République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de Djilali BOUNAAMA Khemis Miliana

Faculté des Sciences et de la Technologie Département de Mathématiques et d'Informatique

Mémoire Présenté

Pour l'obtention d'un diplôme de

Master en Informatique

Option : Ingénierie du logiciel

Thème

Détection de points de vue sur Twitter

Réalisé par :

- RALEM AHMED IBRAHIM
- EZZIANE ABDELHAMID

Encadré par :

Mr. Ilyes mohammed bouziane

Soutenu devant le jury composé de :

Président
Examinateur 1
Examinateur 2

Année Universitaire 2019/2020

Remerciements

On dit souvent que le trajet est aussi important que la destination. Cinq années nous avons permis de comprendre la signification de cette phrase toute simple. Ce parcours ne s'est pas réalisé sans défis et sans soulever de nombreuses questions pour lesquelles les réponses ont suscité de longues heures de travail.

Nous souhaitons vivement que nos parents qui nous ont épaulés moralement, puissent trouver dans ces lignes l'expression de notre profonde gratitude.

Nos remerciements s'étendent à notre promoteur ***Mr. Ilyes mohammed bouziane*** *pour sa patience, sa disponibilité et surtout pour ses judicieux conseils, qui ont contribué à alimenter nos réflexions*

Nous tenons à remercier chacun des membres du jury pour nous avoir fait l'honneur d'examiner et d'évaluer notre travail et de l'enrichir par leurs propositions.

Nous ne saurons épuiser ces remerciements sans gratifier nos enseignants qui ont su nous donner une formation appréciable durant tout notre cursus.

Dédicaces

Je tiens à remercier en premier lieu Dieu le tout Puissant qui nous a donné les capacités physiques et morales ainsi que le courage et la santé pour la réalisation de ce travail.
Je dédie ce travail à :

***A** mes chers parents qui m'ont beaucoup encouragé.*

***A** mon frère abdelkader .*

***A** mon ami Ezziane abdelhamid.*

***A** mon ami younes en particulier qui m'a beaucoup encouragé.*

***A** mes camarades du cycle master.*

***A** tous mes amis (es).*

IBRAHIM

Dédicaces

***A** mes chers parents*

***A** ma mère en particulier qui m'a beaucoup encouragée.*

***A** ma petite famille.*

***A** mes frères et sœurs*

***A** toute ma famille sans exception*

***A** tous mes amis et camarades de la promotion GLSD.*

***A** tous mes amis (es).*

ABDELHAMID

Sommaire

Chapitre 03: Conception

Chapitre 04: Implémentation

Liste des tableaux

Liste des figures

ملخص

أصبحت المدونات الصغيرة الآن أداة اتصال شائعة جدًا مع مستخدمي الإنترنت ، حيث يشارك ملايين المستخدمين الآراء حول جوانب مختلفة من الحياة اليومية.

لذلك ، تعد مواقع المدونات الصغيرة مصادر غنية بالبيانات لاستكشاف الرأي وتحليل المشاعر ، نظرًا لأن المدونات الصغيرة ظهرت مؤخرًا نسبيًا ، فقد تم تخصيص بعض الأبحاث لهذا الموضوع.

في مشروعنا ، نركز على استخدام تويتر ، منصة التدوين المصغر الأكثر شيوعًا لتحليل المشاعر ، نستخدم مجموعة لتحليل المشاعر واستخراج الرأي.

نحن نجري تحليلاً لغويًا لهذه المجموعة ونقوم ببناء مصنف لتصنيف المشاعر، قادر على تحديد المشاعر الإيجابية والسلبية لأي وثيقة أو مجموعة أخرى.

في بحثنا ، عملنا مع مجموعة من اللغة الإنجليزية ، ومع ذلك ، يمكن استخدام التقنية المقترحة مع أي لغة أخرى.

الكلمات المفتاحية : المدونات الصغيرة ,الآراء, وجهات النظر ,مجموعة

Résumé

Le microblogging est devenu aujourd'hui un outil de communication très populaire auprès des internautes, des millions d'utilisateurs partagent des opinions sur différents aspects de la vie de tous les jours.

Par conséquent, les sites Web de microblogage sont des sources riches de données pour l'exploration d'opinion et l'analyse des sentiments, parce que le microblogging est apparu relativement récemment, il y a quelques travaux de recherche qui ont été consacrés à ce sujet. Dans notre projet, nous nous concentrons sur l'utilisation de Twitter, la plateforme de micro-blogging la plus populaire, pour la détection de points de vue, nous utilisons un corpus à des fins d'analyse des sentiments et d'extraction de points de vue.

Nous effectuons une analyse linguistique de ce corpus et nous construisons un classificateur de sentiment, qui est capable de déterminer la polarité positifs, négatifs pour n'importe quel autre document ou corpus.

Dans notre recherche, nous avons travaillé avec un corpus de langue anglaise cependant, la technique proposée peut être utilisée avec n'importe quelle autre langue.

Mots clés : microblogging, opinions, points de vue, un corpus

Abstract

Microblogging has now become a very popular communication tool with Internet users, millions of users share opinions about different aspects of everyday life.

Therefore, microblogging websites are rich sources of data for opinion exploration and sentiment analysis, because microblogging has appeared relatively recently, there is some research that has been devoted to this subject. In our brief, we focus on using Twitter, the most popular micro-blogging platform, for sentiment analysis, we use a corpus for sentiment analysis and opinion extraction.

We perform a linguistic analysis of this corpus and we build a feeling classifier which is able to determine positive, negative feelings for any other document or corpus.

In our research, we worked with a corpus of English language, however, the proposed technique can be used with any other language.

Keywords: microblogging, opinions, points of view, a corpus

Introduction générale

Avec l'avènement du web et l'explosion des sources des données telles que les sites d'avis, les blogs et les microblogs est apparu la nécessité d'analyser des millions des postes de tweets ou d'avis afin de savoir ce que pensent les internautes. L'analyse des sentiments est une technologie d'analyse automatique des discours, écrits ou parlés et d'en faire ressortir les différentes opinions exprimées sur un sujet précis comme une marque, une actualité ou un produit.

L'importance de l'analyse des sentiments est présente dans plusieurs domaines a savoir politique, marketing, gestion de la réputation...ect, l'analyse des sentiments relève de plusieurs disciplines en l'occurrence d'une part du traitement automatique du la langage naturel (Naturel Langage Procession) et d'autre part de l'apprentissage automatique (Machine Learning).

Dans ce mémoire, notre objectif consiste à dévoiler les secrets de l'analyse des sentiments et la détection de points de vue en adoptant une approche d'apprentissage automatique. Pour ce faire, nous avons implémenté la méthode probabiliste Naive Bayes sur le corpus des tweets, le reste du mémoire est organisé en quatre chapitres , nous consacrons un premier chapitre a présenter des généralités sur les réseaux sociaux et spécialement Twitter comme source d'opinions, le second chapitre se focalise sur l'état de l'art de l'analyse des sentiments, et les méthodes de détection de point de vue.

Notre troisième chapitre présente l'expérimentation et la méthode Naive Bayes en considérant les phases d'apprentissage et de test.

Nous concluons avec une description et définition de l'outillage utilisé a l'implémentation de notre système et les étapes d'implémentation et finalement les résultats obtenus.

Chapitre 01:Généralités sur les réseaux sociaux

1.1 Introduction

Les réseaux sociaux sont omniprésents depuis l'apparition d'Internet. Ils permettent aux différents utilisateurs d'interagir en communauté et de se regrouper selon des critères qui leur sont importants. ces réseaux sociaux sont de différents types.

Certains sont connus de tous (*ex. Facebook*[1], *Twitter*[2], *LinkedIn*[3]) et comptent des millions de membres.

D'autres exploitent des niches moins connus et peuvent passer relativement inaperçus ou rester confidentiels, tels les réseaux d'entreprises.

Dans ce chapitre, nous allons présenter les réseaux sociaux, et nous expliquons les types de ces réseaux, leurs impacts sur l'utilisateur et leurs fonctionnements.

1http://www.facebook.com

2 http://www.twitter.com

3http://www.linkedin.com

1.2 Historique

Les moyens de communication développée au fils des ans à apporter sont lot de changement et d'accélération de l'information, toutes les nouveautés technologiques ont eu une réception mitigée lors de leur dévoilement, voici une ligne du temps de l'évolution des communications électroniques.**[1]**

- 1792 : Télégraphe (rapidité d'accès à l'information)
- 1890 : Téléphone (Plusieurs personnes sur une même ligne … on espionne les voisins)
- 1891 : Radio (Nouvelles dictées par le marketing, ligne ouverte avec des personnes opinionnées et des gens qui s'offusquent facilement)
- 1945 : CB et radio amateur (Réglementation très tôt après son invention pour éviter les abus et les mauvaises blagues)
- 1966: 1er courriel envoyé (Ajout de la barrière de l'écran, l'être humain n'a plus accès au ton et au non-verbal qui sont utilisés pour véhiculer le message, début de la période d'interprétation des propos)
- 1995: Ebay (La légitimité du vendeur dépend de la note attribuée par les utilisateurs de la plateforme)
- 1996: ICQ (Communication en direct avec d'autres utilisateurs, possibilité de voir s'ils sont connectés ou non au logiciel)
- 1999: Blogger et Napster (Début du journal personnel et du partage de documents par protocole sécurisé)
- 2000: Trip Advisor (Les usagers donnent une note à l'hôtel, ils peuvent la faire passer ou la casser)
- 2001: Wikipedia (Encyclopédie en ligne écrite par des experts du milieu et éditée par des gens qui suivent l'actualité)
- 2002: LinkedIn (Instauration d'une stratégie de positionnement professionnel et de développement d'expertise)
- 2003: Facebook et Skype (Communication rapide et grand public; on est connecté en tout temps avec le monde en entier)
- 2005: YouTube (Journal personnel en vidéo)
- 2006: Twitter (Courts articles, choix de mots restreints, communication restreinte, les propos doivent être courts et directs)
- 2011: Pinterest (Babillard des objets de ses rêves)
- 2013: SnapChat (Journal personnel éphémère en photos et vidéos).

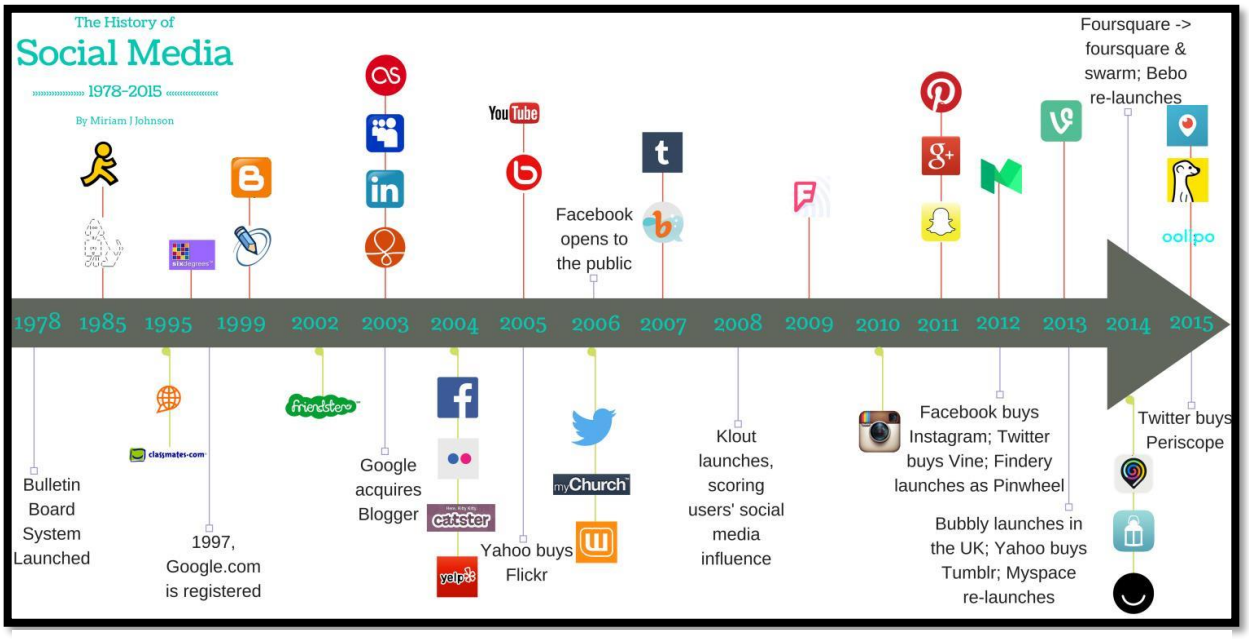

Figure 1-1: Enchainement des réseaux sociaux 1978-2015 [Web1].

1.3 Définition

Dans la littérature, il existe plusieurs définitions de réseaux sociaux, nous présentons quelques-unes :

Un réseau social est constitué à la fois par un ensemble de personnes liées entre elles et par la force de ces liens, on peut aussi dire qu'un réseau social est un ensemble d'individus liés entre eux par des liens caractérisés par un degré de familiarité variable qui va de simple connaissance aux liens familiaux les plus étroits **[2]**.

La figure représente un exemple de réseau social.

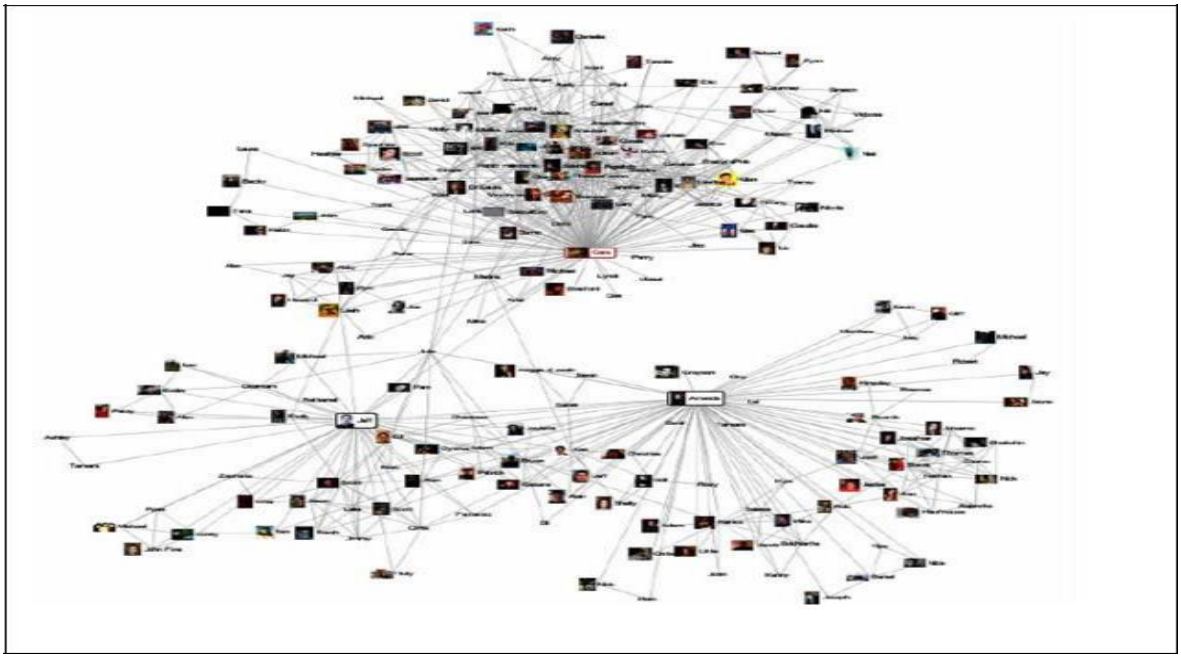

***Figure 1-2*:** Représentation de réseau social [3].

Pour Yahoo (acteur dans le domaine des réseaux sociaux avec son Yahoo!360°), un réseau social est « un terme assez large qui désigne des sites Internet qui aident leurs utilisateurs à créer leur propre profil Internet et partager une partie de leurs contenus préférés, y compris des photos et de la musique » **[4]**.

Une autre définition dans **[5] :** Le réseautage social (distinct du concept de réseau sociale n sociologies) se rapporte à une catégorie des applications d'internet pour aider à relier des amis des associées ou d'autres individus employant ensemble une variété d'outils.
Ces applications connues sous le nom de « service de réseautage social en ligne » (en anglais Social Networking) deviennent de plus en plus populaires, elles peuvent aussi permettre une meilleure distribution artistique, en favorisant la formation de contacts, et en invitant des artistes à assurer une visibilité de leur travail (ex. musique, vidéo, photographie).

