélise son interaction avec le site Web (on parle alors d'analyse du clickstream). Son but est de mieux comprendre et servir les besoins des applications web [32].

❖ **Les phases de WUM**

Le processus général de WUM est basé sur l'analyse des fichiers logs. Il est constitué de trois phases principales : prétraitement de données, découverte du modèle et analyse du modèle, comme le montre le schéma suivant :

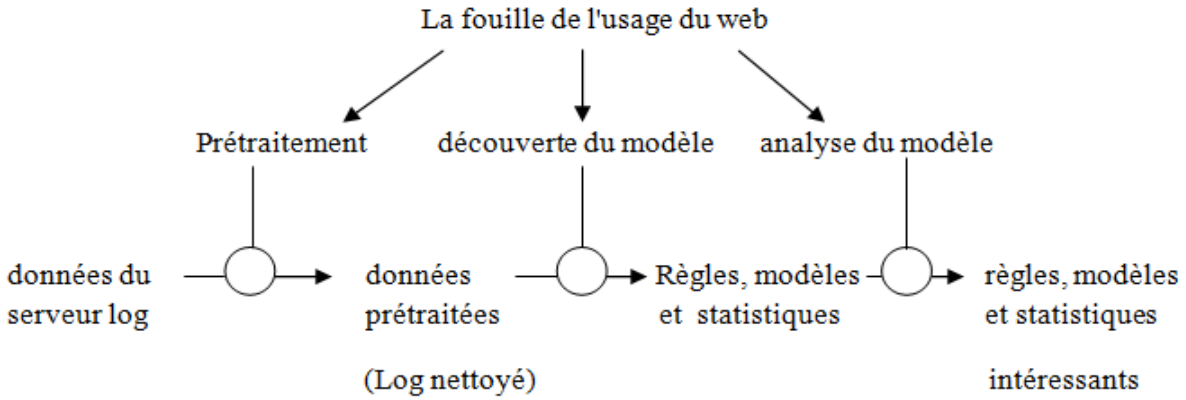

Figure 2.6. Le processus du WUM.

Les fichiers logs : le fichier log est un fichier texte qui contient les requêtes faites au serveur Web enregistrées en ordre chronologique sous les formats les plus utilisées CLF (common log format) et ECLF (Extended CLF) [16].

Le fichier log est formulé comme suit :

1. le nom ou l'adresse IP de la machine appelante.
2. la date et l'heure de la requête.

3. la méthode utilisée par la requête (Get, Post, etc.), l'URL de la requête et le protocole utilisé.
4. le statut de la requête.
5. la taille du fichier envoyé.
6. l'URL qui a référencé la requête.
7. l'Agent (navigateur et le système d'exploitation)

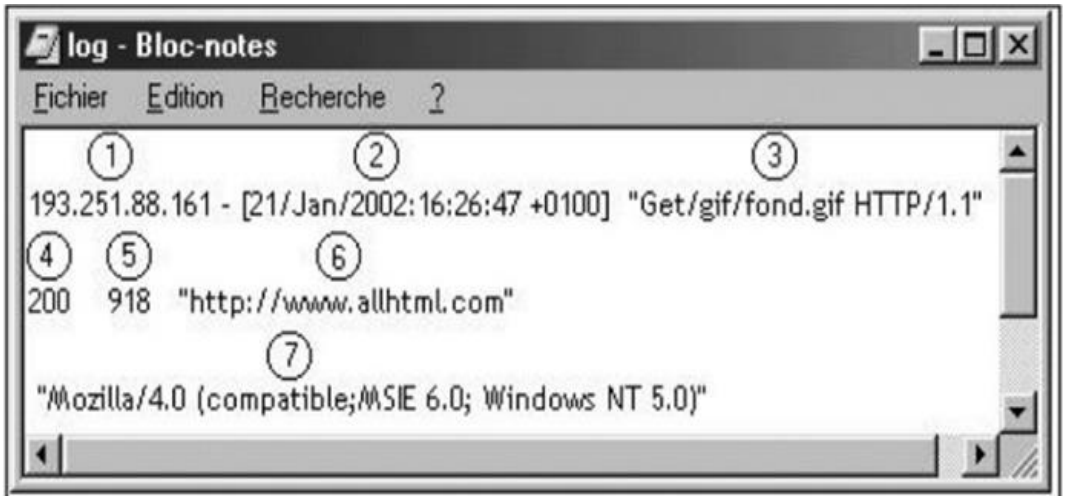

Figure 2.7. Exemple de fichier Log Web. [2]

1) Prétraitement de données / préparation des données : consiste à arranger d'une manière cohérente les données des fichiers logs afin de les nettoyer en enlevant l'information et le bruit non pertinents dans le but d'identifier d'une façon précise les sessions (Une session est composée de l'ensemble de pages visitées par le même utilisateur durant la période d'analyse [9]).
2) Découverte de modèles : Cette étape consiste à appliquer des méthodes statistiques ainsi que des techniques de fouille des données sur le fichier de sessions ou le fichier de navigations afin de détecter des tendances intéressantes.
3) L'analyse de modèles [3]: c'est la dernière étape dans ce processus. Il s'agit ici de voir comment exploiter toutes les informations qui ont été obtenues.

Le WUM peut être encore classé en fonction des types donnés d'utilisations envisagées:

a. **Données du serveur Web** : elles correspondant aux registres des utilisateurs qui sont collectées sur le serveur web.
b. **Serveurs d'applications commerciales** : permettent aux applications e-commerce d'être construites avec peu d'effort.

c. **Données sur le niveau d'application** : de nouveaux types d'événements peuvent toujours être définis dans une application. Cet enregistrement peut être activé pour eux. La génération de l'historique de cette spécialité définie les événements.

2.2.2.3 Web Content Mining (WCM)

Le WCM est le processus d'extraction des informations utiles à partir du contenu Web (données, documents). Le contenu Web se compose de plusieurs types de données telles que du texte brut (non structuré), image, audio, vidéo, métadonnées, ainsi que HTML (semi structuré), des documents dynamiques et des documents multimédias.

2.2.2.3.1 Les approches de WCM

a. Les approches basées sur l'agent (Agent-based) : ces approches impliquent les systèmes de l'IA qui peuvent agir d'une manière autonome ou semi autonome pour le compte d'un utilisateur particulier. Pour découvrir et organiser l'information, elles sont concentrées sur les outils du WM qui sont basés sur la technologie d'agents.

- ✓ Certains agents intelligents Web peuvent utiliser un profil utilisateur pour rechercher des informations pertinentes, puis d'organiser et d'interpréter les informations découvertes. Par exemple: Harvest.
- ✓ Certains utilisent les différentes techniques de recherche d'information et les caractéristiques des documents hypertextes pour organiser et filtrer les informations récupérées. Exemple: Hypursuit.
- ✓ Apprendre les préférences de l'utilisateur et les utiliser pour découvrir les sources d'information pour les utilisateurs particuliers. Exemple: la règle Rminer Xpert.

b. **Les approches de base de données (Data-based) :** focalisé sur l'intégration et l'organisation des données hétérogènes et semi-structurées sur le Web aux données structurées et des collections de ressources de plus haut niveau. Ces ressources organisées peuvent ensuite être consultées et analysées.

2.2.2.3.2 Les techniques du WCM

Sur le web, il existe plusieurs types d'informations :

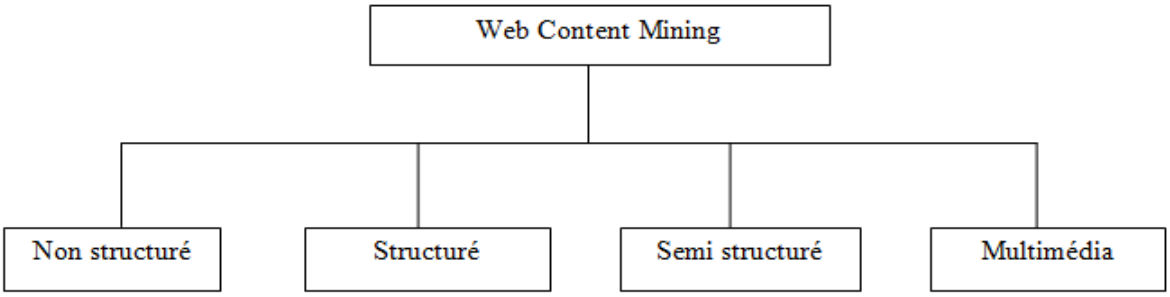

Figure 2.8. Taxonomie du WCM.

- **Les informations structurées** [41] : ces informations sont disposées de façon à être traitées automatiquement et efficacement par un logiciel, et pas nécessairement par un humain. Ils sont trouvés généralement dans les bases de données.
- **Les informations semi-structurées** [41]: dans une page Web, une partie de son contenu s'adresse à l'humain, comme le texte (informations non structurées), alors qu'une autre partie est destinée à la machine, comme les balises (informations structurées). Exemple : XML, HTML.
- **Les informations non structurées** [27]: c'est celles de tous les documents sur support numérique qui ne peuvent être utilisés que par l'homme. Exemple **:** documents textes et multimédias.
- **Multimédia [**34**]:** qualifie l'intégration de plusieurs moyens de représentation de l'information, tels que textes, sons, images fixes ou animées.

2.2.2.3.3 Les tâches du WCM

Quatre tâches sont illustrées sur la figure suivante :

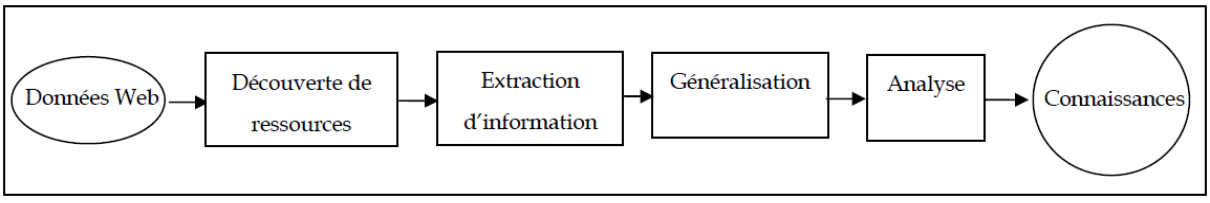

Figure 2.9. Tâches en WCM.

a. Découverte de ressources [15**]**

Cette phase s'intéresse à la recherche automatique de toutes les ressources pertinentes, en faisant recours aux techniques issues du domaine de la recherche d'information.

Cette étape fournit un ensemble de ressources peu ou prou pertinentes.

b. Extraction, sélection et prétraitement

Cette phase s'occupe de l'extraction d'information à partir des ressources obtenues, c.-à-d. de l'identification des fragments spécifiquement intéressants qui constituent le cœur d'un document Web. L'idéal serait de pouvoir développer des outils universels capables d'effectuer l'extraction d'information automatiquement et dynamiquement à partir de n'importe quelle ressource sur le Web [15]. Cette étape repose dans les travaux actuelles sur des outils issus du TAL (traitement automatique des langues) [19].

c. Généralisation [19]

C'est la tâche de fouille proprement dite, ou des techniques du DM sont appliquées pour l'extraction de connaissances à partir des données précédemment préparées et modélisées.

d. Analyse [19]

Dans le cadre du WM, cette tâche se manifeste, grâce à l'interactivité qu'offre le Web, par le travail qu'effectue l'expert pour la validation et l'interprétation des résultats extraits à l'issue de la tâche de généralisation.

2.3 Fouille de Texte (Text Mining)

2.3.1 Bref historique

Le Text Mining (TM) est apparue dans la deuxième moitié des années 90, en écho à des travaux réalisés depuis les années 80 sur des bases de données.

En 1991, Piatetsky-Shapiro introduit comme titre de son ouvrage le terme de Knowledge Discovery from Databases (KDD), en français, Extraction de Connaissances à partir de Bases de Données (ECBD). Ce n'est que vers 1995 que l'usage des termes Knowledge Discovery from Databases et Data Mining se précise [18].

2.3.2 Définition du TM

❖ **L'extraction de connaissances à partir de textes [18]**

L'extraction de connaissances à partir de textes est un processus non trivial qui construit un modèle de connaissances valide, nouveau, potentiellement utile et au final compréhensible à partir de textes bruts.

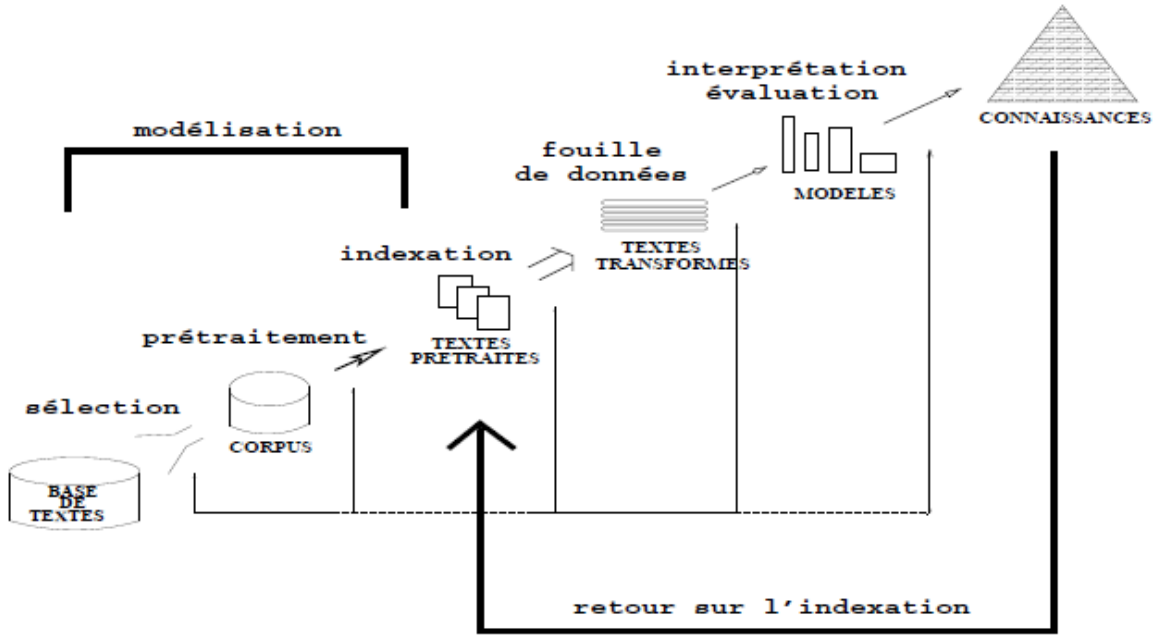

Figure 2.10. Schéma global de l'extraction de connaissances à partir de textes. [6]

Cette figure décrive le processus du TM qui est calculé sur le schéma de l'ECBD présenté précédemment et montre les différentes étapes de traitement dans un processus de TM.

