République **A**lgérienne **D**émocratique et **P**opulaire
Ministère de l'**E**nseignement **S**upérieur et de la **R**echerche **S**cientifique

UNIVERSITE **M**OULOUD **M**AMMERI DE **T**IZI-**O**UZOU

FACULTE DU **G**ENIE **E**LECTRIQUE ET D'INFORMATIQUE
DEPARTEMENT D'INFORMATIQUE

Mémoire de Fin d'Etudes de MASTER ACADEMIQUE

Domaine : **Mathématiques et Informatique**
Filière : **Informatique**
Spécialité : **Systèmes informatiques**

Présenté par
Mouloud HABCHI
Nadia LOUAGUENOUNI

Thème

Analyse des sentiments dans les microblogs

Mémoire soutenu publiquement le 10/07/2016 devant le jury composé de :

Président : Mme F. AMIROUCHE
Encadreur : Mr M. N. AMIROUCHE
Examinateur : Melle S. ILTACHE
Examinateur : Mme L. BELKACEMI

Promotion 2015/2016

Remerciements

Nous remercions le bon Dieu pour le courage, la patience qui nous ont été utiles tout au long de notre parcours.

Nous tenons à remercier notre promoteur M^{r} AMIROUCHE pour nous avoir aidé, guidé, orienté et conseillé durant la réalisation de notre mémoire.

Nous adressons nos remerciements les plus sincères aux membres du jury qui nous font l'honneur de juger notre travail.

Notre profonde gratitude et sincères remerciements vont à tous les enseignants qui nous ont suivis durant notre parcours d'étude.

Ces remerciements ne seraient pas complets si nous n'avons pas pensé à les destiner, avec notre profonde reconnaissance, à nos parents qui nous ont offert un environnement favorable pour mener à terme notre travail.

Dédicace

Je dédie ce travail à mes chers parents pour leurs sacrifices et leurs encouragements que Dieu les protège.

A mes deux sœurs Ania, Kathia et leurs maris.

A mes oncles et mes tantes.

A mon petit neveu Ilan.

A mes cousins et cousines.

A mes chers ami(e)s (Nabil, Mazigh, Said, Menad, Juba, Mounir, Yacine, Karim, Sid Ali, Amar, Massinissa, Amine, Salim, Adlane, Hichem, Mohand, Seddik, Hacene, Samia, Hanane, Meriem, Katia, Daya, Lisa, Nassima, Sarah, Nacera, Zineb, Souhila, Sylia, Souad, Houda, Wassila…etc).

A ma collègue Nadia.

A ma future femme.

Et à tous ceux qui me connaissent de près ou de loin.

Mouloud.

Dédicace

Je dédie ce travail à mes chers parents pour l'éducation qu'ils m'ont prodigué, avec tous les moyens et au prix de tous les sacrifices qu'ils ont consentis à mon égard, pour le sens du devoir qu'ils mon enseigné depuis mon enfance. Que Dieu les gardes et les protège.

A mes chères et adorables sœurs (Sonia, Lynda, Mahnoucha).

A mes chers frères Madjid, Hakim Mohammed Ou Slimane.

A mes neveux et ma nièce (Nounouh, Pouspous, Abderahmane et Aya).

A mes chers amis (e)s (Souhila, Hanane, Samia, Lila, Farida, Mounir, Yacine, Sid Ali, Mouloud…etc).

A mon collègue Mouloud.

A tous ceux qui me connaissent de près ou de loin.

Nadia.

SOMMAIRE

Chapitre 4 : Conception et Réalisation

Liste des figures

Liste des figures

Liste des Tableaux

Introduction générale :

De nos jours, l'Internet se révèle plus que jamais un outil indispensable d'échange d'informations. Il nous offre une quantité considérable d'informations à une vitesse inédite et ses services s'adaptent de plus en plus aux besoins des internautes. Ceux-ci peuvent consulter l'Internet pour trouver de l'information, envoyer des e-mails, acheter des produits, lire des journaux en ligne, etc.

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

La popularité des nouveaux médias est d'autant plus grande que la demande d'informations est devenue plus importante dans notre société. En général, les gens aiment consulter les avis d'autres personnes avant de passer à l'action ou de se faire une opinion. Autrefois, ces avis provenaient surtout de leur environnement social direct. Aujourd'hui, les gens aiment également consulter les opinions objectives de personnes qui leur sont inconnues.

Tous les récents développements dans le domaine d'échange d'informations et d'opinions ont donné naissance aux applications informatiques conçues pour l'analyse et la détection de sentiments exprimés sur Internet. Présentée dans la littérature sous le nom de opinion mining ou sentiment analysis, l'analyse des sentiments s'utilise entre autres pour la détection d'opinions sur des sites web et des réseaux sociaux, l'éclaircissement sur le comportement des consommateurs, la recommandation de produits et l'explication du résultat des élections. Elle consiste à rechercher des textes évaluatifs sur Internet tels que des critiques et des recommandations et à analyser de façon automatique ou manuelle les sentiments qui y sont exprimés afin de mieux comprendre l'opinion publique.

Ce besoin pressant d'informations dans la société n'a pas échappé aux nombreux commerçants qui se réfèrent au Web afin de savoir ce qui est dit à l'égard de leurs produits. De la même façon, la vie politique est de plus en plus dominée par le flux informationnel sur Internet et plus concrètement sur les réseaux sociaux. Depuis quelques années, les hommes politiques étendent leurs campagnes électorales jusqu'à Facebook et Twitter.

Il a déjà été démontré par des études antérieures que l'analyse des sentiments s'avère particulièrement intéressante pour ceux qui ont intérêt à connaître l'opinion publique, que ce soit pour des raisons personnelles, commerciales ou politiques. Ainsi, de nombreux systèmes autonomes ont déjà été développés pour l'analyse automatique des sentiments. Généralement, ces systèmes étaient entraînés aux textes évaluatifs traditionnels tels que les comptes rendus cinématographiques ou les critiques d'un livre.

Par conséquent, il importe de concevoir des systèmes automatiques aptes à rechercher et à analyser les sentiments qui sont exprimés sur les réseaux sociaux. A cet effet, une grande partie de notre étude sera consacrée à l'analyse des sentiments exprimés dans les tweets.

Nous avons réparti notre mémoire en plusieurs chapitres et qui sont :

Chapitre 1 : Les microblogs.

Chapitre 2 : L'analyse des sentiments.

Chapitre 3 : Outils réalisés dans l'opinion mining.

Chapitre 4 : Conception et réalisation.

CHAPITRE 1

1.1 Introduction :

De la famille des médias sociaux, le microblog ou microblogue est un dérivé concis d'un blog du web 2.0 ou web social. Développé à partir de 2006 aux États-Unis, il permet des publications plus courtes que dans les blogs classiques, qu'il s'agisse de textes courts, d'images ou de vidéos embarquées. Les flux d'agrégation sont plus légers que dans les blogs traditionnels et peuvent contenir tout le message. La diffusion peut être restreinte par l'éditeur à un cercle de personnes désirées [1].

Le microblog est un moyen instantané de rester en contact permanent avec ses proches (ou les abonnés à son microblogue), de les tenir au courant de ses moindres faits et gestes.

Les micromessages sont enregistrés et archivés sur le site de microblogage. On peut s'abonner aux archives d'un microblogue (public), selon différentes modalités en fonction des plateformes. Le microblogueur et ses contacts conversent ainsi par microbillets interposés, lesquels peuvent être directement rédigés, publiés et reçus depuis l'interface du microblogue, mais aussi par l'intermédiaire d'un téléphone cellulaire ou autre terminal mobile avec accès à Internet (courriel, messagerie texte, messagerie instantanée, etc.).

Il existe plusieurs plateformes de microblogging. Les 5 plateformes les plus utilisées sont Twitter, Tumblr, Tencent Weibo, Identi.ca et Pinterest. Parmi elles, Twitter est sans conteste la plus utilisée. Twitter est utilisé également comme source d'information.

1.2 Les services de microblogging :

Twitter

Twitter a été créé le 21 mars 2006 par Jack Dorsey, Evan Williams, Biz Stone et Noah Glass, et lancé en juillet de la même année. Le service est rapidement devenu populaire, jusqu'à réunir plus de 500 millions d'utilisateurs dans le monde fin février 2013. Au 30 juin 2015, Twitter compte 316 millions d'utilisateurs actifs par mois avec 500 millions de tweets envoyés par jour et est disponible en plus de 35 langues, Il permet à un utilisateur d'envoyer gratuitement de brefs messages, appelés tweets, sur internet, par messagerie instantanée ou par SMS. Ces messages sont limités à 140 caractères [2]. La figure 1.1 représente le logo de Twitter (2016).

Figure 1.1 : Logo de Twitter à ce jour.

Tumblr

Tumblr est une plate-forme de microblog créée en 2007 par David Karp, 420 millions d'utilisateurs actifs (Janvier 2015) et permet à l'utilisateur de poster du texte, des images, des vidéos, des liens et des sons sur son tumblelog[9]. Elle s'appuie principalement sur le reblogage. Son slogan est « Postez n'importe quoi (de n'importe où), personnalisez tout, et trouvez et suivez ce que vous aimez. Créez votre propre blog Tumblr aujourd'hui » [3]. La figure 1.2 représente le logo de tumblr (2016).

Figure 1.2 : Logo de Tumblr à ce jour.

Tencent Weibo

Tencent Weibo est un site Chinois de microblogging lancé par Tencent en avril 2010, et est toujours en version bêta. Les utilisateurs peuvent publier un message d'au plus 140 caractères chinois par le Web, SMS ou application smartphone.

Tencent Weibo est un réseau social qui connecte tous les utilisateurs ensemble. Les utilisateurs peuvent partager des photos, des vidéos et du texte avec une limite de 140 caractères. La fonction "repost" de Tencent Weibo est similaire à la fonction "retweet" de Twitter, jusqu'au caractère @ [4]. La figure 1.3 représente le logo de Tencent Weibo (2016).

Figure 1.3 : Logo de Tencent Weibo

Identi.ca

Identi.ca est un service de réseautage social et de microblogage fonctionnant sous pump.io, un logiciel libre sous licence Apache.

Le service a reçu plus de 8 000 inscriptions et 19 000 messages durant ses premières 24 heures de fonctionnement, le 1er juillet 2008, et a atteint le million de messages le 4 novembre 2008.

La figure 1.4 représente le logo d'identi.ca (2016).

Figure 1.4 : Logo identi.ca

Les utilisateurs peuvent ajouter des textes d'une longueur maximum de 140 caractères sur le principe du microblog, reprenant la longueur exacte imposée sur Twitter. Bien que similaire à Twitter dans le concept et le mode de fonctionnement, Identi.ca fournit de nombreuses fonctionnalités non disponibles sur Twitter, telles que le support du protocole de messagerie XMPP (***eXtensible Messaging and Presence Protocol*** est un ensemble de protocoles standards ouverts de l'Internet Engineering Task Force (IETF) pour la messagerie instantanée). Identi.ca permet l'export et l'échange de données basé sur le standard FOAF (***Friend Of A Friend*** est un vocabulaire qui permet de décrire des personnes et les relations qu'elles entretiennent entre elles), ainsi les messages peuvent être redirigés vers un compte Twitter ou un autre service [5].

Pinterest

Pinterest est un site web américain mélangeant les concepts de réseautage social et de partage de photographies, lancé en 2010 par Paul Sciarra, Evan Sharp et Ben Silbermann. Il permet à ses utilisateurs de partager leurs centres d'intérêt, passions, hobbies, à travers des albums de photographies glanées sur l'Internet. Le nom du site est un mot-valise des mots anglais pin et interest signifiant respectivement « épingler » et « intérêt ». [6]

Pinterest propose à ses utilisateurs d'épingler des images qui ont pu attirer leur attention dans différentes rubriques. L'ajout d'images peut se faire par l'intermédiaire du bouton pin it, un raccourci à intégrer directement dans le navigateur, ou par l'intermédiaire d'une démarche classique de téléversement via le bouton add du site Pinterest. Une fois l'image sélectionnée, celle-ci peut être catégorisée. Une légende peut aussi être renseignée. La figure 1.5 représente le logo de Pinterest (2016).

Figure 1.5 : Logo de Pinterest

1.3 Concepts et fonctionnement des plateformes de microblogging : cas de twitter

La figure 1.6 montre l'interface de Twitter. L'interface est composée de plusieurs sections. Dans la section Tweets appelée également Timeline, un utilisateur peut voir le flux de ses tweets ainsi que ceux de ses amis, triés par ordre chronologique inverse. On peut remarquer également une section de tendances qui contient les 10 sujets les plus populaires dans Twitter à un moment donné. L'utilisateur peut consulter les tendances du monde entier, comme il peut se focaliser sur un endroit plus spécifique. La plate-forme suggère également des utilisateurs qui ont des centres d'intérêts similaires à l'utilisateur courant dans la section suggestions [FD14].

Figure 1.6 – L'interface graphique utilisateur de Twitter

En s'inscrivant sur une plateforme de microblogging, un utilisateur fournit plusieurs informations telles que sa photo, sa localisation, son site Web et une courte biographie (figure 1.7). Dans la biographie, les utilisateurs décrivent généralement leurs activités et leurs centres d'intérêt. Ces informations sont ensuite probablement utilisées par les plateformes dans la recommandation des utilisateurs.

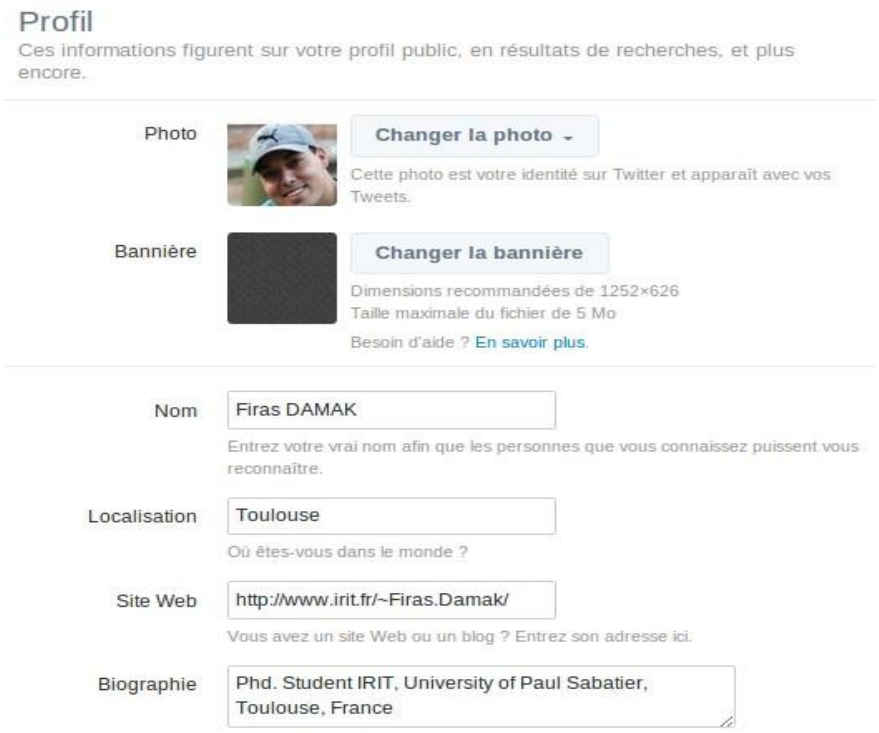

Figure 1.7 : Informations des comptes utilisateurs sur Twitter

La figure 1.8 donne un exemple d'utilisation d'une plateforme de microblogging. Un utilisateur A peut suivre le flux de microblogs envoyés par un utilisateur C sans lui demander la permission (sauf pour les comptes privés que nous ne détaillons pas ici). Les relations entre utilisateurs des réseaux sociaux sont appelées des abonnements. Si A est abonné à C, alors A est appelé abonné (follower) de C (followee) et reçoit automatiquement toutes les publications de C dans sa timeline. Les relations d'abonnement peuvent être unilatérales (dans un seul sens), mais également bilatérales (dans les deux sens) si C s'abonne à son tour à A. On parle dans ce cas d'une relation d'amitié. Si un microbloggeur diffuse un message, tous ses abonnés le reçoivent. Un microbloggeur peut également envoyer un message direct et privé à l'un de ses amis (direct message). Si le microbloggeur partage un message pour la première fois, le message sera un *tweet*, sinon, s'il le rediffuse, le message sera un *retweet* et il va contenir dans ce cas la mention **RT**. En rediffusant un microblog, un microbloggeur peut y ajouter de

l'information complémentaire. Finalement, un utilisateur peut en mentionner un autre dans un message (@mention) [FD14].

Les individus ne sont pas les seuls propriétaires de comptes. Les entreprises ou encore les sites d'information sont aujourd'hui très présents sur les plateformes de microblogging.

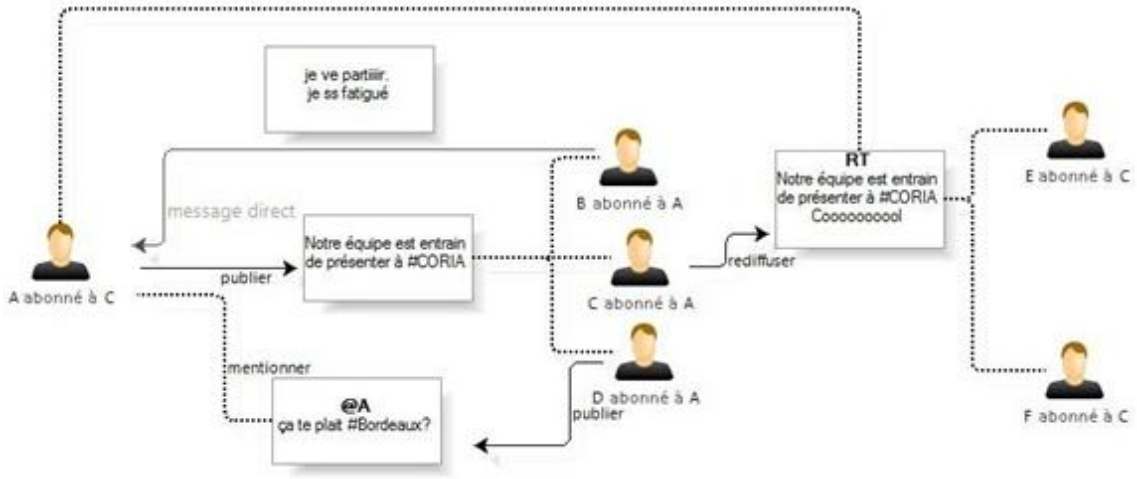

Figure 1.8 : Exemple d'utilisation de Twitter

1.4. Système temps-réel

L'une des spécificités fondamentale des plateformes de microblogging est leur nature temps-réel : la présentation des publications (*timeline*), la présentation des résultats de recherches, le traitement du contenu publié…

Timeline Twitter, comme les autres plateformes de microblogging, est un système temps-réel par excellence dont la fraîcheur est la spécificité la plus importante. Cette spécificité peut être aperçue à plusieurs niveaux :

- Un utilisateur, en accédant à sa page, reçoit en temps-réel les microblogs de ses abonnés. Ces microblogs défilent sur sa page et le plus récent s'affiche au début de la file.
- Pour répondre à un besoin d'information, le moteur de recherche de Twitter affiche les tweets-résultats par ordre chronologique inverse (des plus récents aux plus anciens). Si à un moment donné un nouveau microblog pertinent est publié, l'utilisateur reçoit une notification pour l'afficher (Figure 1.9).

Figure 1.9 : Notification sur l'apparition de nouveaux résultats sur Twitter

1.5. Spécificités des microblogs

Jansen et al. (2009) ont réalisé une étude linguistique sur Twitter. Ils ont trouvé qu'un tweet contient en moyenne 15 mots. Ce chiffre est extrêmement faible comparé aux autres sources d'information du Web. Les articles de Wikipédia, par exemple, possèdent en moyenne 320 termes par article. Cette particularité représente un défi pour les techniques de recherche d'information classiques qui se basent principalement sur les fréquences des termes dans les documents [FD14].