1.4 Différent types de réseaux sociaux

1.4.1 Les réseaux sociaux dits généraliste

C'est les réseaux sociaux les plus connus et les plus utilisés dans le monde entier. Donc c'est des réseaux qui ont un grand public, parmi eux on a :

- **Facebook :** Chaque internaute a la possibilité de créer son profil limité à un réseau d'amis (personnes proches ou inconnues) qu'il a accepté, il permet de partager : statut, photos liens et vidéos, il est aussi utilisé par les entreprises, les artistes pour leur promotion grâce aux pages fans accessibles à tous. Leader dans le monde.
- **Twtter :** outil de microblogging qui permet d'envoyer des messages appelés « tweets » aux internautes qui suivent chaque compte, ce sont les « followers » soient les abonnés.
- **MySpace :** espace web personnalisé, possibilité de présenter des informations personnelles et de faire un blog, ce réseaux est notamment connut grâce aux nombreux groupes musicaux qui ont pris possession de cet espace, sa popularité a baissé ces dernières années.

D'autres réseaux généralistes moins connut : **[6]**

- **Beboomer** : pour les plus de 45 ans.
- **Cafemom** : pour les mamans.
- **Kaboose** : plus accès pour la famille.

1.4.2 Les réseaux sociaux dits professionnelle

- **LinkedIn**: c'est un réseau professionnel, il permet de publier et partager son CV.

- **Viadeo**: il permet d'établir des contacts professionnels, vous faire connaître, en publiant son CV, il propose également des offres d'emploi.

- **Ziki** : il a pour but d'aider les entreprises à trouver le meilleur prestataire de service pour la réalisation d'un projet. **[6]**
- **InterFrench :** réseau francophone mondial pour les projets à l'étranger.
- **Piwie :** le premier chat business.

1.4.3 Les réseaux sociaux dits de service

- **Ma-résidence :** lieu d'échange de bonnes adresses, de services et parler de vos relations entre voisins.
- **Copains d'avant et Trombi :** qui permettent de retrouver des anciens camarades de classes.
- **Réseaux Lycée et Etnoka :** réseaux pour lycéens et étudiants où il est possible de discuter, organiser des soirées et le partage de cours.
- **BeGlob :** dédié aux passionnés de voyages, permet d'échanger les bons plans, conseils expérience…etc. **[6].**

1.4.4 Les réseaux sociaux dits politiques

- **Coolpol :** c'est le réseau social (application française) du parti socialiste de « toutes celles et de tous ceux qui veulent débattre et agir à gauche! » selon le site. c'est un lieu de discussion ou les sympathisants du parti peuvent échanger. on y retrouve les événements débats, partage d'idées de liens, de vidéos…

- **Créateurs de possible**: c'est les réseaux sociaux de l'UMP (Union pour un mouvement populaire) lancé en janvier 2010, propose des fonctionnalités similaires à Coolpol **[6].**

1.5 Impact des réseaux sociaux sur la vie de leurs utilisateurs

Les médias sociaux numériques se sont profondément ancrés dans la vie de leurs utilisateurs et utilisatrices cela comprend bien évidemment les plateformes incontournables telles que Facebook ou Twitter, mais également tous les sites permettant aux utilisateurs de réagir sur leur contenu, en effet, avec l'évolution des technologies liées au Web 2.0, de nombreux sites ont intégré des fonctionnalités sociales, quel que soit leur domaine, peu à peu, les utilisateurs ont eu la possibilité de noter une recette de cuisine sur Marmiton, de partager un livre depuis Amazon ou de commenter un article du monde, avec la croissance rapide des contenus générés par les utilisateurs, de nombreux domaines ont évolué pour tenter de tirer avantage de cette surabondance de données, le nombre de travaux traitant de médias sociaux ou réseaux sociaux a presque décuplé tous les 10 ans depuis 1950 d'après la bibliothèque numérique de l'association for Computing Machinery (ACM) comme indiqué dans la figure suivante.

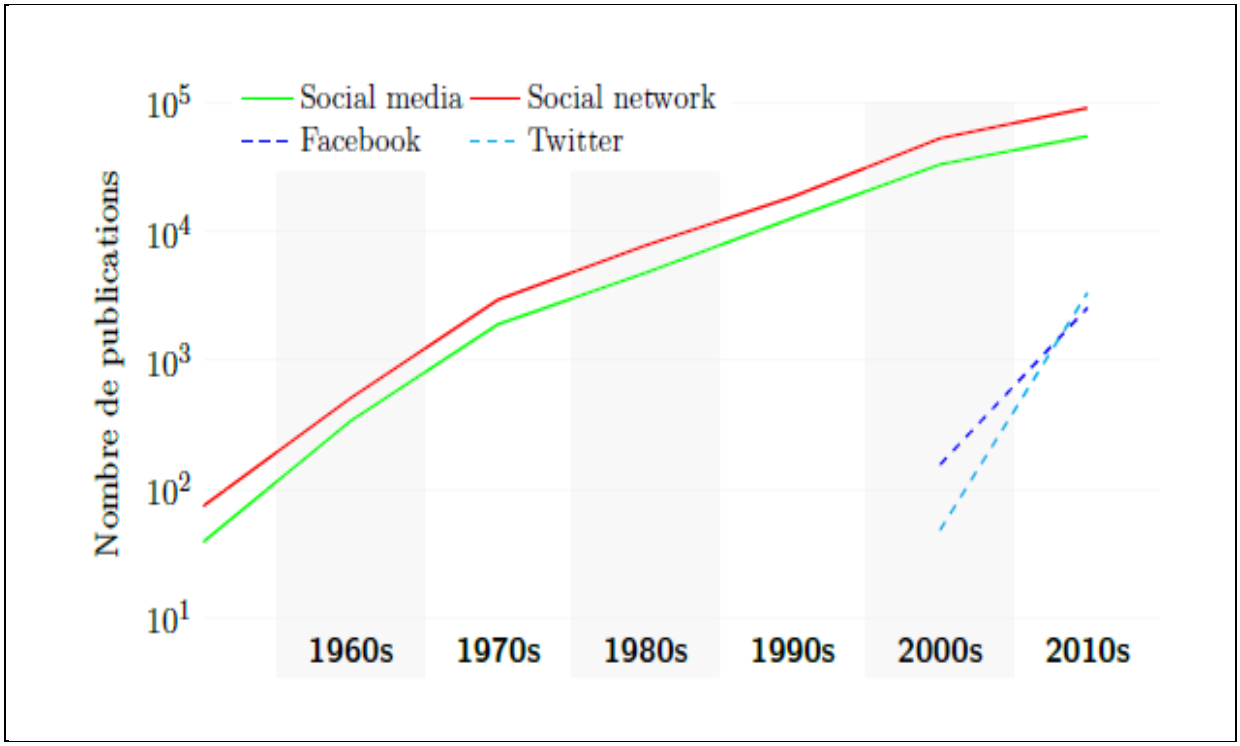

Figure 1-3 : Nombre de publications mentionnant les termes « Social media » « Social network » « Facebook » ou « Twitter » référencés dans la bibliothèque numérique de l'Association for Computing Machinery(ACM).

Dans l'ensemble, les médias sociaux numériques sont devenus, ces dernières années un matériau incontournable de recherches pour les chercheurs en informatique comme en sciences humaines, comme l'illustre la présentation de Twitter par Golder et Macy (2015) :

« Twitter has emerged as the single most powerful "socioscope" available [. . .] for collecting fine-grained time-stamped records of human behavior and social interaction»

Le marketing a tenté de créer des indicateurs pour le marché boursier, de prédire le cours des actions d'estimer la valeur d'organisations caritatives ou de mesurer l'impact de campagnes médiatiques.

D'autres applications se sont révélées utiles pour le cyber sécurité, telles que la découverte de connections entre utilisateurs en temps réel ou encore les mesures de fiabilité.

Les médias sociaux ont également été particulièrement utilisés pour étudier des élections ou référendums, des campagnes de dénigrement et de « fake news» de nombreuses applications mentionnées ci-dessus tirent parti de la fouille d'opinions, qui permet d'automatiquement détecter les états subjectifs partagés par les utilisateurs (Liu, 2012).

Une grande partie de la littérature sur ce sujet se concentre sur les critiques laissées par les utilisateurs sur des plateformes numériques, par exemple des critiques de produits sur Amazon de films sur IMDb, de restaurants sur Yelp ou encore d'hôtels sur TripAdvisor.**[7]**

1.6 Fonctions et fonctionnalités

Le fonctionnement de ces ites est toujours le même, pour créer sa page d'accueil on demande au nouvel arrivant de définir un profil qui constituera la représentation qu'il a de lui-même sous la forme de textes, de photos, de vidéos, de musiques et de liens, des questionnaires lui imposent parfois de préciser ses goûts, ses affinités, ses opinions, etc. Facebook permet aux utilisateurs d'enrichir leur profil avec toutes sortes d'applications, les listes d' « amis » font le lien entre les profils, chaque participant au site peut demander à d'autre participants au même site de faire partie de ses « amis » ; si sa proposition est acceptée, le nom et une photo (ou un avatar) de l' « amis » sont affichés sur le site de l'autre, par « contagion » les participants du site étendent ainsi continuellement leur liste « amis », mais qui dit « amis » sur Facebook ne dit pas forcément camarade dans la vraie vie **[8].**

La liste d' « amis » ne se résume pas à l'exposition des liens forts d'amitié qui ont une réalité dans la « vraie vie », elle permet aussi de mesurer l'audience que chaque participant a auprès d'autres participants, souvent inconnus dans un premier temps, qui partagent les mêmes activités, goûts, opinions, etc. ces liens sont dits faibles, deux pratiques différentes tournent autour de cette distinction : l'accès d'un profil peut être restreint aux « amis », liens forts ou faibles, ou au contraire ouvert à tous, pour faciliter la croissance de la liste.

La recherche à tout prix de nouveaux « amis » peut conduire à accepter des inconnus dans la liste, sans vérifier leur identité, ni s'ils sont véritablement des « amis » d' « amis », c'est un des moyens couramment utilisés pour atteindre les données protégées d'un compte.

La troisième par ordre d'importance est la fonction de revue publique **«Témoignages» « Commentaires », « Panneau d'affichages » [9],** cette fonctionnalité essentielle des réseaux sociaux permet à tout visiteur de déposer des **commentaires** sur une page d'accueil (la sienne ou celle de quelqu'un d'autre), ces commentaires sont lisibles par toutes les personnes autorisées à accéder au profil (Mur sur Facebook), toutes ces possibilités permettent une médiatisation de soi via un profil détaillé et des contenus diversifiés (statuts, photos, liens) ainsi qu'une socialisation active avec les membres de son entourage au moyen des commentaires.

La plupart des sites sociaux propose l'échange de messages privés entre participants. Ils peuvent aussi offrir des services de partage de photos, de vidéos, et offrir des interactions avec les téléphones mobiles, ces trois fonctionnalités - profils, listes d'amis, commentaires constituent la structure de base des réseaux sociaux numériques, même si certains d'entre eux proposent des fonctions complémentaires, les réseaux sociaux numériques permettent à leurs visiteurs de naviguer d'un « Ami » à l'autre et de correspondre avec quiconque possède un profil visible **[9].**

Les réseaux sociaux numériques permettent la navigation d'un profil à l'autre et rendent accessible la communication avec tout participant, le mode d'accès le plus répandu est cependant basé sur l'existence de groupes d' « amis » préexistants, ayant rejoint le site ensemble pour conserver ensuite le contact entre eux (élèves d'une classe, d'une école, etc.) et pouvoir communiquer deux à deux tout en faisant partie d'un groupe **[8].**

1.7 Avantages et inconvénients des réseaux sociaux

Le tableau suivant décrit les avantages et les inconvénients de quelque réseau social :

Nom	Facebook	Twitter	Linkedin
Avantages	- Créer une page Facebook simple et la plateforme offre de plus en plus d'outils qui permettent notamment aux entreprises et marques de suivre la progression de leur nombre d'adaptes ainsi que leur Données démographiques.	- On peut s'informer en temps réel et faire la même chose pour diffuser une Information importante et pertinente. Il n'y a aucun problème à publier fréquemment puisque les tweets se suivent rapidement.	- C'est un réseau social professionnel par excellence, qui permet de bâtir un réseau à partir de ses expériences de travail mais également autour de ses intérêts et compétences.
Inconvénients	- Facebook demande une interaction avec les adeptes.On ne peut pas diffuser une information sur Facebook et ignorer la réaction des adeptes.	- Sur Twitter, on ne peut pas envoyer un message qui contient plus de 140 caractères. Twitter oblige les entreprises à être concises et claires.	- L'interaction sur Linkedin, à part dans certains groupes, est vraiment limitée. Une interface un peu moins accessible ne rend pas les choses faciles.

***Tableau 1-1*:** Tableau comparatif entre les réseaux sociaux

1.8 La plate-forme de microblogging Twitter

1.8.1 Définition

Twitter est un réseau social de microblogage géré par l'entreprise Twitter Inc, il permet à un utilisateur d'envoyer gratuitement de brefs messages, appelés tweets, sur internet, par messagerie instantanée ou par SMS.

À l'origine, les tweets étaient limités à 140 caractères, mais le 7 novembre 2017, cette limite était doublée (280 caractères) pour toutes les langues sauf le japonais, le coréen et le chinois **[10].**

twitter : c'est une sorte de micro blogging. Il permet de suivre l'actualité de l'entourage ou même de personnalité ainsi que d'échanger des avis et des « commentaires » **[11].**

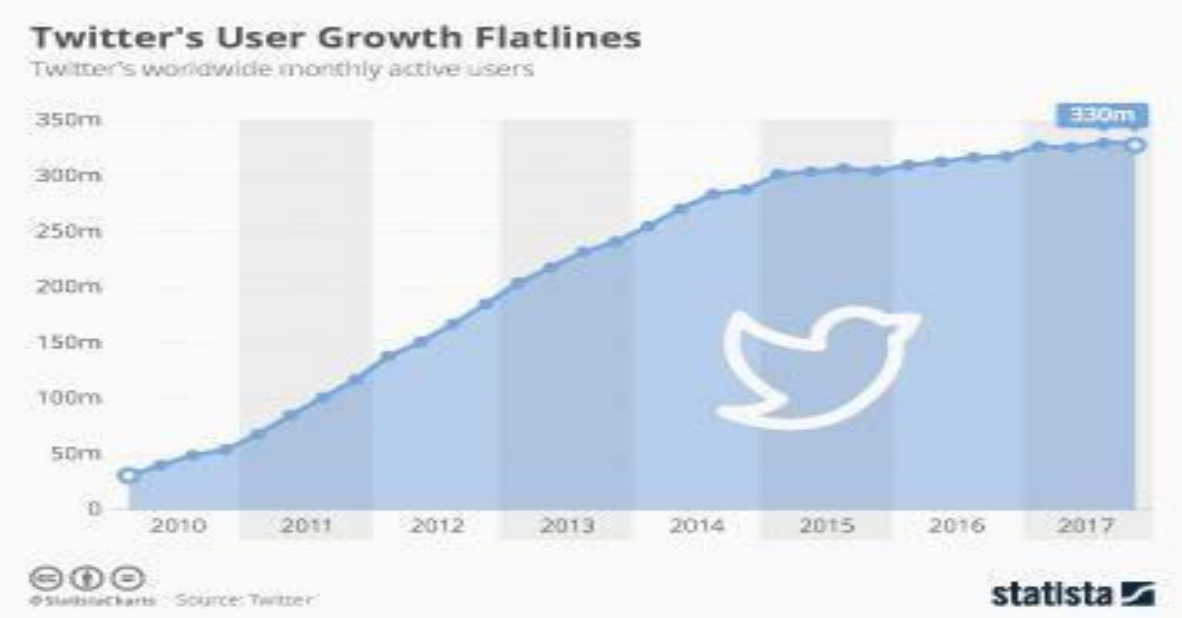

Figure 1-4: Statistique d'utilisateurs twitter [Web2].

1.8.2 Historique de twitter

Twitter a été créé à l'aide de Ruby on Rails, un framework d'application Web spécialisé pour le langage de programmation informatique Ruby, son interface permet une adaptation ouverte et l'intégration avec d'autres services en ligne, le service a été conçu en 2006 par Evan Williams et Biz Stone, qui ont chacun travaillé chez Google avant de partir pour lancer l'entreprise de podcasting Odeo. Williams, qui avait déjà créé l'outil de création Web populaire Blogger, a commencé à expérimenter avec l'un des projets annexes d'Odeo: un service de messages courts (SMS) appelé Twttr, voyant un avenir pour le produit, Williams a racheté Odeo et a lancé Obvious Corp, pour le développer davantage.