Les données traitées sont constituées d'un ensemble de textes. Chaque texte est représenté par un ensemble de mots-clés. Cette représentation est stockée dans une base de données.

Un texte est considéré comme une entité porteuse d'une information qu'il faut préparer, représenter et organiser pour pouvoir utiliser des outils de fouille de données et valider les résultats de la fouille. La transformation des données textuelles en connaissances se compose donc de trois principales étapes :

a. La modélisation du contenu des textes ;
b. Les outils de fouille de données proprement dits ;
c. Le module d'analyse des résultats et leur validation.

2.3.3 Fouille de texte & fouille de données

Bien que le WCM utilise les techniques de fouille de données mais il reste différent du DM car la plupart des données sur le Web sont non structurées et / ou semi-structurés, tandis que le DM permet l'extraction automatique de connaissances à partir de données structurées.

Le WCM est associée au TM parce qu'une grande partie du contenu du Web est du texte. Le contenu du Web peut être semi-structuré, tandis que la fouille de texte permet d'extraire de l'information à partir d'un texte non structurés.

Finalement on peut déduire que la fouille de données est la base de la fouille de texte au sens où celui-ci est l'extension du même but et du même processus vers des données textuelles. En général le DM travaille sur des données structurées et stockées dans des bases de données. En revanche, le TM travaille sur des données textuelles non structurées. Donc l'objectif de la fouille de texte est le traitement de grandes quantités d'information qui sont disponibles sous une forme textuelle et non structurée [17]. Notons, que la fouille d'opinion est un sous domaine de la fouille de texte.

Chapitre

Opinion Mining & Sentiments Analysis

Chapitre 3 : Opinion Mining et Sentiments Analysis

Comme il a été dit précédemment, la fouille d'opinion (Opinion Mining, Sentiment Analysis ou Subjectivity Analysis) est un sous domaine de la fouille de texte. Son but étant de ressortir les marques d'opinions et de sentiments des documents textuels. Une opinion peut être définie comme l'expression des sentiments d'une personne envers une entité. En outre, l'e-commerce devient de plus en plus populaire. Les marchands et les fabricants de produits permettent aux clients de donner leurs avis et opinions sur les produits ou services qu'ils ont vendus (par exemple amazon.com, epinions.com). De plus, les opinions disponibles sur le Web influent sur nos choix et décisions [7].

3.1 Faits & opinions

Il existe deux catégories principales pour classer l'information textuelle : faits et opinions. Dans le premier cas, il s'agit de descriptions objectives (énoncé objectif) sur les entités et les événements dans le monde [23]. Dans l'autre cas, il s'agit d'expressions subjectives d'un individu à propos d'un objet ou d'un sujet particulier.

3.2 Le texte subjectif

La subjectivité est une expression linguistique de quelqu'un : opinions, avis, sentiments, émotions, évaluations, croyances ou jugements personnel [40].

L'étude des textes subjectifs s'est considérablement développée en directe relation avec leur accessibilité sur le web. Pour favoriser l'expression des internautes, il faut générer des textes non formatés, non indexés, non signés, et qui posent des problèmes d'analyse et de formalisation pour pouvoir être traités **[**31**]**.

L'analyse de sentiment a cependant fait le pari de faire la différence au niveau d'un document entre textes subjectifs et textes objectifs. Le problème de classification de textes devient alors crucial [23] :

- ✓ Repère d'abord les phrases porteuses d'opinion (classement objectif/subjectif)
- ✓ Puis, attribuer une polarité (positive, négative ou neutre).

3.3 Opinion Mining et Sentiments Analysis

Ces dernières années marquent une grande augmentation dans le nombre des travaux sur l'opinion Mining, ce qui montre l'importance de l'opinion sur le web.

L'opinion Mining est l'exploitation du Web 2.0. C'est un domaine qui s'occupe de traitement d'opinion, de sentiment, et de la subjectivité dans le texte. Il devrait traiter un ensemble de résultats de recherche pour un cas donné, générer une liste des attributs (qualité, caractéristiques, etc.) et agréger des avis sur chacun d'entre eux (mauvais, modéré, de bonne qualité) **[**5**]**.

Le domaine de la fouille d'opinion peut-être divisé en trois sous-domaines [30] :

- ✓ l'identification des textes d'opinion : consiste à identifier et localiser les parties de textes porteuses d'opinion dans une collection textuelle, c.-à-d. classer les textes ou les parties de texte selon (objectifs ou subjectifs);
- ✓ le résumé d'opinion : consiste à rendre l'information rapidement et facilement accessible en mettant en avant les opinions exprimées et les cibles de ces opinions présentes dans un texte.
- ✓ la classification d'opinion : son objectif est d'attribuer une étiquette au texte selon l'opinion qu'il exprime (positive, négative et neutre).

L'analyse des sentiments est utilisée pour décrire l'analyse automatique de texte évaluatif et pour la recherche de valeur prédictive des jugements **[**5**]**.

Aujourd'hui, l'opinion Mining et l'analyse des sentiments font partie du même domaine de recherche.

3.4 Les approches de détection d'opinions

Il existe trois types d'approches pour la détection d'opinions et l'analyse du sentiment :

3.4.1 Les approches statistiques

(Appelées aussi classification supervisées, approches basées sur l'apprentissage machine ou encore approches basées sur corpus).Elles sont des techniques d'apprentissage automatique [49]. Ces approches regroupent les documents (ou les mots) dans deux axes de classification, soit dans l'opposition (subjectif-objectif), soit dans la distinction des opinions subjectives dans l'opposition (positif-négatif) [49] [26]. Ceci, en utilisant un corpus qui a été

déjà annoté manuellement au préalable, dans le but de le faire apprendre au système [26]. Ces approches consistent à attribuer les données à un classifieur qui génère un modèle qui est utilisé pour la partie test de l'apprentissage. Ce type d'approche comprend deux aspects : extraction de features et apprentissage du classifieur. Les principales features utilisées sont : mots seuls, bigrammes, tri-grammes, part of speech et polarité. Les principaux classifieurs sont les Support Vector Machines (SVM), Naive Bayes, Maximum Entropy et régression logistique [31] [28].

3.4.2 Les approches symboliques

(Appelées aussi classification non supervisées ou encore approches basées sur lexique). Elles utilisent des dictionnaires de mots subjectifs. Ces dictionnaires peuvent être généraux comme par exemple (General Inquirer, Sentiwordnet, Opinion Finder, NTU, etc.) ou construit manuellement, soit généré automatiquement à partir du corpus (les mots qui contiennent une opinion sont extraits directement du corpus). Dans ces dictionnaires, une polarité est associée à priori à chacun de ces mots. Il est donné ensuite au document un score d'opinion égale au nombre total de mots qui contiennent une opinion présente dans le document [31] [28].

3.4.3 Les approches hybrides

(Appelées aussi classification semi-supervisées)**.** Ces approches combinent les points forts des deux approches précédentes. Elles prennent en compte tout le traitement linguistique des approches symboliques avant de lancer le processus d'apprentissage comme dans les approches statistiques [26].

La combinaison des approches symboliques et statistiques a donné des résultats plus précis que chacune des approches employées séparément [49].

3.5 Processus de la fouille d'opinions

La figure suivante montre les étapes de la fouille d'opinions [4].

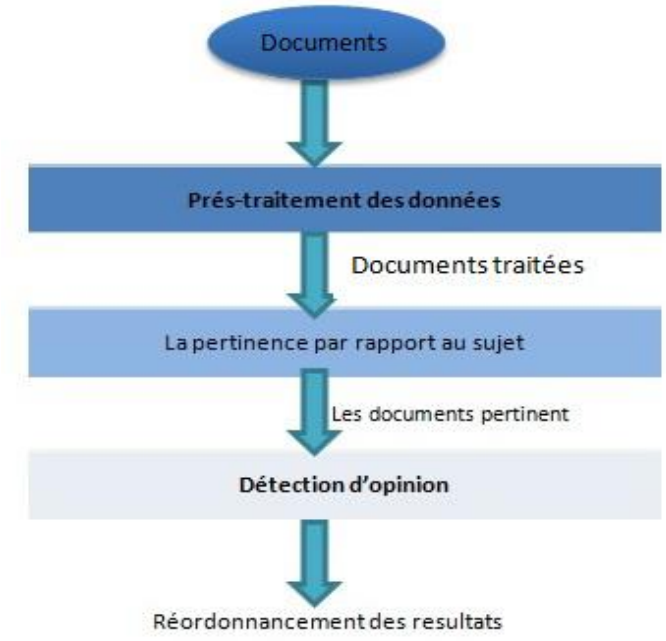

Figure 3.1. Processus de fouille d'opinions.

a. Acquisition et prétraitement des données

Dans cette phase, les textes sont prétraités linguistiquement en éliminant les mots vides et les mots qui n'apportent aucune information importante, ainsi qu'une analyse lexicale pour enlever les mots qui ont le même sens. Dans cette étape, un étiquetage grammatical est fait (pour reconnaître l'adjectif, l'adverbe, le verbe, etc.), les grammaires de dépendances sont utilisées pour structurer la phrase de manière hiérarchique.

b. La pertinence par rapport au sujet

Cette phase consiste à étudier la pertinence des documents par rapport à un sujet donné. Les documents sont classés, et généralement les 1000 premiers documents les plus pertinents sont extraits, et sont utilisés pour l'étape suivante.

c. La détection d'opinions

La détection d'opinions utilise plusieurs méthodes pour le but de réordonner les documents pertinents selon un score d'opinion.

3.6 Domaines d'application

L'importance de la détection d'opinion est présente dans plusieurs domaines, mais la plus grande application de l'Opinion Mining reste dans le monde du business, et du politique.

❖ **Marketing**

- **Coté entreprises**

L'opinion Mining permet au fournisseur d'un produit ou d'un service d'avoir plus de connaissances sur les consommateurs, pour anticiper leurs besoins et leurs attentes afin de tenter d'améliorer la qualité du produit/service et d'augmenter les profits.

Les professionnels du marketing digital, du service client et de la communication cherchent aujourd'hui l'outil idéal leur permettant d'analyser automatiquement la teneur et la tonalité des conversations publiées sur leurs marques sur le web [50].

Le domaine de la fouille d'opinion est devenu un enjeu majeur pour toute entreprise désireuse de mieux comprendre ce qui plait et déplait à ses clients [30].

- **Coté clients**

Le client peut de son côté

- ✓ Donner son opinion.
- ✓ S'inspirer des sentiments et opinions d'autres clients sur le produit auquel il s'intéresse et profiter ainsi d'une aide à la décision [49].
- ✓ Comparer les produits avant de les acquérir [30].
- ✓ Ne pas lire tous les commentaires concernant un produit donné, il suffit de voir le pourcentage positifs associé à ce produit.

❖ **Politique**

Les documents portant des textes rapportant des débats politiques, (un corpus de réactions à des propositions de lois) sont de taille très importante. Ils contiennent environ dix fois plus de textes que celui des critiques d'un produit. Ce qui implique la variété énorme des jugements politiques dans ces articles. Ces derniers se retrouvent au cœur des débats avec l'avènement des médias sociaux (et plus spécialement sur le Web). Les acteurs politiques ont également suivi cette tendance, tel qu'avant de promulguer une nouvelle loi, les politiciens essayent de récolter l'avis des internautes sur cette loi. Il est intéressant de connaître aussi l'avis des internautes sur tel homme politique pour une élection présidentielle [4].

3.7 Difficultés de la fouille d'opinions et de l'analyse de sentiments

Ici la classification de textes d'opinion (positive, négative ou neutre) a pour objectif l'analyse de sentiments exprimés dans les groupes de discussions, les blogs, les forums et autres sites spécialisés dans les critiques de produits.

Cependant, cette classification se heurte à quelques difficultés :

- ✓ Difficulté due à l'ambiguïté de certains mots positifs ou négatifs selon les contextes et qui ne peut pas toujours être levée [31].
- ✓ Difficulté due au contexte : la nécessité d'une bonne analyse syntaxique du texte ; analyse qui peut se révéler particulièrement difficile dans des cas de coordination entre plusieurs parties d'une phrase [49]. Par exemple «les acteurs du film ont bien joué, la musique est bonne mais je n'ai pas aimée l'histoire», l'opinion de la dernière partie de la phrase est la plus importante.
- ✓ Difficulté due au langage naturel pour l'analyse automatique de sentiments selon les contextes intentionnels, pour lesquels l'expression d'opinion n'est pas un vrai sentiment. C'est le cas dans une phrase comme : « Je croyais que la France était un beau pays. »
- ✓ Difficulté due aux structures syntaxiques et sémantiques d'une phrase et l'expression de l'opinion qu'elle véhicule [49]. Par exemple « l'histoire du film est intéressante mais les acteurs étaient mauvais ». Dans ce cas la polarité de la deuxième partie est opposée à la première [4].
- ✓ Difficulté due à l'analyse de la phrase par « paquets de mots ». Les deux phrases suivantes contiennent les mêmes paquets de mots sans pour autant exprimer les mêmes sentiments. La première phrase contient un sentiment positif alors que la deuxième est négative : « Je l'ai apprécié pas seulement à cause de ... », « Je l'ai pas apprécié seulement à cause de ... » [49].