Un microbloggeur peut inclure différents types de signes dans un tweet, en plus du contenu textuel. Ces pratiques ont peu à peu évoluées pour devenir des « normes de balisage » :

- **@** suivi du nom d'utilisateur permet d'indiquer qu'on mentionne ou s'adresse à une personne particulière (représenté par son compte).
- # suivi par un mot est un hashtag. Un hashtag indique un mot important que le système peut utiliser pour permettre une recherche par navigation (figure 1.10). Les hashtags permettent de catégoriser les microblogs selon un contexte (événement, lieu, etc.) : par exemple, certaines émissions télévisées définissent des hashtags spécifiques à utiliser par les microbloggeurs souhaitant exprimer leurs avis sur le sujet de l'émission. Les conférences scientifiques définissent également des hashtags permettant, d'une part, aux participants de partager leurs remarques et, d'autres part, aux gens de l'extérieur de suivre ce qui se passe dans la conférence en temps-réel.
- Les microblogs peuvent également contenir des URL. Ces hyperliens prennent une forme courte en raison du nombre limité de caractères autorisés par microblog. Il existe deux services très connus pour créer la forme réduite des URL : bit.ly et

tinyurl.com. Dans le cas où l'URL correspond à une image, Twitter affiche un aperçu de cette image dans l'interface de l'utilisateur comme le montre la figure 1.6.

- Les internautes peuvent mettre des photos dans leurs microblogs (figure 1.10).
- En cliquant dessus, l'utilisateur pourra voir la photo en taille normale.

Figure 1.10 : Tweet posté par @florencesantrot contenant une image et des hashtags (#Apple #iphone6cost1k). Il a été retweeté sept fois et favori une fois.

Outre les données postées explicitement par les microbloggeurs, les microblogs contiennent également des métadonnées de différentes natures et qui sont :

- **de géolocalisation** : les microblogs publiés à travers les terminaux mobiles équipés de GPS fournissent des informations de géolocalisation. Ces informations permettent de localiser l'endroit duquel le tweet a été publié.
- **d'horodatage** : chaque microblog est caractérisé par sa date de publication. Cette information est utilisée pour mesurer sa fraîcheur s'il fait partie d'une liste de résultats d'une recherche.
- **d'auteur** : Les plateformes de microblogging stockent le compte depuis lequel est publié chaque microblog. Ceci permet aux utilisateurs de trouver les microblogs d'un auteur en particulier.
- **de favoris** : on peut savoir, pour chaque microblog, combien de fois il a été choisi dans les listes de favoris des autres utilisateurs (figure 1.10) ainsi que l'ensemble des

utilisateurs qui l'ont sélectionné.

- **de rediffusion** : Retweet (RT) indique que le message est rediffusé. Le mécanisme de rediffusion permet aux utilisateurs de partager de nouveau tweets qu'ils trouvent intéressants parmi les tweets publiés par leurs amis (par exemple, RT @mashable Top 10 Twitter Trends This Week http://on.mash.to/eA2jY5). Dans Twitter, on peut connaître le nombre de fois qu'un tweet a été retweeté (figure 1.10). On peut également accéder à la liste des utilisateurs qui ont retweeté un tweet donné.

1.6. Spécificités des recherches dans les microblogs

Le moteur de recherche de microblogs est spécifique au niveau des données en entrée ou des résultats. D'une part, outre des mots-clés, un utilisateur peut mélanger des comptes utilisateurs, des hashtags et même des URLs, dans sa recherche. La figure 1.11 montre les suggestions de différents types de données de recherche de Twitter.

Figure 1.11 : Suggestion de différents types de résultats dans le moteur de recherche de Twitter : des mots-clés, des hashtags, des comptes utilisateurs sont présentés.

D'autre part, les résultats affichés diffèrent en fonction du type de données utilisées : si l'utilisateur sélectionne un compte utilisateur parmi la liste des suggestions, l'interface affichera le profil de ce compte (ses informations et ses tweets). Dans les autres cas,

l'interface affichera une liste de tweets contenant les termes, le hashtag ou l'URL recherchée. Les résultats sont présentés par défaut dans l'ordre chronologique inverse.

A quoi sert le microblogging, pourquoi l'utiliser ?

Principalement à 3 choses :

- **La veille.**

Dans la mesure où on peut rechercher des termes (par exemple, WordPress ou Dotclear) dans les solutions de microblogging, beaucoup de bloggeurs se servent donc de ces applications pour faire de la veille technologique, concurrentiel, ou tout simplement pour trouver le Buzz.

- **La promotion**

On peut promouvoir un produit par ce biais, et c'est même devenu pratiquement une recommandation. Le nombre de personnes qui cherchent l'information sur des réseaux de microblogging est manifestement en forte croissance, et on s'en sert finalement comme pour une recherche Google, sauf que là, on a de grandes chances d'avoir l'information avant tout le monde. De plus, les messages sont très courts, ce qui permet d'absorber un maximum d'informations en un minimum de temps.

On note aussi que certains bloggeurs, et non des moindres, publient des informations courtes, sur un sujet qui n'entre pas dans la ligne éditoriale de leur blog. Ils ne veulent donc pas écrire un article sur ce sujet, mais font quand même la promotion du sujet sur leur application de microblogging, laissant alors la liberté à leurs suiveurs de prendre en compte cette information et de la traiter dans leur blog s'ils le souhaitent. Les suiveurs peuvent donc bénéficier de scoop.

Le microblogging n'est pas réservé uniquement aux bloggeurs. Il peut aussi servir de marketing, ou a toutes sociétés (ou marques) qui veulent faire la promotion de leurs produits. En lançant un message court, elles permettent d'informer les bloggeurs qui eux, voient là des buzzs (voir figure 1.12).

Figure 1.12: Exemple de Tweet de promotion

- **L'amusement**

On utilise aussi le microblogging pour s'amuser entre amis en permettant de publier des images, des vidéos.

Figure 1.13: Exemple de Tweet pour amusement

Juin 2016 : Crue de la seine à Paris, ce qui a causé des inondations et des dégâts

1.7. Conclusion :

Nous avons présenté dans ce chapitre, les différentes plateformes de microblogging les plus utilisées ainsi que les caractéristiques et les particularités de la plateforme Twitter. Twitter est sans conteste la plateforme la plus populaire au sein des internautes. Durant ces dernières années, l'utilisation de ces plateformes a pris une autre tournure, où les utilisateurs jouent le rôle de journalistes en véhiculant l'information de manière instantanée, en échangeant des avis entre consommateurs et aussi où les entreprises font du marketing et la promotion de leurs nouveaux produits. Dans le prochain chapitre, nous présenterons le domaine l'analyse des sentiments et plus précisément le domaine de la fouille d'opinion appliqués à des tweets.

CHAPITRE 2

2.1. Introduction

La fouille de données (Data Mining) consiste à rechercher et extraire de l'information (utile et inconnue) de gros volumes de données stockées dans des bases ou des entrepôts de données. Le développement récent de la fouille de données (depuis le début des années 1990) est lié à plusieurs facteurs : une puissance de calcul importante est disponible sur les ordinateurs de bureau ou même à domicile ; le volume des bases de données augmente énormément ; l'accès aux réseaux de taille mondiale, ces réseaux ayant un débit sans cesse croissant, qui rendent le calcul distribué et la distribution d'information sur un réseau d'échelle mondiale viable.

Outre la fouille de données, plusieurs spécialisations de celle-ci ont vu le jour, telles que la fouille d'images (Image Mining), la fouille du web (Web Data Mining), la fouille de flots de données (Data Stream Mining) et la fouille de text (Text Mining). Cette dernière consiste à extraire des connaissances à partir des textes produits par des humains ou pour des humains.

L'analyse de sentiment ou fouille d'opinion (en anglais ***sentiment analysis*** ou ***opinion mining***) est une spécialité du text mining qui essaye de définir les opinions, sentiments et attitudes présente dans un texte ou un ensemble de texte. Développée essentiellement depuis les années 2000, elle est particulièrement utilisée en marketing pour analyser par exemple les commentaires des internautes ou les comparatifs et tests des blogueurs ou encore les réseaux sociaux : une grande part de la littérature sur le sujet concerne par exemple les tweets. Mais elle peut également être utilisée pour sonder l'opinion publique sur un sujet.

La première mention de l'expression « analyse de sentiment » dans la revue de littérature est attribuée à (Nasukawa et Yi 2003). (Pang et Lee 2008) ont apporté une large définition de l'analyse de sentiment définie comme étant le « traitement informatique d'opinion, de sentiment et de subjectivité dans un texte ». Quelques années plus tard, (Liu 2012) a complété la définition du problème d'analyse de sentiment en définissant ce qu'était une opinion de manière plus précise.

Une opinion est un quintuple composé de (e, a, s, h, t) : (e) le nom de l'entité visé par l'opinion, (a) un aspect de l'entité qui est critiqué, (s) le sentiment exprimé sur l'aspect de l'entité, (h) un porteur d'opinion, et (t) le temps auquel l'opinion a été exprimée. Un sentiment peut avoir trois polarités distinctes : il peut être positif, négatif ou neutre. Il est

important de mentionner que l'analyse de sentiment possède plusieurs terminologies dans la revue de littérature, avec le terme « extraction d'opinion » étant le synonyme le plus employé.

2.2. Les besoins de l'analyse des sentiments :

Avant de se concentrer sur l'analyse des sentiments, il importe de définir le mot « sentiment ». Le terme couvre plusieurs acceptions en fonction du domaine dans lequel il est appliqué. Dans la littérature, un sentiment est « un jugement, une opinion qui se fonde sur une appréciation subjective (et non sur un raisonnement logique) → avis, idée, point de vue ».

Connaitre le sentiment des autres personnes a toujours été un élément d'information important durant le processus de décision. Les gens très souvent demandent à d'autres de leur recommander un mécanicien d'automobiles ou d'expliquer leur choix de votes aux élections par exemple. Avant de prendre des décisions, les gens s'intéressent énormément aux avis des autres personnes dans différents domaines. Ils consultent les avis des autres consommateurs avant d'effectuer un achat, ou regardent les avis des autres personnes avant de voir un film au cinéma ou avant d'acheter un disque. Grâce à l'internet nous pouvons découvrir les opinions et les expériences de très grand nombre de personnes qui ne sont ni nos amis, ni les experts de domaines, mais des gens qui peuvent avoir les mêmes goûts que nous, et donc leurs opinions peuvent être très utiles pour nous avant de faire notre choix et d'avoir notre propre idée sur un sujet donné. Aujourd'hui, de plus en plus de personnes donnent leur avis sur différents sujets, ces avis sont à la disposition de tout le monde sur internet.

2.3. Processus de la fouille d'opinion :

Le processus d'un système de fouille d'opinions comprend trois étapes : acquisition et analyse du corpus[2], étude de la pertinence des documents par rapport à un sujet, détection de l'opinion et ré-ordonnancement des documents **[FB10]**. La figure suivante nous montre les étapes de la fouille d'opinions.

Figure 2.1 : Processus de fouille d'opinions

a) Acquisition et prétraitement du corpus

Dans cette phase, les textes sont prétraités linguistiquement. Une élimination des mots vides et des mots qui n'apportent aucune information est faite, ainsi qu'une analyse lexicale pour enlever les mots qui ont un sens commun (redondant).

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 étape consiste à étudier la pertinence des documents par rapport à un sujet donné, appelé «topic» dans TREC. Nous employons indifféremment ces deux termes. Parmi les méthodes les plus utilisées, on peut citer la méthode probabiliste OKAPI et BM25 [JWR00]. Les documents sont classés, et généralement les 1000 premiers documents les plus pertinents sont extraits, et sont utilisés pour l'étape suivante : la détection d'opinions.

c) La détection d'opinions

Plusieurs méthodes ont été utilisées pour la détection d'opinions. Leur but est de réordonner les documents pertinents selon un score d'opinion. Ainsi les documents qui contiennent le plus d'opinions sont classés parmi les premiers.

2.4. Difficultés de la fouille d'opinions et de l'analyse de sentiments

L'opinion et le sentiment sont le plus souvent décrits par la polarité. Cette dernière est en générale, soit positive (opinion favorable), soit négative (opinion défavorable), soit neutre. Nous montrons ci-dessous quelques difficultés de cette fouille d'opinions [PL08] :

- Difficulté due à l'ambiguïté des mots. Par exemple le mot «petit» est un fait dans la phrase suivante «il est petit». Par contre il exprime une opinion dans «c'est un petit».
- Difficulté due à la structuration de la phrase. Par exemple on oppose deux parties d'une phrase avec la conjonction «mais», 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.
- Difficulté due au contexte : par exemple dans la phrase «je trouve que le film est excellent mais ma sœur le trouve mauvais», les deux opinions sont données par des personnes différentes.

  Dans l'exemple suivant «J'ai regardé le film Avatar dans la salle de cinéma 21 qui est très jolie», l'opinion est donnée sur la salle et non sur le film. Dans le dernier exemple «les acteurs du film ont bien joué, la musique est bonne mais je n'ai pas apprécié ce film», l'opinion de la dernière partie de la phrase est la plus importante.

- Difficulté due au vocabulaire qu'on utilise pour exprimer une opinion. Il diffère d'une personne à une autre, comme par exemple un anglo-saxon lorsqu'il exprime ses sentiments utilise des mots bien représentatifs de ce qu'il ressent contrairement aux personnes qui ne connaissent pas ou peu sa langue.
- Difficulté due à l'emploi d'une thématique. Une même thématique peut être utilisée dans différentes classes et peut exprimer une toute autre signification, comme par exemple «un vieux vin», ou un «vieil habit».

- Difficulté due au langage qu'utilisent les internautes pour s'exprimer. Les ponctuations ne sont pas forcément utilisées pour marquer les fins de phrases, des mots spécifiques sont utilisés tel que : «ha ha ha», «Goood», «super».

2.5. Les campagnes d'évaluations :

2.5.1. La campagne TREC (Text REtrieval Conference)

La campagne TREC est financée par la DARPA (Defense Advanced Research Projects Agency) et le NIST (National Institute of Standards and Technology). La première campagne TREC (TREC-1) voit le jour en 1992 avec 25 participants issus du monde académique et industriel. La 17ème édition de TREC est TREC 2008. À chaque session, TREC met à disposition des participants à la campagne un ensemble de documents et de requêtes. Pour chacune des requêtes, l'ensemble des documents pertinents est déterminé par des juges humains. TREC met aussi à disposition des participants un programme nommé trec-eval qui permet de calculer, pour un ensemble de requêtes, les performances des systèmes selon plusieurs critères et mesures. Elle utilise comme principales métriques les taux de rappel et de précision.

Chaque campagne TREC est composée d'un ensemble de tâches focalisées sur un ou plusieurs aspects de la fouille d'opinion ; tel que la recherche de billets de blogs pertinents, la recherche d'opinion dans les blogs, la détection de la polarité.

2.5.2. DEFT (DEfi Fouille de Texte)

Le défi fouille de textes (DEFT) a été créé en 2005 par un groupe de chercheurs (Prince et al, 2007), dans le but d'initier une série de campagnes d'évaluation francophones sur des thématiques relevant de la fouille de textes. La compagne DEFT'07 dont le thème était en 2007 la classification de textes d'opinion, présents dans différents types de textes. Plusieurs groupes de recherche (laboratoires universitaires et entreprises privées) ont pu tester leurs systèmes de classification sur les mêmes textes. Dans la phase initiale chaque groupe inscrit a reçu les deux tiers de chacun des quatre corpus différents qui avaient comme sujet des critiques de films et de livres, des tests de jeux vidéo, des relectures d'articles scientifiques et des notes de débats parlementaires.

Pour les trois premiers corpus une note à trois valeurs (positif, moyen ou négatif) a été attribuée à chaque texte par le comité des organisateurs, une note à deux valeurs seulement (positif ou négatif) pour le dernier corpus. Après un certain temps pendant lequel chaque

groupe a mis au point son ou ses systèmes de classification un troisième tiers de chaque corpus a été envoyé pour faire les tests dont les résultats ont dû être soumis quelques jours plus tard. Dix équipes ont participé à l'édition 2007.

La 11ème édition du défi fouille de textes (DEFT 2015), s'est portée sur l'analyse de l'opinion, des sentiments et des émotions dans des tweets rédigés en français [DEFT15].
Cette édition du défi propose 3 tâches d'analyse des tweets :

a) Classification des tweets selon leur polarité : La première tâche vise à détecter la polarité des tweets parmi trois valeurs possible : positif (+), neutre ou mixte (=), et négatif (-). La catégorie neutre ou mixte renvoie aussi bien aux messages présentant une polarité neutre (ni positif, ni négatif), que ceux présentant les deux polarités en même temps (un sentiment positif et un sentiment négatif).
b) Classification fine des tweets.

   Cette tâche vise une classification fine des tweets. Elle est divisée en deux sous-tâches :

   - Identification de la classe générique de l'information exprimée dans le tweet : Cette première sous-tâche vise l'identification de la classe générique de l'information exprimée dans le tweet, parmi quatre classes : opinion, sentiment, émotion, information.
   - Identification de la classe spécifique de l'opinion, sentiment ou émotion : Cette deuxième sous-tâche vise l'identification de la classe spécifique de l'opinion, du sentiment, ou de l'émotion exprimée, parmi dix-huit classes : accord, amour, apaisement, colère, déplaisir, dérangement, désaccord, dévalorisation, ennui, insatisfaction, mépris, peur, plaisir, satisfaction, surprise négative, surprise positive, tristesse, valorisation.

c) Détection de la source, la cible et de l'expression d'opinion : Cette dernière tâche vise à analyser plus précisément les opinions, du point de vue de l'expression porteuse de l'opinion, de la source (l'émetteur) et de la cible (le récepteur).

2.5.3. NTCIR

L'Institut National en Informatique (NII) du Japon organise chaque année le workshop NTCIR (NII Test Collection for Information Retrieval System) dans le domaine de la recherche d'information. La tâche de l'analyse d'opinions a été l'objet d'étude de NTCIR-6 [YS07] et NTCIR-7. Elle est proposée aussi pour NTCIR-8 qui a eu lieu en juin 2010. La

détection d'opinions se fait au niveau des phrases. Quatre sous-tâches ont été définies dans NTCIR-6 :

- Savoir si une phrase contient une opinion ou pas (opinionated sentence judgement) ;
- Extraire le nom de la personne ou de l'entité qui émet l'opinion (opinion holder extraction) ;
- Savoir si une phrase est pertinente ou non pour un topic (relevance sentence judgement) ;
- Détecter la polarité.

Une cinquième sous-tâche a été ajoutée dans NTCIR-7 et qui consiste à extraire la cible sur laquelle porte l'opinion.

2.5.4. SEMEVAL

SEMEVAL (SEMantic EVALuation) est une série de campagnes organisées chaque année portant sur le thème de l'analyse sémantique [7], la récente campagne SEMEVAL'14 dont la tâche 4 est « l'analyse de sentiments associés aux aspects » c'est-à-dire à détecter les aspects cibles des opinions. Les corpus fournis pour cette tâche étaient constitués de commentaires d'internautes annotés en aspects et polarités. Plus exactement, la tâche était subdivisée en 4 sous-tâches

a) Extraction des termes dénotant les aspects :
   Par exemple, dans la phrase I liked the service and the staff, but not the food, les termes à détecter sont service, staff et food.

b) Extraction des catégories sémantiques dénotant les aspects : étant donné un ensemble prédéfini de catégories par domaine ({"price", "food", "service", "ambience", et "anecdote"} dans le domaine des restaurants), associer ces catégories aux phrases ; leur granularité est moins fine que celle des termes de la sous-tâche précédente et elles ne sont pas nécessairement associées à la présence de termes dans la phrase.
   Par exemple :
   "The restaurant was too expensive"! {price}
   "The restaurant was expensive, but the menu was great"! {price, food}

c) Extraction de la polarité associée aux termes précédemment détectés : la polarité prend ici 4 valeurs :{positif, négatif, neutre et conflit}.
   Exemple :
   "I hated their fajitas, but their salads were great"!{fajitas : negative, salads : positive}
   "The fajitas are their first plate"!{fajitas : neutral}
   "The fajitas were great to taste, but not to see"! {fajitas : conflit}

d) Extraction de la polarité associée aux catégories précédemment détectées :
   Par exemple :
   "The restaurant was too expensive"!{price : negative}
   "The restaurant was expensive, but the menu was great"! {price : negative, food : positive}

Depuis 2013, SemEval proposent chaque année une tâche pour l'analyse des sentiments sur Twitter.