L'ingénieur Jack Dorsey s'est joint à l'équipe de direction, et la version complète de Twitter a débuté à la conférence South by Southw est à Austin, au Texas, en mars 2007, le mois suivant, Twitter Inc. a été créé en tant qu'entité corporative de capital-risque **[12].**

Comme la base d'utilisateurs de Twitter a commencé à croître, une chose drôle a commencé à se produire. Les utilisateurs créaient un nouveau jargon et différentes façons d'utiliser le service. Pensez-y comme une innovation née par nécessité.

Au départ, les utilisateurs n'avaient aucun moyen de répondre ou de se crier sur Twitter, certains utilisateurs incluraient un symbole @ avant leur nom d'utilisateur pour identifier un autre utilisateur dans un Tweet, c'est devenu un moyen si courant de reconnaître un autre utilisateur que l'équipe de Twitter a ajouté la fonctionnalité nativement à la plate-forme Twitter.

La même chose s'est produite avec les hashtags, qui font maintenant partie intégrante de l'écosystème Twitter **[13].**

Cette fonctionnalité axée sur l'utilisateur est également vraie pour la façon dont les retweets sont créés, les utilisateurs voulaient un moyen de republier un message d'un utilisateur de Twitter tout en créditant l'utilisateur qui l'avait initialement tweeté.

1.8.3 Les concepts de base sur twitter

Le fonctionnement de ce réseau social est assez simple, un compte peut représenter une personne, une entreprise, un département, …. lorsque vous avez un compte, vous pouvez faire les actions suivantes **[14].**

- **Écrire un « tweet » :** c'est un message de 280 caractères qui peut contenir des images, liens, vidéos …etc.

- **Le fil d'actualité :** c'est la page d'accueil de votre compte twitter. les tweets apparaissent dans un ordre chronologique.

- **Être suivi par d'autres comptes :** ce que l'on appelle les « **Followers** », ou les abonnés en Français, ces personnes verront vos tweets apparaître sur leurs fils d'actualité.

- **Suivre un compte :** ce que l'on appelle le **« Following »,** ou abonnement en Français, vous verrez les tweets de ces comptes apparaître sur votre fil d'actualité.

- **Reprendre le tweet d'un autre compte :** ce que l'on appelle « Retweet » c'est à dire le transmettre à vos « followers » ou bien le « Citer, en y ajoutant un message, c'est un peu l'équivalent de transférer pour un mail.

- **Ajouter un « Hashtag » dans vos tweets:** c'est à dire mettre un mot clé dans votre tweet. Votre tweet sera ajouté à une liste avec tous les tweets ayant le même Hash Tag.

Twitter a inventé ce concept mais il a été popularisé par d'autres réseaux sociaux. comme Facebook ou Pinterest par exemple.

- **Faire une « Mention » :** c'est à dire ajouter le nom d'un utilisateur dans un tweet, ce dernier sera avertit de la chose.

- **Messages privés :** comme la plupart des réseaux sociaux, Twitter vous offre la possibilité d'envoyer des messages privés à vos amis ! vous pouvez profiter de cette fonctionnalité sur la version mobile et sur la version de bureau en appuyant sur l'onglet messages en bas à droite de votre application (mobile) ou en cliquant sur le même bouton dans le coin supérieur gauche de votre page Twitter **[15].**

- **Favoris ou Like :** le like de tweet est la dernière action proposée par le réseau social sur les statuts d'autres utilisateurs, action qui était il y a peu nommée "ajouter au favoris". Cette action montre à l'utilisateur qui a émis le tweet que l'on aime on message, l'émetteur reçoit là encore une notification similaire à la notification de retweet.

Figure 1-5 : Exemple d'un profil Twitter.

1.9 Conclusion

La plupart des réseaux sociaux se développent tellement qu'il est difficile d'avoir un point de vue critique, en y réfléchissant, personne n'a forcément besoin d'avoir 500, 800 ou même 5000 amis...

L'amitié se construit dans la réalité avec des personnes en qui on a confiance, le plus important est la qualité du lien social et non pas la quantité, la vie professionnelle, pour de nombreux employeurs, ces réseaux sont une véritable mine d'informations pour recruter ou congédier du personnel, ou pour vendre un produit ou présenter un service.

Les avis et les points de vue de utilisateurs est un moyen pour mesurer l'importance de tel service vis-à-vis un autre.

Collecter et analyser ces opinions est devenue une tâche primordiale et c'est l'objectif de l'analyse des sentiments et l'opinion mining qui sera détaillé dans le prochain chapitre.

Chapitre 02: Etat de l’art

2.1 Introduction

Les utilisateurs des réseaux sociaux sont envahis par le flux d’information diffusé continuellement et ils sont généralement influencés par ce dernier. ainsi, ils rediffusent l’information, commentent des publications, achètent des produits ou des services et parfois organisent des évènements, signent des pétitions, etc.

Dans ce chapitre, nous allons diffuser notre travail sur deux approches qui nous intéressent qui sont l'analyse de sentiment et la fouille de points de vue.

2.2 L’analyse des sentiments

Dans le domaine de l'analyse des sentiments, les gens confondent souvent un point de vue avec un sentiment, c'est pour cela que nous avons découvrir c'est quoi l'analyse de sentiment.

2.2.1 Sentiment

Le sentiment est la composante de l'émotion qui implique les fonctions cognitives de l'organisme, la manière d’apprécier, le sentiment est à l'origine d'une connaissance immédiate ou d'une simple impression. il renvoie à la perception de l'état physiologique du moment.

Le sens psychologique de sentiment qui comprend un état affectif est à distinguer du sens propre de la sensibilité **[16].**

Le dictionnaire Larousse définit le sentiment comme étant un état affectif complexe et durable lié à certaines émotions ou représentations.

2.2.2 C’est quoi l'analyse des sentiments

L'analyse des sentiments est un type d'exploration de données qui mesure l'inclination des opinions par le biais du traitement du langage naturel, de la linguistique computationnelle et de l’analyse de texte utilisées pour extraire et analyser des informations subjectives sur le Web. Les données analysées quantifient les sentiments ou réactions du grand public envers certains produits personnes ou idées et révèlent la polarité contextuelle de l'information.

L'analyse des sentiments utilise des processus et des techniques d'exploration de données pour extraire et capturer des données à analyser afin de discerner l'opinion subjective d'un document ou d'une collection de documents, comme des articles de blog, des critiques, des articles de presse et des mises à jour **[17].**

L'analyse des sentiments peut cependant également reposer sur d'autres éléments que les données textuelles, elle peut par exemple être basée sur l'usage des émoticônes, sur les « émotions » facebook, sur l'analyse de la voix ou même sur le facial coding / decoding. **[18].** Les sentiments sont généralement classés en trois types : négatifs, neutres ou positifs.

2.2.3 Les problèmes de l'analyse des sentiments

Toute personne qui a étudié la linguistique ne vous dirait que les langues sont complexes, il serait trop naïf de simplifier à l'excès le langage en pensant que son sentiment sous -jacent peut toujours être examiné avec précision par une machine ou un algorithme.

Il y a cinq facteurs principaux qui nous empêchent actuellement de compter aveuglément sur des outils pour l'analyse des sentiments **[19]:**

- **Contexte**

Un mot positif ou négatif peut avoir un sens inverse en fonction du contexte. "J'ai fait un excellent travail" peut être interprété comme une affirmation positive.

Cependant, dans "mon fournisseur d'Internet fait un excellent travail quand il s'agit de me voler de l'argent", faire un bon travail n'est plus une chose positive, basée sur le contexte ("me voler de l'argent") **[20].**

- **Ambiguïté de sentiment**

Une phrase avec un mot positif ou négatif n'exprime pas nécessairement un sentiment. par exemple, "pouvez-vous recommander un bon outil que je pourrais utiliser?" n'exprime aucun sentiment, bien qu'il utilise le mot positif "bon".

De même, les phrases sans mots de sentiment peuvent exprimer le sentiment aussi.

Ainsi, "ce navigateur utilise beaucoup de mémoire", ne contient aucun mot de sentiment, bien qu'il exprime clairement un sentiment négatif.

➢ Sarcasme

Un mot de sentiment positif ou négatif peut changer polarité s'il y a du sarcasme dans la phrase. "Bien sûr, je suis heureux que mon navigateur plante au milieu de mes cours" est évidemment une déclaration sarcastique (et négative), même si elle a le mot positif "heureux".

Nous pouvons détecter le sarcasme principalement à partir de la façon dont la phrase commence par «sûr», et le contexte (nous savons pour un fait qu'un navigateur se bloque est négatif).

➢ Comparatifs

Les outils d'écoute sociale se méprennent souvent sur les énoncés comparatifs. Par exemple, quel est le sentiment de "Pepsi est beaucoup mieux que Coca-cola"? si vous faites des rapports pour Pepsi, alors c'est certainement une déclaration positive.
Cependant, si vous travaillez pour Coca-Cola et que vous faites rapport à l'entreprise cette déclaration serait négative, la plupart des outils d'écoute sociale ne sont pas assez intelligents pour «choisir des côtés» quand ils trouvent des affirmations comparatives comme ci -dessus, les laissant choisir le sentiment basé sur des mots-clés.

Ainsi, l'exemple précédent serait étiqueté comme "positif" car il contient un mot-clé positif, "beaucoup mieux", peu importe pour qui vous faites des rapports.

➢ Variations régionales

Un mot peut changer le sentiment et la signification en fonction de la langue utilisée. Ceci est souvent vu dans l'argot, les dialectes, et les variations de langue.
Un exemple est le mot «sick», qui peut changer le sens en fonction du contexte, du ton et de la langue, bien que clair pour le public cible («that is a sick song!» Versus «I'm not feeling well at all, I might be sick"), un exemple d'une variation régionale peut être trouvé entre l'anglais britannique et américaine pour des mots comme «quite», «rather», «pretty»: en anglais britannique ces mots prennent le sens de «fairly» tandis qu'en anglais américain ils prennent le sens de "very".

Cela peut parfois être mal compris dans les conversations quotidiennes. Il n'est donc pas étonnant que les outils puissent trouver cela problématique.

2.2.4 Solutions proposées pour faciliter l'analyse

Suite à l'expression de multiples difficultés rencontrées lors de l'analyse des sentiments certaines solutions ont été proposées, dans un premier temps, afin de faciliter ce processus et d'assurer la crédibilité des résultats obtenus **[21]**:

- Utilisation de dictionnaires de lexique.
- Utilisation de dictionnaires d'émoticônes et d'émojis.
- Utilisation de dictionnaires d'acronymes.
- Utilisation de dictionnaires de mots vides pour mieux cibler le traitement.

2.2.5 Domaines d'application d'analyse des sentiments

La communauté de traitement du langage naturel montre beaucoup d'intérêt dans l'analyse des sentiments, l'explosion d'internet a changé le style de vie des gens, maintenant ils sont plus expressifs sur leurs avis et opinions, et cette tendance a aidé les chercheurs à obtenir du contenu généré par l'utilisateur facilement.

La figure ci-dessous **[22]**, présente quelques domaines d'application de l'analyse des sentiments.

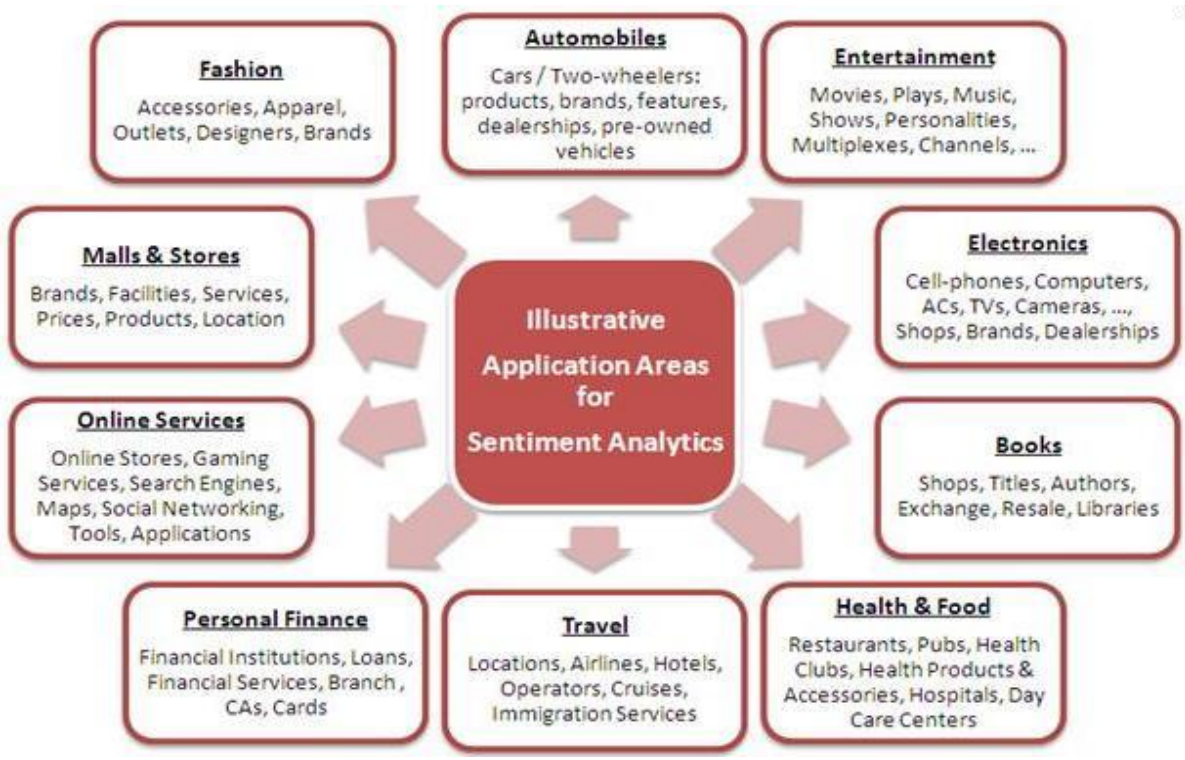

Figure 2-1: Domaines d'application d'analyse des sentiments [Web3].

Dans ce qui suit on expliquerons les domaines de l'analyse des sentiments **[23]** :

➢ Achat d'un produit ou service

L'achat d'un produit ou d'un service, en prenant une bonne décision n'est plus une tâche aussi difficile, mais par cette technique, les gens peuvent facilement évaluer les opinions et expériences des autres concernant n'importe quel produit ou service et ils peuvent aussi facilement comparer les marques concurrentes, maintenant, les gens ne veulent pas se fier à un conseiller externe.

L'analyse des sentiments extrait les opinions des gens à partir de l'immense collection de contenu non structuré, l'internet, les analyse et les présente de façon très structuré et compréhensible.

➢ Marketing

Le marketing a rapidement compris l'intérêt de l'analyse des sentiments.

L'analyse des sentiments vous aide à compléter votre étude de marché en apprenant à connaître l'opinion de vos clients sur vos produits / services et comment vous pouvez aligner la qualité et les fonctionnalités de vos produits / services.

Vos produits et services sont jugés non seulement en fonction de leur performance fonctionnelle, mais aussi en termes de présentation, par exemple, d'un beau design, de promotions irrésistibles, de prix raisonnables et même d'une décoration impressionnante d'un magasin **[24]**.

- ## Le politique

La publication croissante sur internet de textes à teneur politique (lois, rapports, billets de blogs politiques, etc.) et le constat que la politique ne se fait plus seulement dans les hémicycles mais aussi dans les débats en ligne, a conduit certains chercheurs à utiliser les techniques d'analyse des sentiments pour déterminer l'accord ou le désaccord des commentateurs avec telle ou telle proposition de loi.

Les acteurs politiques ont également suivi cette tendance, tel qu'avant de promulguer une nouvelle loi, les politiciens essayent de récolter l'avis des internautes sur cette loi, il est intéressant de connaître aussi l'avis des internautes sur tel homme politique pour une élection présidentielle par exemple, l'analyse des médias sociaux a indiqué que le républicain Donald Trump gagnerait les élections américaines, avant que les sondages confirmé que ce soit vrai, ceci selon les données de médias sociaux analysées par BrandsEye[4], qui a pointé vers une victoire de Trump avant même que les votes aient été jetés.

4 BrandsEye est une solution de surveillance des médias sociaux prise en charge par Crowd avec des fonctionnalités de BI. BrandsEye fournit une analyse précise des sentiments, ainsi que des données et des informations pertinentes et structurées sur les médias sociaux et en ligne, sur les marques de votre entreprise.

➢ La publicité en ligne

Si la publicité en ligne, ciblée et contextuelle, pourrait bénéficier des recherches en analyse des sentiments, car si une annonce publicitaire est d'autant plus efficace qu'elle apparaît au bon endroit et au bon moment, elle pourrait l'être encore plus si elle s'adaptait au ressenti des consommateurs vis -à-vis d'un produit ou d'un service.