3.8 L'Opinion Mining & le e-commerce

Depuis l'apparition du commerce en ligne (e-commerce), la plupart des entreprises l'intègrent. Ceci explique la grande augmentation de chiffre d'affaires des sites de vente en ligne. Le e-commerce a permis à nombre de sociétés de toutes tailles d'explorer de nouvelles voies commerciales. En effet, l'entreprise présente sur Internet a la possibilité de mieux appréhender les besoins de ses clients.

L’intérêt massif pour la fouille d’opinion est directement lié à une demande sociale forte, à savoir l’expansion fulgurante du commerce en ligne. Comment accéder aux jugements privés relatifs à tel ou tel produit afin de mieux anticiper les besoins et mieux évaluer l’impact du marketing visant tel ou tel segment de consommateurs ? Quelle que soit la qualité du produit ou service, le fait est que personne ne l'achètera si le client ne le veut pas ou pense qu'il n'en a pas besoin. Donc connaître et comprendre les besoins du client est au centre de toute entreprise fructueuse. Lorsque l’entreprise possède cette connaissance, elle peut l'utiliser pour persuader les clients éventuels et existants qu'il est dans leur intérêt d'acheter chez elle. Ainsi que chaque entreprise a besoin d'avoir une raison pour laquelle ses clients achètent chez elle et non chez ses concurrents [38] [33].

Chapitre

Présentation des outils réalisés dans le domaine de l'Opinion Mining

Chapitre 4 : Présentation des outils réalisés dans le domaine de l'Opinion Mining

Ici, nous présenterons et comparerons les principaux travaux réalisés dans le domaine de la fouille du Web et spécialement sur le réseau social Twitter.

4.1 Analyse de la tonalité sur Twitter

Twitter est un réseau social, avec un nombre moyen de 200 millions de messages courts (tweets) qui ne peuvent pas dépasser 140 caractères envoyés par jour [35]. Twitter permet de recueillir des opinions spontanées sur une variété de sujets.

Dans le cadre de l'analyse des sentiments, la petite taille de message formule l'hypothèse que ce message ne renferme pas a priori plus d'une seule idée, ce qui facilite l'identification de la cible d'une opinion. Mais certains tweets apparaître aux non-initiés comme des messages codés tant l'usage des hashtags, abréviations en tout genre, argot, anglicismes, et autres émoticons y est répandu [31].

Nous allons citer quelques services proposant d'analyser la tonalité des messages partagés sur Twitter :

4.1.1 Sentiment140

Sentiment140 (anciennement connu sous le nom "Twitter sentiment") est un outil en ligne gratuit qui a été créé par trois étudiants en computer science de Stanford, donc c'est un projet académique. Cet outil, contrairement à la plupart des autres sites d'analyse de sentiments, n'utilise pas de listes de mots positifs ou négatifs mais est fondé sur les algorithmes d'apprentissage automatique [21].

Sentiment140 permet de découvrir des sentiments des tweets d'une marque, un produit ou un sujet sur Twitter.

❖ **Utilisation de l'outil Sentiment140**

- ✓ Accédez au lien http://www.sentiment140.com/ ;
- ✓ Tapez dans la boîte de recherche le nom d'un produit, d'une célébrité ou d'une marque exemple « iPhone 4s ».
- ✓ Cliquez sur le bouton «Search» comme dans la figure suivante.

Figure 4.1. L'interface de Sentiment140.

Les résultats détaillés pour la requête « iPhone 4s » s'affiche (figure suivante) :

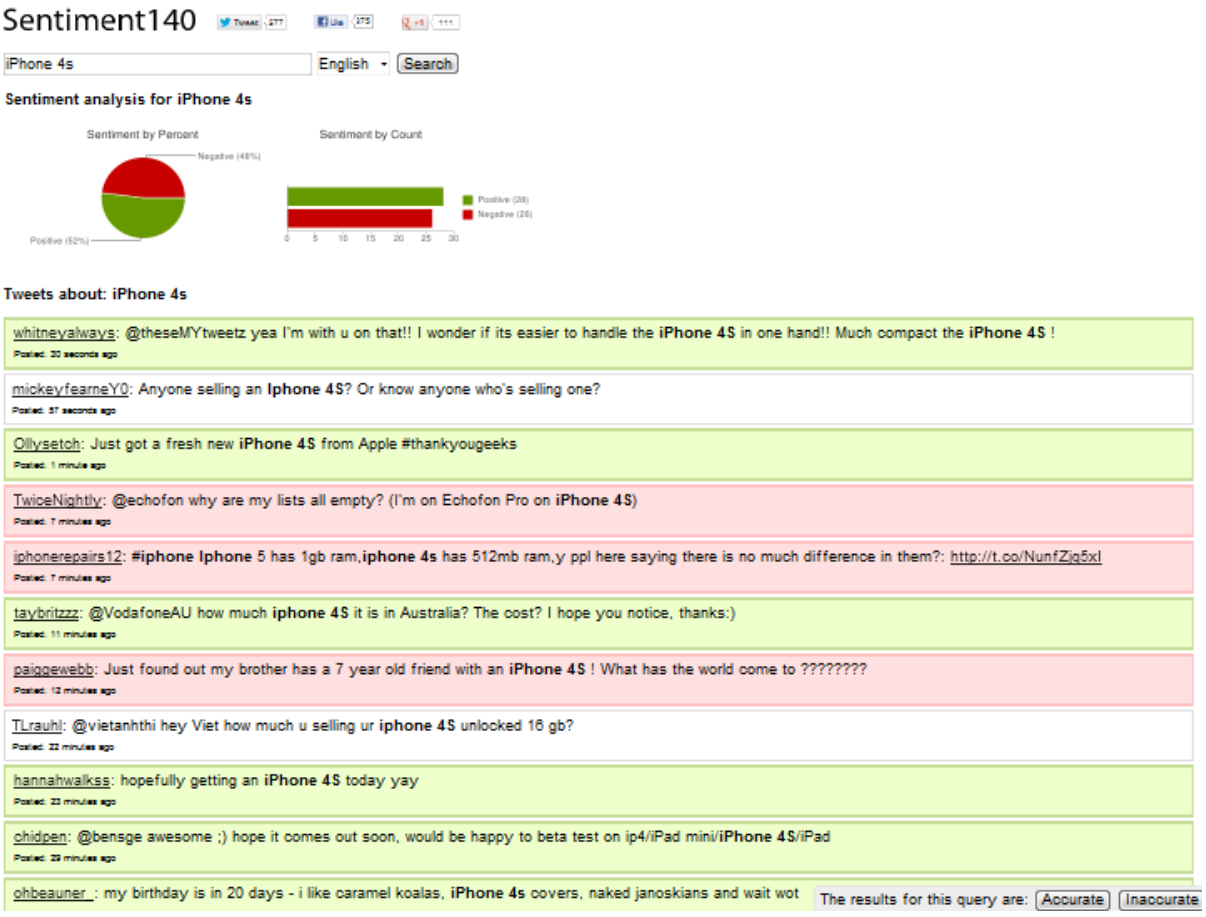

Figure 4.2. Résultats de la requête « iPhone 4s ».

4.1.2 Tweetfeel

Le service Tweetfeel est un outil en ligne d'analyse du sentiment sur Twitter. Il propose une version gratuite et une version payante. Il s'appuie sur les capacités temps réels de Twitter qui donne des sentiments positifs et négatifs des tweets sur des choses comme les films, musiciens, émissions de télévision et de marques populaires [39].

L'évaluation de TweetFeel se fait sur la base de présence de mots clés précis dans les tweets tels que Good, Bad, etc. Ensuite un pourcentage est calculé selon le nombre de tweets positifs ou négatifs un sentiment global de Twitter sur la marque [52].

❖ **Utilisation de l'outil TweetFeel**

- ✓ Accédez au lien http://www.tweetfeel.com/;
- ✓ Tapez dans la boîte de recherche le nom d'un produit, d'une célébrité ou d'une marque exemple « iPhone 4s » ;
- ✓ Cliquez sur le bouton «Search» comme sur la figure suivante.

Figure 4.3. L'interface de TweetFeel.

Les résultats des tweets portant sur « iPhone 4s » (figure suivante) :

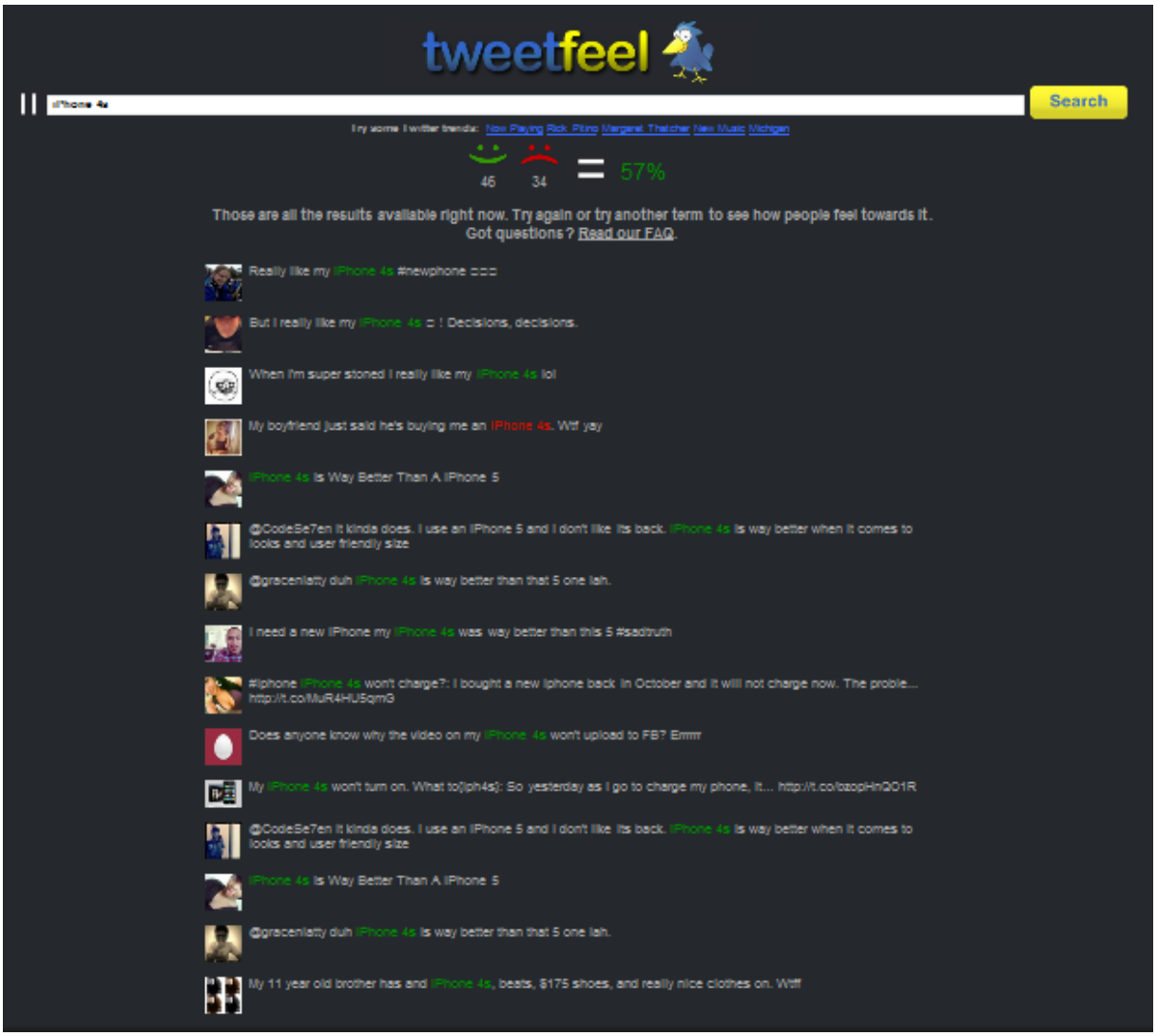

Figure 4.4. Résultats de la requête « iPhone 4s ».

4.1.3 twitrratr

twitrratr est un outil en ligne gratuit, qui a émergé à partir d'un projet Startup Weekend. Twitrratr fonctionne à partir d'une liste de mots positifs et d'une liste de mots négatifs [43].

Cet outil classe une opinion sur le mot clé de la requête s'il est capable de le croiser avec un mot d'une des deux listes. Les mots positifs et négatifs qui servent à classer les tweets sont surlignés dans l'interface [43].

❖ **Utilisation de l'outil twitrratr**

- ✓ Accédez au lien http://twitrratr.com/;
- ✓ Tapez dans la boîte de recherche le nom d'un produit, d'une célébrité ou d'une marque exemple « iPhone 4s » ;
- ✓ Cliquez sur le bouton «Search» comme sur la figure suivante

Figure 4.5. L'interface de twitrratr.

Les résultats des tweets sur « iPhone 4s » apparaissent sur la figure suivante:

Figure 4.6.Résultats de la requête « iPhone 4s ».

4.1.4 Tweet Sentiments Analysis

Tweet Sentiments Analysis est un outil en ligne gratuit et open source d'analyse du sentiment sur Twitter. Il peut donner des sentiments positifs, négatifs et neutres des tweets sur le mot clé lancé dans la requête. Il peut travailler sur 12 langues. Il donne les résultats sous forme graphique.

❖ **Utilisation de l'outil Tweet Sentiments Analysis**

- ✓ Accédez au lien http://smm.streamcrab.com/
- ✓ Tapez dans la boîte de recherche le nom d'un produit, d'une célébrité ou d'une marque exemple « iPhone 4s » ;
- ✓ Cliquez sur le bouton «Search» comme sur la figure suivante.