SemEval 2013 tâche 2 [SEV13]:

a) Classification de polarité dans un contexte : Étant donné un message contenant une instance marquée d'un mot ou une phrase, déterminer si cette instance est positive, négative ou neutre dans ce contexte.
b) Classification de polarité d'un message : Étant donné un message, classer si le message est du sentiment positif, négatif ou neutre. Pour les messages transportant à la fois un sentiment positif et négatif, le sentiment le plus fort doit être choisi.

SemEval 2014 tâche 9 [SEV14]:

Cette tâche est une rediffusion de la tâche 2 de l'édition 2013, avec de nouvelles données de test à partir de Twitter.

SemEval 2015 tâche 10 [SEV15]:

Cette tâche comprend en plus des 2 tâches des précédentes éditions sur l'analyse des sentiments sur twitter, 3 nouvelles sous tâches qui sont :

a) Classification de messages par polarité et par sujets : Étant donné un message et un sujet, classer si le message est du sentiment positif, négatif ou neutre vers le sujet donné. Pour les messages transportant à la fois un sentiment positif et négatif vers le sujet, le sentiment le plus fort doit être choisi.

b) Détecter les tendances d'un sujet : Étant donné un ensemble de messages sur un sujet donné à la même période de temps, déterminer si le sentiment dominant envers le sujet cible dans ces messages est (a) fortement positif, (b) faiblement positif, (c) neutre, (d) faiblement négatif, ou (e) fortement négatif.

c) Détermination de la force de l'association des termes Twitter avec le sentiment positif : Étant donné un mot ou une phrase, fournir une note comprise entre 0 et 1 qui est une indication de sa force d'association avec un sentiment positif. Un score de 1 indique une association maximale avec un sentiment positif et un score de 0 indique moins association avec le sentiment positif. Si un mot est plus positif que l'autre, alors il devrait avoir un score plus élevé que l'autre.

SemEval 2016 tâche 4 [SEV16]:

Cette édition est une rediffusion de SemEval-2015 Tâche 10 avec deux changements importants:

- Mettre l'accent sur les nouveaux problèmes d'apprentissage de la machine: la quantification et la classification ordinale.
- 2 points et échelle de 5 points (contre 3 points qu'ils ont utilisés dans le passé).

Nous présentons la liste des sous-tâches de cette nouvelle édition :

a) Classification de polarité d'un message (Rediffusion).

b) Classement de tweets sur une échelle de 2 points (Partiellement nouvelle): Étant donné un tweet connu sur un sujet donné, classer si le tweet transmet un sentiment positif ou négatif envers le sujet. (C'est une simplification de la sous tâche 3 de SemEval 2015 tâche 10 (Classification de message par polarité sur des sujets basés)).

c) Classement des tweets sur une échelle de 5 points (Nouvelle) : Étant donné un tweet connu sur un sujet donné, estimer le sentiment véhiculé par le tweet vers le sujet sur une échelle de cinq points.

d) Quantification des tweets sur une échelle de 2 points (Nouvelle) : Étant donné un tweet connu sur un sujet donné, estimer la distribution des tweets dans les classes positives et négatives.

e) Quantification des tweets sur une échelle de 5 points (Nouvelle) : Étant donné un tweet connu sur un sujet donné, à travers les cinq classes d'une échelle de cinq points.

2.5.5. ROMIP

ROMIP est une campagne internationale d'évaluation annuelle en recherche d'information qui a débuté en 2002 [ROMIP]. Pour la campagne de 2011 [ROMIP11], les organisateurs ont ajouté une piste sur l'analyse de sentiments dont le but était le classement en opinion des textes écrits par des consommateurs. Un jeu de données composé de critiques de produits commerciaux issues de services en ligne de recommandation Imhonet et de l'agrégateur de produits Yandex.Market a été fourni aux participants pour entraîner leurs systèmes.

Le jeu de données contenait des critiques pour trois types de produits : les appareils photo numériques, les livres et les films.

2.6. Les approches de détection d'opinions

Il existe trois types d'approches pour la détection d'opinions et l'analyse des sentiments :

2.6.1. L'approche symbolique :

(Appelées aussi classification non supervisées ou encore approches basées sur lexique). L'approche symbolique se base sur une analyse syntaxique du texte faite par un analyseur fonctionnel et relationnel. Cet analyseur traite un texte donné en entrée phrase par phrase et en extrait, pour chaque phrase, les relations syntaxiques présentes. Une majorité de relations d'opinions positives détermine une polarité positive du texte, tandis qu'une majorité de relations d'opinions négatives provoque une polarité négative du texte entier. L'équilibre entre relations d'opinions positives et négatives entraine une classification moyenne du texte [MCD].

2.6.2. L'approche statistique :

(Appelées aussi classification supervisées, approches basées sur l'apprentissage machine ou encore approches basées sur corpus). Ce sont des techniques d'apprentissage automatique. Cette approche regroupe 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) [MCD] [CH10]. Ceci, en utilisant un corpus qui a été déjà annoté manuellement au préalable, dans le but de le faire apprendre au système [CH10]. Ces approches consistent à attribuer les données à un classifieur qui génère un modèle qui est utilisé pour la partie test de l'apprentissage.

2.6.3. L'approche hybride

(Appelées aussi classification semi-supervisées)**.** Cette approche combine les points forts des deux approches précédentes. Elle prend en compte tout le traitement linguistique des approches symboliques avant de lancer le processus d'apprentissage comme dans les approches statistiques [CH10].

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

2.7. Etat de l'art sur l'analyse des sentiments

2.7.1 Introduction

Nous nous intéressons dans cette partie aux travaux relatifs à la détection d'opinions et de la polarité (La classification des sentiments). La détection d'opinions est une tâche qui permet d'extraire les opinions d'un ensemble de documents pertinents pour un sujet donné. Elle est confrontée à des problèmes qui la distinguent de la recherche traditionnelle thématique dont les sujets sont souvent identifiés par des mots clés seulement. L'opinion (ou le sentiment) peut être exprimée de manière très variée et subtile et donc il est difficile de la déterminer. La classification du sentiment (polarité) est une sous-tâche de la détection d'opinions. Elle consiste de façon générale à déterminer si l'opinion du document sur le sujet est positive ou négative. La détection d'opinions se fait au niveau du document, du paragraphe ou de la phrase.

Nous présentons dans cette première partie les travaux relatifs à la détection d'opinions et dans la deuxième partie ceux relatifs à la détection de la polarité.

2.7.2. Détection d'opinion

Les premiers travaux évoqués dans la littérature s'intéressent à la classification de textes suivant des genres, dont certains tels que « éditorial » sont subjectifs (Karlgren et Cutting (1994)). En 1994, Wiebe [WJ94] s'intéressait plus précisément à l'idée de subjectivité, en cherchant à détecter des private states, définit comme des états (dans ce cas des parties de textes) qui ne peuvent pas être associés à des observations objectives et vérifiables. À cette époque, les données disponibles devaient être annotées manuellement, ce qui rendait la tâche assez laborieuse. Avec l'essor de l'internet social, la disponibilité de données annotées par le critique (en associant une note à son texte) a explosée. En 2002,

Turney [PT02], Pang et al. [PLV02] encouragent la recherche dans le domaine de sentiment analysis en classant des critiques de cinéma. Dans les travaux de Turney, la classification dans ce domaine donnait les moins bons résultats parmi les autres domaines (automobiles, banques et tourisme). Pang parvient tout de même à de bons résultats en utilisant des techniques d'apprentissages simples et adaptables aux autres domaines.

La principale idée derrière la fouille d'opinions est de guider le processus de décision d'un utilisateur en proposant un résumé des évaluations sur un concept donné. Dans leur livre, Pang et Lee (2008) exposent les intérêts socio-économiques d'un hypothétique moteur de recherche qui répondrait à la requête « Que pensent les gens de .. ? ». Ce problème est approché par Vernier et al. (2009) qui tentent de répondre à la requête « Que pensent les gens de Sarah Palin ? ».

En général, il existe deux types d'approches utilisées, l'une basée sur un lexique et l'autre sur l'apprentissage machine.

2.7.2.1.Approches basées sur le lexique (Lexicon-Based Approach)

Ces approches utilisent des dictionnaires de mots subjectifs, Ces dictionnaires peuvent être généraux comme le General Inquirer, Sentiwordnet[7], Opinion Finder, NTU Sentiment Dictionary (NTUSD), Wilson lexicon ...etc. Ils peuvent également être construits à partir des corpus. Dans ces dictionnaires, une polarité est associée a priori à chacun des mots. Quel que soit le contexte dans lequel il sera inséré, le mot devrait ainsi avoir toujours la même polarité. On donne ensuite au document un score d'opinion en fonction de la présence de mots issus de ces dictionnaires dans le texte.

Ces dictionnaires ont été constitués de différentes façons :

- à la main ;
- à partir de corpus ;
- à partir de dictionnaires existants.

Ces dictionnaires sont utilisés pour classifier des textes dont on sait qu'ils parlent de l'entité nommée qui nous intéresse, par exemple pour les critiques de cinéma, le titre d'un film. La méthode consiste à détecter le terme (adjectif ou l'expression qualifiante) qui est en co-occurrence avec l'entité nommée, souvent au sein de la même phrase. Le dictionnaire

fournit ainsi la tonalité affectée à l'entité concernée par ces termes qualifiants. De nombreux problèmes, notamment syntaxiques, rendent problématique cette relation entre une entité et ses qualificatifs, c'est pourquoi d'autres traitements doivent être effectués pour obtenir un résultat satisfaisant. Aussi standardisée que puisse paraître cette approche par les dictionnaires, c'est elle qui est le plus souvent utilisée dans les services qui sont mis sur le marché et, en première approximation, elle donne des résultats intéressants, selon le niveau d'exigence que l'on se fixe. L'approche par les dictionnaires a donc des limites évidentes, mais elle a surtout l'avantage de permettre des calculs rapides sur de grands corpus.

L'apprentissage peut être utilisé pour régler une partie des défauts de l'approche par lexique, elle peut être efficace pour générer des scores individuels pour les documents et aussi pour l'apprentissage d'une fonction de classement qui combine le score d'opinion et le score de pertinence dans une seule fonction de classement [FB10].

Nous exposons brièvement dans ce qui suit les travaux de quelques universités et laboratoires de recherche qui se sont basés sur le lexique pour déterminer l'opinion. Le résultat de ces approches est comparé à ceux jugés corrects (on nomme ces derniers «baseline»).

a) L'université de Glasgow (Écosse)

Le travail de cette université a été exposé dans l'article [HMP+07] et présenté à TREC 2007. Deux approches pour la recherche d'opinion ont été proposées. La première est basée sur un lexique de termes pondérés et la deuxième utilise OpinionFinder[5] [Opi05].

Dans la première approche, un dictionnaire de 12000 mots anglais dont les termes sont pondérés a été utilisé. Ce dictionnaire est construit à partir de la collection de documents.

Le poids des termes est calculé en utilisant un modèle de pondération des termes. En plus, la proximité des termes de la requête par rapport aux phrases avec opinion est prise en considération pour améliorer la performance dans la recherche d'opinion. À cet effet, le modèle pBiL2 «terme proximité» est utilisé. Ce dernier est un modèle binomial aléatoire qui calcule le score d'une paire de termes de la requête dans le document. Dans la deuxième approche, les auteurs utilisent OpinionFinder. Le score d'opinion du document est calculé en fonction de celui fourni par OpinionFinder. Les résultats montrent que la première approche améliore les performances dans la recherche d'opinion de 15,8 % de celles du « baseline » et de 8,96 % de celles obtenues par OpinionFinder.

b) L'université d'Indiana (États-unis)

Les travaux de cette université sur la fouille d'opinions sont exposés dans l'article [YYZ07] et présentés à TREC 2007. Les auteurs utilisent cinq lexiques. Le premier HF «High Frequency» regroupe les termes avec opinion, les plus couramment utilisés. Ce lexique est construit semi automatiquement. Le deuxième lexique WL «Wilson's Lexicon» est construit à partir des termes de Wilson's subjectivity. Il est divisé en trois collections : subjectivité forte, subjectivité faible et les termes qui amplifient l'intensité de l'opinion. Le troisième lexique LF «Low Frequency» est construit à partir de mots créés par les gens pour exprimer leur forte opinion (exemple «soooo», «goood») et sont des mots rares. Le quatrième lexique est appelé «IU» lexique (I and You). Les auteurs remarquent que les pronoms (I, you, my, your, our, me) sont souvent utilisés pour exprimer des opinions. Ce lexique est constitué par les n-grammes[3] qui commencent et qui se terminent par un terme IU. Le dernier lexique dit «OA» lexique (Opinion Acrononym) est construit manuellement.

Le score d'un document est calculé selon une pondération[6] des scores des différents lexiques.

Cette approche améliore les performances du «baseline» de 14,07 %.

c) Le laboratoire DUTIR (Information Retrieval laboratory of Dalian University of Technology) (Chine)

Ce laboratoire a participé à toute les tâches de TREC 2007 Blog Track [STS+07]. Dans la recherche d'opinion, les auteurs créent leur propre lexique composé de 2000 mots subjectifs qui sont fréquents dans le corpus étudié, au lieu d'utiliser les lexiques tels que General Inquiry ou SentiWordNet. Ces derniers, d'après eux, contiennent certains mots qui ne sont pas fréquemment utilisés dans les blogs. La requête est étendue par des mots pris du descriptif ou de la narration du topic, ou bien de Wikipedia. L'expression d'opinion dans le blog est détectée par une simple recherche de mots subjectifs du lexique considérée autour des mots de la requête. L'évaluation de cette approche a donné une amélioration de 10,38 % par rapport à la «baseline».

d) L'université d'Amsterdam (Pays-bas)

Cette université a participé à la campagne de TREC 2007 dans plusieurs tâches notamment celle de la recherche d'opinion [MdR06]. Les auteurs utilisent dans ce contexte là une approche basée sur le lexique «General Inquiry» qui contient 10000 mots et est construit

manuellement. Il est divisé en 5 catégories. Catégorie «positifs-négatifs», catégorie «émotionnelle» (pleasure, pain, feel, etc.), catégorie «pronom» (self, our, you), catégorie «adjectif» (adjectif relationnel, adjectif indépendant) et catégorie «respect» (une liste complète des catégories de General Inquiry est donnée dans http://www.wjh.harvrd.edu/~inquirer/homecat.htm). Pour chaque blog post, deux valeurs «sentimentales» sont calculées, en utilisant les mots du dictionnaire de chaque catégorie : la valeur sentimentale au niveau du post «post opinion level» et la valeur sentimentale au niveau du blog «feed opinion level». Dans les deux cas, la valeur d'opinion est le nombre d'occurrences des mots de chaque catégorie, normalisé par le nombre total de mots. La différence est le texte utilisé pour compter le nombre d'occurrences. Pour le «post opinion level», les phrases pertinentes sont extraites à partir du post, et utilisées pour le calcul de la valeur d'opinion.

Le score final d'opinion d'un document est une combinaison linéaire des scores pondérés, obtenus par les différentes méthodes : pertinence, opinion (blogpost et feed blog), autorité.

e) L'université de Pohang (république de la Corée) : KLE (Knowledge and Langage Engineering)

Pour la détection d'opinions, les auteurs proposent une approche basée sur le lexique.

Ce dernier est créé en utilisant SentiWordNet et «Amazon's review corpus». L'approche est basée sur le modèle «passage level» qui a été aussi utilisé pour la tâche de recherche de documents pertinents. Cette approche consiste à segmenter le document en passages, et à calculer le score d'opinion de chaque passage. Le passage ayant le score maximal est sélectionné. Les auteurs considèrent ensuite la partie du document qui contient ce passage et les mots avant et après ce passage dans le document. Le score est calculé pour cette partie et représente le score du document. Le modèle OKAPI est utilisé pour la normalisation des blogposts selon leur longueur. Le score d'opinion est calculé en utilisant une extension de la requête initiale par un ensemble représentatif de tous les mots subjectifs. Cet ensemble est appelé POW (Pseudo Opinionated Word). Cette université a participé à TREC 2008.

2.7.2.2.Approches basées sur l'apprentissage machine (Machine Learning)

Ces approches utilisent des classifieurs. Des données sont fournies au classifieur pour l'apprentissage.

Les données représentent des phrases subjectives (ou des documents avec opinion).

Le classifieur génère un modèle, qui sera utilisé dans la partie test. Des «features» sont utilisées pour l'apprentissage tels que les bigrammes, les n-grammes, POS (étiquettes morphosyntaxiques) etc. Plusieurs types de classifieurs ont été utilisés [PLV02] : SVM, Multiples Classifieur, Naïf de Bayes, Classifieur d'entropie maximale (MaxEnt), ainsi que la régression logistique.

a) Naïf Bayes est une approche probabiliste qui utilise une loi de Bayes où les probabilités sont fonction des mots contenus dans les documents :

$$P(c/d) = P(c) * \frac{P(d/c)}{P(d)}$$

Avec d est un document et c la classe du document. «P(c/d)» est déterminé par le classifieur Bayésien naïf [YH03].

b) L'approche SVM repose sur la notion d'hyperplan séparateur et de marge maximale. Un hyperplan séparateur entre deux ensembles de points (ensemble de documents de polarité positive et l'ensemble de document de polarité négative) est la frontière entre ces deux ensembles. La marge représente la distance entre un de ces ensembles et cet hyperplan.

c) La régression logistique est une méthode statique permettant de produire un modèle pour décrire des relations entre une variable catégorielle et un ensemble de variables de prédiction.

On présente ci-dessous quelques travaux utilisant l'apprentissage machine :

a) Université d'Arkansas à Little Rock (États-unis)

Cette université a participé à la tâche de détection d'opinions et de polarité dans les blogs [ZJB07]. Trois approches différentes ont été proposées. La première est basée sur l'hypothèse que la présence d'un indicateur de subjectivité près du topic ou des mots de la

requête permet de dire que l'opinion du document est sur le topic donné. Les mots indicateurs de subjectivité considérés sont : «I, You, We, Me, They, He, She» nommés IU model, et les mots «Like, Feel, Think, Love, hate, Suck, Nice, Good, Bad, awesome, awful, never, think, feel» nommé IU2. La recherche de ces indicateurs de subjectivité se fait sur une fenêtre de 20 mots. Deux exécutions ont été soumis, l'un utilisant seulement le titre du topic, l'autre le TDN (titre, narration, descriptif) et l'expansion de la requête avec des mots indicateurs d'opinion. La deuxième approche utilisée est basée sur l'apprentissage machine, de type SVM. Dans cette approche les topics sont divisés en 6 catégories : thing, compagny, food, events, location, person. Un SVM est utilisé pour chaque catégorie de topic. La troisième approche est basée sur le traitement du langage naturel (NLP) et dite «one-pass-processing approach» et elle traite la pertinence et la détection d'opinion en une seule étape. Les documents sont analysés et segmentés en passages. La première méthode qui est basée sur les IU mots donne de meilleurs résultats, elle améliore la «baseline» de 14,07 %, le CML de 2,96%, le IU2 de 3,2%, le TDN de 6,8% et le NPL de 5,3 %.

b) L'université de Neuchâtel (Suisse) (Computer Science Department)

Cette université a participé aux campagnes d'évaluation de TREC et NTCIR à travers deux travaux différents :

(a) Les travaux de cette université présentés à TREC 2008 sont exposés dans l'article [FS08]. L'opinion et la polarité sont traitées en une seule étape. Le document est classé en positif, négatif, mixte ou neutre. Si le document est classé neutre alors il est considéré sans opinion autrement il est considéré comme contenant une opinion. Deux approches, basées sur la méthode de Muller sont utilisées pour cette classification. La première dite «Additive Model» utilise les statistiques d'un terme pour calculer un score de polarité pour chaque document. Ce score est basé sur le Z-score[15] de Muller. La deuxième approche utilise le classifieur de la régression logistique, douze features et le Z-score de Muller. La requête est étendue de deux manières. La première est basée sur la méthode de Rocchio où les m termes les plus importants extraits des k premiers documents retrouvés pour la requête initiale, sont rajoutés à la requête étendue. La deuxième extension utilise Wikipedia. Le titre du topic est soumis à Wikipedia et les dix premiers mots les plus fréquents du premier article retrouvé sont ajoutés à la requête.