C'est ce que propose par exemple la stratégie DASA (Dissatisfaction-oriented Advertisingbased on Sentiment Analysis), qui a pour but de détecter et de prendre en compte les points d'insatisfaction des consommateurs afin d'adapter encore mieux les annonces publicitaires à leurs cibles.

➢ Systèmes de recommandation

Un système de recommandation permet de prédire un choix sans aucune connaissance personnelle des alternatives.

Les Algorithmes des moteurs de prédiction sont basés sur l'expérience et l'avis des autres utilisateurs. il est utile de trouver des recommandations de personnes qui ont les mêmes goûts que nous, qui sont familiers avec le problème, ou qui sont des experts reconnus ».

Le but des systèmes de recommandation est de prédire l'affinité entre un utilisateur et un article en se fondant sur un ensemble d'informations déjà acquises sur cet utilisateur et sur d'autres, ainsi que sur cet article et sur d'autres, il existe plusieurs familles de systèmes de recommandation, en fonction de la manière dont la recommandation est effectuée et de la nature des données.

2.2.6 Les approches utilisés dans l'analyse des sentiments

Les travaux existants dans le domaine de l'analyse des sentiments sont très nombreux les méthodes employées sont diverses et variées, c'est pourquoi dans cette section, nous nous intéressons aux travaux traitant les points de vue ordinaires, qui sont classées par polarité à savoir positive, négative ou neutre, nous détaillerons celles-ci dans la suite du rapport.

Les trois approches qui existent sont:

- Approche basée apprentissage automatique.
- Approche basée lexique.
- Approche hybride.

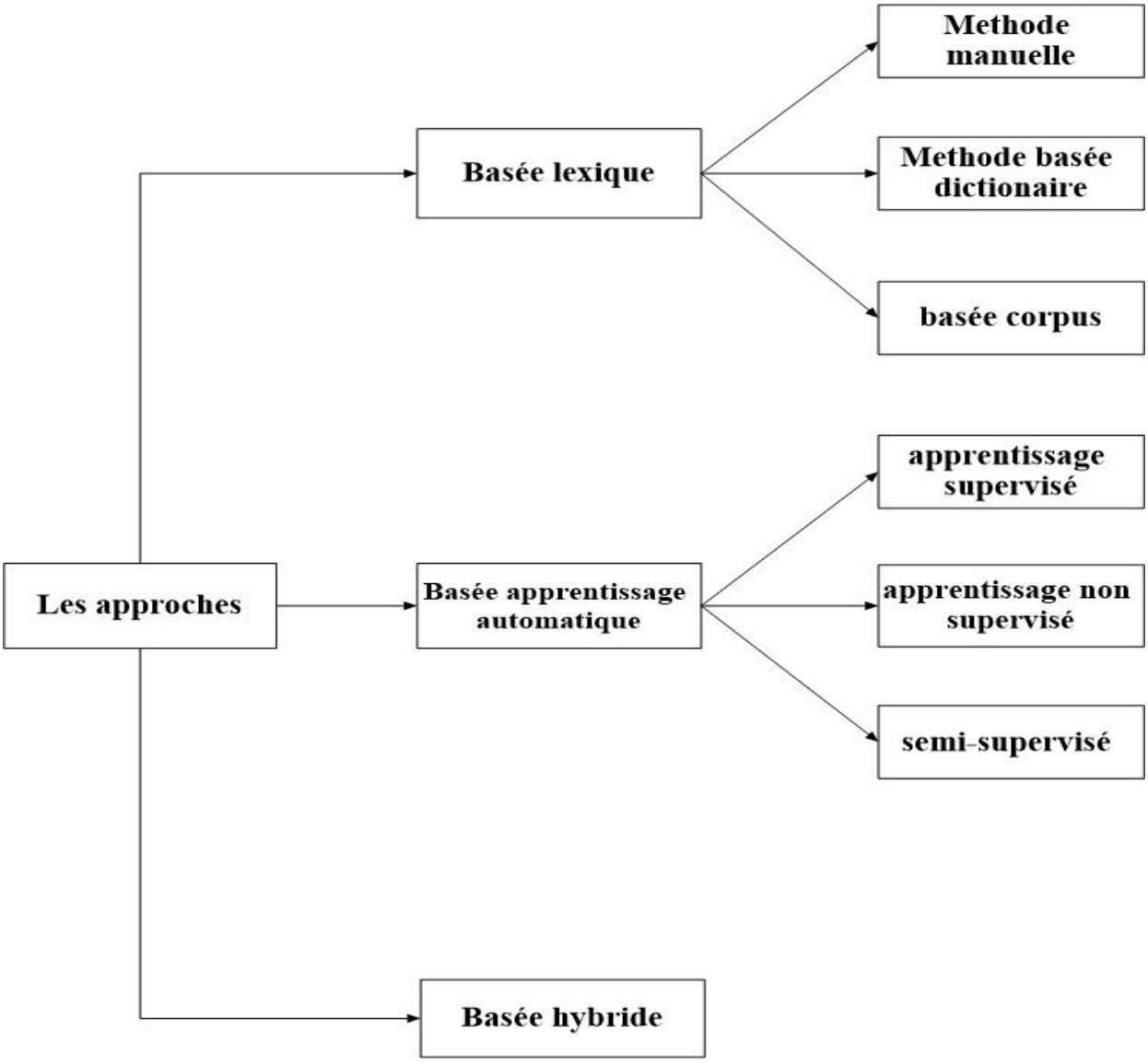

Figure 2-2: Hiérarchie des approches d'analyse.

2.2.6.1 Apprentissage automatique

L'apprentissage automatique est une tentative de comprendre et reproduire la faculté de l'apprentissage humain dans des systèmes artificiels.

Il s'agit de concevoir des algorithmes capables, à partir d'un nombre important d'exemples, d'en assimiler la nature afin de pouvoir appliquer ce qu'ils ont ainsi appris aux cas futurs, ainsi le but essentiel de l'apprentissage automatique est de déterminer la relation entre les objets et leurs catégories pour la prédiction et la découverte des connaissances.

On distingue ainsi trois types d'apprentissage automatique : l'apprentissage supervisé l'apprentissage non supervisé et l'apprentissage semi-supervisé.

- **Apprentissage supervisé**

L'apprentissage supervisé consiste à construire un modèle basé sur un jeu d'apprentissage et des labels (nom des catégories sou des classes) et à l'utiliser pour classer des données nouvelles **[25].**

Cette technique est utilisée dans une variété d'applications telles que les diagnostics médicaux, la prédiction des pannes et la détection des opinions trompeuses dans les réseaux sociaux.

Il existe plusieurs algorithmes et techniques de classification supervisée, tels que:

- Classification Bayésienne: C'est une méthode de classification statistique qui se base principalement sur le théorème de Bayes, elle est utilisée dans plusieurs applications telles que les applications de détection de courriels (ou Spams) pour séparer les bons courriels des mauvais **[25]**. il y a beaucoup d'études qui ont été utilisées avec cette méthode certains utilisent un modèle multivarié de Bernoulli, c'est-à-dire un réseau bayésien sans dépendance entre les mots et les caractéristiques de mots binaires (par exemple Larkey et Croft 1996, Koller et Sahami 1997), d'autres utilisent un modèle multinomial, c'est-à-dire un modèle de langage uni-gramme avec des nombres de mots entiers (par exemple Lewis et Gale 1994, Mitchell 1997).

- Machine à vecteurs de support (SVM) : Il s'agit d'un ensemble de techniques destinées à résoudre des problèmes de discrimination (prédiction d'appartenance à des groupes prédéfinis) et de régression (analyse de la relation d'une variable par rapport à d'autres) **[25].**

Diverses études ont montré que Naive Bayes et SVM sont les meilleurs pour les microblogs, par exemple l'étude de Fang en 2015 **[26]**, il montré un classificateur SVM allant de 0,61% à 0,94% que ses données d'entraînement a augmenté de 180 à 1,8 millions de critiques de produits.

- Réseau de neurones: c'est une technique de type induction c'est-à-dire que par le biais d'observations limitées, elle essaye de tirer des généralisations plausibles.

Elle est basée sur l'expérience qui se constitue une mémoire lors de la phase d'apprentissage appelée entraînement **[25].**

Essam Al-Mansouri et Sean Amos **[27]** ils ont fait un projet d'utilisation de réseaux neuronaux artificiels et analyse de sentiment pour prédire vers le haut mouvement en Stock Prix.

Le but de son projet était de concevoir et de mettre en place un système d'apprentissage automatique permettrait de prédire avec précision si le prix d'une action serait plus élevé 65 minutes dans le futur.

- Forêts d'arbres décisionnels (Random Fortress): C'est une application de graphe en arbres de décision permettant ainsi la modélisation de chaque résultat sur une branche en fonction des choix précédents, on prend en suite la meilleure décision en fonction des résultats qui suivront, on peut considérer ceci comme une forme d'anticipation **[25].**

En particulier dans la fouille de point de vue, ils ont l'avantage d'être lisibles et très faciles à comprendre et à interpréter ce qui est une des raisons de leurs succès.

- Le Boosting: Il s'agit d'une méthode de classification émettant des hypothèses qui sont au départ de moindre importance, plus une hypothèse est vérifiée, plus son indice de confiance augmente, ce qui prend de l'importance dans la classification **[25].**

- **Apprentissage non supervisé**

L'apprentissage non supervisé (en anglais clustering) vise à construire des groupes (clusters) d'objets similaires à partir d'un ensemble hétérogène d'objets, chaque cluster issu de ce processus doit vérifier les deux propriétés suivantes **[25]** :

- La cohésion interne (les objets appartenant à ce cluster sont les plus similaires possibles).
- L'isolation externe (les objets appartenant aux autres clusters sont les plus distincts possibles).

Le processus de «clustering » repose sur une mesure précise de la similarité des objets qu'on veut regrouper. Cette mesure est appelée distance ou métrique.

Le « clustering » est utilisé dans plusieurs applications telles que le traitement d'images, les études démographiques, la recherche génétique, le forage des données et l'analyse des opinions.

On distingue plusieurs algorithmes de clustering, exemple:

- K-moyennes (KMeans) : un algorithme de partitionnement des données en K groupes ou clusters. chaque objet sera associé à un seul cluster. Le K est fixé par l'utilisateur.
- FuzzyK Means : Il s'agit d'une variante du précédent algorithme proposant qu'un objet ne soit pas associé qu'à un seul groupe.
- Espérance-Maximisation (EM) : Cet algorithme utilise des probabilités pour décrire qu'un objet appartient à un groupe, le centre du groupe et ensuite recalculé par rapport à la moyenne des probabilités de chaque objet du groupe.
- Regroupement hiérarchique : deux sous-algorithmes en découlent : le «bottom up » qui a pour fonction d'agglomérer des groupes similaires, donc en réduire le nombre (les rendre plus lisibles) et d'en proposer un ordre hiérarchique et le «top down » qui fait le résonnement inverse en divisant le premier groupe récursivement en sous-ensembles **[25].**

- **Apprentissage semi-supervisé**

L'apprentissage semi-supervisé utilise un ensemble de données étiquetées et non-étiquetés, il se situe ainsi entre l'apprentissage supervisé qui n'utilise que des données étiquetées et l'apprentissage non-supervisé qui n'utilise que des données non-étiquetées.

L'utilisation de données non-étiquetées, en combinaison avec des données étiquetées, permet d'améliorer de façon significative la qualité de l'apprentissage.

Un autre avantage vient du fait que l'étiquette de données nécessite l'intervention d'un utilisateur humain, lorsque les jeux de données deviennent très grands, cette opération peut s'avérer fastidieuse **[25].**

Dans ce cas, l'apprentissage semi-supervisé, qui ne nécessite que quelques étiquettes, revêt un intérêt pratique évident et indiscutable.

2.2.6.2 Approche basée lexique

- **Hypothèse :** Dans cette approche, les chercheurs partent du principe que le corpus n'est pas annoté au préalable, dans cette approche, les chercheurs construisent alors le vocabulaire de mots d'opinion initial, et utilisent des méthodes pour l'enrichir.

Pour ensuite arriver à l'étape de catégorisation des messages en classe positive négative ou neutre, cela est fait de manière générale, en calculant le score d'un message par rapport aux mots d'opinions présents. il y a beaucoup d'études qui utilisent cette approche à partir de laquelle, par exemple Samhaa R. El-Beltagy de l'université du Nil **[28]** a fait une étude en ce qui concerne lexique du sentiment et des mots au niveau du langage pour l'arabe égyptien et l'arabe moderne.

Approche basée sur le lexique est fondamentalement divisé en trois méthodes :

- **Méthode manuelle**

Une annotation de sentiment manuelle prend beaucoup de temps et le travail est intensif et est donc généralement utilisé en combinaison avec des approches automatisées telles que l'approche basée sur un dictionnaire et sur le corpus, elle peut être exécutée après une approche automatisée pour vérifier les résultats et corriger les erreurs possibles.

- **Méthode basée dictionnaire**

Cette approche consiste à établir un ensemble de dictionnaires comportant des mots d'opinion avec la valeur qu'ils expriment suivis de leurs synonymes et antonymes.

L'utilisation de ces dictionnaires ne permet néanmoins pas de traitements orientés vers des contextes spécifiques **[29].**

Il existe des outils permettant d'identifier le sentiment dégagé par un texte, voici une liste non exhaustive des outils les plus connus :

- Werfamous **:** outil d'analyse en ligne gratuit, donnant un score de sentiment sur une échelle de -100 à 100, ainsi qu'un niveau de confiance lié à ce score.
- AFINN**:** évalue la positivité/négativité d'un mot à l'aide d'un dictionnaire contenu dans une archive.
- General Inquirer **:** lemmatise les mots, effectue une analyse graphique et statistique et produit un rapport contenant des phrases avec les mots les plus significatifs.
- SenticNet**:** L'objectif principal de SenticNet est de rendre l'information conceptuelle et affective véhiculée par le langage naturel (destiné à la consommation humaine) plus facilement accessible aux machines.**[30].**
- WordNet: permet de savoir à l'aide de groupe de synonymes si un mot est positif ou non.
- SentiWordNet: Il s'agit d'une extension à WordNet; il attribue à chaque

groupe de synonymes provenant de WordNet, trois scores de sentiment : la positivité, la négativité , l'objectivité.

- SentiSense **[31]:** Il s'agit également d'un travail basé sur WordNet permettant de polariser les mots de façon plus précise.
- Subjectivity Lexicon **[32] :** permet d'obtenir la polarité d'un mot tout en prenant soin d'analyser son contexte d'utilisation.
- MicroWNOp **:** se base à la fois sur General Inquirer et WordNet ; il fournit des ensembles de mots positifs, négatifs et objectifs qui sont synonymes.

Dans le domaine de l'analyse de sentiment, une étude comparative **[33]** a été effectuée afin de déterminer quels étaient les avantages et inconvénients de chaque source de données, dans le cadre d'analyse de tweets relatifs à des évènements majeurs, l'étude met en avant le fait que plusieurs de ces tweets n'ont pas pu être reconnus par les sources de données.

On peut y voir que SentiWordNet, SenticNet et SentiStrength semblent couvrir un plus grand nombre de tweets, cependant l'article met également en évidence que le taux de couverture n'est pas synonyme de reconnaissance efficace et que la polarité d'un mot donné n'est pas fiable, c'est pourquoi l'article se propose de combiner plusieurs de ces méthodes afin d'exploiter les avantages de chacun et d'obtenir le résultat le plus proche

possible de la réalité.

- **Méthode basée corpus**

Cette approche consiste à établir un ensemble de dictionnaires comportant des mots d'opinion avec la valeur qu'ils expriment suivis de leurs synonymes et antonymes.

L'utilisation de ces dictionnaires ne permet néanmoins pas de traitements orientés vers des contextes spécifiques **[34],** voici la différence principale entre méthode basée corpus et méthode basée lexique **[35]:**

Un corpus est un grand corps de texte en langage naturel utilisé pour accumuler des statistiques sur le texte en langage naturel, le pluriel est le corpus.

Les corpus incluent souvent des informations supplémentaires comme une étiquette pour chaque mot indiquant sa partie de discours, et peut-être l'arbre d'analyse pour chaque phrase.

Un lexique est une collection d'informations sur les mots d'une langue à propos des catégories lexicales auxquelles ils appartiennent, un lexique est généralement structuré comme une collection d'entrées lexicales.