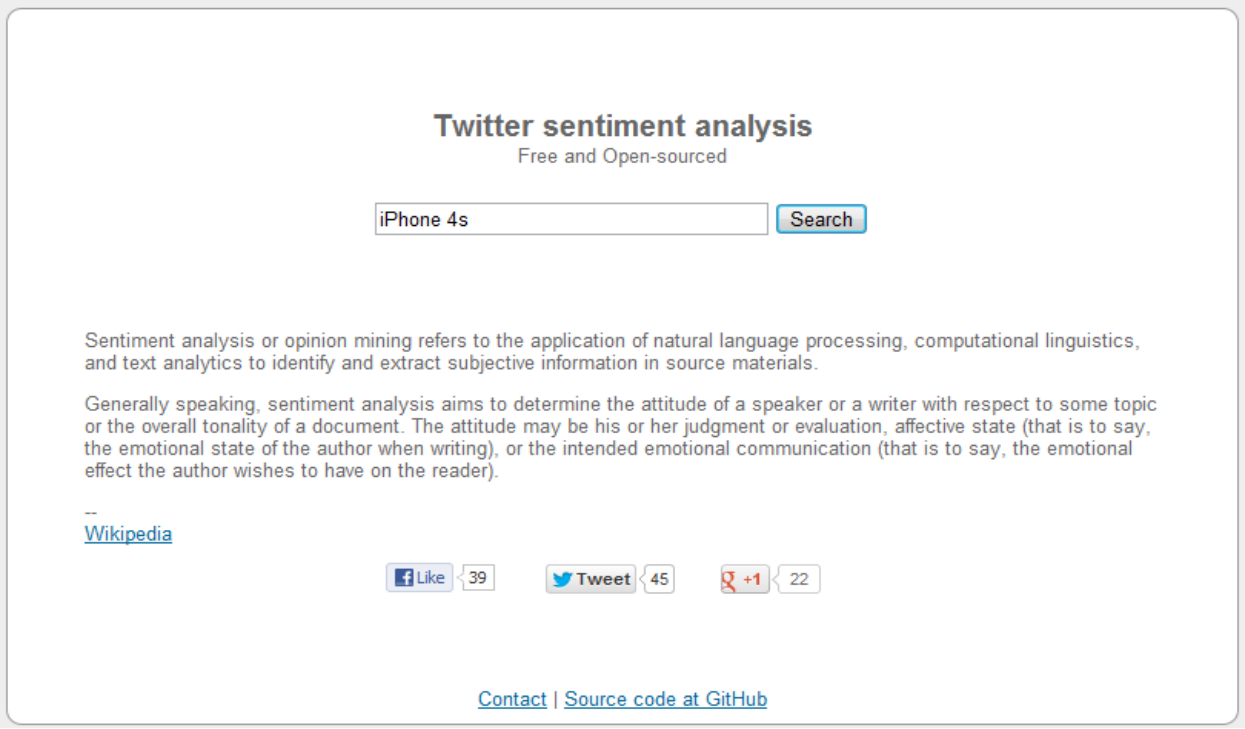

Figure 4.7. L'interface Twitter Sentiment Analysis.

Les résultats des tweets sur « iPhone 4s » sont montrés sur les figures suivantes :

Figure 4.8. L'analyse de sentiment.

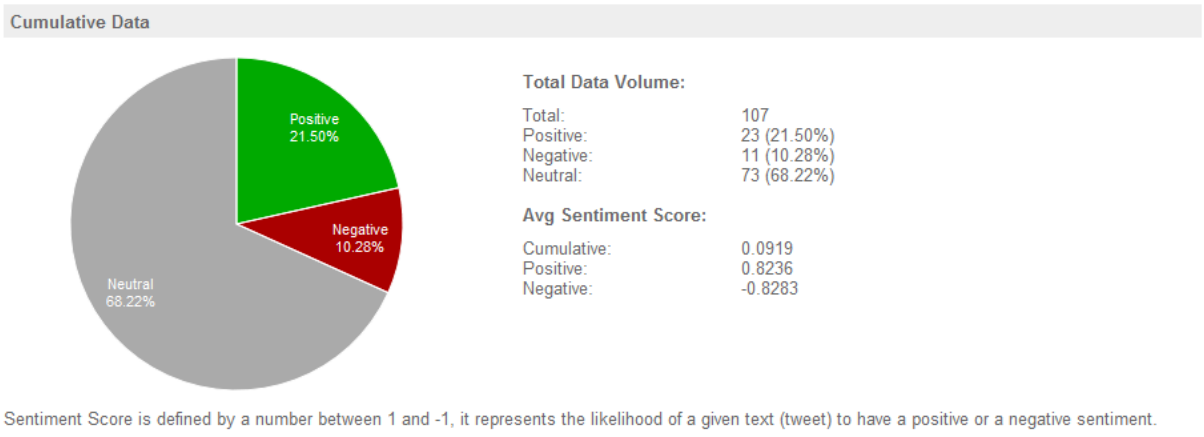

Figure 4.9. Les données cumulatives.

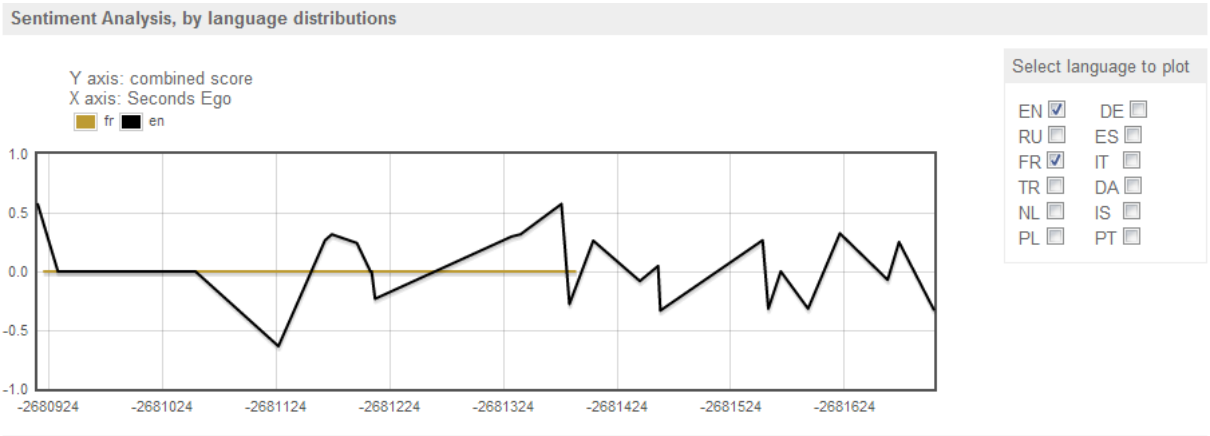

Figure 4.10. L'Analyse de Sentiment, par des distributions de langue.

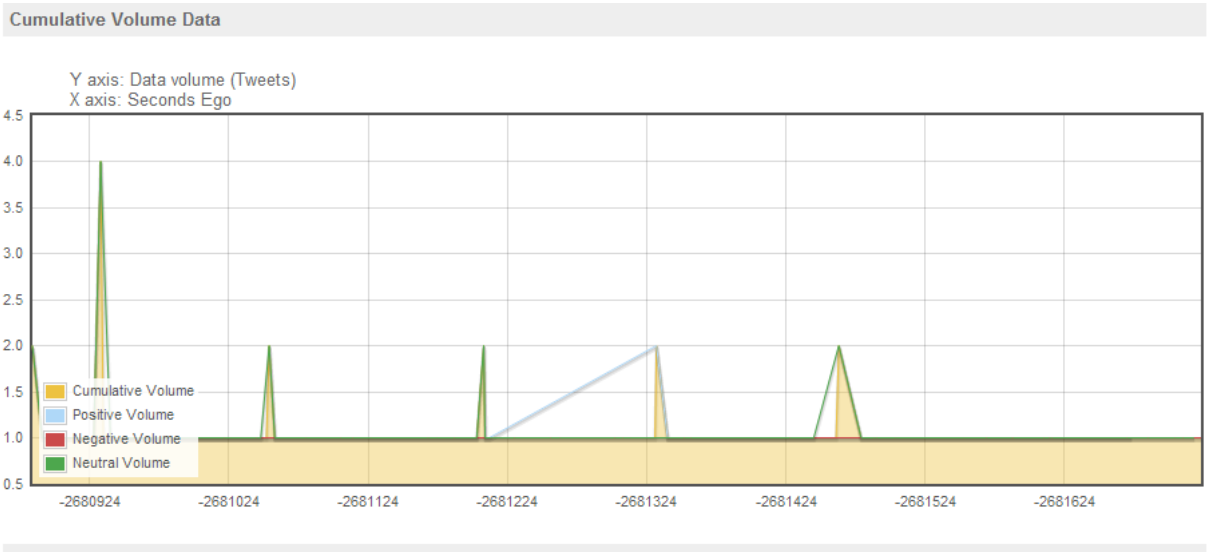

Figure 4.11. Les données de volume cumulatif.

Figure 4.12. Les tweets.

4.2 Tableau comparatif

Le tableau suivant représente les différents outils présentés et leurs caractéristiques :

L'application / Caractéristiques	Sentiment 140	TweetFeel	Twitterratter	Twitter Sentiment Analysis
Version gratuite	Oui	Oui	Oui	Oui
Version payante	Non	Oui	Non	Non
Open Source	Oui	Oui	Oui	Oui
On ligne	Oui	Oui	Oui	Oui
Simple à utiliser	Oui	Oui	Oui	Oui
Analyse des sentiments	Oui	Oui	Oui	Oui
Axé au twitter	Oui	Oui	Oui	Oui
Avoir un compte twitter	Oui	Non	Non	Non
Avec une démo	Oui	Oui	Oui	Oui
Temps réel	Non	Oui	Non	Oui
Mots clé lancé dans la requête	Marque, produit, célébrité, un sujet sur twitter	Film, célébrité, entreprise, produit, marque.	Produit, célébrité, un sujet sur twitter.	Produit, célébrité, un sujet sur twitter.
Classification	Positive, négative.	Positive, négative.	Positive, négative et neutre.	Positive, négative et neutre.
Langues disponibles	Anglais, espagnol.	Anglais.	Anglais.	Anglais, français, russe, allemand, néerlandais, turc, polonais, espagnol, italien, islandais, danois et portugais.
Approche utilisée	Apprentissage automatique.	la version gratuite (basé sur dictionnaire), la version payante (apprentissage automatique)	Basé sur dictionnaire.	Apprentissage automatique.

Tableau 4.1 . Tableau comparatif.

Chapitre

Réalisation

Chapitre 5 : Réalisation

Avant d'aborder la réalisation proprement dite de notre application, nous allons d'abord présenter les principaux outils utilisés, en général, pour le développement dans le domaine de l'Opinion Mining. Ceci nous conduira à la motivation quant au choix de l'environnement Python pour l'implémentation de notre système.

5.1 Principaux outils de développement pour l'Opinion Mining

Il existe plusieurs outils de Data Mining et de Text Mining utilisés dans le domaine d'Opinion Mining. Ces outils peuvent être intégrer comme des APIs Java. Ci-dessous, une brève présentation de ces outils.

5.1.1 GATE

- **Présentation générale**

GATE (General Architecture for Text Engineering) est une boîte à outils logicielle open source et gratuite développée en Java par l'université de Sheffield en Grande-Bretagne. Elle est utilisée pour le développement et le déploiement des technologies du traitement automatique de la langue à grande échelle [11]. Ces outils sont utilisés par une large communauté, autant scientifique que professionnelle [48].

- GATE offrent une variété d'outils très variés :
  - ✓ **GATE Developers** : est un environnement de développement avec une interface graphique qui sert principalement à annoter des documents (extraction d'informations).
  - ✓ **GATE Cloud** : une solution de cloud-computing pour l'hébergement à grande échelle de traitement de texte.
  - ✓ **GATE Embedded** : est la librairie permettant d'utiliser tous ces outils dans une application.
  - ✓ **GATE Teamware** : est une annotation collaborative basée sur le Web et de conservation de l'environnement ; ce qui permet aux annotateurs non qualifiés de se former et ensuite l'utiliser pour abaisser le coût des projets d'annotation de corpus.
  - ✓ **GATE Mimir** : peut être utilisé pour indexer et effectuer une recherche sur un texte, des annotations, des schémas sémantiques (ontologies) et des méta-données sémantiques (données d'instance).

❖ Les ressources de GATE, sauf celles de bases, sont sous la forme de plugins. Ces derniers sont faciles à créer et à mettre en place. Certains sont présents par défaut dans GATE :

✓ **ANNIE** (Nearly-New Information Extraction System) : un système d'extraction d'informations permettant d'extraire des informations tels que les entités nommées de temps (des dates), de personnes, de lieux (des adresses).

✓ **CREOLE** (gestion de certaines langues, travail sur les ontologies, etc.) : il est une sorte de "super-plugin" qui fournit de nombreuses ressources.

❖ **Installation**

GATE s'exécutera sous n'importe quel système d'exploitation (Solaris, Linux, Mac OS X et Windows) supportant Java 6 ou une version ultérieure.