(b) Les travaux présentés à NTCIR-7 (MOAT : Multilingual Opinion Analysis Task) sont exposés dans l'article [OZ08]. Ils consistent à étudier l'opinion et la polarité au niveau

d'une phrase dans un contexte multi-langues (anglais, japonais et chinois). L'approche utilisée est basée sur l'apprentissage machine. Les auteurs utilisent la méthode de Muller pour calculer les poids des termes, et le classifieur de type régression logistique pour déterminer la catégorie de la phrase (positive, négative, neutre ou sans opinion). Pour la langue anglaise, les auteurs étendent la requête initiale avec 500 mots identifiant les événements du discours (« explained», «commented», etc.) ou des expressions subjectives («sympathized», «accused», etc.). Les auteurs remarquent que les résultats ne sont pas performants pour certaines requêtes.

c) Le laboratoire «National Laboratory of Pattern Recognition», Chinese Academy of Sciences, Pékin de Chine

Ce laboratoire a participé à NTCIR-7 pour la tâche MOAT (Multilingual Opinion Analysis Task), plus particulièrement aux sous-tâches de détection de phrases subjectives et d'extraction du porteur d'opinion (opinion holder extraction), pour la langue chinoise simplifiée. L'approche utilisée pour la détection d'opinions est basée sur l'apprentissage machine. Les auteurs utilisent un classifieur subjectif et un certain nombre de features tels que les adjectifs et verbes apparaissant dans la phrase, les entités nommées, la structure de la phrase lors de l'analyse, les idiomes[4] apparaissant dans la phrase, et les mots de deux dictionnaires. Le premier dictionnaire «opinion operator lexicon» est constitué de verbes qui peuvent signaler un événement d'opinion (par exemple, «believe», «say», «persist», etc.) et est constitué à partir du corpus NTCIR-6. Le deuxième dictionnaire (opinion word lexicon) contient les mots (verbes, adjectifs et adverbes) qui expriment une opinion. Ce dictionnaire est construit à partir du corpus de NTCIR-6, du dictionnaire de sentiment de l'université nationale de Taiwan (il contient 11,088 mots) et du dictionnaire HowNet (qui contient 8,938 mots subjectif). Les auteurs concluent que les résultats ne sont pas satisfaisants et que d'autres features doivent être ajoutées.

Il existe d'autres travaux importants concernant le domaine de la détection d'opinions. Dans [ABM09], les auteurs proposent une approche pour l'analyse d'opinions dans les textes, basée sur une analyse lexicale-sémantique des expressions et sur les relations rhétoriques entre ces expressions. Les auteurs proposent une catégorisation des expressions d'opinions en quatre catégories «reporting expressions», «judgment expressions», «advise expressions», «sentiments expressions». Chaque catégorie est divisée en sous-catégories. Le deuxième apport de ce travail est l'utilisation des relations rhétoriques entre les expressions contenant

des opinions. Cinq types de relations rhétoriques ont été définies : «contrast», «correction», «explanation», «result», «continuation». L'approche proposée consiste à segmenter le texte en segments contenant des opinions, et à représenter chaque segment par une structure contenant la catégorie à laquelle appartient le segment, la modalité associée, l'entité qui a émis l'opinion, le sujet de l'opinion, le mot d'opinion et le contenu de l'opinion. Cette représentation permet en utilisant un ensemble de règles de calculer l'opinion associée à un texte. La validation de cette approche s'est faîte sur trois types de corpus : un corpus sur des critiques de film (pris de Telerama et AlloCine.fr), un corpus sur des lettres aux éditeurs (The San Francisco Chronicle et la Depeche du Midi) et un corpus sur des nouvelles (Le monde, 20 Minutes et MUC6).

2.7.3. Classification de la polarité d'opinion :

En plus de la détection d'opinions (subjectif, objectif), des travaux ont consisté à la classification de ces opinions par (positif, négatif, neutre). La classification des sentiments est un raffinement de la détection d'opinions dans la mesure où elle permet de classifier les documents ayant une opinion sur un sujet en classes. Il existe deux types de classifications : binaire ou multi-classes. La classification binaire définit deux classes : positive et négative. Par contre la classification multi-classes définit cinq classes : fortement positive, positive, neutre, négative, fortement négative. La plupart des travaux se sont focalisés sur la classification binaire mais la classification multi-classes peut être utile dans certaines applications où on veut faire une meilleure classification. Les auteurs de [KS05] montrent qu'il est crucial d'utiliser des exemples neutres dans l'apprentissage de la polarité pour diverses raisons.

L'apprentissage des exemples positifs et négatifs seuls ne permet pas de bien classifier les exemples neutres. De plus, l'apprentissage d'exemples neutres facilite la distinction entre les exemples positifs et les exemples négatifs. Comme pour la détection d'opinions, il existe deux approches pour la polarité, l'une basée sur le lexique, l'autre sur l'apprentissage machine. Nous présentons dans ce qui suit quelques travaux pour cette tâche.

2.7.3.1.Approches basées sur le lexique

a) L'université de Glasgow (Écosse)

Le calcul de la polarité a été fait en utilisant leur première approche de la détection d'opinions. Cette approche est basée sur le même lexique interne de 12000 mots anglais.

Une mesure de divergence est calculée des mots du dictionnaire qui sont distribués dans les documents pertinents. Le résultat a été exprimé par une mesure R_ACC qui est de 0,2295 et la valeur de la précision à 10 documents est de 0,37. Les résultats obtenus sont parmi les meilleurs [OMS08].

b) L'université d'Indiana (États-unis)

La même approche est utilisée pour la polarité que celle qui a été utilisée pour la détection d'opinions et les mêmes lexiques. Cette approche est basée sur l'utilisation du lexique WL «Wilson's Lexicon» et d'autres lexiques construits.

c) L'université de Pohang (république de la Corée) : KLE (Knowledge and Langage Engineering)

La même approche pour la détection de l'opinion est utilisée pour la polarité sauf qu'au lieu de considérer la subjectivité du mot, ils considèrent l'orientation sémantique (polarité).

Un score de polarité du document est calculé, et qui est la différence entre le score positif et le score négatif. Si cette différence dépasse un seuil, alors la polarité du document est considérée comme positive, si elle est inférieure au seuil, elle est considérée comme négative, autrement elle est neutre. Les auteurs considèrent seulement le titre du topic pour la requête.

2.7.3.2.Approches basées sur l'apprentissage machine (Machine Learning)

a) L'institut «Institute of Computing Technology» de l'académie chinoise des sciences de Pékin de Chine

Cet institut a participé à TREC 2007 dans la recherche d'opinion. Le travail est exposé dans l'article [LCW+07]. Les auteurs utilisent le classifieur Drag-push pour obtenir la polarité des blogs pertinents. L'apprentissage de ce classifieur a été fait en utilisant les résultats de TREC 2006. Ils proposent aussi de filtrer les spams. Un blog est classifié comme un spam si le nombre de ses liens dépasse un seuil prédéfini. Ce seuil est calculé par une approche heuristique. Les SVM sont utilisés comme outil pour classifier les blogs spam et les blogs non spam.

b) Le laboratoire DUTIR (Information Retrieval laboratory of Dalian University of Technology) (Chine)

Concernant la polarité, les chercheurs utilisent une méthode mixte basée sur un lexique de 2000 mots subjectifs et sur un classificateur de type SVM. Ce dernier prend comme «features» les mots du lexique et aussi des mots sélectionnés par une méthode basée sur le gain de l'information. L'ensemble d'apprentissage n'est pas un ensemble de documents mais un ensemble de phrases contenant les mots du topic. Ces phrases sont extraites des documents. Les auteurs soulignent que les meilleurs résultats sont obtenus en utilisant seulement le titre du topic.

Remarque :

Il existe d'autres travaux proposés par les auteurs [BMCB], ils proposent une approche mixte pour la détection de la polarité, une approche basée sur le lexique et l'apprentissage machine du type régression logistique, Une étude de l'impact de la catégorisation des topics de TREC Blog 2007 a été effectuée. Il en résulte que la polarité est très sensible à la catégorisation : une amélioration de 98,18% pour la polarité positive et de 85,24% pour la polarité négative a été constatée et ceci comparativement aux performances obtenues sans classification.

Les auteurs [BPR14] ont participé à la campagne SEMEVAL 2014, pour la tâche 4 concernant l'analyse des sentiments associés aux aspects, il se sont fondé sur l'utilisation d'un composant symbolique permettant de détecter les termes du domaine et leur polarité et d'un

composant de classification, qui classe les phrases selon leurs catégories sémantiques (aspects) et dans un second temps, leur associe une polarité. Ils montrent également comment les résultats de la classification de la polarité des catégories sémantiques peuvent améliorer à posteriori la polarité des termes. Les auteurs ont cherché à améliorer leur système, en particulier pour certaines polarités, (neutre et conflit). Pour cela, ils ont réalisé quelques expériences préliminaires qui montrent qu'une amélioration est possible.

2.8. Les techniques de classification :

Les approches basées sur l'apprentissage machine consistent à attribuer des données à un classificateur pour l'apprentissage. Le rôle d'un classifieur est de classer dans des groupes (des classes) les échantillons qui ont des propriétés similaires, mesurées sur des observations. Ce dernier 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 classificateur. Les principales features utilisées sont : mots seuls, bigrammes, tri-grammes, part of speech et polarité. Les principaux classificateurs sont les SVM, Naïf Bayes, Réseaux de neurones, Maximum Entropy et régression logistique. La plupart des chercheurs considèrent que seuls les adjectifs sont porteurs de subjectivité, alors que d'autres pensent que certains adverbes, noms et verbes peuvent aussi contenir de la subjectivité.

2.8.1. Classifieur Naïf Bayes :

La classification naïve bayésienne est un type de classification Bayésienne probabiliste simple basée sur le théorème de Bayes avec une forte indépendance (dite naïve) des hypothèses [8].

Le classifieur bayésien naïf est une méthode d'apprentissage supervisé qui repose sur une hypothèse simplificatrice forte : les descripteurs (Xj) sont deux à deux indépendants conditionnellement aux valeurs de la variable à prédire (Y).

Malgré leur modèle de conception « naïf » et ses hypothèses de base extrêmement simplistes, les classifieur bayésiens naïfs ont fait preuve d'une efficacité plus que suffisante dans beaucoup de situations réelles complexes.

Modèle de Bayes :

De manière abstraite, le modèle probabiliste pour un classificateur bayésien est un modèle conditionnel. Il se base sur la règle de bayes qui s'énonce de la manière suivante :

$$P(A|B_1, B_2, \dots, B_n) = \frac{P(B_1, \dots, B_n|A) * P(A)}{P(B_1, \dots, B_n)}$$

La probabilité d'avoir l'évènement A étant donné B_1,..., B_n est donné par le rapport entre la probabilité d'avoir les évènements B_1, .., $B_n$étant donné A et la probabilité que B_1, .., B_n se soient produits. Tant que le dénominateur ne dépend pas de l'évènement A, on peut considérer la probabilité $P(B_1, \dots, B_n)$ comme étant constante.

Le caractère "naïf" de ce théorème vient du fait qu'on suppose l'indépendance des différentes classes $B_i, \dots, B_j$. Ce qui en d'autres termes se traduit par :

$$P\left(B_i \middle| A, B_j\right) = P(B_i|A)$$

Cette hypothèse permet également d'écrire :

$$P(A, B_1, \dots, B_n) = P(A) * P(B_1|A) * \dots * P(B_n|A) = P(A) \prod_{i=1}^{n} P(B_i|A)$$

Ce théorème a beaucoup d'applications dans le domaine du traitement de l'information notamment en traitement de la parole, traitement des images et bien d'autres.

2.8.2. Classifieur Support Vector Machine (SVM)

Les machines à vecteurs de support (Support Vector Machine, SVM) appelés aussi séparateurs à vaste marge sont des techniques d'apprentissage supervisées destinées à résoudre des problèmes de classification. Les machines à vecteurs supports exploitent les concepts relatifs à la théorie de l'apprentissage statistique et à la théorie des bornes de Vapnik et Chervonenkis. La justification intuitive de cette méthode d'apprentissage est la suivante : si l'échantillon d'apprentissage est linéairement séparable, il semble naturel de séparer parfaitement les éléments des deux classes de telle sorte qu'ils soient le plus loin possibles de la frontière choisie. Ces fameuses machines ont été inventées en 1992 par Boser et al. ,mais leur dénomination par SVM n'est apparue qu'en 1995 avec Cortes et al. [Ma14]. Depuis lors, de nombreux développements ont été réalisés pour proposer des variantes traitant le cas non-

linéaire. Le succès de cette méthode est justifié par les solides bases théoriques qui la soutiennent. Elles permettent d'aborder des problèmes très divers dont la classification. SVM est une méthode particulièrement bien adaptée pour traiter des données de très haute dimension.

Principe de la technique SVM

Cette technique est une méthode de classification à deux classes qui tente de séparer les exemples positifs des exemples négatifs dans l'ensemble des exemples. La méthode cherche alors l'hyperplan qui sépare les exemples positifs des exemples négatifs, en garantissant que la marge entre le plus proche des positifs et des négatifs soit maximale. Cela garantit une généralisation du principe car de nouveaux exemples pourront ne pas être trop similaires à ceux utilisés pour trouver l'hyperplan mais être situés d'un côté ou l'autre de la frontière. L'intérêt de cette méthode est la sélection de vecteurs supports qui représentent les vecteurs discriminant grâce auxquels est déterminé l'hyperplan. Les exemples utilisés lors de la recherche de l'hyperplan ne sont alors plus utiles et seuls ces vecteurs supports sont utilisés pour classer un nouveau cas, ce qui peut être considéré comme un avantage pour cette méthode.

a) Classifieur linéaire :

Un classifieur est dit linéaire lorsqu'il est possible d'exprimer sa fonction de décision par une fonction linéaire en *x*. On peut exprimer une telle fonction par:

$$h(x) = < w, x > + b = \sum_{i=1}^{n} w_i \, x_i + b$$

Où $w(\in R^n)$ est le vecteur de poids et $b(\in R^0)$ le biais, alors x que est la variable du problème. X est l'espace d'entrée et qui correspond à R^n, où n est le nombre de composantes des vecteurs contenant les données. Notons que l'opérateur < > désigne le produit scalaire usuel dans R^n. w et b sont les paramètres à estimer de la fonction de décision $h(x)$.

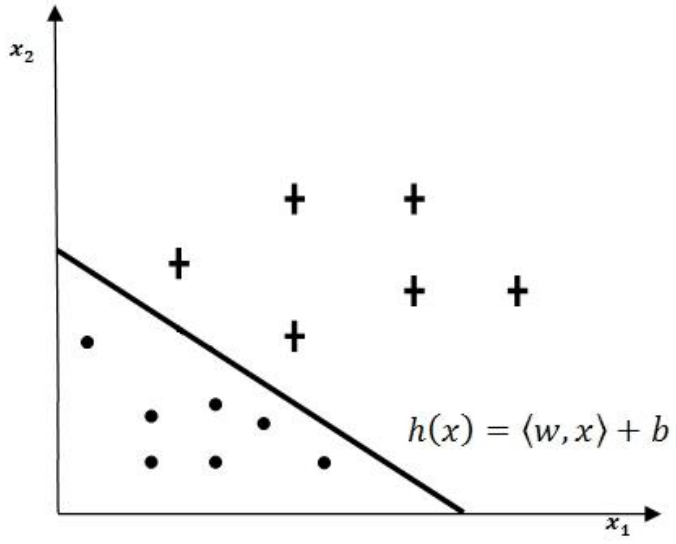

Figure 2.2 : Exemple d'un hyperplan séparateur

b) Classifieur non-linéaire :

Le paragraphe précédent décrit le principe des SVM dans le cas où les données sont linéairement séparables. Cependant, dans la plupart des problèmes réels, ce n'est pas toujours le cas et il est donc nécessaire de contourner ce problème (difficile de séparer n'importe quel jeu de données par un simple hyperplan). Si par exemple les données des deux classes se chevauchent sévèrement, aucun hyperplan séparateur ne sera satisfaisant.

Dans ce but, l'idée est de projeter les points d'apprentissage x_i dans un espace T de dimension q, plus élevée que n grâce à une fonction non-linéaire θ qu'on appelle fonction noyau. L'espace ainsi obtenu est appelé espace des caractéristiques ou aussi espace transformé.

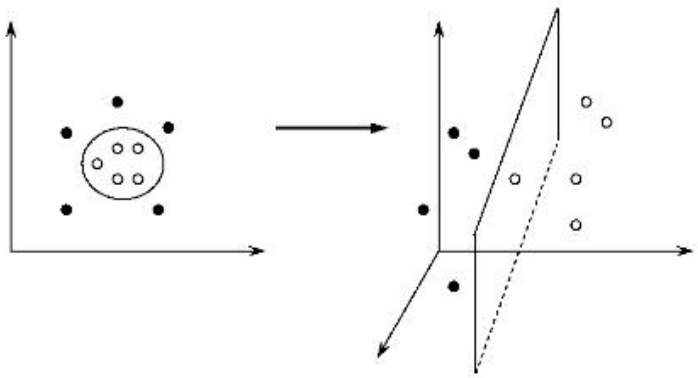

Figure 2.3 : Transformation d'un problème de séparation non-linéaire en un problème de séparation linéaire

2.8.3. Réseaux de neurones

Avec l'avancée dans le domaine de la neurobiologie concernant le fonctionnement du cerveau et des neurones, des mathématiciens ont essayé de modéliser le fonctionnement du cerveau en intégrant ces connaissances en biologie dans des programmes informatiques pour leur donner la possibilité d'apprendre : c'est la naissance des réseaux de neurones.

Qu'est-ce qu'un réseau de neurones ? [RFRN]

Tout d'abord, ce que l'on désigne habituellement par réseau de neurones. Est en fait un réseau de neurones artificiels basé sur un modèle simplifié de neurone. Ce modèle permet certaines fonctions du cerveau, comme la mémorisation associative, l'apprentissage par l'exemple, le travail en parallèle, mais le neurone artificiel est loin de posséder toutes les capacités du neurone biologique. Les réseaux de neurones biologiques sont ainsi beaucoup plus compliqués que les modèles mathématiques et informatiques.

Il n'y a pas de définition universellement acceptée de « réseau de neurones ». On considère généralement qu'un réseau de neurones est constitué d'un grand ensemble d'unités (ou neurones), ayant chacune une petite mémoire locale. Ces unités sont reliées par des canaux de communication (les connexions, aussi appelées synapses d'après le terme biologique correspondant), qui transportent des données numériques. Les unités peuvent uniquement agir sur leurs données locales et sur les entrées qu'elles reçoivent par leurs connexions.

Certains réseaux de neurones sont des modèles de réseaux biologiques, mais d'autres ne le sont pas. Historiquement l'inspiration pour les réseaux de neurones provient cependant de la volonté de créer des systèmes artificiels sophistiqués, voire intelligents, capables d'effectuer des opérations semblables à celles que le cerveau humain effectue de manière routinière, et d'essayer par-là d'améliorer la compréhension du cerveau.

La plupart des réseaux de neurones ont une certaine capacité d'apprentissage, cela signifie qu'ils apprennent à partir d'exemples. Le réseau peut ensuite dans une certaine mesure être capable de généraliser, c'est-à-dire de produire des résultats corrects sur des nouveaux cas qui ne lui avaient pas été présentés au cours de l'apprentissage.

Neurone formel

Les réseaux de neurone formels ou artificiels sont des réseaux dont l'architecture est inspirée de celle des réseaux de neurones biologiques (naturels), ils sont généralement optimisés par des méthodes d'apprentissage de type statistique. Leur modélisation revient à décrire le modèle du neurone (unité de base) et le modèle des connexions entre les neurones. Le premier neurone formel est apparu en 1943, introduits par MacCulloch et Pitts (unité à seuil). La figure ci-après montre un schéma d'un neurone formel.