Une entrée lexicale inclura d'autres informations sur les rôles joués par le mot tels que les informations sur les caractéristiques - par exemple, si un verbe est transitif intransitif etc…, quelle forme prend le verbe (par exemple participe présent, passé…etc)

2.2.6.3 Approche hybride

Cette approche combine entre l'approche basée lexique et l'approche basée apprentissage automatique et tente de corriger l'inconvénient du l'approche basée lexique d'être indépendante du domaine et du contexte et l'annotation manuelle pour l'approche basée apprentissage automatique, l'utilisation de l'approche hybride permet d'annoter automatiquement le corpus d'apprentissage avec la méthode basée lexique, et ensuite entrainer le classificateur sur ce corpus avec une méthode issue des méthodes de l'apprentissage automatique**.** Narayanan et al en 2009 **[36]**, ont effectué une fouille d'opinions au niveau phrase, ils déterminent la polarité des phrases par la méthode basée lexique en utilisant les mots d'opinions positifs et négatifs, ensuite, ils appliquent SVM sur les phrases annotées automatiquement par la méthode basée lexique, cela a donné une exactitude de 75.6%.

Ils ont travaillé sur un corpus de Tweets avec différentes requêtes **[37]**, Li et Xu en 2011 ont annotés le corpus d'apprentissage avec la méthode basée lexique en utilisant POS et les mots d'opinions ensuite, entrainent le classificateur sur ce corpus avec SVM. Cette méthode a donné de bons résultats avec une exactitude de 85.4%.

2.3 Fouille de point de vue

2.3.1 Définition

À l'heure actuelle, la terminologie du domaine de la fouille d'opinion et de sentiments dont fait partie la fouille de points de vue, est encore loin d'être stabilisée.

Un même terme peut donc recouvrir plusieurs notions proches mais néanmoins différentes, nous présentons ici la définition de point de vue que nous employons dans le reste de ce mémoire le point de vue peut être présenté comme un type d'opinion particulier.

En effet, l'une des définitions courante d'opinion est :

Ensemble des idées d'un groupe social sur les problèmes politiques, économiques moraux, etc. [Exemple] : L'opinion française.

Par conséquent, un point de vue traduit le positionnement social d'une personne, un positionnement bien pensé, un ensemble de valeurs et de croyances qui le justifie, et est lié à d'autres points de vue existants sur un sujet donné **[7].**

2.3.2 Points de vue et réseaux sociaux numériques

Nous avons défini les concepts de points de vue et de réseaux sociaux, mais nous n'avons pas prouvé qu'il est raisonnable de les utiliser ensemble. d'un point de vue pratique, cet usage peut s'expliquer simplement par la prolifération de vues facilement accessibles dans les réseaux sociaux. Cependant, les données collectées sur les réseaux sociaux ont des limites connues: leur qualité est difficile à mesurer et manque souvent de profondeur d'argument, il est donc raisonnable de poser les questions suivantes: la pertinence de l'utilisation de ces données et les points de vue détectées sur ces plateformes, cependant, bien que la similitude soit importante :

- Sur les réseaux sociaux, nous avons constaté que les expressions vocales produites par de nombreuses personnes sont liées à des thèmes fixés dans une culture commune et sont importantes pour leurs auteurs, représentant généralement l'approbation ou la désapprobation de l'objet commun en question.

- Les utilisateurs de réseaux sociaux sont conscients que d'autres peuvent réagir à propos de ces sujets et de leurs publications sans être nécessairement en contact direct avec ces personnes.

- Ils / elles savent également que leur comportement peut leur permettre d'atteindre un but (partager le plus largement possible une information, convaincre quelqu'un, etc.).

- Lorsque les utilisateurs n'ont pas le même point de vue, bien que ces comportements aient des limites, ces comportements provoquent souvent des conflits interpersonnels, de sorte que les réseaux sociaux semblent être des supports d'information intéressants pour la recherche de vues, de distribution et d'évolution. **[7].**

2.3.3 Détection de points de vue

2.3.3.1 Les Travaux connexes :

Avec l'émergence massive de blogs et de réseaux sociaux, l'exploration de points de vue et l'analyse des sentiments sont devenues des domaines d'intérêt pour de nombreuses études. (Pang et Lee 2008) ont donné un aperçu très large des travaux existants dans leur enquête, les auteurs ont décrit les techniques et les méthodes de recherche d'informations orientées vers l'opinion..

Cependant, peu de recherches en exploration d'opinion blogs et encore moins de microblogging. dans (Yang et al., 2007), les auteurs utilisent des blogs Web pour construire un corpus pour l'analyse des sentiments et utilisent des icônes d'émotions affectées aux articles de blog comme indicateurs de l'humeur des utilisateurs. Les auteurs ont appliqué les apprenants SVM[5] et CRF[6] pour classer les sentiments au niveau de la phrase, puis a étudié plusieurs stratégies pour déterminer le sentiment général du document.

---

5 Machine à vecteurs de support (SVM)

6 **Conditional Random Fields** ou en français **(**champs aléatoires conditionnel) , leur intérêt est sur les tâches d'étiquetage supervisées, pour calculer indirectement ces similarités sur des séquences de textes.

En conséquence, la stratégie gagnante est définie en considérant le sentiment de la dernière phrase du document comme le sentiment au niveau du document. J. Read in (Read, 2005) a utilisé des émoticônes telles que «:-)» et «: -(" Pour former un ensemble de formation pour la classification des sentiments. Pour à cet effet, l'auteur a collecté des textes contenant des émoticônes des groupes de discussion Usenet.

L'ensemble de données a été divisé en «positif» (textes avec des émoticônes joyeuses) et «négatif» (textes avec émoticônes tristes ou en colère) échantillons.

Classificateurs à émoticônes: SVM et Naïve Bayes, ont pu obtenir jusqu'à 70% de précision sur l'ensemble de test.

Dans (Go et al., 2009), les auteurs ont utilisé Twitter pour collecter des données d'entraînement, puis pour effectuer une recherche de points de vue L'approche est similaire à (Read, 2005). Les auteurs construisent des corpus en utilisant des émoticônes pour obtenir des échantillons «positifs» et «négatifs», puis utilisent divers classificateurs. Le meilleur résultat a été obtenu par le classifieur Naïve Bayes avec une mesure d'information mutuelle pour la sélection des caractéristiques. Les auteurs ont pu obtenir jusqu'à 81% de précision sur leur ensemble de test. Cependant, la méthode a montré une mauvaise performance avec trois classes («négative», «positive» et «neutre») **[37].**

Nous allons présenter dans cette section les travaux de détection des points de vue sur les réseaux sociaux.

Le tableau (2.1) synthétise les plateformes utilisées par les modèles introduits et permet de constater que Twitter est largement majoritaire pour cette tâche, bien que certains travaux utilisent d'autres médias sociaux traditionnels, tels que Facebook ou Youtube, ou une plate-forme peu connue. La plupart des modèles sont évalués exclusivement ou partiellement sur Twitter.

Plateforme	Travaux
Twitter	BARBERA(2015),BOIREAU(2014),CERON(2017), CHEREPNALKOSKI **ET** MOZETIC(2015),CONOVER **ET** AL.(2011B),DAVID ET AL.(2016),FANGETAL.(2015),GUERRERO-SOLE(2017),MAGDY **ET** AL.(2016),MA-KAZHANOV,RAFIEI **ET** WAQAR(2014),MOHAMMAD, SOBHANI **ET** KIRITCHENKO(2017),PENNACCHIOTTI **ET** POPESCU(2011),RABELO,PRUDENCIO **ET** BAR-ROS(2012),RAJADESINGAN **ET** LIU(2014),RIZOS,PA-PADOPOULOS **ET** KOMPATSIARIS(2017),THONET **ET** AL.(2017),VOLKOVA,BACHRACH **ET** DURME(2016), WONG **ET** AL.(2013)ETZUBIAGA **ET** AL.(2018)
Facebook	ABBASI **ET** AL.(2014), CERON(2017)
CreateDebate	HASAN **ET** NG(2013) TRABELSI **ET** ZAIANE(2018)
Youtube	RIZOS,PAPADOPOULOS **ET** KOMPATSIARIS(2017)
Flickr	RIZOS,PAPADOPOULOS **ET** KOMPATSIARIS(2017)
CNN News	DONGETAL.(2017)
Forum	DONG **ET** AL.(2017),TRABELSI **ET** ZAIANE(2018), ZHANG **ET** AL.(2017)
Blog	CERON(2017)
Données synthétiques	AKOGLU(2014)

Tableau2-1: Les travaux de détection de points de vue effectués sur les réseaux sociaux **[7].**

Nous nous concentrerons ici sur la détection de points de vue nous présenterons les modèles de détection en fonction des informations utilisées par ceux-ci.

Nous nous attarderons tout d'abord sur les méthodes fondées exclusivement sur le contenu textuel, puis sur les méthodes se distinguant par leur usage des interactions sociales présentes sur les réseaux sociaux et, pour finir, sur les méthodes mixtes exploitant à la fois le texte et l'aspect social.

2.3.3.2 Méthodes fondées sur le contenu textuel

Pendant longtemps, le contenu textuel a été la seule source d'information disponible et de nombreux modèles ont été développés pour détecter les points de vue basés uniquement sur le texte.

Par exemple, nous pouvons citer le modèle thématique de Thonet et al. (2016) Distinguer les mots sujets et les mots d'opinion pour identifier la tendance de réflexion de l'auteur en fonction du choix de vocabulaire de l'auteur.

D'autre part, sur les réseaux sociaux, le contenu textuel est rarement utilisé seul, principalement en raison de son importance pour l'interaction sociale sur ces plateformes , aussi parce que le langage utilisé sur les réseaux sociaux se développe très rapidement, Eisenstein (2013) a spécifiquement déclaré que la proportion des inadéquations de vocabulaire dans les systèmes de traitement automatique du langage augmente avec le temps, et si elles ne sont pas mises à jour pour les prendre en compte, elles deviendront bientôt obsolètes. Le nouveau formulaire apparaît sur la plateforme. Certains travaux sont également consacrés à la découverte et à la mise à jour du vocabulaire utilisé pour discuter de certains sujets.

Néanmoins, il existe des modèles de détection de points de vue qui restent concentrés sur ce type d'information, par choix ou en raison de contraintes liées à la tâche considérée, ceci est particulièrement vrai pour les travaux s'intéressant à cette tâche dans le cadre de plateformes sur lesquelles les interactions sociales sont limitées ou difficiles à extraire, telles que les sites de débats ou les forums.

Certains modèles tentent de prédire le point de vue des données personnelles en modélisant le contenu spécifique de leur discussion, en supposant que ces modèles varient en fonction du point de vue, tandis que ces modèles s'appuient sur les aspects séquentiels de la discussion pour en déduire, déterminez les vues en fonction de la chaîne d'arguments et utilisez les vues et les sentiments du modèle sur les problèmes sous-jacents pour essayer d'agréger ces éléments afin de déterminer les vues correspondantes. **[7].**

2.3.3.3 Méthodes fondées sur les interactions sociales

Avec l'essor des médias sociaux, de nombreux modèles tentent d'utiliser les nouvelles informations représentées par l'interaction sociale entre les données personnelles sous le prétexte d'être caractérisées par l'environnement social des données personnelles.

Ces travaux présentent également l'avantage de ne pas être sensible à l'ironie, très fréquemment utilisée sur les réseaux sociaux, et qui peut grandement handicaper les modèles fondés sur le contenu textuel.

L'utilisation des interactions sociales a été de plus justifiée par les travaux portant sur l'homogénéité et les chambres d'écho sur les médias sociaux, particulièrement pour les points de vue en politique.

Le phénomène d'homogénéité exprime le fait que si deux personnes sont similaires, la probabilité de contact entre les deux personnes est plus élevée, et il existe deux types d'homogénéité: l'homogénéité du statut basée sur le statut formel, informel ou déduite Et l'homogénéité de la valeur établie sur les médias sociaux numériques en fonction de la valeur, des croyances et des comportements, il peut être compliqué de déterminer l'identité des membres, nous pouvons donc supposer que l'homogénéité de la valeur est plus importante que l'homogénéité du statut.

Il existe des modèles tentant de prédire les points de vue des profils fondés sur les interactions sociales de ces derniers, les premiers modèles présentés propagent les points de vue le long des liens inter-profils et, les modèles utilisant des communautés pour déterminer les points de vue, et ceux qui exploitant des représentations alternatives de la similarité inter-profils **[7].**

2.3.3.4 Méthodes mixtes

La plupart des modèles actuels tentent d'utiliser le contenu textuel et l'interaction sociale. De nombreuses combinaisons texte / interactions sociales sont possibles, dans cette partie nous avons catégorisé les modèles en deux familles, selon leur fonctionnement :

Ceux qui utilisent du texte et des interactions sociales en même temps, et ceux qui ont deux modules indépendants qui se complètent pour produire des prédictions plus fiables.

Le tableau suivant présente une synthèse des travaux présentés dans tous les différents types de modèles de détection de points de vue, en résumant quelles informations sont utilisées, le nombre de plateformes et de points de vue gérés et la quantité d'annotations nécessaire à leur bon fonctionnement **[7].**

Remarque: Les colonnes plus d'une plateforme gérée et plus de deux points de vue gérés indiquent si le modèle proposé est défini pour théoriquement prendre en compte ces paramètres mais n'impliquent pas que ces cas de tableau aient été évalués par les auteurs.

Le symbole.indique les éléments présentant des limites ☑

	INFORMATION UTILISATEUR					
	Contenu Textuel	Interactions social	Autre Informations	Peu d' annotations	Plus d' une plateforme gérée	Plus de deux points de vue gérés
ABBASIETAL.(2014)	X	✓	X	X	✓	✓
AKOGLU(2014)	✓	X	X	X	✓	X
BARBERA(2015)	X	✓	X	✓	X	☑
BOIREAU(2014)	✓	X	X	✓	✓	☑
CERON(2017)	✓	X	X	✓	✓	☑
CONOVER ET AL.(2011B)	X	✓	X	✓	X	✓
DAVID ET AL.(2016)	✓	X	X	✓	X	☑
DONG ET AL.(2017)	✓	✓	X	✓	✓	X
FANG ET AL.(2015)	✓	X	X	✓	✓	✓
HASANETNG(2013	✓	X	X	✓	✓	X
MAGDY ET AL.(2016)	✓	✓	✓	X	X	X
MKAZHANOV,RAFIEI ET WAQAR(2014)	✓	✓	X	X	X	✓
MOHAMMAD, SOBHANI ET KIRITCHENKO(2017)	✓	X	✓	X	X	X
PENNACCHIOTTI ET POPESCU(2011)	✓	✓	✓	X	X	X
RABELO,PRUDENCIO ET BARROS(2012)	✓	✓	X	✓	X	✓
RAJADESINGANETLIU(2014)	✓	✓	X	✓	X	X
RIZOS,PA-PADOPOULOSETKOMPATSIARIS (2017)	X	✓	X	✓	✓	✓
THONETET AL.(2017)	✓	✓	X	✓	✓	✓
VOLKOVA,BACHRACHETDURM E(2016)	X	✓	X	X	X	X
WONGETAL.(2013)	✓	✓	X	✓	X	X
WONGETAL.(2016)	✓	✓	X	✓	X	X
ZHANGETAL.(2017)	✓	X	X	X	X	X
ZUBIAGAETAL.(2018)	✓	✓	✓	X	X	✓

Tableau 2-2: Synthèse de la littérature existante sur la détection des points de vue.

2.4 Discussion

Suivons de tous les travaux effectués dans la détection de points de vue dans les réseaux sociaux et spécialement les méthodes fondées sur le contenu textuel qu'il convient à notre titre de recherche et que la plupart des modèles existants de fouille de points de vue sont des modèles supervisés, ces derniers obtiennent généralement des performances supérieures aux modèles non supervisés dû à la difficulté de catégorisation .

Et vu que les données annotées soient essentiels pour améliorer et mesurer de manière fiable les performances des modèles ,et que les modèles non supervisés présentent eux l'inconvénient de nécessiter une étape supplémentaire d'interprétation par l'utilisateur final après la classification, nous choisissant de travailler avec l'apprentissage supervisé, et dans la construction de notre modèle de classification nous choisissant l'algorithme naïves bayes.

Naive bayes est très connu pour son efficacité, sa facilite d'implémentation et ses résultats considérables.