✓ Il suffit de télécharger l'installeur de puis cette adresse : http://gate.ac.uk/download/

✓ Puis il faut exécuter et suivre les instructions à l'écran suivant :

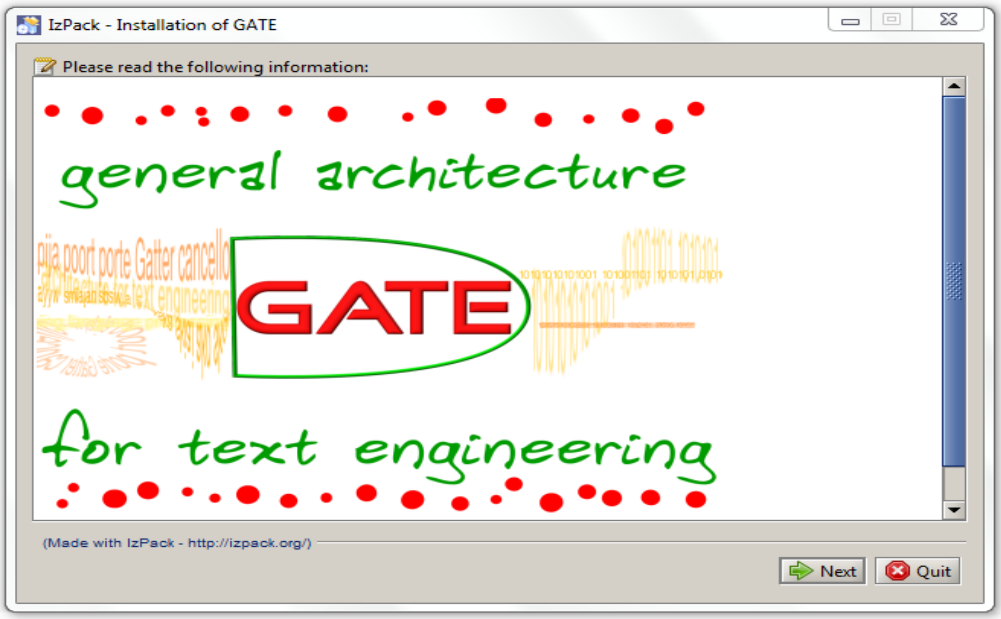

Figure 5.1. IzPack- L'installation de GATE.

5.1.2 LingPipe

❖ **Présentation générale**

LingPipe est une bibliothèque logicielle de traitement de langage naturel implémentée en Java. LingPipe fournit un ensemble d'outils pour le traitement de texte, il est utilisé pour effectuer des tâches telles que [22]:

✓ Trouvez les noms des personnes, des organisations ou des lieux.

- ✓ Classer automatiquement les résultats de recherche sur Twitter dans des catégories.
- ✓ Proposer une orthographe correcte des requêtes.

LingPipe propose des versions commerciales payantes, et une version gratuite, qui ne permet qu'une utilisation en production limitée. Il est disponible sous les plates-formes (Linux, Windows et de Mac OS X) [22].

❖ **Installation**

- ✓ Télécharger LingPipe sous le lien: http://alias-i.com/lingpipe/web/download.html ;
- ✓ Décompresser le fichier dans un nouveau répertoire appelé *LingPipeDir ;*
- ✓ LingPipe nécessite une installation de Java 5 ou une version ultérieure pour l'exécuter et le JDK pour compiler.

❖ **Exécution de l'interface Demos**

Les démos de LingPipe sont disponibles sur le web, comme ils sont aussi disponible localement grâce à une interface graphique d'utilisateur (GUI). Chaque démonstration est appelée à partir d'une ligne de commande spécifique. Le texte est entré dans le cadre Input. Le texte peut être fourni de trois façons différentes:

- ✓ **Text Input** : le texte peut être saisi directement dans le champ de texte, ou par le moyen couper-coller.
- ✓ **File Selection** : un clic sur le bouton « Select File » puis choisir le fichier voulu.
- ✓ **Drag and Drop** : fichiers (ou d'autres sélections) peuvent être glissés et déposés dans la fenêtre de saisie.

Un clique sur le bouton « Analyze », et le résultat s'affichera dans le cadre à droite.

- o **Les données entrantes (Input)** : peuvent être sous les trois formats : texte brute, HTML, ou XML.
- o **Les données sortantes (output)** : le format de sortie est XML.

Ci-dessous, la liste des démos :

- **Echo Demo**: retourne le texte donné.
- **Sentence Demos** : permet d'extraire des phrases à partir du texte.
- **Part of Speech Demo**: attribuer les parties du discours à des mots.
- **Named Entity Demo**: permet d'extraire l'entité mentionné à partir du texte.
- **Within-Document Coreference Demo**: une coréférence détermine quand deux entités mentionnées dans un texte réfèrent à la même entité dans le monde.

- **Chinese Word Segmentation Demo**: permet de découper le texte chinois en mots.

❖ Par exemple la démo de l'écho :

- ✓ hébergé sur le web à l'adresse suivante: http://lingpipe-demos.com:8080/lingpipe-demos/echo/textInput.html
- ✓ Pour lancer Echo Demo sous Windows :

  >cd % LINGPIPE_HOME% \demos\generic\bin

  > gui_echo.bat
- ✓ Pour lancer Echo Demo sous autre plateforme :

  >cd $LINGPIPE_HOME/demos/generic/bin

  >sh gui_echo.sh

La figure suivante s'affiche:

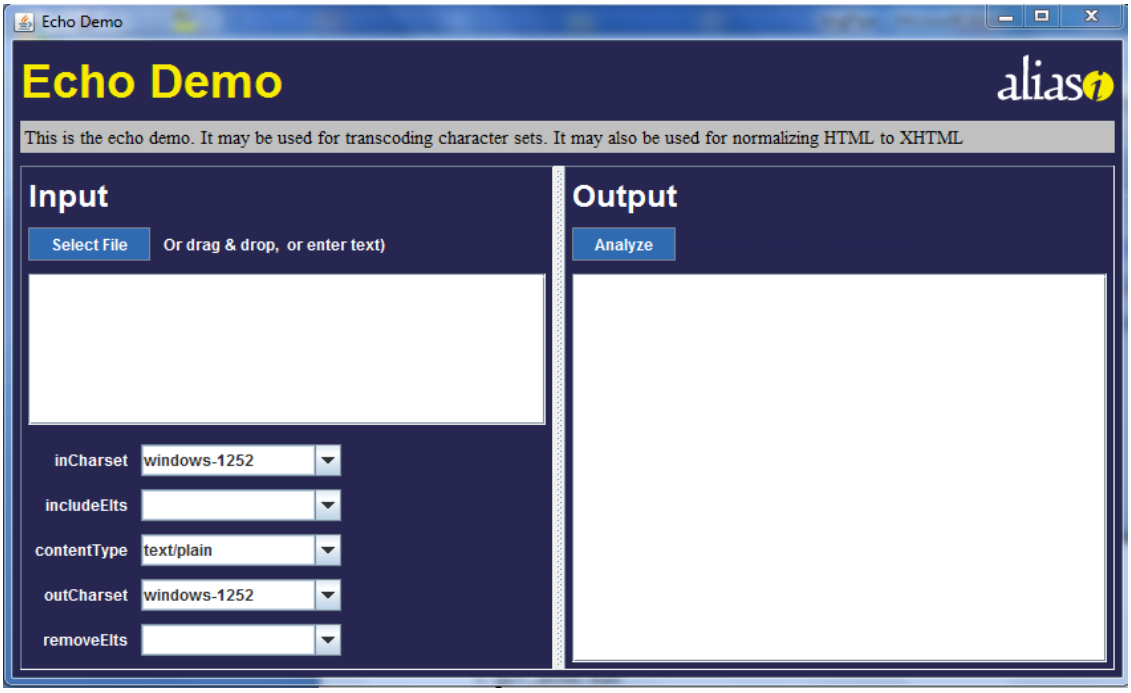

Figure 5.2. L'interface Echo Demo.

Ensuite, il suffit d'entrer le texte puis cliquer sur « Analyze », le résultat s'affichera comme sur la figure suivante :

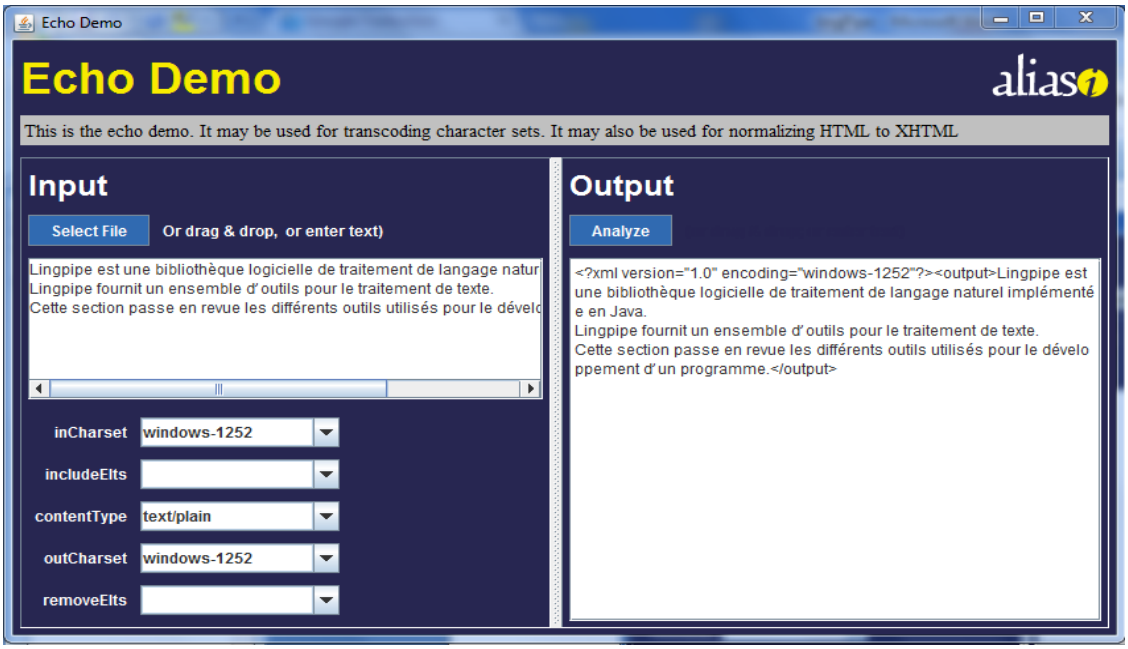

Figure 5.3. Le résultat Echo Demo.

5.1.3 RapidMiner

❖ **Présentation générale**

RAPIDMINER est un logiciel gratuit et open-source issu du projet YALE de l'équipe « Intelligence Articificielle » de l'Université de Dortmund. Ce projet a été repris par la Société Rapid-I qui maintient deux versions en parallèle [45]:

- ✓ Une version distribuée gratuitement « Community Edition » ;
- ✓ Une version commerciale « Enterprise Edition ».

Ce logiciel est dédié au Data Mining. Il contient de nombreux outils pour l'analyse et le traitement des données. Il est aussi bien adapté pour une variété de tâches d'extraction de texte [45].

❖ **Installation**

- **Pour Windows**
  - ✓ Il suffit de télécharger l'installeur depuis cette adresse : http://rapid-i.com/;
  - ✓ Puis, il faut exécuter et suivre les instructions à l'écran suivant :

Figure 5.4. Setup de RapidMiner.

- **Pour les autres plateformes**

  RapidMiner s'exécutera sous les autres systèmes d'exploitation supportant Java 7 ou une version ultérieure.

  - ✓ télécharger le fichier Zip RapidMiner ;
  - ✓ puis extraire ce fichier.

❖ **Pour ouvrir RapidMiner**

- **Pour Windows**

  Un double clic sur l'icône sur le bureau ou effectuez un clic sur le menu démarré.

- **Pour les autres plateformes**

  Il existe deux façons pour ouvrir RapidMiner :

  a) Définir la quantité maximale de mémoire pouvant être utilisée par Java et l'emplacement de Java lui-même via les variables d'environnement JAVA_HOME et MAX_JAVA_MEMORY. Ensuite, un double-clic sur *RapidMinerGUI.bat* dans le sous-répertoire scripts dans le répertoire RapidMiner ou le lancer à travers la console commande.

  b) certaines plateformes fonctionnent tout simplement en effectuant un double-clic sur le fichier lib/rapidminer.jar.

L'interface principale de RapidMiner est le suivant :

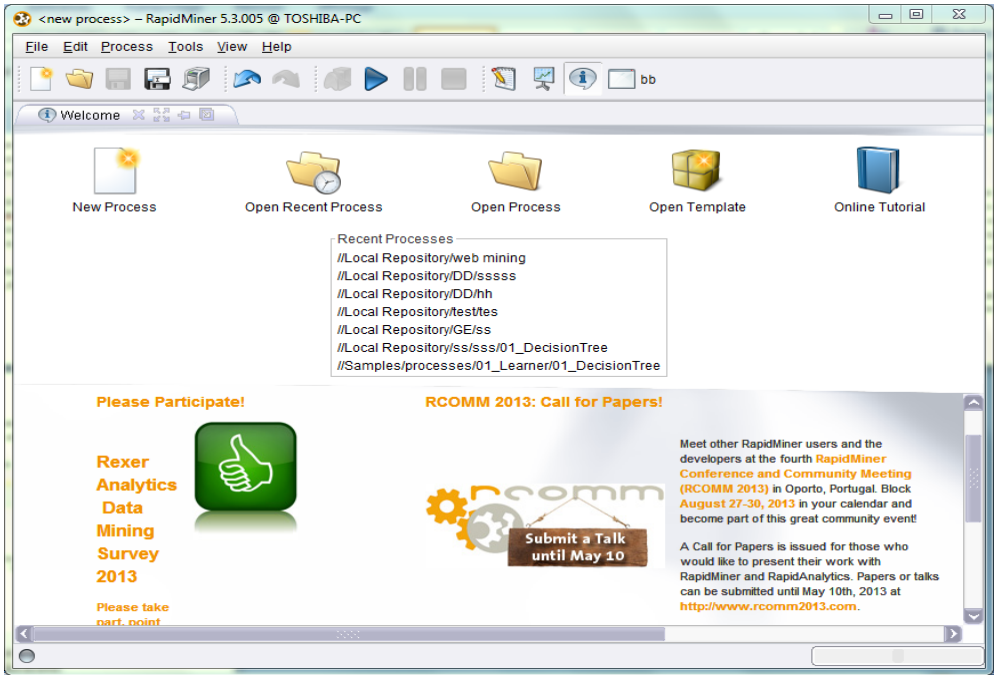

Figure 5.5. L'interface de RapidMiner.

Sur cette interface, il existe les icônes suivantes :

- ✓ **New Process** : il démarre un nouveau processus d'analyse.
- ✓ **Open Recent Process** : il ouvre un processus déjà existant à partir de la liste « Recent Processes ».
- ✓ **Open process** : il permet d'ouvrir « Repository browser » l'emplacement des processus existants.
- ✓ **Open templates** : il affiche une liste des processus prédéfinis.
- ✓ **Online Tutorial** : il démarre un tutoriel utilisé directement dans RapidMiner.