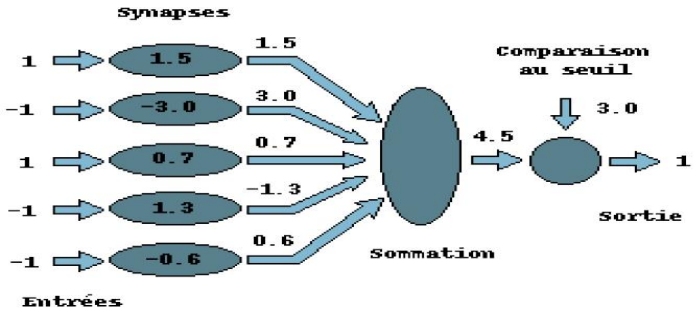

Figure 2.4 : Neurone formel (Artificiel)

2.9. L'analyse des sentiments appliquée sur les données issues de Twitter

(Go, Huang, et Bhayani 2009) furent les premiers (selon leur connaissance) à travailler sur l'analyse de sentiments sur les messages de Twitter. Ils ont comparé différents classificateurs (algorithmes qui performent une classification de sentiment sur un texte ou phrase donnée), basés sur différentes approches bien connue dans la revue de littérature de l'analyse sémantique telles que la méthode naïve bayésienne, le « Support Vector Machine » (SVM), et l'approche commune de base qui attribue une classification de polarité selon le nombre de mots contenant une certaine polarité (ex : si une phrase contient plus de mots positifs que de mots négatifs alors la phrase est positive). À la suite de leur travaux, ils ont conçu une application nommée Sentiment140 qui est considérée être la meilleure application publiquement disponible pour l'analyse de sentiment sur Twitter selon (Barbosa and Feng 2010).

(Pak et Paroubek 2010) ont construit un classificateur basé sur un algorithme de type Naïves Bayes qui utilisent des options tels que les n-grams, les POS tags. Cependant, dans une autre étude (Efthymios Kouloumpis 2011) mentionnent que les tags POS ne semblaient pas très utiles dans leur approche d'analyse de sentiment. L'objectif de recherche de (Pak et Paroubek 2010) était d'évaluer l'utilité d'options linguistiques tels que les tags POS, les

hashtags ou les émoticônes dans la tâche de classification de la polarité des tweets. Ils indiquent que les émoticônes représentent une excellente option linguistique dans la classification de la polarité des tweets.

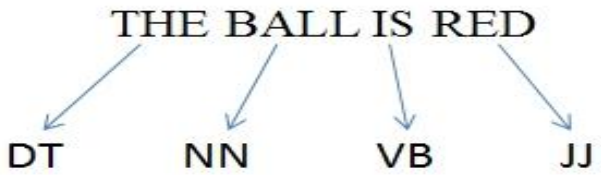

Figure 2.5 : étiquetage morpho-syntaxique (POSTag)

DT : déterminant, NN : Nom, VB : Verbe, JJ : Adjectif.

(Agarwal et al. 2011) ont introduit un nouveau modèle d'analyse de sentiment basé sur un « arbre kernel ». L'arbre kernel (appelé aussi Tree Kernel) a été défini pour le calcul de similarité entre les arbres grammaticaux (ou arbre syntaxiques). Un arbre syntaxique est obtenu à partir d'une phrase en la décomposant en groupe grammatical [AV07]. Ce nouveau modèle semble avoir de meilleurs résultats que les modèles de base introduits par les précédents chercheurs tels que (Go, Huang, et Bhayani 2009; Pak et Paroubek 2010). Un exemple du modèle d' « arbre kernel » est représenté à la figure ci-après.

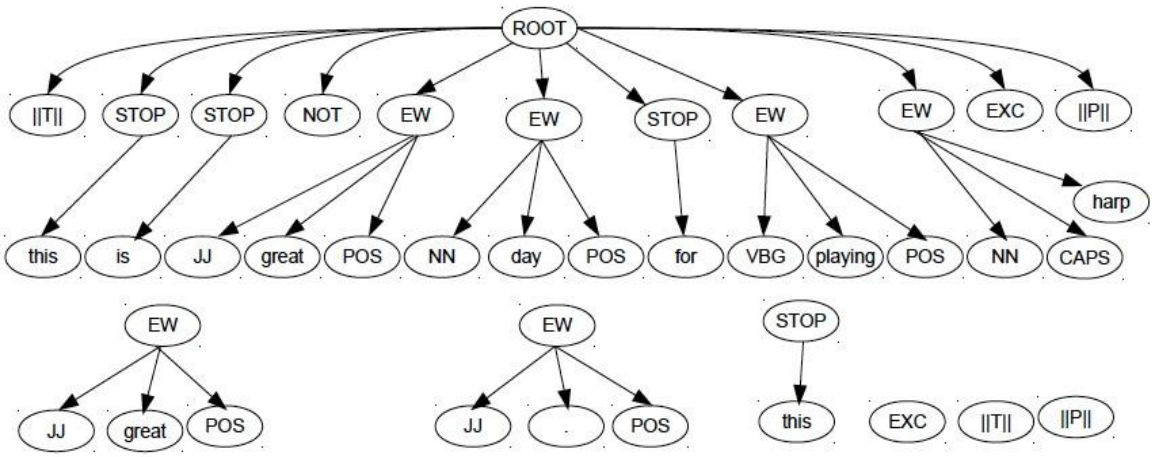

Figure 2.6 : Arbre Kernel pour le Tweet « @Fernando this isn't a great day for playing the HARP! :) »

Dans la figure ci-dessus, on peut voir un exemple d'un arbre kernel pour le Tweet suivant "@Fernando this isn't a great day for playing the HARP! :).". A) Tout d'abord initialisation de l'arbre par le mot "ROOT", ajout du tag ||T|| (target) pour "@Fernando", ajout du tag "NOT" pour le token "n't", ajout du tag "EXC" pour le point d'exclamation à la fin de la phrase et ||P|| pour l'émoticône. B) Si un token est un mot vide (Stop Word), nous allons

simplement ajouter le sous arbre "(STOP('stop-word')) à la racine (ROOT). Pour l'exemple qu'on vient de voir, nous allons ajouter un sous arbre pour chacun des mots vide suivants (this, is et for). C) Si le token est un mot de l'Anglais (English Word), ajouter à ce mot son POS Tag et sa polarité, ici pour le mot great, nous allons ajouter le sous arbre (EW (JJ great POS)).

(Brun et Roux 2014) [BR14] ici ont procédé à plusieurs expériences de classification sur le corpus Imagiweb de tweets annotés, qui consiste à étudier l'image des personnalités politiques ou de compagnies telle qu'elles apparaissent sur la Toile. Ce corpus comprend donc 3920 tweets dont les opinions sont annotées en polarité (ici positive ou négative), comprenant un total de 392 hash tags décomposés. Ici les auteurs essayent de voir si la méthode d'intégration de hashtags proposé par (Wang et Al. 2011) améliore la détection de la polarité.

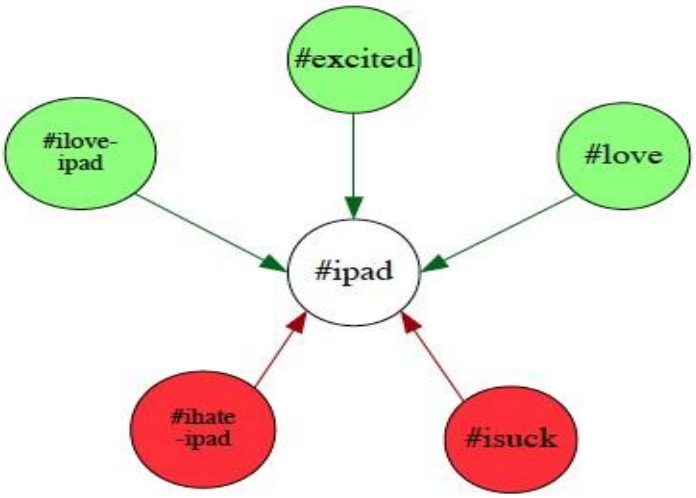

Figure 2.7 : Le modèle graphique de Hashtag représentant les liens avec le hashtag #ipad (adapté de Wang et al. 2011)

Pour cela les auteurs ont réalisé une série d'expériences :

- Expérience 1 (référence) est basée sur l'approche en sac de mots[8].
- Expérience 2 (hash tags) rajoute les hashtags et leur polarité.
- Expérience 3 (opinions) rajoute les opinions détectées sans les hashtags.
- Expérience 4 (opinions + hashtags) intègre les opinions détectées et les hashtags.

Le tableau suivant représente le résultat des quatre expériences :

Expériences	Exactitude %
Expérience 1 : sac de mots	80.1%
Expérience 2 : hashtag	82.6 %
Expérience 3 : opinions	80.2 %
Expérience 4 : opinions + hashtags	82.2 %

Tableau 2.1 : Tableau comparatif des différentes expériences

Alors que l'utilisation des relations d'opinion dans l'entrainement de la classification n'a pas d'impact significatif, l'utilisation des hash tags améliore l'exactitude de la classification d'environ 2%.

Pour confirmer ce résultat, Ils ont effectué les mêmes expériences avec un sous-ensemble du corpus initial, dans lesquels ne sont conservés que les tweets contenant des hash tags :

Expériences	Exactitude %
Expérience 1 : sac de mots	79.9 %
Expérience 2 : hashtag	84.6 %
Expérience 3 : opinions	80.1 %
Expérience 4 : opinions + hashtags	84.7 %

Tableau 2.2 : Tableau comparatif des différentes expériences sur le sous-ensemble du corpus Imagiweb

Dans ce dernier cas, l'amélioration sur la tâche de classification est passée à 4,8% par rapport à la référence. Ces résultats prouvent sans ambigüité l'impact significatif de l'intégration des hash tags et de leur polarité dans la classification des tweets.

(Karoui et al. 2015) ont présenté une approche par apprentissage automatique pour la détection de l'ironie dans les tweets. Leur objectif était de tester si la négation est un bon indice pour la détection de tweets ironiques. Pour cela, ils ont constitué un corpus de tweets ironiques et non ironiques contenant ou non des mots de négation. Pour mesurer l'effet de la négation sur la tâche de détection de l'ironie, ils ont divisé le corpus récoltés en 3 corpus : les tweets avec négation (NegOnly), les tweets sans négation (NoNeg), et un corpus regroupant l'ensemble des tweets (All). Ils ont vu que les traits de surface[12] traditionnellement utilisés

pour cette tâche dans d'autres langues sont aussi efficaces pour le français sur le corpus NegOnly. Les auteurs ont introduit de nouveaux traits tels que les traits de sentiments[10], traits modifieurs de sentiments[11], traits d'opposition[13] et trait de contexte[14]. Les traits de sentiment sont meilleurs pour le corpus NoNeg, et les traits pour modifieurs et opposition sont les meilleurs pour le corpus All [KZMAH15].

(Wang et al. 2015) ont montré que les émoticônes sont largement utilisés par les utilisateurs de Twitter. En particulier, les émoticônes exprimant un sentiment positif, comme :) et ;), et ceux exprimant un sentiment négatif, alors que beaucoup d'autres sont relativement rare. Pour analyser l'impact des émoticônes sur la détection de la polarité, les auteurs ont effectué trois autres analyses sur des données issues Twitter et ont étudié la relation entre les émoticônes et l'expression du sentiment sur les média sociaux. La première analyse a illustré le fait que les émoticônes sont utilisés régulièrement dans des contextes similaires en démontrant que des émoticônes et des mots exprimant un sentiment similaire sont regroupés dans un même groupe. Ça a pu montrer que le sens complexe véhiculé par certaines émoticônes, tels que :| et :'), peut être compris à travers les mots qui apparaissent dans les mêmes contextes. La deuxième analyse, en comparant le sentiment de tweets avec et sans émoticônes, fournit des preuves directes sur l'importance des émoticônes pour exprimer le sentiment sur les questions sociales médias. Dans près de la moitié des cas, les émoticônes ont été les seuls composants dans le texte qui exprime un sentiment positif ou un sentiment négatif. Lorsque les émoticônes ont été enlevés, la polarité de ces tweets est devenue neutre ou pas clair. En dernière analyse, ils ont évalué l'impact de la suppression des émoticônes du texte pour les classifieurs d'apprentissage automatique. Ils montré que les classifieurs sont devenus moins précis lorsque les émoticônes ont été enlevés. D'après toutes ces analyses, les émoticônes sont un bon moyen pour la détection de sentiments et de polarité [WAC15].

La majorité des méthodes proposées pour l'extraction d'opinions et la classification par polarité ont été créées pour l'Anglais, quelques travaux plus récents existent aussi pour le Français (Grouin et al, 2009) et pour les émotions (Mohammad, 2012). (Abdaoui et al. 2015) ont participé à la 11ème édition du défi fouille de textes (DEFT 2015), ils s'intéressent particulièrement à la classification des tweets en langue française. Pour les 3 tâches auxquelles ils ont participé, ils utilisent des méthodes d'apprentissage supervisé, comme les machines à vecteur de support SVM avec la méthode SMO (Sequential Minimal Optimization) (Platt, 1999) implémenté dans WEKA. D'après l'état de l'art, cet algorithme

d'apprentissage s'est avéré efficace sur des tâches de catégorisation de textes et spécifiquement d'analyse de sentiments et d'émotions [ANABLMP15].

2.10. Utilité de l'application de l'analyse de sentiment avec Twitter

La recherche sur l'analyse de sentiment au niveau de granularité de la phrase appliquée aux messages de Twitter s'étend sur différents champs d'application tels que la politique, la finance de marché, ou l'industrie du cinéma.

(Tumasjan et al. 2010) ont conduit des travaux de recherche dans le contexte des élections nationales de 2009 en Allemagne. Le premier objectif de la recherche était d'évaluer si Twitter était utilisé seulement comme un moyen pour une personne d'exprimer son opinion politique ou si Twitter était aussi un moyen d'entretenir une conversation politique. Le second objectif de leur recherche était de déterminer si Twitter avait un certain pouvoir prédictif d'élections électorales en utilisant le volume des tweets reliés aux six partis en compétition durant ces élections nationales. Les résultats de leur recherche indiquent que Twitter est utilisé dans 30% des cas pour avoir une conversation politique et dans 70% des cas pour exprimer une opinion sur un parti. Pour la tâche d'analyse de sentiment, ils ont utilisé un logiciel nommé LIWC. Après avoir déterminé pour chaque parti politique le nombre de tweets positifs et le nombre de tweets négatifs publiés sur Twitter, ils ont compilé les scores en classifiant les partis selon le nombre de tweets publiés et le ratio de tweets positifs sur tweets négatifs. Leurs résultats indiquent que le volume de tweets permet de prédire les résultats des élections avec une erreur absolue moyenne (MAE) de 1.65%, erreur similaire à celles des autres sources de prédictions électorales basées sur des sondages.

(O'Connor et al. 2010) indiquent que Twitter possède un potentiel de prédiction d'élections politique aux Etats-Unis. Leurs travaux ont été conduits après les élections présidentielles américaines opposant le démocrate et actuel président Barack Obama et le républicain John McCain. L'analyse de sentiment a été réalisée à l'aide d'OpinionFinder, une liste de mots avec leur polarité (négative ou positive), similaire à WordNet mentionné précédemment. Ainsi chaque mot d'un tweet est comparé aux mots contenus dans OpinionFinder et caractérisé comme positif ou négatif selon la classification d'OpinionFinder. La polarité des tweets a été évaluée en comptant le nombre de mots positifs et négatifs présent dans un tweet. Si le nombre de mots positifs est supérieur au nombre de mots négatifs alors le tweet est positif. Par la suite, un score de sentiment a été mesuré chaque jour pour chacun des candidats en calculant le ratio de tweets positifs sur les tweets négatifs. Les résultats de

l'analyse indiquent les données de twitter sont corrélés à ceux des sondages et de plus que Twitter est un indicateur précurseur comparé aux sondages : la détection du sentiment de la population est détectable avant les sondages.

Dans le domaine de la finance, l'analyse de sentiment a été utilisée pour développer des stratégies d'investissements sur les marchés boursiers. (Sprenger et Welpe 2010) ont analysé la polarité d'environ 250 000 tweets afin de détecter le caractère à la hausse de certains échanges sur des actions définies et en déduire le volume d'échange du jour suivant. Les chercheurs soulèvent la difficulté de comprendre comment l'information est considérée au sein des microblogs tels que Twitter. Ils mentionnent que « choisir les bons tweets est aussi difficile que d'effectuer les bonnes transactions » sur les marchés financiers.

(Smailovic, Grcar, et Znidarsic 2012) ont utilisé l'analyse de sentiment sur les tweets et le volume de tweets pour identifier d'important évènement et prédire le prix de l'action d'Apple. Pour la tâche d'analyse de sentiment, ils ont utilisé un classificateur de type « Supervised Vector Machine » (SVM) fonctionnant avec la base de donnée de (Go, Huang, et Bhayani 2009). Leurs résultats indiquent qu'il est possible de prédire les changements du prix de l'action d'Apple deux jours à l'avance en se basant sur les variations du sentiment positif des tweets indiquées par l'analyse de sentiment.

(Asur et Huberman 2010) ont développé un modèle basé sur les messages de Twitter afin de prédire les ventes de billets de cinéma. Ils ont conçu un premier modèle basé sur le taux de tweets par minute portant sur un film donné présentement à l'affiche dans les salles de cinéma. Le second modèle construit, se base sur l'analyse de sentiment des tweets du film à l'étude. L'analyse de sentiment a été réalisée avec un classificateur DynamicLCM. Ils ont ensuite calculé pour un film donné, un ratio de polarité (PNratio) calculé comme le quotient du nombre de tweets positifs (numérateur) sur le nombre de tweets négatifs (dénominateur). Les résultats de leur analyse indique le ratio de tweet par minute est un meilleur indicateur afin de prédire les ventes de film que la polarité des tweets. Cependant, combiner le ratio de tweet par minute et la polarité des tweets dans le modèle améliore les performances par rapport à utiliser le ratio de tweet par minute individuellement.

2.11. Conclusion :

Dans ce chapitre, nous avons présenté le domaine de la fouille de textes et plus précisément le domaine de la fouille d'opinions. Nous nous sommes intéressés aux différentes approches de l'analyse des sentiments, les différents travaux sur la détection d'opinion et de polarité. Nous avons aussi abordé les différents travaux de recherche effectués sur l'analyse des sentiments appliqués à des tweets. Dans le prochain chapitre nous présenterons quelques outils réalisés dans le domaine de la fouille d'opinions.

CHAPITRE 3

3.1. Introduction :

L'analyse du sentiment est extrêmement utile en veille des medias sociaux car elle permet d'obtenir une vue d'ensemble sur l'opinion du public au sujet de certains thèmes. Dans ce chapitre, nous allons présenter les outils réalisés dans ce domaine qui rendent ce processus facile et rapide grâce à leurs capacités d'analyse en temps réel.

3.2. 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 [9]. 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 apparaissent aux non-initiés comme des messages codés tant l'usage des hashtags, abréviations en tout genre, argot, anglicismes, et autres émoticons y est répandu [10].

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

3.2.1. Sentiment140

Sentiment140 (anciennement connu sous le nom "Twitter sentiment") est un outil en ligne gratuit qui a été créé par trois étudiants en informatique à 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 [11].

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 3.1 : L'interface de Sentiment140.