David.D Lewis dans (Lewis, 2004) **[38]** et Hassane Hilali dans (Hilali, 2009)**[39].** listent un ensemble d'avantages du classifieur bayésien naïf, parmi lesquelles :

- Algorithme facile et simple à implémenter.
- Basée sur une théorie mathématique précise.
- Efficacité et rapidité dans l'apprentissage et la classification.
- Facile à mettre à jour avec de nouveaux exemples d'apprentissage.
- Equivalent à un classifieur linéaire, dans sa rapidité d'application.
- Utile pour la classification déterministe comme pour le Ranking puisque il ordonne les classes par degré d'appartenance pour un texte donné.
- Requiert une petite quantité de données d'apprentissage pour estimer les paramètres
- Enfin, le plus important c'est que les méthodes Naïve Bayes donnent de bons résultats.

Tous les avantages cités auparavant et particulièrement la simplicité des calculs la qualité des résultats et la facilité de l'implémentation de cette méthode, au contraire à d'autres techniques plus sophistiquées gourmandes en ressources (gestion de mémoire vive) et en temps d'exécution avec des taux d'amélioration des résultats très minimes, ont stimulé et justifier le choix du modèle d'indépendance conditionnelle (Naïve Bayes classifier) pour notre travail.

2.5 Conclusion

Nous avons présenté dans ce chapitre plusieurs travaux sur la fouille de point de vue et énuméré toutes les approches existantes. Dans le chapitre suivant, nous allons décrire, processus par processus, la conception de notre projet sur la plate-forme de microblogging Twitter.

Chapitre 03: Conception

3.1 Introduction

Pour la conception du modèle de détection de points de vue nous nous avons choisi l'utilisation de l'approche d'apprentissage supervisé et étudier les points de vue sur le réseau social Twitter. en d'autres termes, nous utilisons un modèle pour déterminer le positionnement général d'une population envers un sujet quelconque

ce positionnement est traduit par une polarité positive ou négative des points de vue exprimées dans des tweets.

3.2 Architecture du système

Les processus composant ce système sont présentés dans la figure 3-1 représentant le processus général de la méthodologie du système.

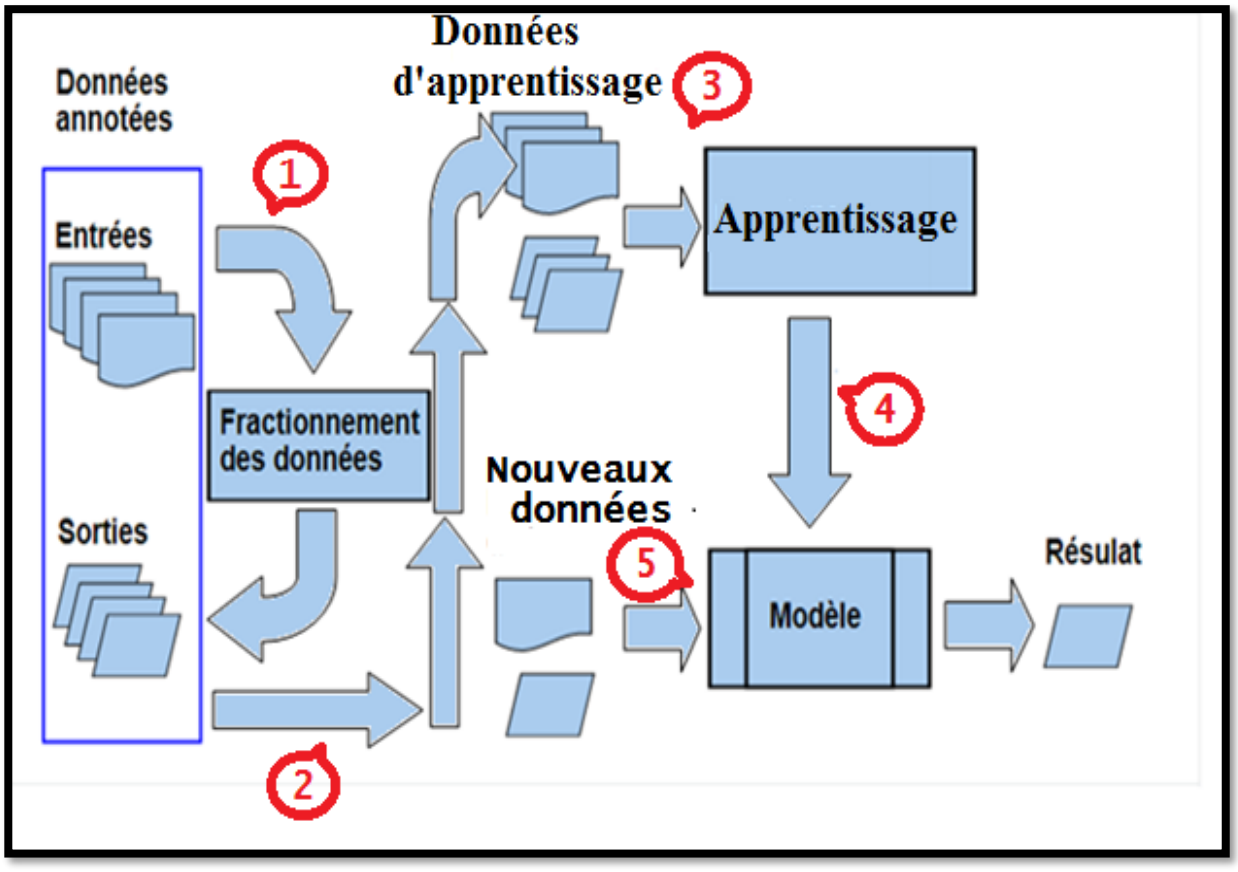

Figure 3-1: Processus général de la méthodologie du système.

Dans ce qui suit nous avons expliqué chaque étape de cette méthodologie:

3.2.1 Jeu de données annotées ou dataset

Il s'agit de l'ensemble de données sentiment140. Il contient 1 600 000 tweets extraits à l'aide de l'API Twitter, les tweets ont été annotés (0 = négatif, 4 = positif) et ils peuvent être utilisés pour détecter le sentiment, il se présente sous forme d'un fichier d'extension (.csv).

Il contient les 6 champs suivants:

Cible: la polarité du tweet (0 = négatif, 4 = positif)

Id: l'identifiant du tweet (2087)

Date: la date du tweet (sam 16 mai 23:58:44 UTC 2009)

Flag: la requête (lyx). S'il n'y a pas de requête, cette valeur est NO_QUERY.

Utilisateur: l'utilisateur qui a tweeté (robotickilldozr)

Texte: le texte du tweet (Lyxis cool)

Selon les créateurs de l'ensemble de données:

"Notre approche était unique car nos données d'entraînement ont été crées automatiquement, par opposition aux tweets d'annotation manuelle des humains, dans notre approche, nous supposons que tout tweet avec des émoticônes positives, comme ":)" était positif et que les tweets avec des émoticônes négatives, comme ":(", étaient négatifs.

Nous avons utilisé l'API de recherche Twitter pour collecter ces tweets à l'aide de la recherche par mot clé ".

Note : nous n'avons pas pu trouver un bon dataset en français pour faire l'apprentissage de notre classifieur, c'est pour ca en quelque sorte nous étions obligés de faire tous en anglais.

D'ailleurs même dans nos tests, nous essayerons de rechercher des termes en anglais, par ce que tous le traitement des tweets suppose que la langue utilisé est bien l'anglais.

Ce dataset contient 1.6 millions de tweets.

Trouvons cette source de donnée (dataset) sur :

Lien : https://www.kaggle.com/kazanova/sentiment140/data

3.2.2 Fractionnement des données ou la préparation du dataset

Le dataset est un ensemble de tweets (1.6 millions) contenant des tweets marquées négative (0) ou bien positive (4), la tache consiste à isoler les tweets (création de deux fichiers),le premier est negative_tweets.csv pour les tweets négatives et postive_tweets.csv pour les tweets positives ,ces deux fichiers appelés les données d'entrainement ou d'apprentissage.

Dans cette préparation et fractionnement nous écrivons seulement le contenu du tweet et on ignore les autres informations tel que qui a tweeter le tweet, la date de rédaction de ce tweet, l'id …etc.

Le processus de suppression de ces informations facilitera le processus d'apprentissage que nous aborderons plus tard, et cette suppression n'affecte pas l'efficacité de notre système ou notre modèle.

3.2.3 Apprentissage

La phase d'apprentissage comporte le prétraitement des données d'apprentissage ainsi que la préparation des ces données prétraiter pour le modèle et finalement la construction et test du modèle, la figure suivante explique la hiérarchie des tous les étapes composé la phase d'apprentissage :

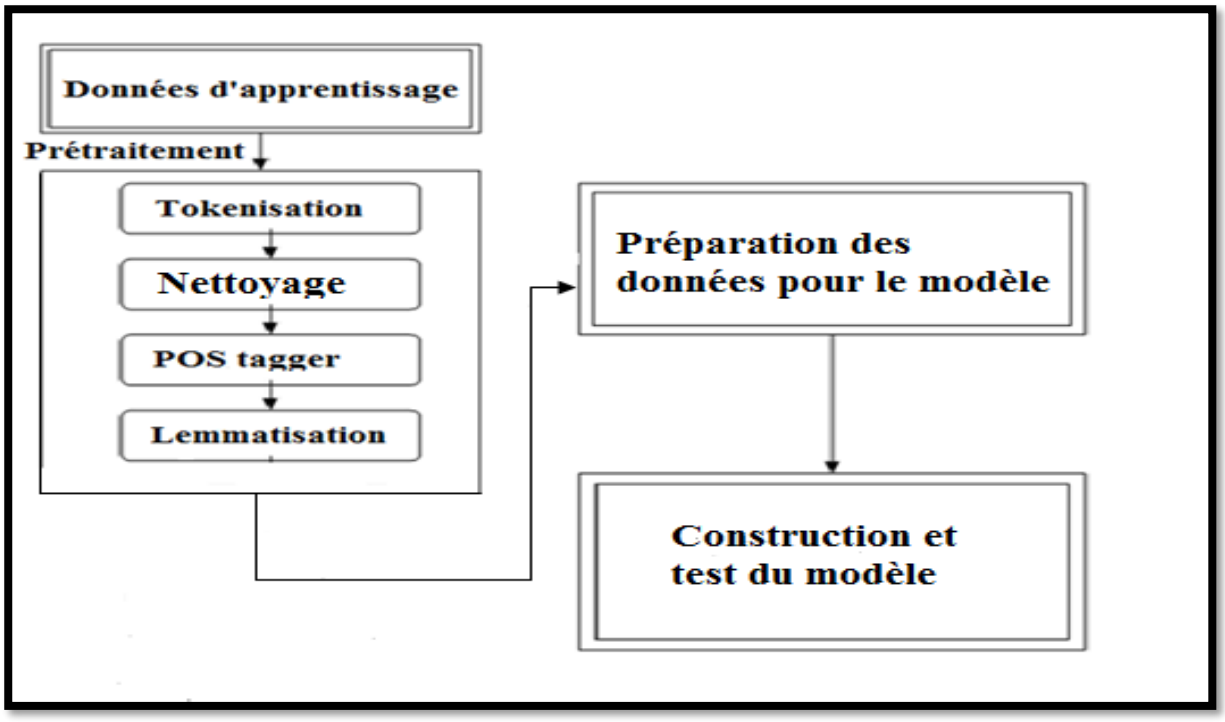

Figure 3-2: La hiérarchie de la phase d'apprentissage.

3.2.3.1 Prétraitement

Prétraitement (preprocessing) est une étape très importante qui consiste en plusieurs techniques visant à traiter les messages pour les structurer et faciliter leurs utilisations.
Elles convertissent les données textuelles originales dans une structure d'exploration de données prêtes et en passant par les étapes détaillées ci-après en dessous des objectifs **[31].**

Le prétraitement des données est le processus de nettoyage et de préparation du texte pour la catégorisation.

Les textes en ligne contiennent généralement beaucoup de bruit et des parties non informatives telles que des balises HTML, des scripts et des publicités, de plus, au niveau des mots, de nombreux mots dans le texte n'ont pas d'impact sur l'orientation générale de celui-ci. Garder ces mots rend la dimensionnalité du problème élevée et donc la classification plus difficile puisque chaque mot dans le texte est traité comme une dimension.

Voici l'hypothèse d'un prétraitement correct des données: réduire le bruit dans le texte devrait aider à améliorer les performances et accélérer le processus de classification, aidant ainsi à l'analyse des sentiments en temps réel. **[36]**

Après avoir lu beaucoup d'articles et des rapports concernant les problèmes de la fouille d'opinion, nous sommes arrivés à la conclusion que ces étapes de prétraitement que nous avons choisies ont le plus d'impact et d'influence sur l'obtention d'un meilleur résultat.

a) Tokenisation

La langue dans sa forme originale ne peut pas être traitée avec précision par une machine, nous devons donc la traiter pour la rendre plus facile à comprendre, pour donner un sens aux données, la première étape consiste à utiliser un processus appelé ***tokenization*** ou à scinder des chaînes en parties plus petites appelées ***tokens ou jetons*.**

Un jeton est une séquence de caractères dans le texte qui sert d'unité, selon la manière dont nous créons les jetons, ceux-ci peuvent être constitués de mots, d'émoticônes, de hashtags de liens ou même de caractères individuels, un moyen simple de diviser le langage en jetons consiste à scinder le texte en fonction des espaces et de la ponctuation, par exemple

```
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

text = 'In this tutorial, I\'m learning NLTK. It is an interesting platform.'
stop_words = set(stopwords.words('english'))
words = word_tokenize(text)
```

***Figure 3-3*:** Exemple de tokenisation d'une chaine de caractères [Web4].

Le résultat sera :

```
['In', 'tutorial', ',', 'I', "'m", 'learning', 'NLTK', '.', 'It', 'interesting', 'platform','.']
```

b) Nettoyage

Dans cette étape, nous allons supprimer le bruit du jeu de données, le bruit est une partie du texte qui n'ajoute pas de signification ou d'information aux données, parmi les éléments constituant un bruit dans les tweets:

- **Les Hyperliens**

Tous les hyperliens de Twitter sont convertis en raccourcisse d'URL.
Par conséquent, les conserver dans le traitement de texte n'ajouterait aucune valeur à l'analyse.

Pour supprimer les liens hypertexte, vous devez d'abord rechercher une sous-chaîne correspondant à une URL commençant par + http: // + ou + https: // +, suivie de lettres, de chiffres ou de caractères spéciaux.

- **Stop words**

Dans ce cas, nous effectuons une suppression des mots qui n'influent pas sur le point de vue exprimée dans le message et qui, de plus, augmente considérablement et inutilement le nombre de mots dans le vocabulaire, ces mots sont :

- Les conjonctions de coordination (for, and, nor, but,or ,yet ,so).
- Les déterminants (*a/an, the, this, that, these, those*).
- Les prépositions (at, in, to).

c) POS tagger

En linguistique de corpus, le balisage d'une partie du discours (balisage POS ou balisage PoS ou POST), également appelé balisage grammatical ou désambiguïsation de catégorie de mots, est le processus de balisage d'un mot dans un texte (corpus) comme correspondant à une partie particulière de la parole, basée à la fois sur sa définition et son contexte, c'est-à-dire sa relation avec les mots adjacents et apparentés dans une phrase, une phrase ou un paragraphe.
Une forme simplifiée de ceci est communément enseignée aux enfants d'âge scolaire, dans l'identification des mots comme noms, verbes, adjectifs adverbes……. etc.

En anglais, il y a huit parties du discours qui sont connues comme le verbe, le nom le pronom l'adjectif, l'adverbe, la préposition, la conjonction et l'interjection.

```
>>> text = word_tokenize("They refuse to permit us to obtain the refuse permit")
>>> nltk.pos_tag(text)
[('They', 'PRP'), ('refuse', 'VBP'), ('to', 'TO'), ('permit', 'VB'), ('us', 'PRP'),
('to', 'TO'), ('obtain', 'VB'), ('the', 'DT'), ('refuse', 'NN'), ('permit', 'NN')]
```

Figure 3-4: Exemple de POS tagging [Web5].

d) Lemmatisation

Les mots ont des formes différentes, par exemple, «run», «runs» et «running» sont différentes formes du même verbe, «run», selon les exigences de notre analyse, il peut être nécessaire de convertir toutes ces versions dans le même formulaire.

La lemmatisation est l'étape qui désigne l'analyse lexicale chargée de faire regrouper les mots d'une même famille qui partagent le même suffixe lexical, chacun des mots du texte se trouve ainsi réduit en une entité appelée « Lemme », ce lemme désigne la forme canonique des mots.
La lemmatisation regroupe les différentes formes que peut avoir un mot, par exemple, un nom en pluriel va être réduit au singulier, un verbe à son infinitif etc.