5.2 Motivations pour l'environnement de développement

Comme nous l'avons vu précédemment, les outils présentés sont difficiles à utiliser pour les raisons suivantes :

- ✓ Ce sont de nouveaux logiciels, ce qui implique un manque de documentation.
- ✓ Leurs apprentissage est souvent long et difficile.
- ✓ Certains parmi ces outils nécessitent en entrée des données bien structurées.

- ✓ Enfin, pour pouvoir les utiliser dans une application donnée, il faut les intégrer uniquement sous Java du moment où leurs implémentation est faite en Java.

Cependant, le langage Python est un langage de programmation de haut niveau, facile à apprendre, orienté objet, totalement libre et terriblement efficace. Il est conçu pour produire du code de qualité, portable et facile à intégrer. Ainsi la conception d'un programme Python est très rapide et offre au développeur une bonne productivité. En tant que langage dynamique, il est très souple d'utilisation et constitue un complément idéal à des langages compilés. Contrairement à des langages spécifiques comme PHP qui se focalise sur un domaine précis, Python est universel. Il peut être utilisé dans un grand nombre de contextes. Un autre avantage de Python est la richesse de ses bibliothèques ainsi que la qualité du développement des interfaces graphiques offerte par le module wxPython. C'est toutes ces nombreuses et importantes caractéristiques qui font la force de ce langage. Contrairement à certains langages comme Java, Python est facile à manipuler et peut s'apprendre sans formation. Il est aussi portable que Java. Il est plus général que Java ; il est utilisé dans une grande variété de domaines. Il possède aussi des bibliothèques beaucoup plus riches que celle de Java.

5.2.1 Généralités sur Python

Python est un langage développé depuis 1989 par Guido van Rossum et de nombreux contributeurs bénévoles. Il permet -sans l'imposer- une approche modulaire et orientée objet de la programmation [37].

❖ Caractéristiques

- ✓ Python est gratuit, mais on peut l'utiliser sans restriction dans des projets commerciaux.
- ✓ Python est portable, non seulement sur les différentes variantes d'Unix, mais aussi sur les OS propriétaires: MacOS, BeOS, NeXTStep, MS-DOS et les différentes variantes de Windows. Un nouveau compilateur, baptisé JPython, est écrit en Java et génère du bytecode Java.
- ✓ Python convient aussi bien à des scripts d'une dizaine de lignes qu'à des projets complexes de plusieurs dizaines de milliers de lignes.
- ✓ La syntaxe de Python est très simple et, combinée à des types de données évolués (listes, dictionnaires,...), conduit à des programmes à la fois très compacts et très lisibles.

- ✓ Python gère ses ressources (mémoire, descripteurs de fichiers, ...) sans intervention du programmeur, et ce, par un mécanisme de comptage de références (proche, mais différent, d'un garbage collector).
- ✓ Il n'y a pas de pointeurs explicites en Python.
- ✓ Python est (optionnellement) multi-threadé.
- ✓ Python est orienté-objet. Il supporte l'héritage multiple et la surcharge des opérateurs.
- ✓ Python intègre, comme Java ou les versions récentes de C++, un système d'exceptions, qui permettent de simplifier considérablement la gestion des erreurs.
- ✓ Python est dynamique (l'interpréteur peut évaluer des chaînes de caractères représentant des expressions ou des instructions Python), orthogonal (un petit nombre de concepts suffit à engendrer des constructions très riches), réflectif (il supporte la métaprogrammation ; par exemple la capacité pour un objet de se rajouter ou de s'enlever des attributs ou des méthodes, ou même de changer de classe en cours d'exécution) et introspectif (un grand nombre d'outils de développement, comme le debugger ou le profiler, sont implantés en Python lui-même).
- ✓ Comme Scheme ou SmallTalk, Python est dynamiquement typé. Tout objet manipulable par le programmeur possède un type bien défini à l'exécution, qui n'a pas besoin d'être déclaré à l'avance.
- ✓ Python possède actuellement deux implémentations. L'une, interprétée, dans laquelle les programmes Python sont compilés en instructions portables, puis exécutés par une machine virtuelle (comme pour Java, avec une différence importante: Java étant statiquement typé, il est beaucoup plus facile d'accélérer l'exécution d'un programme Java que d'un programme Python). L'autre génère directement du bytecode Java.
- ✓ Python est extensible : comme Tcl ou Guile, on peut facilement l'interfacer avec des bibliothèques C existantes. On peut aussi s'en servir comme d'un langage d'extension pour des systèmes logiciels complexes.
- ✓ La bibliothèque standard de Python, et les paquetages contribués, donnent accès à une grande variété de services : chaînes de caractères et expressions régulières, services UNIX standards (fichiers, pipes, signaux, sockets, threads...), protocoles Internet (Web, News, FTP, CGI, HTML...), persistance et bases de données, interfaces graphiques.
- ✓ Python est un langage qui continue à évoluer, soutenu par une communauté d'utilisateurs enthousiastes et responsables, dont la plupart sont des supporters du logiciel libre. Parallèlement à l'interpréteur principal, écrit en C et maintenu par le créateur du langage, un deuxième interpréteur, écrit en Java, est en cours de développement.

✓ Enfin, Python est un langage de choix pour traiter le XML.

5.2.2 Les outils utilisés pour le développement de notre système

✓ **Python à la version 2.7** : est un langage de programmation évolué, libre et gratuit, disponible sur toutes les plateformes (windows, Linux, Unix ou Mac OS X).

❖ Installation de python sous Windows :

1. Télécharger Python depuis le lien : http://www.python.org/download/
2. Ensuite l'installer par un double clic sur l'installeur et suivre les étapes d'installation.
3. En effet, dans le menu démarrer, le programme IDLE - (Python GUI) apparaîtra.

✓ **WingIDE :** est un environnement de développement professionnel (IDE : Integrated Development Environment) pour développer des logiciels programmés en python. Il est disponible sous licence professionnelle, personnelle et open-source. Il est multiplateforme et disponible pour GNU/Linux, Windows, Mac OS X.

❖ Installer Wing IDE

1. Télécharger l'installeur depuis le lien : http://wingware.com/downloads
2. Ensuite un double clic sur l'installeur et suivre les étapes d'installation.

✓ **Boa Constructor**: est un éditeur pour python et un constructeur de GUI sur le toolkit wxPython. Il permet la création et la manipulation visuelle de fenêtres graphiques. Il est écrit en python et utilise la bibliothèque qui interface la bibliothèque wxWidgets.

❖ Installer Wing IDE

1. Télécharger l'installeur depuis le lien : http://sourceforge.net/projects/boa-constructor/files/
2. Ensuite un double clic sur l'installeur et suivre les étapes d'installation.

5.2.3 Les modules python utilisés

Un module est un fichier contenant des définitions et des instructions Python. Le nom du fichier est le nom du module avec le suffixe .py ajouté. Les modules sont gratuits et disponibles **[**44**]**.

Voici une description de quelques modules utilisés dans notre application **[**44**]** [20]:

a. wxPython : est une collection de modules Python réalisée sur la base des wxWidgets de wxWindows ; un framework multi-plateformes écrit en C++. Elle met à la disposition des développeurs un nombre impressionnant de classes permettant de

réaliser des interfaces homme machine (IHM) complètement indépendantes de l'Operating System sur lequel ils sont mis en œuvre.

b. JSON : (JavaScript Object Notation – Notation Objet issue de JavaScript) est un format léger d'échange de données. C'est un format texte complètement indépendant de tout langage, mais les conventions qu'il utilise seront familières à tout programmeur habitué aux langages descendant du C ; dans notre cas c'est le langage Python.
c. Urllib : est un module Python pour récupérer les URL (Uniform Resource Locators). Il dispose d'une interface très simple, sous la forme de la fonction « urlopen ». Ceci est capable de récupérer des URL en utilisant une variété de protocoles différents.
d. NLTK : est une plate-forme qui fournit des interfaces faciles à utiliser pour plus de 50 corpus et les ressources lexicales. telles que WordNet, avec un ensemble de bibliothèques de traitement de texte pour la classification, tokenization, l'analyse et le raisonnement sémantique, etc. NLTK est une bibliothèque incroyable de traitement automatique du langage naturel avec Python.
e. Pycurl : est une interface Python pour libcurl. Pycurl peut être utilisé pour aller chercher des objets identifiés par une adresse URL à partir d'un programme Python, similaire au module Python urllib. Pycurl est mature, très rapide, et prend en charge un grand nombre de fonctionnalités.
f. codecs (String encoding and decoding) : est un module python qui fournit des interfaces de flux et de fichier pour le transcodage des données dans un programme. Il est le plus couramment utilisé pour travailler avec du texte Unicode, mais d'autres encodages sont également disponibles pour d'autres fins.
g. cStringIO : est une version plus rapide de StringIO. Ce dernier est un module Python qui permet de lire et écrire des chaînes sous forme de fichiers. Ce module implémente une classe de type fichier, StringIO, qui lit et écrit un tampon de chaîne (également connu sous le nom des fichiers de la mémoire).
h. Ast : est un module Python qui permet l'analyse du code python dans son arbre de syntaxe abstraite ainsi que la manipulation de l'arbre.
i. Math : est un module python qui donne accès à des fonctions mathématiques définies par la norme C.
j. re : Ce module fournit des opérations d'appariement d'expressions régulières similaires à ceux trouvés dans Perl. Les deux modèles et les chaînes à rechercher peuvent être des chaînes Unicode ainsi que des chaînes de 8 bits.

❖ **Installation des modules sous Windows**

1. Télécharger le lien sous le lien : http://www.lfd.uci.edu/~gohlke/pythonlibs/
2. Exécuter le fichier s'il est exécutable sinon le décompresser puis taper l'instruction suivante depuis la console de commande :

```
$_home> python setup.py install
```

5.3 Présentation de l'application

5.3.1 Présentation de l'outil

L'application est nommée « Wech'Rayek! » en français « votre opinion ». Elle s'intéresse à la classification des opinions des consommateurs sur tel ou tel produit sur les sites de l'e-commerce (Amazone, Cent et Trustedreviews) sous les trois classes (positive, négative ou neutre). Cette application travaille en ligne, et est développée entièrement en Python.

5.3.2 Architecture générale de Wech'Rayek!

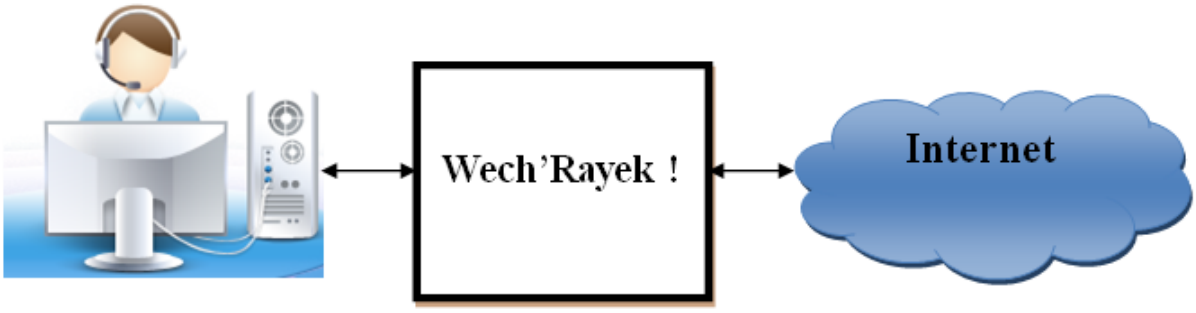

Figure 5.6. L'architecture générale de Wech'Rayek!

5.3.3 Fonctionnement de l'application

Cette application est composée de trois fonction principales (extraction(), pretraitement() et compter_sentiments()).Ci-dessous, une brève description de ces fonctions, en montrant les entrées, les sorties, les modules utilisés et le fonctionnement de chaque fonction :

a) La fonction extraction ()

1. Entrées : le nom du produit et les noms des sites.
2. Sorties : un ensemble de fichiers.txt (contient les contenus des pages Web à l'état brut).
3. Modules utilisés : json, urllib, nltk, codecs.

4. Fonctionnement : A l'aide du moteur de recherche « google », cette fonction permet de récupérer les URLs des pages Web des sites e-commerce contenant des opinions sur le produit saisie dans le champ en question. Ensuite, à partir de ces URLs, la fonction met les contenus des pages Web dans des fichiers textes (.txt).

b) La fonction prétraitement ()

1. Entrées : l'ensemble des fichiers.txt sortis de la fonction précédente.
2. Sorties : un ensemble de fichiers .txt (ne contenant que les commentaires)
3. Modules utilisés : re.
4. Fonctionnement : le rôle de cette fonction est de nettoyer les fichiers.txt résultants de l'étape précédente, tout en éliminant les lignes vides et les textes qui ne portent aucune opinion.