Les résultats détaillés pour la requête « iPhone 4s » s'affiche (figure 3.2) :

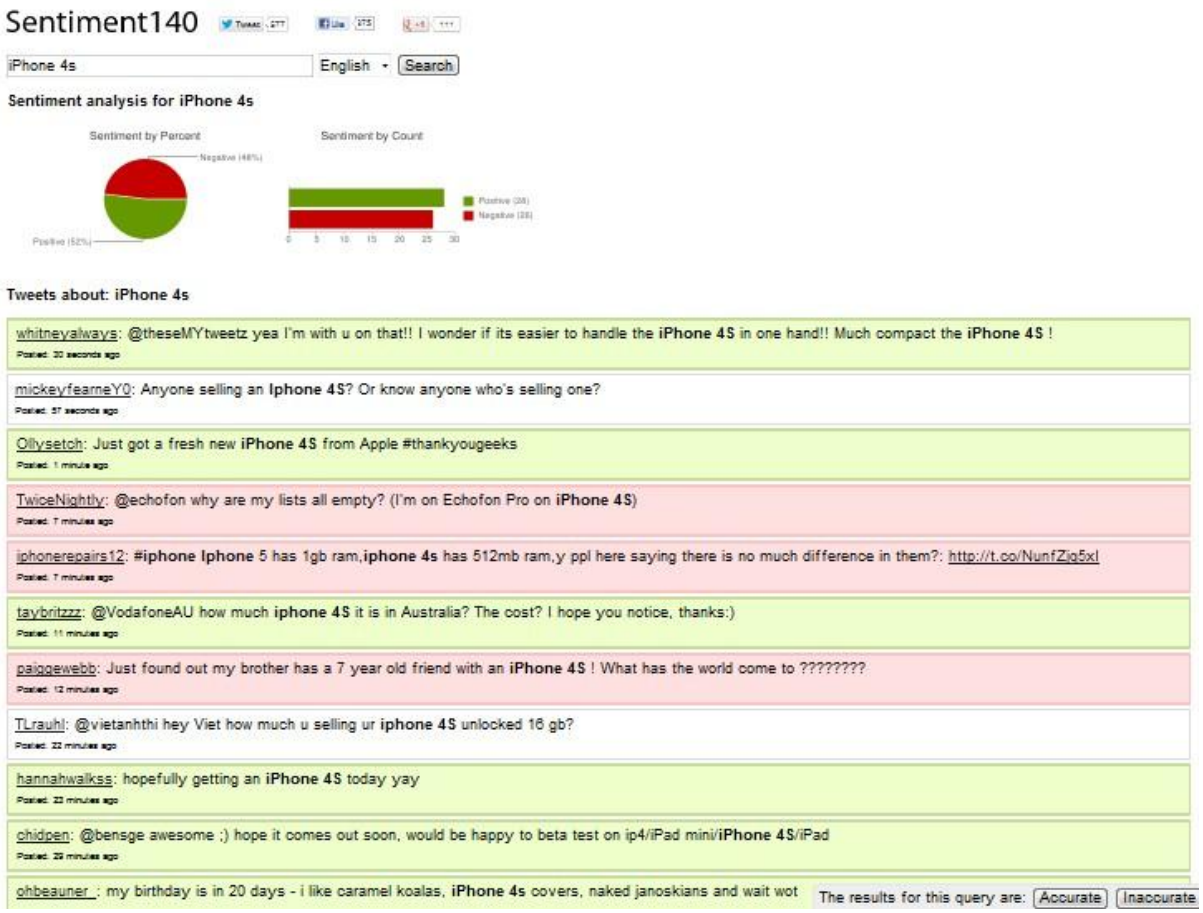

Figure 3.2 : Résultats de la requête « iPhone 4s ».

3.2.2. 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 [12].

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 donnant un sentiment global de Twitter sur la marque [13].

- 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 3.3 : L'interface de TweetFeel.

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

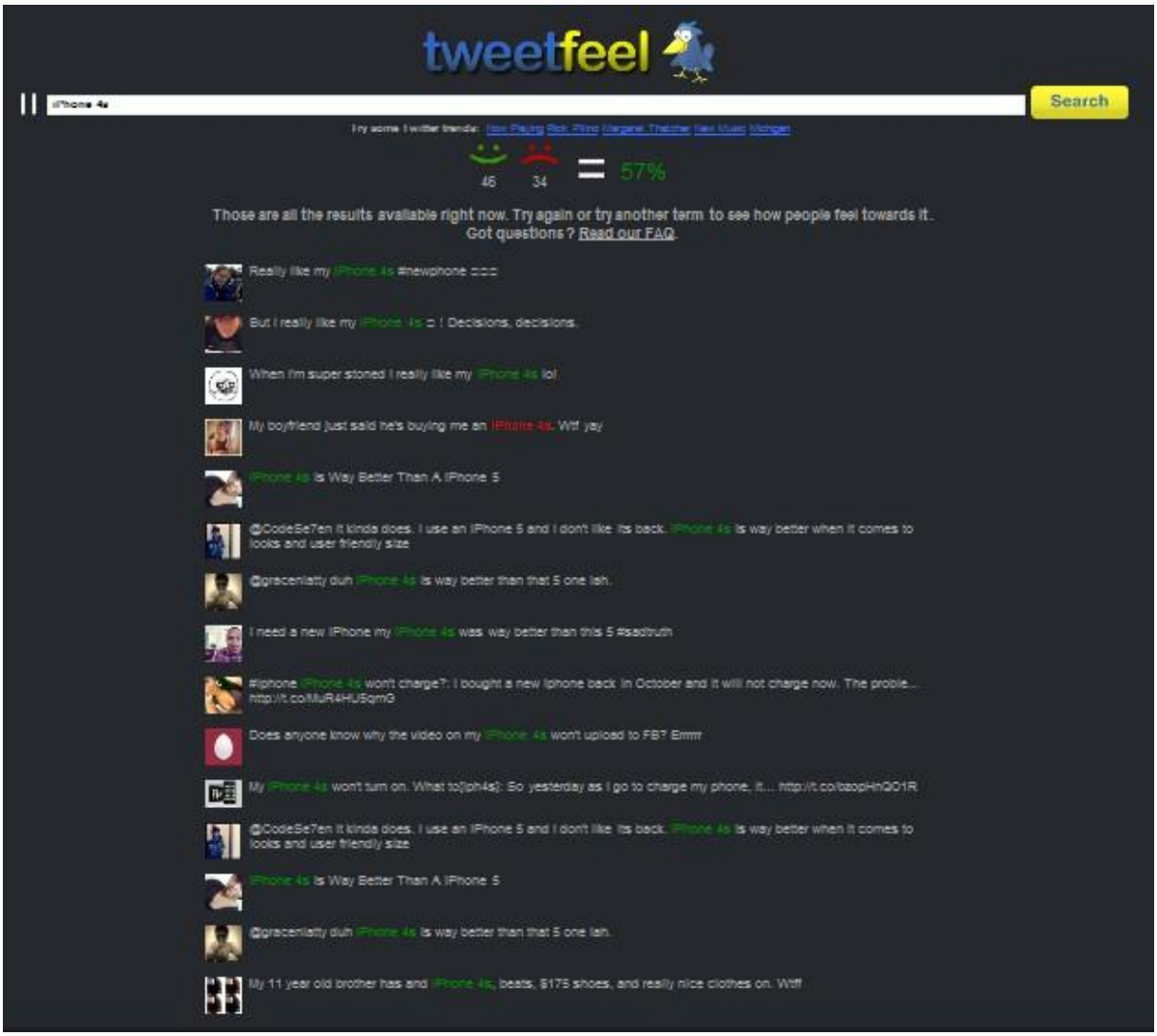

Figure 3.4 : Résultats de la requête « iPhone 4s ».

3.2.3. 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 [14]. 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 [14].

- 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 3.5 : L'interface de twitrratr.

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

Figure 3.6 : Résultats de la requête « iPhone 4s ».

3.2.4. 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.

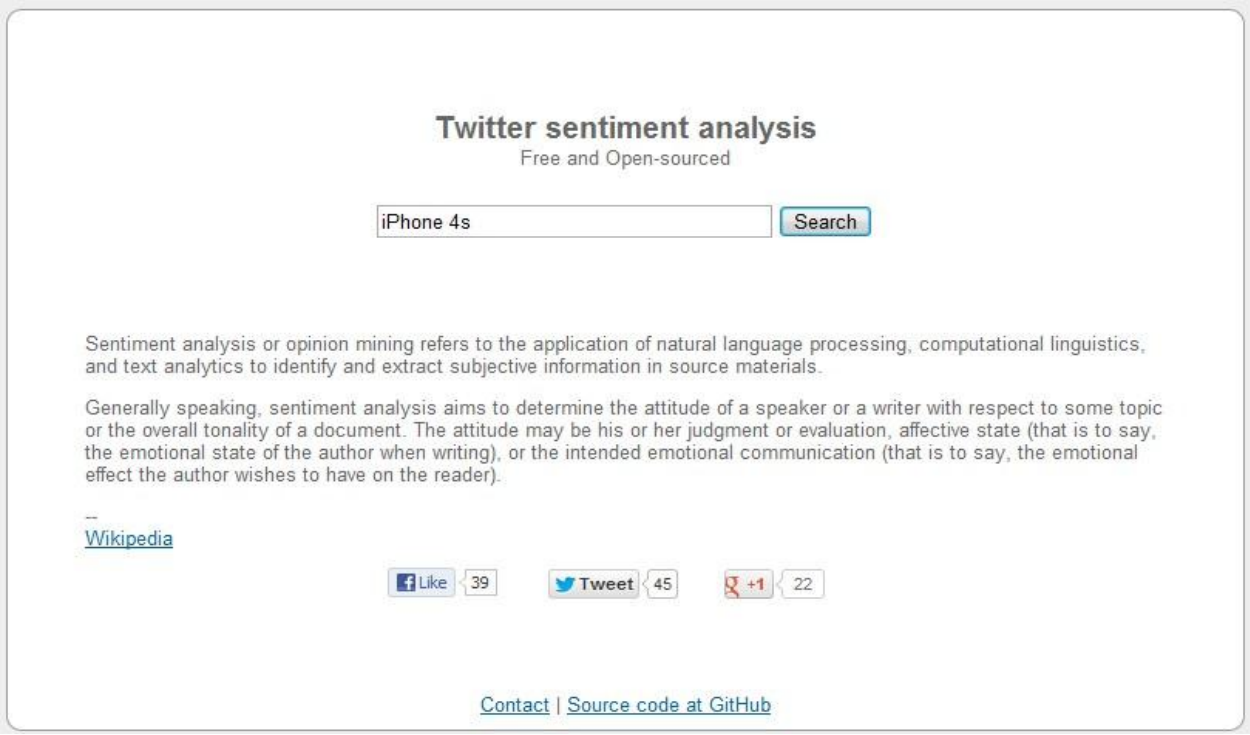

Figure 3.7 : L'interface Twitter Sentiment Analysis.

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

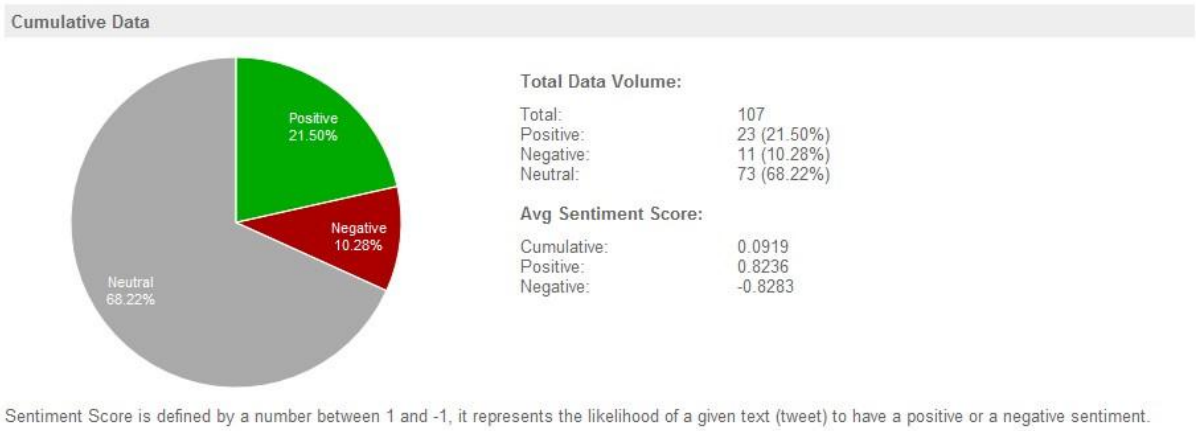

Figure 3.8 : Les données cumulatives.

3.3. Tableau comparatif des outils d'analyse des sentiments :

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

L'application / Caractéristiques	Sentiment 140	TweetFeel	twitrratr	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 3.1 : Tableau comparatif des outils d'analyse des sentiments.

3.4. Conclusion :

Dans ce chapitre, nous avons présenté les services en ligne qui proposent des outils pour analyser les données générées via le service de microblogging Twitter. On a vu que certaines plateformes proposent en plus de l'analyse des sentiments, un monitoring sur un sujet donné. Dans le prochain chapitre, nous présenterons notre application qui se porte sur le domaine de la fouille d'opinion sur Twitter.

CHAPITRE 4

4.1. Introduction :

Dans les chapitres précédents nous avons présenté les différents travaux de recherche effectués dans le domaine de l'analyse des sentiments et plus précisément l'analyse des sentiments sur des Tweets. Nous avons aussi présenté quelques outils réalisés dans ce domaine-là. Dans ce chapitre nous allons présenter la conception de notre application, ce qui nous permettra en suite de passer à la réalisation de cette dernière.

4.2. Conception :

Dans cette partie, on va s'intéresser à la conception de notre application, ce qui consistera à trouver des solutions techniques pour nous guider dans la réalisation de l'application.

4.2.1. Phase 1 : Phase d'apprentissage

Cette phase permet de classifier un ensemble de documents selon des règles déterminées par un classifieur. Cette phase doit être entrainée sur un jeu de données d'apprentissage (qu'on appellera par la suite corpus d'entrainement), pour élaborer un modèle de classification sur ce jeu de données, qui sera utilisé à chaque fois dans la classification d'un jeu de données prédictive.

Entrée :

- Un ensemble de classes C = $\{c_1, c_2, .. , c_n\}$ = {Positive, Négative}
- Un ensemble de documents déjà annotés par leur classe $\{(d_1, c_1), (d_2,c_2),..,(d_n,c_n)\}$

  = {(corpus annoté positivement, Positive), (corpus annoté négativement, Négative)}

Sortie :

- Un classifieur modèle des termes pondérés à utiliser pour la phase de prédictive.

Pour le calcul de la pondération des mots, on utilise les probabilités de bayes suivantes :

$$\boldsymbol{P}(\boldsymbol{W_i/Positif}) = \frac{Le\ nombre\ de\ Tweets\ positifs\ avec\ le\ mot\ W_i}{Le\ nombre\ de\ Tweets\ positifs}$$

$$\boldsymbol{P}(\boldsymbol{W_i/N\acute{e}gatif}) = \frac{Le\ nombre\ de\ Tweets\ n\acute{e}gatifs\ avec\ le\ mot\ W_i}{Le\ nombre\ de\ Tweets\ n\acute{e}gatifs}$$

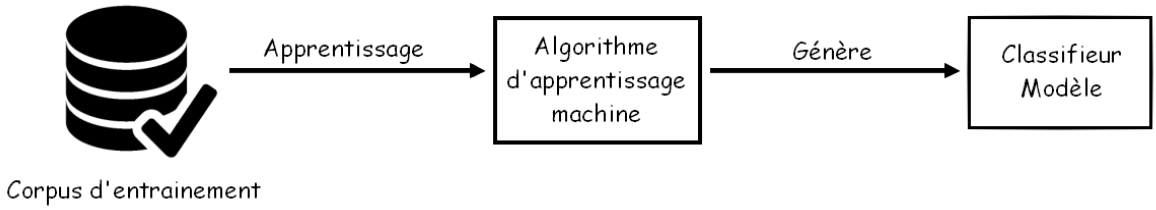

Figure 4.1 : Processus de la phase d'apprentissage

Pour cette phase, on a utilisé un corpus de tweets annotés manuellement qu'on a téléchargé sur le site de l'institut Max-Planck d'informatique (Allemagne) [15].

4.2.2. Phase 2 : Phase prédictive.

Durant cette phase, le programme utilise le classifieur modèle généré par la phase d'apprentissage pour prédire la classe de chaque tweet.

4.2.2.1. Architecture de l'application :

a) Fonctionnement en ligne

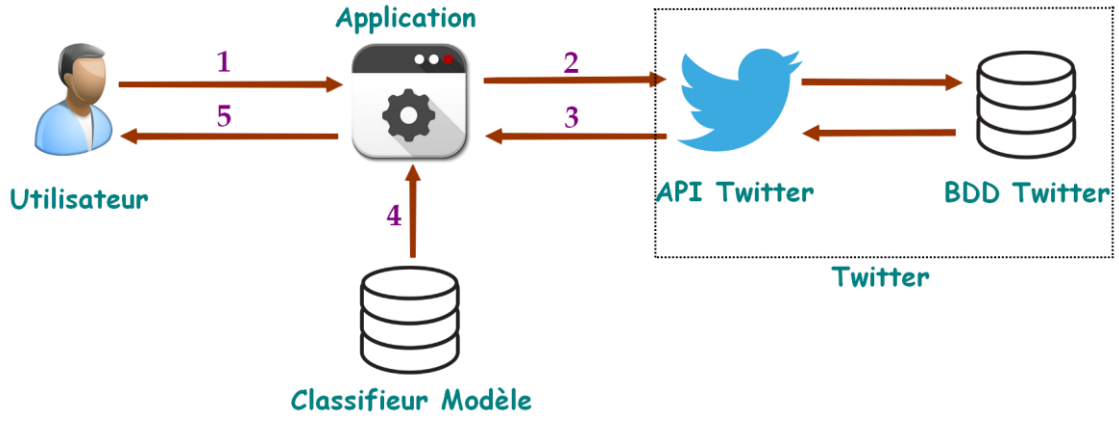

Figure 4.2 : Architecture de l'application en mode en ligne

Etape 1 : L'utilisateur effectue une recherche d'un sujet sur l'interface de l'application.

Etape 2 : Formulation de la requête pour contacter l'API Twitter.

Etape 3 : Retour des Tweets spécifique à la recherche de l'utilisateur.

Etape 4 : Utilisation du classifieur modèle généré par la phase d'apprentissage pour l'analyse des sentiments sur les tweets retournés par l'API Twitter.

Etape 5 : Affichage des résultats pour l'utilisateur.

Exemple de requête pour l'API Twitter :

```
GET

https://api.twitter.com/1.1/search/tweets.json?q=iphone&lang=en
```

q (obligatoire) : ce paramètre contiendra les mots à rechercher, limité à 500 caractères et peut inclure des opérateurs (@, #, …).

lang (optionnel) : ce paramètre nous permettra de récupérer que les tweets d'une langue spécifique.

Pour voir l'ensemble des paramètres proposés par l'API, consulter la documentation officielle de l'API Twitter [16].

b) Fonctionnement en hors ligne

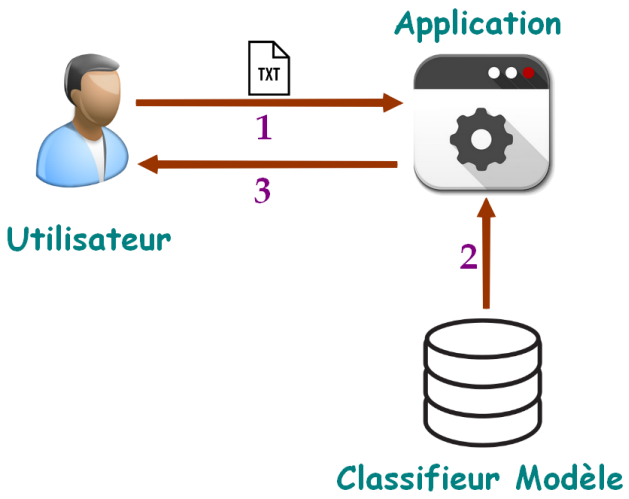

Figure 4.3 : Architecture de l'application en mode hors ligne

Etape 1 : L'utilisateur soumet un fichier de tweets à analyser.

Etape 2 : Utilisation du modèle généré par la phase d'apprentissage pour l'analyse des sentiments sur le fichier soumis par l'utilisateur

Etape 3 : Affichage des résultats pour l'utilisateur.

4.2.2.2. Fonctionnement de l'application :

Nous avons décomposé notre application en trois sous programmes et qui sont : extraction(), pretraitement() et analyse_sentiments().Ci-dessous, une brève description de ces fonctions, en montrant les entrées, les sorties, les modules utilisés et le fonctionnement de chaque sous-programme :

a) Le sous-programme extraction() :

1. Entrées : Le sujet à rechercher OU le fichier de tweets à analyser.
2. Sorties : Un ensemble de Tweets qui ont été retournés par l'API Twitter OU l'ensemble des tweets contenues dans le fichier texte.
3. Fonctionnement : Ce sous-programme permet de retourner un ensemble de Tweets OU un l'ensemble des tweets contenus dans le fichier texte à l'état brute pour être utilisé dans la prochaine étape.

b) Le sous-programme prétraitement() :

1. Entrées : L'ensemble des tweets retournés par le sous-programme précédent.
2. Sorties : Un ensemble de tweets prétraités.
3. Fonctionnement : Ce sous-programme nous permet de faire la tokenisation (retourne l'ensemble des mots contenus dans un tweet ou une phrase), l'étiquetage morpho-syntaxique (permet d'étiqueter chaque mot avec son étiquette grammaticale), permet de supprimer les mots vides (stop words) et la détection de la négation.

c) Le sous-programme analyse_sentiment() :

1. Entrées : L'ensemble des tweets prétraités retournés par le sous-programme précédent.
2. Sorties : Retourne le pourcentage et le nombre de tweets (positifs, négatifs et neutres), et les tweets classifiés.
3. Fonctionnement : Ce sous-programme utilise l'approche supervisée avec le classifieur Naïf Bayes pour la classification des sentiments.