3.2.3.2 Préparation des données pour le modèle

Un modèle est une description d'un système utilisant des règles et des équations. Cela peut être aussi simple qu'une équation qui prédit le poids d'une personne, compte tenu de sa taille, un modèle d'analyse des sentiments que nous construirons associerait les tweets à une polarité positif ou négatif, nous devrons diviser notre jeu de données en deux parties. Le but de la première partie est de construire le modèle, tandis que la prochaine partie est consacré pour tester les performances du modèle.

Au cours de l'étape de préparation des données, nous allons préparer les données pour l'analyse des sentiments en convertissant les jetons en formulaire de dictionnaire, puis en les scindant à des fins de formation et de test.

a) Conversion de jetons en dictionnaire

Tout d'abord, nous allons préparer les données à intégrer au modèle, nous utiliserons le classifieur Naive ayes pour effectuer l'exercice de modélisation, notez que le modèle nécessite non seulement une liste de mots dans un tweet, mais également un dictionnaire.

Les **dictionnaires** se révèlent très pratiques lorsqu'il sagit de manipuler des Structures complexes à décrire et que les listes présentent leurs limites, les dictionnaires sont des collections non ordonnées d'objets, c'est-à-dire qu'il n'y a pas de notion d'ordre (i.e. *pas* d'indice). On accède aux **valeurs** d'un dictionnaire par des **clés**, les clés sont les jetons ou tokens et les valeurs sont par défaut "True".

b) Division du jeu de données pour la formation et le test du modèle

Ensuite, nous devons préparer les données pour la formation de la classification dans ce contexte vous avons créer le dataset des tweets positifs et le dataset des tweets négatifs, pour cela il suffit juste de remplacer ou d'associer une étiquette (Positive) a la place de la constante **True** dans le dictionnaire précédent, même opération pour le dataset du tweets négative donc avec l'étiquette (négative), nous créons ensuite le dataset global en joignant les tweets positifs et négatifs.

Par défaut, les données contiennent tous les tweets positifs suivis de tous les tweets négatifs en séquence, lors de la formation du modèle, nous devons fournir un échantillon de nos données qui ne contient aucun biais, pour éviter les biais, nous avons organisé les données de manière aléatoire.

Enfin, en divise les données mélangées dans un rapport de 70:30 pour la formation et les tests, respectivement, étant donné que le nombre de tweets est 10 000, nous pouvons utiliser les 7 000 premiers tweets du jeu de données brassé pour la formation du modèle et les 3 000 derniers pour le test du modèle.

Au cours de cette étape, nous avons converti les jetons nettoyés en une forme de dictionnaire, mélangé de manière aléatoire l'ensemble de données et divisons en données d'apprentissage et de test.

3.2.3.3 Construction et test du modèle

La construction des modèles impose l'utilisation d'algorithmes d'apprentissage ou le paramétrage d'un algorithme si nécessaire. Ce processus est généralement itératif et interactif qui vise à obtenir le meilleur modèle à la fin.

Nous utilisons l'algorithme naïve bayes pour construire notre modèle, donc nous avons détailler cet algorithme avec un exemple sur des tweets .

Naïve Bayes est un classifieur probabiliste, pour un document d, sur toutes les classes $c \in C$ le classifieur renvoie la classe $\hat{c}$ qui a la probabilité a posteriori maximale données du document, dans l'équation **(3.1).** Nous utilisons le $\hat{c}$ pour designer l'estimation de la bonne classe.

$$\hat{c} = \arg\max_{c \in C} P(c/d) \qquad (3.1)$$

L'intuition de la classification bayesienne est d'utiliser la règle de Bayes pour transformer, l'équation **(3.1)** en d'autres probabilités qui ont des propriétés utiles.

La règle de Bayes est exprimée dans l'équation (**3.2)**, elle nous donne un moyen de briser toute probabilité conditionnelle P (A/B) en trois autres probabilités :

$$P(A/B) = \frac{P(B/A) * P(A)}{P(B)} \qquad (3.2)$$

On peut alors remplacer l'équation **(3.2)** dans **(3.1)** pour obtenir **(3.3)** :

$$\hat{c} = \arg\max_{c \in C} P(c/d) = \arg\max_{c \in C} \frac{P(d/c) * P(c)}{P(d)} \qquad (3.3)$$

Nous pouvons facilement simplifier l'équation **(3.3)** en supprimant le dénominateur **P(d).**

Cela est possible parce que nous allons calculer $\frac{\boldsymbol{P(d/c)*P(c)}}{\boldsymbol{P(d)}}$ pour chaque classe possible, mais **P(d)** ne change pas pour chaque classe, ainsi, nous pouvons choisir la classe qui maximise cette formule simple :

$$\hat{c} = \arg\max_{c \in C} P(c/d) = \arg\max_{c \in C} P(d/c) * P(c) \qquad (3.4)$$

Nous calculons donc la classe la plus probable $\hat{c}$ étant donné un document **d** en choisissant la classe qui a le plus grand produit de deux probabilités : la probabilité a priori de la classe **P(c),** et la probabilité conditionnelle **P(d=c)** du document.

Nous pouvons représenter un document d comme un ensemble des descripteurs **F** $= f_1, f_2, \ldots\ldots, f_n$ ainsi l'équation **(3.4)** devient comme suit:

$$\hat{c} = \arg\max_{c \in C} P(f_1, f_2, \ldots, f_n/c) * P(c) \qquad (3.5)$$

Malheureusement l'équation **(3.5)** est encore trop difficile à calculer directement. Le classifieur Naïve Bayes fait donc deux hypothèses simplificatrices.

La première est hypothèse du sac de mots qui suppose que la position du mot n'a pas d'importance. ainsi, nous supposons que les descripteurs $\boldsymbol{f_1, f_2, \ldots\ldots, f_n}$ encodent seulement les identités des mots et non pas leurs positions.

Le second est appelle l'hypothèse Naïve Bayes : ceci est l'hypothèse d'indépendance conditionnelle que les probabilités $\mathbf{P}(\boldsymbol{f_i/c})$ sont indépendants étant donnée la classe c et donc peuvent être "naïvement" multipliées comme suit :

$$P(f_1, f_2, \ldots, f_n/c) = P(f_1/c).P(f_2/c).\ldots.P(f_n/c) \qquad (3.6)$$

L'équation finale pour la classe $\boldsymbol{C_{NB}}$choisie par le classifieur Naïve Bayes est la suivante :

$$c_{NB} = \arg\max_{c \in C} P(c) \prod_{f \in F} P(f/c) \qquad (3.7)$$

Pour appliquer le classifieur Naive Bayes au texte, il suit de considérer les mots comme des descripteurs, ainsi F =$w_1, w_2, \ldots\ldots, w_n$.

$$c_{NB} = \arg\max_{c \in C} P(c) \prod_{w_i \in F} P(w_i/c) \qquad (3.8)$$

Le calcul de Naïve Bayes, comme les calcule pour la modélisation du langage sont faites de logarithme afin d'éviter les problèmes de calcul des petites valeurs (underflow), et pour augmenter la rapidité du calcul. Ainsi, l'équation **(3.8)** devient :

$$c_{NB} = \arg\max_{c \in C} logP(c) + \sum_{w_i \in F} logP(w_i/c) \qquad (3.9)$$

Il est claire clair que le calcul de la classe estimée est une fonction linéaire des descripteurs d'entrée.

Le classifieur qui utilise une combinaison linéaire des entrées pour prendre une décision par classification comme Naïve Bayes sont appelés des classifieurs linéaires.

Pour calculer l'équation 3.9 nous avons besoins de calculer les valeurs **P(c)** et **P(w_i/c)**

P(c): est la probabilité a priori de la classe **c** sans connaitre les données.

$$P(c) = \frac{N_c}{N_{doc}} \qquad (3.10)$$

N_c: est le nombre des documents ayant comme classe c.

N_{doc}: est le nombre total des documents.

P(w_i/c) est la proportion du nombre du mot w_i parmi tous les mots dans tous les documentsde la thématique (classe) c.

$$P(w_i/c) = \frac{count(w_i/c)}{\sum_{w \in F}(count(w,c))} \qquad (3.11)$$

Mais Naive Bayes multiplie naïvement toutes les probabilités, la probabilité zéro dans le terme de la probabilité conditionnelle cause à la probabilité zéro de la classe, et cela est inacceptable.

La solution la plus simple est l'ajout de 1 (Laplace) lissage, bien que le lissage de Laplace est généralement remplacé par des algorithmes de lissage plus sophistiqués dans la modélisation du langage, il est couramment utilisé dans Naive Bayes texte catégorisation :

$$P(F_i/c) = \frac{count(F_i/c)+1}{\sum_{f \in V}(count(f,c)+1)} = \frac{count(F_i/c)+1}{\sum_{f \in V}(count(f,c))+|V|} \qquad (3.12)$$

$|V|$: le nombre des termes de tout les documents (sac de mot).

Ceci étant, L'algorithme suivant résume la démarche de la méthode Naïve Bayes pour la classification des textes:

Algorithm 1 Algorithme Train Naïve Bayes(D,C)

INPUT: D: labeled training documents

C : Classes

N_{doc} = number of document in D

V = ExtractUniqueTerms(D) // V is vector of all the Bag-of-Words in the corpus

$V \leftarrow$ vocabulary of D

for each class c in C

  $logprior[c] \leftarrow \log \frac{N_c}{N_{doc}}$

  $bigdoc[c] \leftarrow$ **append**(d) **for** d in D **with** class c

  for each word w in V

    $count(w, c) \leftarrow$ number of occurences of w in $bigdoc[c]$

    $loglikelihood[w, c] \leftarrow \log \frac{count(w,c)+1}{\sum_{w' in V}(count(w',c)+1)}$

return $logprior : logP(c)$, $logliklihood : logP(w, c)$, V

Algorithm 2 Algorithme Test Naïve Bayes($testdoc$, $logprior$, $liklihood$, C, V)

INPUT C: Classes, d: test document, D: Document corpus,

$logP(c)$, $logP(w, c)$, V

for each class c in C # Calculate $P(c)$ terms

  $sum[c] \leftarrow logprior$

  for each w in $testdoc$

    if $word[w] \in V$

      $sum[c] \leftarrow sum[c] + logliklihood$[word[w],c]

return $argmax_c\ sum[c]$

Exemple d'application

L'exemple ci-dessous présente l'apprentissage et le test du classifieur Naive Bayes.

Nous utilisons un domaine d'analyse de sentiment avec deux classes positives (+) et négatives(-), et prenons l'apprentissage miniature suivante et testons des simples documents des avis d'un film.

Ensemble	Classe	Tweets
Apprentissage	Négative (−)	**Just plain boring**
	Négative (−)	**Entirely predictable and lacks energy**
	Négative (−)	**No surprises and very few laughs**
	Positive (+)	**Very powerful**
	Positive (+)	**The most fun film of the summer**
Test	**?**	**Predictable with no originality**

Tableau 3-1 : Exemple d'application de l'algorithme Naïve Bayes.

La probabilité à priori P(c) pour les deux classes se calcule via l'équation **(3.10)** :

$$\mathrm{P}(négative) = \frac{3}{5} \qquad \mathrm{P}(positive) = \frac{2}{5}$$

Les probabilités conditionnelles de l'ensemble d'apprentissage prévu pour les quatre mots "Prédicable", "with", "no", "originality", sont les suivants, à partir de l'équation **(3.12):**

$$\mathrm{P}(\text{"Predictable"}\mid-) = \frac{1+1}{14+20} \qquad \mathrm{P}(\text{"Predictable"}\mid+) = \frac{0+1}{9+20}$$

$$\mathrm{P}(\text{"with"}\mid-) = \frac{0+1}{14+20} \qquad \mathrm{P}(\text{"with"}\mid-) = \frac{0+1}{9+20}$$

$$\mathrm{P}(\text{"no"}\mid-) = \frac{1+1}{14+20} \qquad \mathrm{P}(\text{"no"}\mid-) = \frac{0+1}{9+20}$$

$$\mathrm{P}(\text{"originality"}\mid-) = \frac{0+1}{14+20} \qquad \mathrm{P}(\text{"originality"}\mid-) = \frac{0+1}{9+20}$$

Pour la phrase de test S = "Prédicable with no originality", la classe choisie, se calcule par l'équation **(3.8)** ou l'équation **(3.9)**

$$P(-)P(\text{"}S\text{"}\mid-) = \frac{3}{5} * \frac{2*1*2*1}{34^4} = 1.8 * 10^{-6}$$

$$P(+)P(\text{"}S\text{"}\mid+) = \frac{2}{5} * \frac{1*1*1*1}{29^4} = 5.7 * 10^{-7}$$

ou:

$$\log(P(-)) + \log(P("S"|-)) = \log\left(\frac{3}{5}\right) + \left(\log\left(\frac{2}{34}\right) + \log\left(\frac{1}{34}\right) + \log\left(\frac{2}{34}\right) + \log\left(\frac{1}{34}\right)\right)$$

$$= \mathbf{-5.75}$$

$$\log(P(+)) + \log(P("S"|+)) = \log\left(\frac{2}{5}\right) + \left(\log\left(\frac{1}{29}\right) + \log\left(\frac{1}{29}\right) + \log\left(\frac{1}{29}\right) + \log\left(\frac{1}{29}\right)\right)$$

$$= \mathbf{-6.75}$$

Le modèle estime une classe négative pour la phrase de test S.

Donc nous avons utilisé l'algorithme de naïve bayes pour la construction de notre modèle avec une méthode pour former notre modèle qui prend en entrées les données utilisé pour la formation et une autre méthode pour tester le modèle sur les données de test.

La précision est définie comme le pourcentage de tweets dans le jeu de données de test pour lequel le modèle a pu correctement prédire le sentiment, une précision de 99,5% sur l'ensemble de test est plutôt bonne.

3.2.4 Les nouveaux données

Nous avons choisi un ensemble de tweets collectés a partir Twitter, ces tweets seront choisis selon le terme saisie par l'utilisateur, ensuite on va faire un nettoyage sur ces tweets c'est le même traitement ou le prétraitement fait dans la phase d'apprentissage .

Finalement les tweets seront testées par le modèle en vue de leur classification en tweets positifs et négatifs.

3.3 Conclusion

Nous nous sommes concentrés principalement dans ce chapitre sur la conception système et nous l'avons décomposé en détails en expliquant le rôle et les étapes de chaque processus.

Nous avons utilisé une combinaison de différentes méthodes de prétraitement pour réduire le bruit dans le texte, nous avons choisi les plus communs pour avoir une meilleure et plus efficace analyse. Après avoir prétraité les tweets, l'étape qui suit serait la catégorisation qui est essentiellement le classement des tweets en fonction du sentiment exprimé: positif, négatif.

Le chapitre suivant va porter sur les outils et bibliothèques utilisés pour la mise en œuvre ou réalisation de du système.

Chapitre 04: Implémentation

4.1 Introduction

Dans ce chapitre nous présentons la liste des outils utilisés dans l'implémentation et les différentes phases d'implémentation de notre algorithme, ainsi que les résultats obtenus.

4.2 Environnement de Travail

D'abord, dans cette section nous allons présenter et décrire les outils et l'environnement utilisés dans notre implémentation:

4.2.1 Environnement matériel

Quand il s'agit de l'exigence matériel, nous devons utiliser un ordinateur puissant et rapide en raison de la grande capacité de la base de données utilisée pour l'apprentissage (dataset), un ordinateur qui marche correctement est suffisant qui marche sous n'importe quel système d'exploitation (Windows, Linux, Mac Os), vu que le langage utilisé est un langage interprété. Pour notre cas, nous avons utilisé un ordinateur portable Dell équipé d'un processeur **I3**.

4.2.2 Environnement logiciel

4.2.2.1 Le langage de programmation

Nous avons utilisé le langage de programmation Python la version 3.6.5.

Python est un langage de programmation, interprété car, avant de pouvoir les exécuter, un logiciel spécialisé se charge de transformer le code du programme en langage machine, multi-paradigme et multiplateformes, est placé sous une licence libre, ce qui nous permet de travailler rapidement et d'intégrer les systèmes plus efficacement.

Python peut être utilisé pour gérer des données volumineuses et effectuer des calculs complexes. Il existe ce qu'on appelle des bibliothèques qui aident le développeur à travailler sur des projets particuliers, plusieurs bibliothèques peuvent ainsi être installées pour, par exemple, développer des interfaces graphiques en python.

Figure 4-1: Logo en langage python.

- Python est interprété : Python est traité à l'exécution par l'interpréteur. Vous n'avez pas besoin de compiler votre programme avant de l'exécuter. Ceci est similaire à PERL et PHP.

- Python est interactif : Il est possible d'utiliser une invite Python et interagir avec l'interpréteur directement pour écrire vos programmes.

- Python est orienté objet : Python prend en charge le style orienté objet ou la technique de programmation qui encapsule le code dans les objets.