   Exemple : la figure suivante représente un fichier récupéré par la fonction précédente depuis le site « Cent » sur le produit « Iphone 5 » :

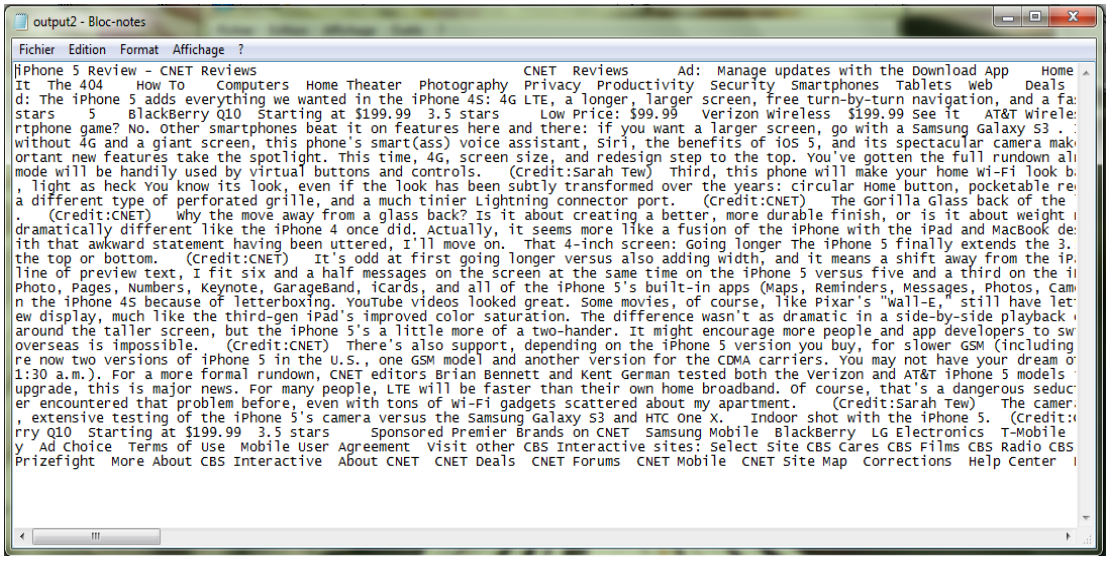

Figure 5.7. Le fichier récupéré avant le nettoyage.

Voici le fichier nettoyé par la fonction « pretraitement ()» :

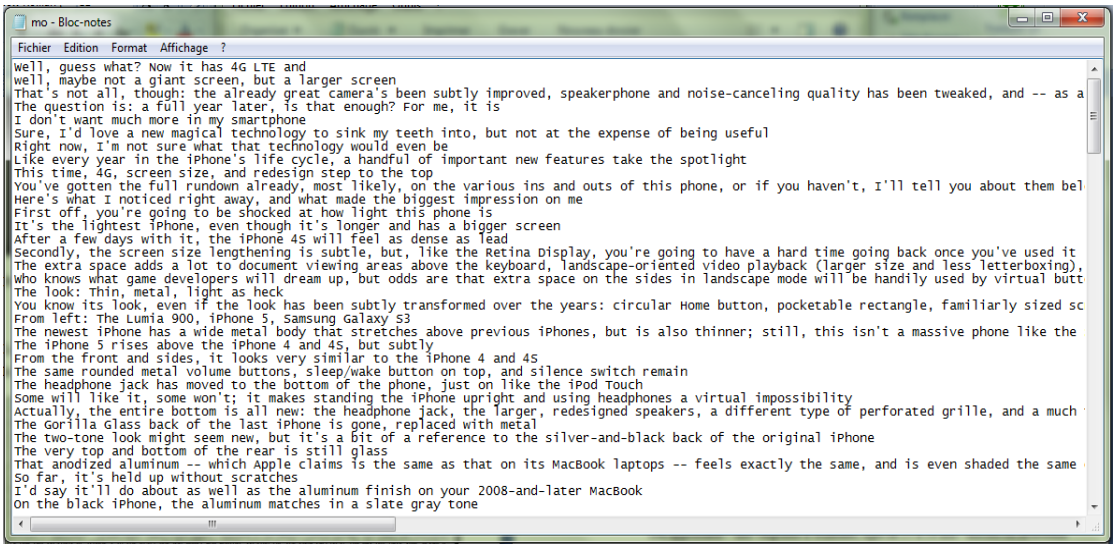

Figure 5.8. Le fichier récupéré après le nettoyage.

c) La fonction compter_sentiments()

1. Entrées : l'ensemble des fichiers.txt sortis de la fonction pretraitement().
2. Sorties : le nombre total des commentaires, les nombres des commentaires et les pourcentages (positifs, négatifs et neutres) pour chaque site choisi, ainsi que le pourcentage total et une représentation graphique.
3. Modules utilisés : ast, cStringIo, pycurl.
4. Fonctionnement : cette fonction permet de compter et afficher le nombre des commentaires positifs, négatifs et neutres. Ensuite, elle affiche le pourcentage et les commentaires pour chaque site sélectionnés, et enfin elle affiche le pourcentage général et une représentation graphique sous forme d'un secteur.

« Wech'Rayek ! » est basée sur le processus de Texte Mining et de Traitement du langage naturel, voici une projection du processus de Text Mining sur cette application :

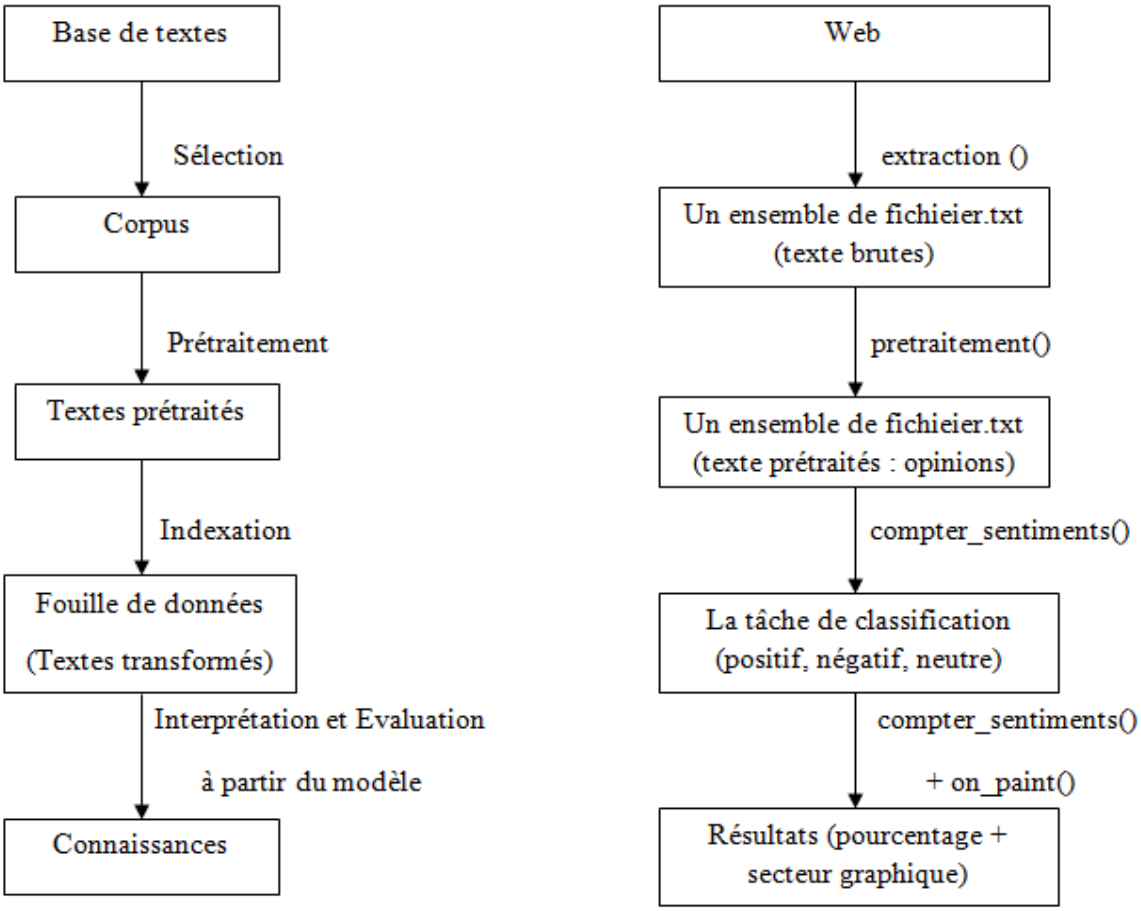

Figure 5.9. Un schéma qui projette le processus de Text Mining sur « Wech'Rayek ! ».

5.3.4 La valeur ajoutée de notre application

L'Opinion Mining et de Sentiments Analysis est un domaine nouveau. Cependant, et comme il a été souligné dans le chapitre 4, il y a déjà plusieurs outils (tweetfeel, etc.) conçus pour la classification des opinions des internautes sur une marque, un produit, une célébrité, etc. Bien que ces applications soient gratuites, open-source, en ligne, rapides et simples à utiliser, néanmoins elles sont limitées sur plusieurs plans.

Le but de notre système est d'améliorer les fonctionnalités des applications déjà existantes tout en ajoutant d'autres options. Ci-dessous les points positifs ajoutés par notre application :

Les autres applications	**Notre application**
Une seule source d'information : 'twitter'	Plusieurs sources d'informations (plusieurs sites e-commerce)
La petite taille du commentaire (tweet) : un tweet ne peut pas dépasser 140 caractères.	La taille du commentaire n'est pas limitée
La plupart des applications classent les commentaires en deux polarités (positive et négative).	Classification sous les trois polarités (positive, négative et neutre).
La plupart des applications donnent le résultat en pourcentage, sans une représentation graphique.	Le résultat : pourcentage + un secteur graphique.

Tableau 5.1. La valeur ajoutée de Wech'Rayek !.

5.3.5 Utilisation de « Wech'Rayek ! »

L'interface principale de « Wech'Rayek ! » est la suivante :

Figure 5.10.L'interface principale de « Wech'Rayek! ».

Dans le champ de texte taper le nom d'un produit, d'une célébrité ou d'une marque exemple « iPhone 5 », et valider par le bouton « Search ». La figure suivante s'affiche :

Figure 5.11.Les sites disponibles dans « Wech'Rayek! ».

Puis sélectionner les sites voulus. Par exemple, nous avons choisi (Cent, Amazon, Trustedreviews) ; le résultat pour chaque site sélectionné s'affiche comme suit :

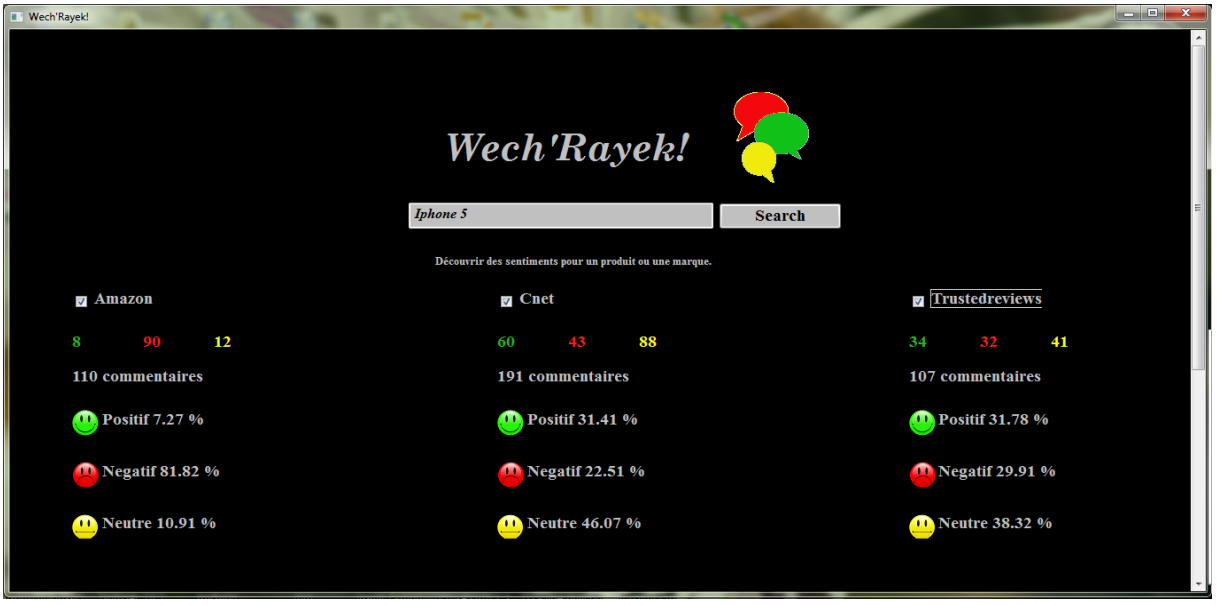

Figure 5.12. Les résultats pour chaque site sélectionné.

Ainsi que le résultat général sous forme de pourcentage et d'une représentation graphique:

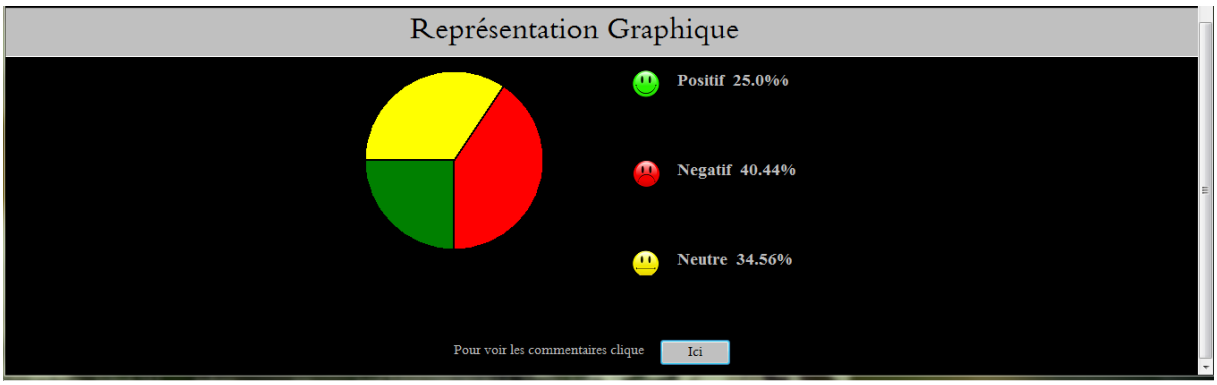

Figure 5.13. La représentation graphique.

Pour voir les commentaires, il suffit de cliquer sur le bouton « Ici ».