La figure suivante représente le processus de l'analyse des sentiments :

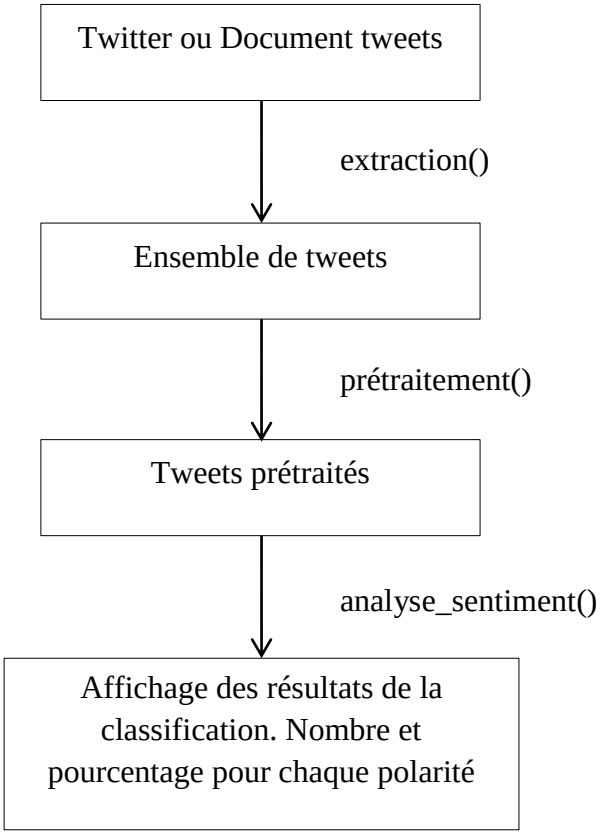

Figure 4.4 : Processus d'analyse des sentiments sur les tweets.

Utilisation du classifieur naïf bayes pour la classification d'opinions :

Ce classificateur se base sur le théorème de Bayes permettant de calculer les probabilités conditionnelles. Dans un contexte général, ce théorème fournit une façon de calculer la probabilité conditionnelle d'une cause sachant la présence d'un effet, à partir de la probabilité conditionnelle de l'effet sachant la présence de la cause ainsi que des probabilités a priori de la cause et de l'effet. On peut résumer son utilisation lorsqu'il est appliqué à la classification de textes ainsi : on cherche la classification qui maximise la probabilité d'observer les mots du document. Lors de la phase d'entraînement, le classificateur calcule les probabilités qu'un nouveau document appartienne à telle catégorie à partir de la proportion des documents d'entraînement appartenant à cette catégorie. Il calcule aussi la probabilité qu'un mot donné soit présent dans un texte, sachant que ce texte appartient à telle catégorie. Par la suite, quant un nouveau document doit être classé, on calcule les probabilités qu'il appartienne à chacune des catégories à l'aide de la règle de Bayes et des chiffres calculés à l'étape précédente.

Pour construire notre classifieur Naïf Bayes, nous nous sommes basé sur les modèles qui ont été proposées par ces deux références wébliographiques [17] [18].

Dans notre cas, chaque tweet que nous classerons contient des mots notés avec $W_{i\ (i=1..n)}$. Pour chaque mot W_i, nous pouvons extraire les probabilités suivantes (noté P).

$$\boldsymbol{P}(\boldsymbol{W_i/Positif}) = \frac{Le\ nombre\ de\ Tweets\ positifs\ avec\ le\ mot\ W_i}{Le\ nombre\ de\ Tweets\ positifs}$$

$$\boldsymbol{P}(\boldsymbol{W_i/N\acute{e}gatif}) = \frac{Le\ nombre\ de\ Tweets\ n\acute{e}gatifs\ avec\ le\ mot\ W_i}{Le\ nombre\ de\ Tweets\ n\acute{e}gatifs}$$

Pour le calcul de la probabilité d'un tweet étant positif ou négatif, nous utilisons le théorème suivant :

$$\boldsymbol{P}(\boldsymbol{Positif/Tweet}) = \frac{P(Tweet/Positif) \times P(Positif)}{P(Tweet)}$$

$$\boldsymbol{P}(\boldsymbol{N\acute{e}gatif/Tweet}) = \frac{P(Tweet/N\acute{e}gatif) \times P(N\acute{e}gatif)}{P(Tweet)}$$

Avec

$$\boldsymbol{P}(\boldsymbol{Positif}) = \frac{Le\ nombre\ de\ Tweets\ positifs}{Le\ nombre\ total\ de\ Tweets}$$

$$\boldsymbol{P}(\boldsymbol{N\acute{e}gatif}) = \frac{Le\ nombre\ de\ Tweets\ n\acute{e}gatifs}{Le\ nombre\ total\ de\ Tweets}$$

Et

$$\boldsymbol{P}(\boldsymbol{Tweet}) = 1$$

On a donc :

$$\boldsymbol{P}(\boldsymbol{Positif/Tweet}) = P(Tweet/Positif) \times P(Positif)$$
$$= P(W_1/Positif) \times P(W_2/Positif) \times .. \times P(W_i/Positif) \times P(Positif)$$

$$\boldsymbol{P}(\boldsymbol{N\acute{e}gatif/Tweet}) = P(Tweet/N\acute{e}gatif) \times P(N\acute{e}gatif)$$
$$= P(W_1/N\acute{e}gatif) \times P(W_2/N\acute{e}gatif) \times .. \times P(W_i/N\acute{e}gatif) \times P(N\acute{e}gatif)$$

A la fin on compare **P (positif / tweet)** et **P (négatif / tweet).** Si les deux probabilités se rapprochent, alors le tweet est classé comme neutre, sinon la probabilité la plus élevée décidera si le tweet est positif ou négatif.

Exemple d'application du classifieur naïf bayes :

Nous utiliserons le classifieur naïf bayes pour prédire la classe pour le tweet suivant : « Have a good day ».

Après avoir éliminé les stops words contenus dans le tweet, nous allons calculer les probabilités des mots restants « Have good day ».

Le tableau suivant représente le nombre de tweets contenant les occurrences des mots du tweet dans le corpus d'entrainement qu'on a téléchargé précédemment.

Mots	Positif	Négatif
have	309	413
good	371	183
day	703	300

Tableau 4.1 : Tableau du nombre d'occurrences des mots have, good et day.

On a :

Nombre de Tweets positifs : 7355

Nombre de Tweets négatifs : 7866

$$P(Positif) = \frac{7355}{(7355 + 7866)} = 0.4832$$

$$P(Négatif) = \frac{7866}{(7355 + 7866)} = 0.5168$$

On calcule :

$$\boldsymbol{P(Positif/Tweet)}$$
$$= P(have/positif) \times P(good/positif) \times P(day/positif) \times P(Positif)$$
$$= \left(\frac{309}{7355}\right) \times \left(\frac{371}{7355}\right) \times \left(\frac{703}{7355}\right) \times 0.4832 = 9.7874 \times 10^{-5}$$

$$\boldsymbol{P(Négatif/Tweet)}$$
$$= P(have/négatif) \times P(good/négatif) \times P(day/négatif)$$
$$\times P(négatif) = \left(\frac{413}{7866}\right) \times \left(\frac{183}{7866}\right) \times \left(\frac{300}{7866}\right) \times 0.5168 = 2.4075 \times 10^{-5}$$

On a P(Positif / Tweet) > P(Négatif/Tweet), donc le tweet est positif.

4.3. Réalisation :

4.3.1. Environnement de travail :

4.3.1.1. Généralités sur le JAVA

Le langage Java est un langage de programmation informatique orienté objet créé par James Gosling et Patrick Naughton, employés de Sun Microsystems, avec le soutien de Bill Joy (cofondateur de Sun Microsystems en 1982). La société Sun a été ensuite rachetée en 2009 par la société Oracle qui détient et maintient désormais Java [19].

Java est à la fois un langage de programmation et une plateforme d'exécution. Le langage Java a la particularité principale d'être portable sur plusieurs systèmes d'exploitation tels que Windows ou Linux. C'est la plateforme qui garantit la portabilité des applications développées en Java [20].

Le choix du langage JAVA :

Le choix de java comme langage objet est facile à expliquer. Bien que très récent, car il a été introduit par la société Sun en 1995, il connaît un engouement extraordinaire. Son immense succès s'explique par ses nombreuses caractéristiques intéressantes :

- ✓ Java est un langage de programmation orienté objet.
- ✓ Il est portable sur la plus part des plates-formes.
- ✓ C'est un langage généraliste ayant un très vaste champ d'application (réseau, base de données, calcul scientifique, etc.). Il permet de développer des applications professionnelles de grande taille.
- ✓ Il emprunte sa syntaxe, en grande partie, aux langages C et C++ qui sont les références actuelle en matière de langages professionnels. Toutefois il est beaucoup plus simple à apprendre et à utiliser. Il est facile de passer de Java aux langages C et C++ et réciproquement.
- ✓ Il intègre une interface graphique de haut niveau.
- ✓ Il est sûr, car de nombreuses vérifications sont faites, aussi bien à la compilation qu'à l'exécution, pour limiter le nombre et la gravité des erreurs. De ce fait, le développement de grosses applications en java est plus rapide qu'avec des langages comme C, C++.
- ✓ Il existe de nombreuses bibliothèques de programmes dans des domaines très variés, de sorte que ce langage devient un langage professionnel de premier plan.

- ✓ Le code produit (il s'agit d'un pseudo-code ou byte-code) est indépendant de la plateforme utilisée.

4.3.1.2. NetbeansIDE :

NetBeans est à l'origine un EDI Java. NetBeans fut développé au départ par une équipe d'étudiants à Prague (République Tchèque), racheté ensuite par Sun Microsystems. En 2002, Sun a décidé de rendre NetBeans open-source. L'IDE NetBeans est donc un environnement de développement intégré permettant d'écrire, compiler, déboguer et déployer des programmes Java. En plus de Java, il supporte différents autres langages, comme Python, C, C++, XML, Ruby, PHP et HTML[21].

Le choix de NetbeansIDE :

- ✓ Netbeans est très léger et très rapide.
- ✓ Il est open-source.
- ✓ Possède une riche communauté de développeurs de plugins.
- ✓ Support de plusieurs langages.
- ✓ Peut-être installé sur plusieurs systèmes d'exploitation supportant le JAVA.

4.3.2. Les bibliothèques JAVA utilisées :

Appelée bibliothèque logicielle (Library en Anlgais) est un ensemble de fonctions utilitaires, regroupées et mises à disposition afin de pouvoir être utilisées sans avoir à les réécrire. Les fonctions sont regroupées de par leur appartenance à un même domaine conceptuel (mathématique, graphique …etc). L'intérêt des bibliothèques réside dans le fait qu'elles contiennent du code utile que l'on ne désire pas avoir à réécrire à chaque fois [22].

Voici une description des bibliothèques utilisées dans notre application :

- **Twitter4J**

Twitter4J est une bibliothèque non-officielle de JAVA permettant d'intégrer facilement l'API de Twitter dans toute l'application JAVA, la librairie propose différentes classes et méthodes permettant de manipuler les méthodes qu'offre l'API Twitter.

Les caractéristiques de Twitter4J :

- Fonctionne sur toutes les versions JAVA Plateform 5 ou version ultérieure.
- Fonctionne sur les plateformes Android et Google App Engine.
- Zéro dépendance : Aucune autre bibliothèque n'est requise.
- Utilise le support d'authentification OAuth.

- **Apache OpenNLP**

La librairie Apache OpenNLP est une boite à outils pour le traitement automatique de textes en langage naturel, basé sur l'apprentissage machine. Elle fournit aussi un grand nombre de modèles prédéfinis pour une variété de langues, ainsi que les ressources de texte annotées.

Les outils proposés par Apache OpenNLP :

- La tokenisation.
- Détection de phrases.
- Etiquetage morpho-syntaxique (aussi appelé étiquetage grammatical, POS tagging (part-of-speech tagging) en anglais).
- Extraction des entités nommées.
- Catégorisation de documents.
- Support du classifieur Entropie Maximum pour l'apprentissage machine.

4.3.3. Utilisation de l'application :

Figure 4.5 : Interface de lancement de l'application.

La figure 4.5 s'affiche lors du lancement de l'application :

Lors du clique sur le bouton « Analyse Twitter », vous serez redirigez sur la page pour l'analyse des Tweets (Figure 4.8).

Pour le bouton « Analyse Document », vous serez redirigez sur la page pour l'analyse des documents (Figure 4.9).

- **Page Analyse Twitter :**

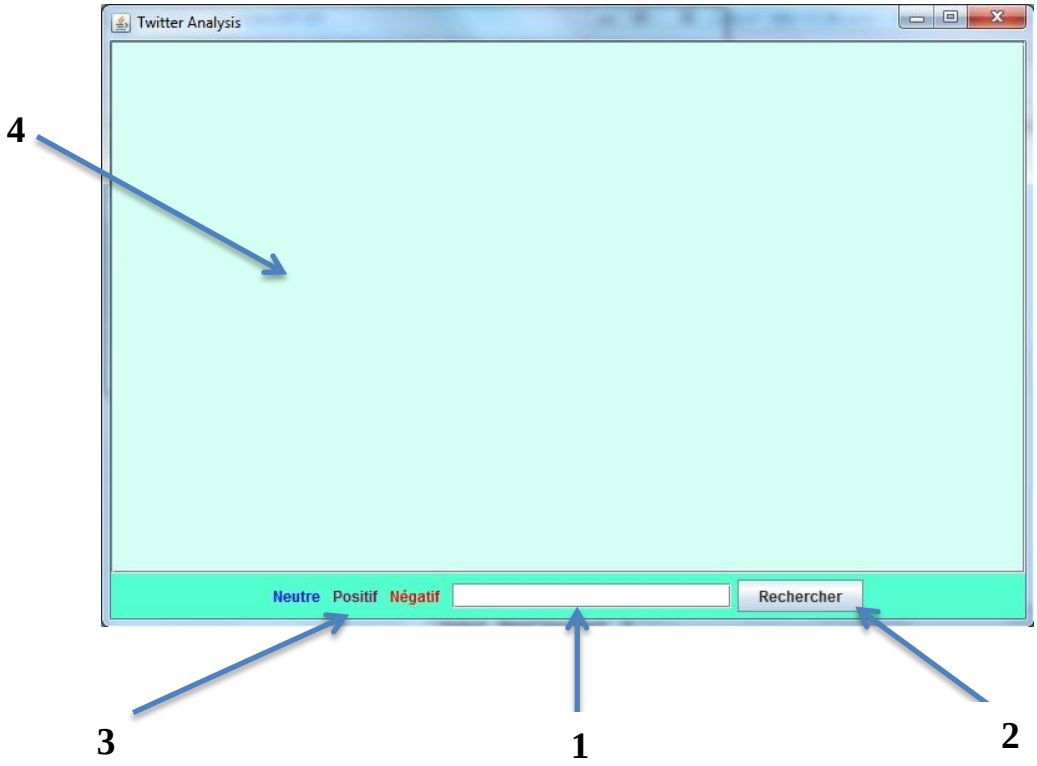

Figure 4.6 : Interface de la page Analyse Twitter

1 : Champs de texte pour rechercher les tweets.

2 : Bouton pour lancer la recherche et l'analyse des tweets.

3 : Affichage des résultats.

4 : Affichage des Tweets retournés.

- **Page Analyse Document :**

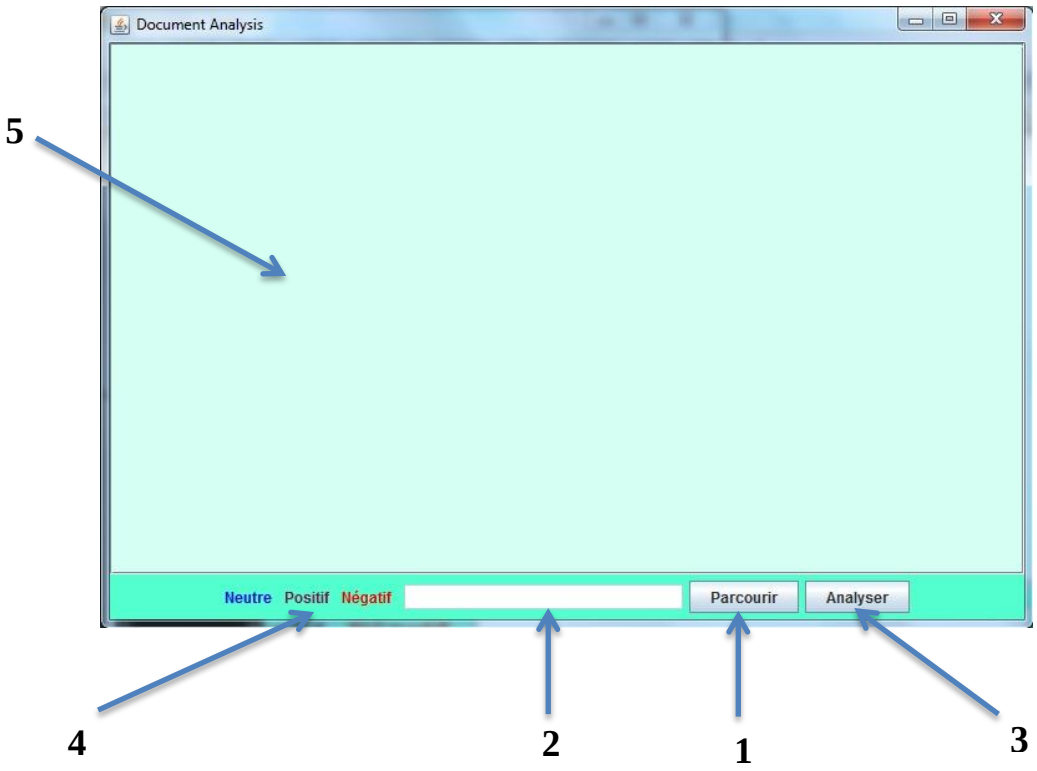

Figure 4.7 : Interface de la page Analyse Document

1 : Bouton pour parcourir et ouvrir un fichier de tweets à analyser.

2 : Affichage du chemin du fichier sélectionné.

3 : Bouton pour lancer l'analyse du document.

4 : Affichage des résultats.

5 : Affichage des Tweets classifiés du document.

- **Test**
  - **Test de l'analyse en ligne :**

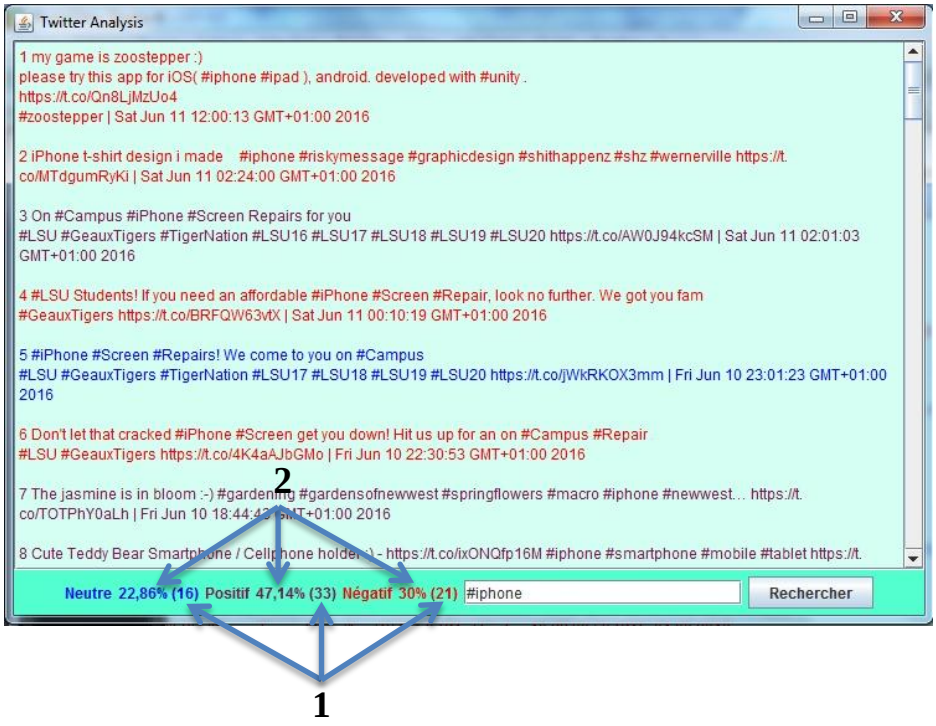

Figure 4.8 : Test de recherche pour la requête #iphone

1 : Nombre de Tweets positifs, négatifs et neutre.

2 : Pourcentage des Tweets positifs, négatifs et neutre.

- **Test de l'analyse hors ligne :**

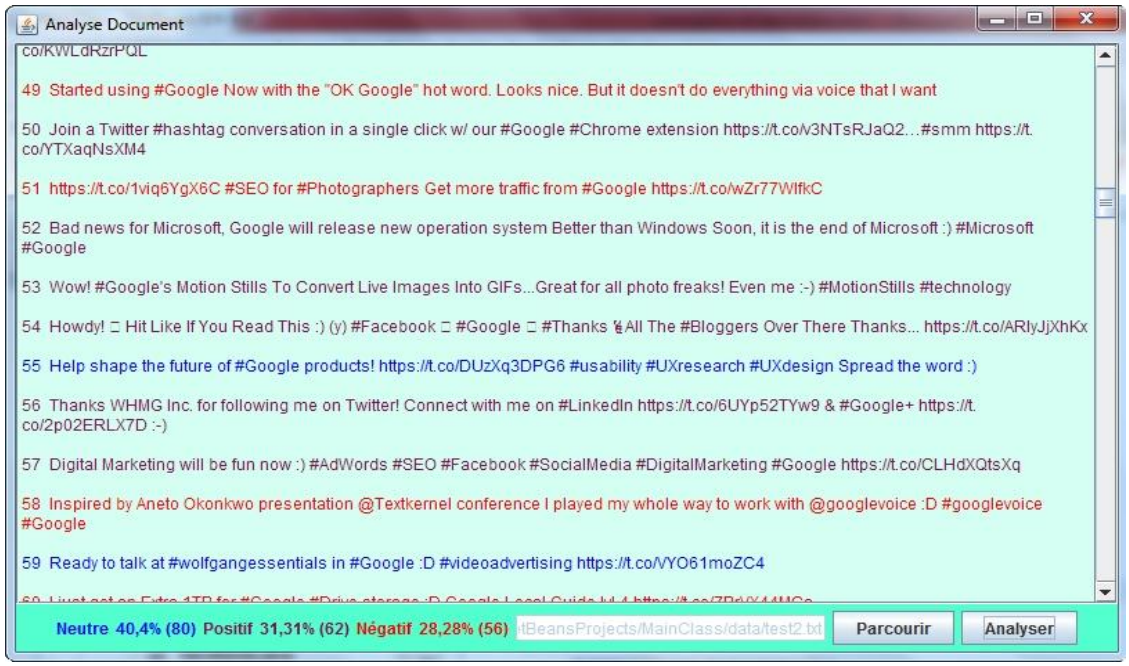

Figure 4.9 : Test de l'analyse avec un document

4.3.4. Expériences :

Nous avons effectué deux expériences pour voir l'impact des hashtags et des émoticônes sur la classification de polarité. Pour cela, nous avons téléchargé et utilisé un corpus de test qui contient 1654 tweets annotés. Ce corpus a été utilisé pour la sous-tâche B de la tâche 2 « Analyse des sentiments sur Twitter » de SemEval 2013 [23].

Un Hashtag (mot-dièse) est un marqueur de métadonnées couramment utilisé sur Internet où il permet de marquer un contenu avec un mot-clé plus ou moins partagé. Composé du signe typographique croisillon « # » (appelé hash en anglais), suivi d'un ou plusieurs mots accolés (le tag, ou étiquette), il est particulièrement utilisé sur les réseaux sociaux [24].

Exemple de hashtags positifs : #good, #cool, #fun

Exemple de hashtags négatifs : #bad, #dislike

Une émoticône est une courte figuration symbolique d'une émotion, d'un état d'esprit, d'un ressenti, d'une ambiance ou d'une intensité, utilisée dans un discours écrit. La tendance est née de la combinaison de plusieurs caractères typographiques, comme le smiley :-) [25].

Exemple d'émoticônes positives : :) :D :P ;-)

Exemple d'émoticônes négatives : :(:'(:'\

Expériences	Exactitude %
Expérience 1 : sans hashtags et émoticônes	58.1 %
Expérience 2 : avec hashtags et émoticônes	61.55 %

Tableau 4.2 : Tableau comparatif des deux expériences.

Expérience 1 : 961 Tweets / 1654 classés correctement.

Expérience 2 : 1018 Tweets / 1654 classés correctement.

Amélioration de 3.45 % lors de l'intégration des hashtags et des émoticônes.

4.4. Conclusion :

Ce chapitre nous a permis de présenter la conception et la réalisation de notre application. Ainsi, nous y avons présenté les étapes de traitements par lesquelles passe l'application. A la fin, on a testé notre application en effectuant deux expériences qui nous a montré que les hashtags et les émoticônes permettent d'améliorer l'exactitude de la classification des tweets.

Conclusion générale :

La fouille d'opinion est un domaine d'études vaste et qui se développe très rapidement. Il permet de faciliter et améliorer la vie quotidienne en analysant et en classifiant les sentiments des internautes. Au cours de notre travail, nous nous sommes intéressés à ce domaine et plus particulièrement au domaine de la fouille d'opinion sur les Tweets. Et à la fin nous avons pu concevoir et réaliser une application qui permet d'analyser les messages sur twitter basée sur une approche statistique en utilisant le classifieur naïf bayes qui est entrainé sur un corpus de tweets annotés positifs et négatifs. Ce classifieur se révèle efficace pour la tâche d'analyse des sentiments et qui est facile à intégrer. Notre application permet d'extraire des tweets, les analyser et les classifier en trois polarités (Positives, Négatives et Neutres).

Notre application présente certaines limites, qui sont :

- Apprentissage du classifieur sur deux classes seulement (Positive et négative).
- Utilise un seul classifieur pour la classification des sentiments.
- Une seule source d'information (Twitter).

De ce fait, nous envisageons comme perspectives d'amélioration de notre travail, les actions suivantes :

- Plusieurs sources d'informations.
- Intégration de la classe neutre pour l'apprentissage automatique.
- Intégrer d'autres types de classifieurs afin d'effectuer une étude comparative.

[ABM09] Asher N., Benamara F. et Mathieu Y. : *Appraisal of Opinion Expressions in Discourse.* Lingvisticæ Investigationes, 2009.

[ANABLMP15] Amine Abdaoui, Mike Donald Tapi Nzali, Jérôme Azé, Sandra Bringay, Christian Lavergne, Caroline Mollevi, Pascal Poncelet. ADVANSE : Analyse du sentiment, de l'opinion et de l'émotion sur des Tweets Français, 2015.

[AV07] Sujeevan Aseervatham, Emmanuel Viennet, *Méthodes à noyaux appliquées aux textes structurés, 2007.*

[BMCB] Faiza BELBACHIR, Malik Muhammad Saad MISSEN, Guillaume CABANAC, Mohand BOUGHANEM. *EXPÉRIMENTATION D'APPROCHES POUR LA DÉTECTION D'OPINIONS ET DE LEUR POLARITÉ DANS LES BLOGS.*

[BPR14] Caroline Brun, Diana Nicoleta, Popa Claude Roux. *Un système hybride pour l'analyse de sentiments associés aux aspects*, 2014.

[BR14] Caroline Brun, Claude Roux. *Décomposition des « hash tags » pour l'amélioration de la classification en polarité des « tweets »*, 2014.

[CH10] Carol Hermann. *Entre Web 2.0 et 3.0: opinion mining*, (2010).

[DEFT] https://deft.limsi.fr/

[DEFT15] https://deft.limsi.fr/2015/

[FB10] Faiza BELBACHIR. *Expérimentation de fonctions pour la détection d'opinions dans les blogs*. Master's thesis, Université de Toulouse, 2010.

[FD14] Firas Damak : *Etude des facteurs de pertinence dans la recherche de microblogs.* Thèse de Doctorat à l'université de Toulouse (UT3 Paul Sabatier), 2014.

[HMP+07] Hannah D., Macdonald C., Peng J., He B. et Ounis I. : University of Glasgow at TREC 2007: *Experiments in Blog and Enterprise Tracks with Terrier*. Dans TREC: Proceedings of the Text Retrieval Conference, 2007.

[JWR00] Jones K. S., Walker S. et Robertson S. E. : A probabilistic model of information retrieval: development and comparative experiments - part 1. *Inf. Process. Manage,* 2000.

[KZMAH15] Jihen Karoui, Farah Benamara Zitoune, Véronique Moriceau, Nathalie Aussenac-Gilles, Lamia Hadrich Belguith : *Détection automatique de l'ironie dans les tweets en français*, 2015.

[LCW+07] Liao X., Cao D., Wang Y., Liu W., Tan S., Xu H. et Cheng X. : Experiments in trec 2007 blog opinion task at cas-ict. Dans TREC, 2007.

[Ma14] Mahamat Atteib, Mahamat Adoum : *INTRODUCTION AUX SVM (SUPPORT VECTOR MACHINES). APPLICATIONS AUX DONNEES MEDICALES,* 2014.

[MCD] Sigrid Maurel, Paolo Curtoni et Luca Dini. *L'analyse des sentiments dans les forums.*

[MdR06] Mishne G. et de Rijke M. : A study of blog search. *Dans ECIR*, p. 289–301, 2006.

[OMS06] Ounis I., Macdonald C. et Soboroff I. : Overview of TREC-2008 Blog Track. Dans TREC:Proceedings of the Text Retrieval Conference, 2008.

[Opi05] *Opin, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*. The Association for Computationa Linguistics, 2005.

[PLV02] Pang B., Lee L. et Vaithyanathan S. : *Thumbs up*: *Sentiment Classification using Machine Learning Techniques*. Dans EMNLP'02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing, p. 79–86, Morristown, NJ, USA, 2002. ACL.

[PT02] P. Turney, *"Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews,"* in Proceedings of the Association for Computational Linguistics (ACL), 2002.

[PLV02] B. Pang, L. Lee, and S. Vaithyanathan, *"Thumbs up? Sentiment classification using machine learning techniques,"* in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002.

[PL08] Pang B. et Lee L., éditeurs. *Opinion Mining and sentiment analysis Foundations and Trends in Information Retrieval*, 2008.

[RFRN] Timothée COUR, Guillaume GIRAUD, Antoine KODSI, Tuan-Anh LUONG, Rémy LAURANSON, Clémentine MARCOVICI et Kolia SADEGHI. *Reconnaissance de formes par réseau de neurones*. Ecole Polytechnique.

[RHB] R. H. Baayen, *Word frequency distributions* vol. 18: Springer Science & Business Media, 2001.

[ROMIP] B. Dobrov, I. Kuralenok, N. Loukachevitch, I. Nekrestyanov, and I. Segalovich. Russian Information Retrieval Evaluation Seminar. In Proceedings of the Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal, May 2004.

[ROMIP11] Chetviorkin I. I., Braslavski P. I., Loukachevitch N. V**.,** *Sentiment analysis track at ROMIP* 2011.

[RRJGH12] Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, Sanda M. Harabagiu. EmpaTweet: Annotating and Detecting Emotions on Twitter. *Human Language Technology Research Institute Université du Texas Dallas*, 2012.

[SEV13] https://www.cs.york.ac.uk/semeval-2013/task2/

[SEV14] http://alt.qcri.org/semeval2014/task9/

[SEV15] http://alt.qcri.org/semeval2015/task10/

[SEV16] http://alt.qcri.org/semeval2016/task4/

[STS+07] Song R., Tang Q., Shi D., Lin H. et Yang Z. : DUTIR at TREC 2007 Blog Track. Dans TREC:Proceedings of the Text Retrieval Conference, 2007.

[WAC15] Hao Wang, Jorge A. Castanon. *Sentiment Expression via Emoticons on Social Media*. Silicon Valley Lab IBM, 2015

[WJ94] Wiebe, J. *Tracking point of view in narrative*. Computational Linguistics, 1994.

[YYZ07] Yang K., Yu N. et Zhang H. : WIDIT in TREC 2007 Blog Track: *Combining Lexicon- Based Methods to Detect Opinionated Blogs*. *Dans TREC: Proceedings of the Text Retrival Conference*, 2007.

[YH03] Yu H. et Hatzivassiloglou V. : *Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences*. Conference on Empirical Methods in Natural Language Processing *(EMNLP)*, Sapporo, Japan, July 2003.

[YS07] Yohei Seki, D Lun-wei Ku H. C.-N. k. C. L. : Overview of Opinion Analysis Pilot tas kat NTCIR-6. *Dans NTCIR-6 Workshop Meeting, Tokyo, Japan*, 2007.

[ZJB07] Zhou G., Joshi H. et Bayrak C. : Topic categorization for relevancy and opinion detection. *Dans TREC:Proceedings of the Text Retrieval Conference*, 2007.

[1] Wikipedia – Microblog : https://fr.wikipedia.org/wiki/Microblog

[2] Wikipedia – Twitter : https://fr.wikipedia.org/wiki/Twitter

[3] Wikipedia – Tumblr : https://fr.wikipedia.org/wiki/Tumblr

[4] Wikipadia – Tencent Weibo : https://fr.wikipedia.org/wiki/Tencent_Weibo

[5] Wikipedia – Identi.ca : https://fr.wikipedia.org/wiki/Identi.ca

[6] Wikipedia – Pinterest : https://fr.wikipedia.org/wiki/Pinterest

[7] Wikipedia – SEMEVAL : https://en.wikipedia.org/wiki/SemEval

[8] Wikipedia – Naïf Bayes : https://fr.wikipedia.org/wiki/Classification_na%C3%AFve_bay%C3%A9sienne

[9] 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/

[10] 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

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

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

[13] 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

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

[15] Insitut Max-Planck d'informatique : http://people.mpi-inf.mpg.de/~smukherjee/data/

[16] API Twitter –Search/tweets : https://dev.twitter.com/rest/reference/get/search/tweets

[17] http://cloudacademy.com/blog/naive-bayes-classifier/

[18] http://technobium.com/sentiment-analysis-using-mahout-naive-bayes/

[19] Wikipedia – JAVA : http://fr.wikipedia.org/wiki/Java_(langage)

[20] All2All – JAVA : http://www.all2all.org/fr/informations/glossary/Java/

[21] Developpez.com – Netbeans : http://netbeans.developpez.com/faq/?page=Introduction

[22] Techno-Science – Bibliothèque Logicielle : http://www.techno-science.net/?onglet=glossaire&definition=1470

[23] SEMEVAL 2013 – Données Tâche 2 : https://www.cs.york.ac.uk/semeval-2013/task2/index.php%3Fid=data.html

[24] Wikipedia – Hashtag : https://fr.wikipedia.org/wiki/Hashtag

[25] Wikipedia – Émoticône : https://fr.wikipedia.org/wiki/Emoticone

ANNEXE

Nous donnons, dans cette annexe, les définitions de quelques termes utilisés dans ce mémoire :

1. **API** : Une API (Application Programmable Interface en anglais, « interface de programmation » ou « interface pour l'accès programmé aux applications) est un ensemble de fonctions permettant d'accéder aux services d'une application, par l'intermédiaire d'un langage de programmation. Elle permet de fournir un certain niveau d'abstraction au développeur, c'est-à-dire qu'elle lui masque la complexité de l'accès à un système ou à une application en proposant un jeu de fonctions standard dont seuls les paramètres et les valeurs retournées sont connus.

2. **Corpus** : ensemble limité d'éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique, ensemble de textes réunis à des fins de comparaison, servant de base à une étude quantitative.

3. ***n-gramme*** : Succession de N éléments de même type extraits d'un texte, d'une séquence ou d'un signal, les éléments pouvant notamment être des mots ou des lettres. Les N-grammes sont beaucoup utilisés en traitement automatique du langage naturel mais aussi en traitement du signal.

4. **Idiome** : Chaque langue possède ses propres expressions idiomatiques. Il s'agit d'expressions qui ne peuvent être prises mot pour mot. Même en sachant la signification de chaque mot qui la compose, il demeure que la signification de l'expression idiomatique est très différente. Par exemple: "The red car caught my eye." Nous savons parfaitement qu'une voiture ne peut attraper quelque chose et qu'un œil ne peut être lancé. Nous devons comprendre la signification de l'idiome "caught my eye" afin de comprendre ce qui est dit.

5. **OpinionFinder** : est un système qui traite les documents et identifie les phrases subjectives, ainsi que divers aspects de la subjectivité dans les phrases, y compris les agents qui sont sources d'opinion, les expressions subjectives directs et des événements de la parole, et les expressions de sentiment automatiquement. OpinionFinder a été développé par des chercheurs de l'Université de Pittsburgh, l'Université Cornell, et l'Université de l'Utah.

6. **Pondération** : Relation entre des poids ou des puissances qui s'équilibrent mutuellement, Balancement des masses, équilibre des figures, Juste équilibre; Caractère de ce qui est pondéré, bien équilibré.

7. **SentiWordNet** : est une ressource lexicale pour l'extraction d'opinion basée sur WordNet. SentiWordNet attribue à chaque synset de WordNet trois scores de sentiment : la positivité, la négativité et l'objectivité.

8. **Sac de mots (bag of words en Anglais)** : un document particulier est représenté par l'histogramme des occurrences des mots le composant: pour un document donné,

chaque mot se voit affecté le nombre de fois qu'il apparaît dans le document (voir la notion de multi-ensemble, bag en anglais). Un document est donc représenté par un vecteur de la même taille que le dictionnaire, dont la composante i indique le nombre d'occurrences du i-ème mot du dictionnaire dans le document.

9. **Tumblelog** : est un blog qui prend en charge la forme abrégée du contenu multimédia mixte. Les messages de Tumblelog sont des brefs commentaires qui peuvent contenir de l'audio, des citations, des liens, des images ou des vidéos. Parce que le mot tumblelog est si étroitement associé au site de réseau social gratuit Tumblr, le terme a généralement été remplacé dans les médias par son synonyme, microblog.

10. **Traits de sentiment** : ce sont les traits qui indiquent la présence de mots ou d'expressions d'opinion positive ou négative.

11. **Traits pour les modifieurs de sentiment** : ils regroupent deux nouveaux traits qui indiquent si un tweet contient un mot d'opinion dans la portée d'une modalité ou d'un adverbe d'intensité. Les traits pour les modifieurs vérifient aussi si un tweet contient : un intensifieur, une modalité, un mot de négation ou un verbe de discours rapporté.

12. **Traits de surface** : le premier est la longueur du tweet en nombre de mots. Les autres sont tous binaires et indiquent la présence ou non de ponctuation, mots en lettres majuscules, interjections, émoticônes, citation, argot, mots d'opposition tels que "mais" et "bien que" séquence de points d'exclamation ou d'interrogation, combinaison de points d'exclamation et d'interrogation.

13. **Traits d'opposition** : ils indiquent la présence d'opposition explicite grâce à des patrons lexico-syntaxiques spécifiques. Ces traits ont été partiellement inspirés de (Riloff et al., 2013) qui a proposé une méthode par bootstrapping pour détecter les tweets sarcastiques correspondant à une opposition entre un sentiment/opinion positif et une situation négative.

14. **Traits de contexte** : ces traits indiquent la présence/absence d'éléments de contexte tels que les pronoms personnels, les mots-clés d'un thème donné et les entités nommées donnés par l'analyseur syntaxique. Par exemple, l'ironie dans le tweet « Elle nous avait manqué » ! est difficile à détecter car il ne contient pas d'élément contextuel.

15. **Z-score** : est une mesure numérique de la relation entre une valeur à la moyenne dans un groupe de valeurs. Si un Z-score est de 0, le score est identique à la note moyenne. Le Z-scores peut également être positif ou négatif, avec une valeur positive indiquant le score est supérieur à la moyenne et un score négatif indiquant qu'il est en dessous de la moyenne.