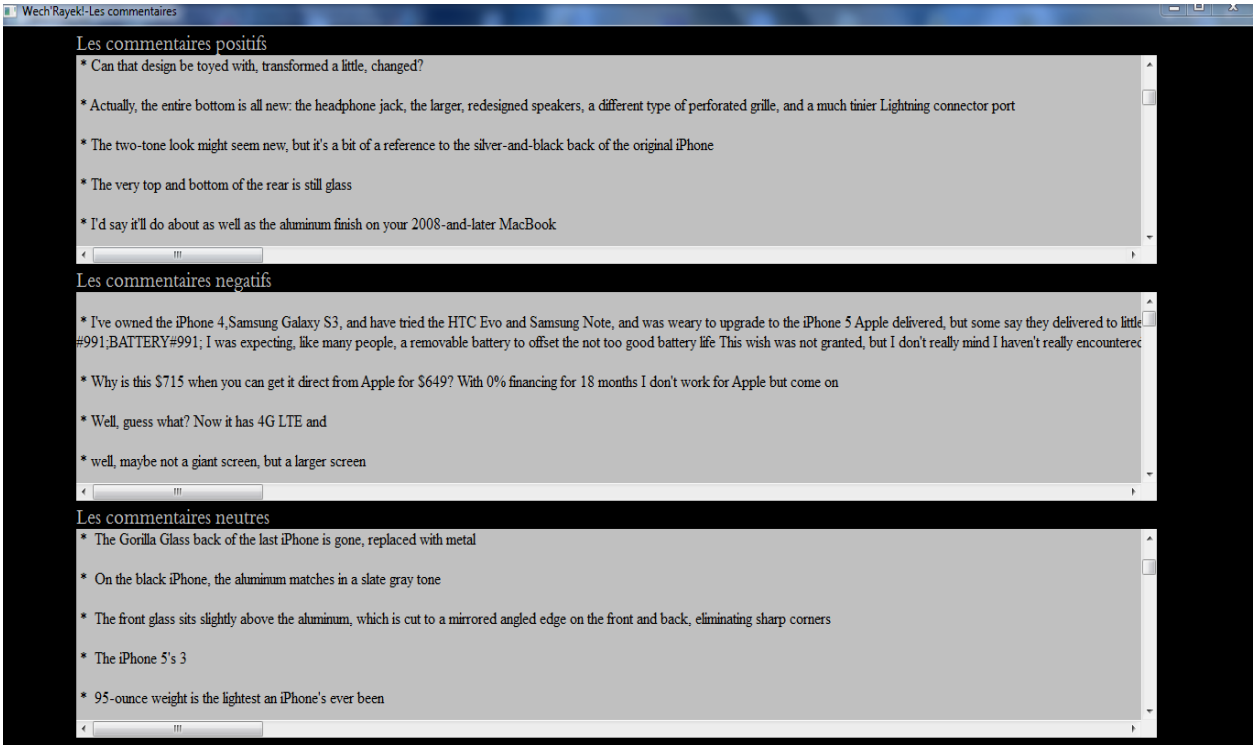

Figure 5.14. Les commentaires.

Conclusion générale

Conclusion générale

Le domaine d'opinion Mining et du Sentiments analysis est un nouvel axe de recherche permettant de faciliter et améliorer la vie quotidienne en analysant et classifiant les opinions et les sentiments des internautes. Malgré l'existence de quelques travaux dans ce domaine, ils comportent tous des limites que notre application a pu lever :

- Ils travaillent sur une seule source d'information : 'twitter',
- Ils travaillent sur des commentaires de petite taille (tweet) qui ne peut pas dépasser 140 caractères,
- La plupart des applications classent les commentaires en deux polarités (positive et négative),
- La plupart des applications donnent le résultat en pourcentage, sans une représentation graphique,
- Etc.

Au terme de ce projet, nous pouvons affirmer que nous sommes arrivés à répondre aux objectifs qui nous ont été fixés.

Si le système, à l'état actuel, est intéressant du point de vue richesses en informations et innovation, Il demeure malgré tout intéressant de revoir certains aspects de l'application après la réalisation de l'étape de l'expérimentation et l'évaluation d'usage.

Bibliographie

Références bibliographiques

[1] Abdelhamid DJEFFAL, (2012). *Fouille de données avancée.* Cours. Université Mohamed Khider de Biskra.

[2] Antonio Machado, (2010). *Chapitre2 : Le Web Usage Mining*. thèse, INRIA.

[3] Cheikh B., (2002). *Etat de la connaissance sur le web usage mining* Diplôme d'Etudes Approfondies. Université de Montpellier II.

[4] Faiza Belbachir, (2010). *Expérimentation de fonctions pour la détection d'opinions dans les blogs*. Mémoire de master. Université de Paul Sabatier, Toulouse.

[5] Grzegorz DZICZKOWSKI, (2008). *Analyse des sentiments : système autonome d'exploration des opinions exprimées dans les critiques cinématographique.* Thèse de doctorat. L'ECOLE NATIONALE SUPERIEURE DES MINES, PARIS.

[6] HACÈNE CHERFI, (2004). Etude et réalisation d'un système d'extraction de connaissance à partir de textes. Thèse de doctorat. Université de Henri Poincaré Nancy I.

[7] Jmal, J., (2012). ResTS : Système de Résumé Automatique des Textes d'opinions basé sur Twitter et SentiWorldNet. In Actes de la conférence conjointe JEP-TALN-RECITAL, volume 3 : RECITAL, pages 233-246, Gronoble, France.

[8] Ladeg Hamza, (2012). *Modélisation de connaissances du décideur pour le renforcement du processus KDD*. Mémoire de magister, Ecole nationale supérieure d'informatique (ESI), Alger.

[9] Nabila Merzoug et Hanane Bessa, (2009), *Application du processus de fouille de données d'usage du web sur les fichiers logs du site cubba*. Mémoire d'ingéniorat. Université de Bordj Bou Arréridj Algerie.

[10] Nicolas BÉCHET, (2009). *Extraction et regroupement de descripteurs morpho-syntaxiques pour des processus de Fouille de Textes.* Thèse de doctorat. Univérsité Montpellier II.

[11] Noureddine MOKHTARI, (2006). *Extraction et exploitation*

d'annotations sémantiques contextuelles à partir de texte. Thèse de doctorat. L'Université de Nice-Sophia Antipolis.

[12] PERIGNON Xavier, SOH KAMLA Rodrigue, (2008). LE DATA MINING, Travaux de recherches.

[13] Poirier, D., (2011). Des textes communautaires à la recommandation. Thèse de doctorat de l'université d'Orléans & Pierre Marie Curie – Paris 6.

[14] Rabah Rahmani, (). *Découverte d'associations sémantiques dans les bases de données relationnelles par des méthodes de Data Mining*, Mémoire de magister, Université Mouloud Mammeri de Tizi-ouzou.

[15] Slimane OULAD NAOUI,(2009). *Prétraitement & Extraction de Connaissances en Web Usage Mining*. Université Kasdi Merbah d'Ouargla.

[16] SLIMANI Yacine, MOUSSAOUI Abdelouahab, (2010). *La fouille des usagers du web par application de l'algorithme Apriori sur les fichiers logs.* Université Ferhat Abbas de Sétif.

[17] Xavier Polanco, (2001). *TEXT MINING ET INTELLIGENCE ECONOMIQUE AUJOURD'HUI ET DEMAIN.* Université Catholique de Louvain-la-Neuve.

[18] Yannick Toussaint, (2004). *Fouille de textes et organisation de documents : Extraction de connaissances à partir de textes structurés*. LIRIS/INSA de Lyon.

[19] Yannick Toussaint, (2011). *Fouille de texte : des méthodes symboliques pour la construction d'ontologie et l'annotation sémantique guidée par les connaissances.* Projet de recherche. Université de Henry Poincaré Nancy 1.

Références hypertextes

[20] Alain Delgrange, (2005). *Programmer des interfaces graphiques avec le framework open source wxPython*. URL :http://alain72.developpez.com/tutos/wxPython/

[21] Alec Go, Richa Bhayani, et Lei Huang, (2013). *Sentiment140*. URL : http://help.sentiment140.com/

[22] Alias-i, (2011). *LingPipe*. URL: http://alia-i.com/lingpipe/

[23] Bing Liu, (2010). *Sentiment analysis and subjectivity*. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.216.5533

[24] Bo Pang et Lillian Lee, (2008). *Opinion Mining and Sentiment Analysis*. URL :http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf

[25] CALIS, (2008). *Internet et Web*. URL: http://campus.hesge.ch/calis/m1/l3/

[26] Carol Hermann, (2010). *Entre Web 2.0 et 3.0: opinion mining*. URL :http://doc.rero.ch/record/22375/files/TB_Hermann_Carol.pdf

[27] Christian Fauré, (2005). L'information non-structurée. Mémoire. URL : http://www.christian-faure.net/2005/07/17/linformation-non-structure/

[28] Claude MARTINEAU et all, (2013). *DÉTECTION FINE D'OPINIONS ET SENTIMENTS : ATTRIBUTION DE POLARITÉ ET CALCUL INCRÉMENTAL DE L'INTENSITÉ*. URL : http://hal-enpc.archivesouvertes.fr/docs/00/79/02/53/PDF-/ARTICLE_LG_2011_Martineau_Voyatzi_Varga_Brizard_Migeotte.pdf

[29] Cristina OPREAN, (2010). *État de l'art sur les aspects méthodologiques et processus en Knowledge Discovery in Databases*. URL: ftp://ftp.irisa.fr/local/caps/DEPOTS/BIBLIO2011/Oprean_Cristina.pdf

[30] Damien Poirier et al, (2010). *Approches Statistique et Linguistique Pour la Classification de Textes d'Opinion Portant sur les Films*. URL: http://hal.archives-ouvertes.fr/docs/00/46/64/12/PDF/rnti09-poirier_et_al.pdf

[31] Dominique Boullier et Audrey Lohard, (2012). *« Chapitre 5. Détecter les tonalités : opinion mining et sentiment analysis », de Opinion mining et Sentiment analysis*. URL : http://books.openedition.org/oep/214

[32] Doru Tanasa, Brigitte Trousse et Florent Masséglia, (2004). *Application des techniques de fouille de données aux logs web : Etat de l'art sur le Web Usage Mining.*URL *:* http://www-sop.inria.fr/axis/personnel/Florent.Masseglia/cmi_03.pdf

[33] EVELYNE BOURION et all, (2011). *De la fouille de données à la fabrique de l'opinion. Enjeux épistémologiques et propositions.* URL : http://halshs.archives-ouvertes.fr/docs/00/65/92/18/PDF/Eensoo-Ramdani_et_al._2011_VERSION_SOUMISE_La_fabrique_de-_l_Opinion_pour_les_Cahiers_du_NumA_riques.pdf

[34] Futura-Sciences. (2013). *Multimédia*. URL: http://www.futura-sciences.com/fr/definition/t/high-tech-1/d/multimedia_1257/

[35] Futura-Sciences, (2013). *Suivez Futura-Sciences sur Twitter !*. URL : http://www.futura-sciences.com/fr/news/t/vie-du-site/d/suivez-futura-sciences-sur-twitter_19218/

[36] Geniva Lab, (2009). Découverte de connaissances dans les données (Datamining ou KDD). URL :http://cui.unige.ch/AI-group/teaching/dmc/09-10/cours/dm01-intro.pdf

[37] Gérard swinnen, (2008). *Apprendre à programmer avec Python*. URL: http://www.framasoft.net/IMG/pdf/python_notes-2.pdf

[38] Info Entrepreneurs, (2009). *Connaître les besoins de vos clients*. URL: http://www.infoentrepreneurs.org/fr/guides/connaitre-les-besoins-de-vos-clients/

[39] TweetFeel, (2013). *TweetFeel an analytical look at Twitter's feeling*. URL: http://jour2722.jacdigital.com.au/tag/tweetfeel/

[40] Jan Wiebe, (2008). *Subjectivity and Sentiment Analysis*. URL :http://videolectures.net/icwsm08_wiebe_ssa/

[41] Jian-Yun Nie, (20??). *La recherche d'information*. Cours, Université de Montréal. URL: http://benhur.teluq.uqam.ca/SPIP/inf6104/article.php3?-id_article=17&id_rubrique=4&sem=2

[42] Le Crosnier, H., (1995). *L'hypertexte en réseau : repenser la bibliothèque*.

Bulletin des bibliothèques de France 1995 – Numéro 2. URL: http://www.enssib.fr/bbf/bbf-95-2/lecrosni.doc

[43] Mike Luby et al, (2008). *Twitrratr*. URL : http://twitrratr.com/about/

[44] Python Software Foundation, (2013). *The Python Tutorial.* URL: http://docs.python.org/2/tutorial/

[45] rapid-I, (2012). RapidMiner. URL: http://rapid-i.com/content/view/181/190/

[46] Sandrine Curtet, (2005). *Modules de prétraitement de données dans le cadre du Data Mining*. URL :http://wod.heig-vd.ch/documents/pdf/a2005_scurtet.pdf

[47] Sandy RIHANA, (2012). *Introduction à l'Internet*. URL: http://www.nachez.info/meth21f/1coursInternet.pdf

[48] Sébastien Paumier, (2012). *Comparaison d'outils d'informatique linguistique pour l'extraction d'information*. URL: http://mai21.free.fr/wp-content/uploads/rapport_tal_annotation_url_mai.pdf

[49] Sigrid Maurel, Paolo Curtoni et Luca Dini, (). *L'analyse des sentiments dans les forums.* URL: http://www2.lirmm.fr/~mroche/FODOP08/ArticlesFODOP08/Article2.pdf

[50] Synthesio, (20 ??). *La vérité sur la tonalité & l'analyse*. URL: http://synthesio.com/corporate/wp-content/uploads/2010/11/SYNTHESIO-TALN.pdf

[51] Talbi E-G.(2012). *Fouille de données (Data Mining) : Un tour d'horizon*,Laboratoire d'Informatique Fondamentale de Lille. URL :http://www.lifl.fr/~talbi/Cours-Data-Mining.pdf

[52] TWITTERMAN, (2009). *TweetFeel Prenez le pouls de Twitter envers votre Marque*. URL: http://twitteradar.com/tweetfeel-prenez-le-pouls-de-twitter-envers-votre-marque/applications-twitter