- Python est un langage de débutant : Python est un excellent langage pour les programmeurs de niveau débutant et prend en charge le développement d'un large éventail d'applications, du simple traitement de texte aux navigateurs WWW en passant par les jeux.

Ce choix a été motivé par les raisons suivantes :

- L'une des principales langues parmi les langues appropriées pour la programmation de problèmes d'apprentissage profond.

- Il dispose un grand nombre de bibliothèques pour le traitement du langage naturel telles que NLPnet, NLTK, … .

- Un langage simple, productif et utilisable dans presque tous les domaines et systèmes.

4.2.2.2 bibliothèque et packages utilisé dans l'environnement logiciel

Nous avons utilisé des différentes packages et bibliothèques :

a -Pip

Pip est un gestionnaire de paquets utilisé pour installer et gérer des paquets écrits en Python, de nombreux paquets peuvent être trouvés sur le dépôt Python Package Index, il existe différents scripts d'installer ce gestionnaire et tous dépend du système d'exploitation utilisé comme linux ou Windows.

b- La bibliothèque Nltk

NLTK est une plate-forme leader pour la construction de programmes Python pour travailler avec des données de langage humain, il fournit des interfaces faciles à utiliser à plus de 50 corpus et ressources lexicales telles que WordNet, ainsi qu'une suite de bibliothèques de traitement de texte pour la classification, la tokenisation, le stemming, le balisage, l'analyse et le raisonnement sémantique, des wrappers pour les bibliothèques NLP de qualité industrielle, et un forum de discussion actif.

Grâce à un guide pratique présentant les principes fondamentaux de la programmation ainsi que des sujets en linguistique informatique, ainsi qu'une documentation API complète, NLTK convient aux linguistes, ingénieurs, étudiants, éducateurs, chercheurs et utilisateurs de l'industrie. NLTK est disponible pour Windows, Mac OS X et Linux, mieux encore, NLTK est un projet gratuit, open source et piloté par la communauté.

NLTK a été appelé «un merveilleux outil pour enseigner et travailler en linguistique informatique en utilisant Python» et «une bibliothèque incroyable pour jouer avec le langage naturel».

c- Package re (Regular expressions)

Une expression régulière est une séquence spéciale de caractères qui nous aide à faire correspondre ou à trouver d'autres chaînes ou ensembles de chaînes, à l'aide d'une syntaxe spécialisée contenue dans un modèle.
Les expressions régulières sont largement utilisées dans le monde unix, le module Python fournit une prise en charge complète des expressions régulières de type Perl en Python, le module **re** lève l'exception re.error si une erreur se produit lors de la compilation ou de l'utilisation d'une expression régulière.

d- Le module random :

Random est un module Python regroupant plusieurs fonctions permettant de travailler avec des valeurs aléatoires.

La distribution des nombres aléatoires est réalisée par le générateur de nombres pseudo-aléatoires Mersenne Twister, l'un des générateurs les plus testés et utilisés dans le monde informatique.

Le module comprend plusieurs fonctions travaillant chacune avec un type défini de variables, ces fonctions peuvent être séparées en trois groupes :

- Celles qui travaillent avec des nombres entiers
- Celles qui travaillent avec des nombres réels
- Celles qui travaillent avec des séquences (par exemple des listes).

e- Le module csv :

CSV est un format ouvert très populaire d'importation et d'exportation de données tel que des feuilles de calcul, les données d'un fichier .csv sont sous forme textuelles séparées par des virgules d'où son nom Comma Separeted Values.

f- La bibliothèques Matplotlib

Matplotlib est une bibliothèque du langage de programmation Python destinée à tracer et visualiser des données sous formes de graphiques. Elle peut être combinée avec les bibliothèques python de calcul scientifique NumPy et SciPy.

Matplotlib est distribuée librement et gratuitement sous une licence de style BSD. sa version stable actuelle (la 2.0.1 en 2017) est compatible avec la version 3 de Python.

g- Json :

JSON est un format de fichier pratique pour stocker des données structurées qui peuvent être traitées un enregistrement à la fois, il fonctionne bien avec les outils de traitement de texte. c'est un format idéal pour les fichiers journaux, c'est aussi un format flexible pour transmettre des messages entre les processus de coopération.

h- Twitter API

Twitter offre une série d'API pour fournir un accès programmatique aux données de twitter, y compris lire des tweets, accéder à des profils d'utilisateurs et publier du contenu au nom d'un utilisateur authentifié, dans notre projet, l'utilisation de Twitter API est indispensable pour la récupération d'un ensemble de tweets qui parle sur un sujet que nous avons saisie avec le clavier, afin de permettre à accéder aux données de Twitter, il faut d'abord enregistrer une demande d'accès.

L'étape de l'enregistrement prendra quelques minutes, après la connexion au compte twitter, il faut pointer sur le manageur d'applications et créer une nouvelle application.

Une fois l'application enregistrée, sous l'onglet Clés et jetons d'accès, on peut trouver les informations nécessaires pour authentifier la demande :
la Clé du Consommateur et le Secret Consommateur (aussi appelé clé API et secret API, respectivement) est un paramètre de l'application, le jeton d'accès et le Secret du jeton d'accès sont à la place des paramètres du compte utilisateur.

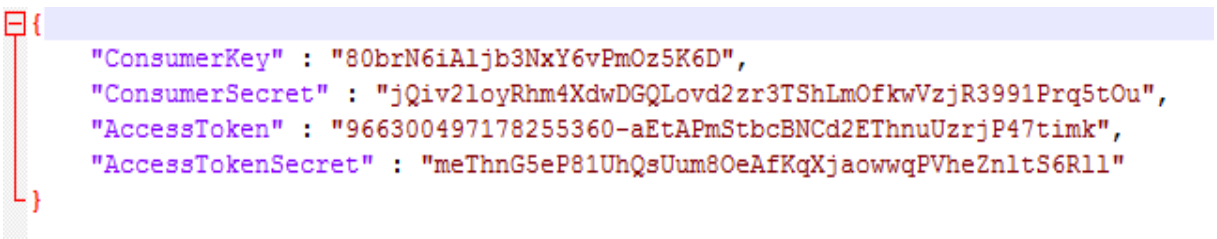

Figure 4-2: Clés de twitter-API de notre système.

Le niveau d'accès à ces paramètres définit ce que l'application peut faire pendant l'interaction avec Twitter au nom d'un utilisateur, l'option en lecture seule est l'option la plus conservatrice, car l'option ne sera pas autorisée à publier quoi que ce soit ou à interagir avec d'autres utilisateurs par l'intermédiaire d'une application directe.

4.3 Exécution du code

Etape 01: préparation du dataset.

Le fichier training.1600000.processed.noemoticon.csv contient 1.6 millions de tweets (mélange des tweets positifs et négatifs).

Pour cela il faut exécuter le script prepare_dataset.py en premier, ce script permet de séparer les tweets et de créer deux fichiers csv, un fichier pour les tweets positifs et un autre pour les tweets négatifs.

Et voici le code utilisé dans le fichier prepare_dataset.py pour faire cette préparation.

```
import csv

# ouverture du dataset ...

with open('training.1600000.processed.noemoticon.csv', 'r', encoding='latin-1') as dataset:
    reader = csv.reader(dataset)

    # on parcours le fichier ligne par ligne (trés grand)
    # environ 240 MB.

    for row in reader:
        # test ...
        # si la ligne commence par '0' c'est une tweet negative

        if row[0] == '0':
            # création d'un ficher csv pour stocker dedans les
            # tweets negatives.
            with open('negative_tweets.csv', 'a',encoding='latin-1', newline='') as negative_tweets:
                writer = csv.writer(negative_tweets)
                # on ecrit seulement le contenu du tweet
                # on ignore les autres informations.
                # tel que qui a tweeter la tweet.

                writer.writerow(['{ text :',row[5], '}'])
        else:
            # création d'un ficher csv pour le stockage du tweets
            # positives.
            with open('positive_tweets.csv', 'a',encoding='latin-1', newline='') as positive_tweets:
                writer = csv.writer(positive_tweets)
                # Idem ici , on ne garde que le contenu
                # du tweets positives.
                writer.writerow(['{ text :',row[5], '}'])
```

Figure 4-3: Le code écrit dans prepare_dataset.py.

Etape 02: prétraitement des tweets résultant après l'exécution du prepare_dataset.py.

Dans cette étape, on supprime le bruit de l'ensemble de données, le bruit est une partie du texte qui n'ajoute pas de sens ou d'information aux données utilisant la fonction suivante pour faire le nettoyage.

```
def remove_noise(tweet_tokens, stop_words = ()):

    # resultat final (la liste qui sera retourné)
    cleaned_tokens = []

    # on parcour la tweet ... mot par mot
    for token, tag in pos_tag(tweet_tokens):
        # on commence par les liens
        # tous qui commence par http:.... ou bien https:: ...
        # c'est un lien et ne va pas infulencer notre classification
        # donc on le supprime.
        token = re.sub('http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+#]|[!*\(\),]|'\
                       '(?:%[0-9a-fA-F][0-9a-fA-F]))+','', token)
        token = re.sub("(@[A-Za-z0-9_]+)","", token)

        # deteminer s'il s'agit d'un not ou un verbe ou autre.
        if tag.startswith("NN"):
            pos = 'n'
        elif tag.startswith('VB'):
            pos = 'v'
        else:
            pos = 'a'

        # trouver la racine du mot ...
        # ex : is -> be.
        lemmatizer = WordNetLemmatizer()
        token = lemmatizer.lemmatize(token, pos)

        # on ajout à la liste final ce mot seulement
        # s'il ne sagit pas d'une poncuation (, ; ? !)
        # et n'appartient pas a stop_words (ex : a , an , the ...)
        if len(token) > 0 and token not in string.punctuation and token.lower() not in stop_words:
            cleaned_tokens.append(token.lower())
    return cleaned_tokens
```

Figure 4-4: La fonction utilisée pour la suppression du bruit.

Après l'utilisation de cette fonction dans le main() nous avons produit deux liste ,une pour contient les jetons nettoyés des tweets positifs et l'autre les jetons nettoyés des tweets négatifs.

^

Etape 03: préparation des données pour le modèle

a) Conversion de jetons en dictionnaire

Dans cette étape, nous avons convertie les jetons des deux listes en dictionnaire, avec les clés qui sont les **jetons** et les valeurs sont **(True)**, la fonction utilisée dans cette conversion est la suivante :

```
def get_tweets_for_model(cleaned_tokens_list):
    # il suffit juste de parcourir la liste.
    for tweet_tokens in cleaned_tokens_list:
        # on retourne mot par mot
        # associe a True (sera utilise comme entre pour la classification
        # dans la bib ntlk.)
        yield dict([token, True] for token in tweet_tokens)
```

Figure 4-5: La fonction utilisée pour la conversion des jetons en dictionnaire.

b) Division du jeu de données pour la formation et le test du modèle

nous devons préparer les données pour la formation de la classe **(NaiveBayesClassifier)** Ajoutons le code suivant au fichier pour préparer les données

```
# la creation du dataset du tweet positive
# il suffit juste de remplacer la constante True
# dans la dict précedente par le mot "positive"
positive_dataset = [(tweet_dict, "Positive")
                    for tweet_dict in positive_tokens_for_model]

# ici la creation du dataset du tweet negative
# comme dans le cas du tweet postives
# il suffit juste de remplacer la constante True
# par le mot "negative"
negative_dataset = [(tweet_dict, "Negative")
                    for tweet_dict in negative_tokens_for_model]

# le dataset global
# c'est l'union des tweets positives et negatives.
dataset = positive_dataset + negative_dataset

# on va mélanger tous les tweets.
random.shuffle(dataset)

# on divise le dataset en deux
# la 1 ère partie sera utilisée pour l'apprentisage
# la 2 ème pour le test.(calcul de la précision).
train_data = dataset[:1120000]
test_data = dataset[48000:]
```

Figure 4-6: Le code utilisé pour la préparation des données avant de les intégrées au modèle.

Ce code associe une étiquette (Positive) ou (Negative) a chaque tweet, il crée ensuite un (dataset) en joignant les tweets positifs et négatifs.

Et pour organiser les données de manière aléatoire en utilisant la méthode shuffle() de (random),ensuite le code divise les données mélangées dans un rapport de 70:30 pour la formation et le test.

Étape 04 : Construction et test du modèle

Enfin, nous pouvons utiliser la classe (NaiveBayesClassifier) pour construire le modèle.

Utilisez la méthode .train () pour former le modèle et la méthode .accuracy () pour tester le modèle sur les données de test.

```
# classifieur de Bayes
from nltk import classify
from nltk import NaiveBayesClassifier
   classifier = NaiveBayesClassifier.train(train_data)
   print("Accuracy is:", classify.accuracy(classifier, test_data))

   print(classifier.show_most_informative_features(10))

   print('Fin d\'apprentissage. ')
```

Figure 4-7: Le code utilisé pour la construction et test de module.

Etape 05: Initialisation du twitter API

Nous devons accéder au twitter pour récupérer des tweets selon le terme saisie par nous ou par un utilisateur de cette application, pour cela il faut d'abord enregistrer une demande d'accès , l'étape de l'enregistrement prendra quelques minutes, après la connexion au compte Twitter il faut pointer sur le manageur d'applications et créer une nouvelle application.

Une fois l'application enregistrée, twitter nous donne des paramètres d'authentification en tant que programmeur ces données sons les clés utilisé pour initialiser le twitter api comme le montre cette figure:

```
# importer twitter API.

import twitter

# pour parser le fichier key.json.

import json

# pour 'arreter' l'exécution temporairement.

import time

# ouvrir le fichier key.json
# assuerer que le fichier key.json se trouve dans le meme dossier.

with open('./key.json') as file:
  keys = json.load(file)

# initialiser l'api twitter
# en utilisant les données du fichier key.json
#

twitter_api = twitter.Api(
    consumer_key=keys['ConsumerKey'],
    consumer_secret=keys['ConsumerSecret'],
    access_token_key=keys['AccessToken'],
    access_token_secret=keys['AccessTokenSecret']
    )
```

Figure 4-8: Le code utilisé pour l'initialisation de l'api twitter.

Etape 06: Téléchargement des tweets .

Nous téléchargeons l'ensemble de tweets contenant le terme saisie par clavier , le code suivant montre comment faire ca

```
# téléchargement du tweets.
from get_tweets import *
# le terme à chercher
# pour récupere les tweets contenant ce terme
# ex: #Corona, #Covid_19, #REAL_MADRID , etc...
search_keyword = input('Entrer un terme à chercher ... ')

# le nombre de tweets à recupérer ...
# multiple de 100.
# example 400, 1000, 2000 , etc
number_of_tweets = input('Entrer le nombre de tweets \'multiple de 100\'... ')
tweets = getDataSetFromTweets(search_keyword, int(number_of_tweets))
```

Figure 4-9: Le code utilisé pour téléchargement des tweets.

Etape 07: Le nettoyage et la classification des tweets téléchargées en tweets positifs et négatifs:

Après le téléchargement des tweets correspond au terme saisie, on va procéder a les nettoyés du bruit c'est le même traitement qu'on a déjà appliqué pour l'apprentissage.

Ensuite on va les classifier en tweets positifs et tweets négatifs comme illustré dans cette figure.

```
# compteur pour les tweets pos.
number_of_postive_tweets = 0
# compteur pour les tweets nega.
number_of_negative_tweets = 0

# ouvrir un fichier csv pour le stockage des resultat
import csv

# ouverture d'un ficher csv
with open('resultats_final.csv', 'w',encoding='utf-8', newline='') as file:
    #
    writer = csv.writer(file)
    # ecriture de l'entete du fich. csv
    writer.writerow(["Tweet", "Classification"])

# on parcours les tweets obtenus une par une ...

for tweet in tweets:
    # ici on faire un nettoyage pour nos tweets.
    # on supprime les liens ...etc
    # le meme traitement qu'on appliquer pour l'apprentissage.

    custom_tokens = remove_noise(word_tokenize(tweet['text']))

    # on va calssifier la tweet ...

    if classifier.classify(dict([token, True] for token in tweet['text'])) == 'Positive':
        # dans le cas où elle est postive
        # on incremente le nombre des tweets postives.
        number_of_postive_tweets = number_of_postive_tweets + 1
    else:

        # dans le cas où la tweet est negative
        # on incremente le nombre des tweets negatives.

        number_of_negative_tweets = number_of_negative_tweets + 1

    #print(custom_tokens, classifier.classify(dict([token, True] for token in tweet['text'])))
```

```
# stockage du reultat
# sous la forme d'un ficher csv
# le fichier contient les tweets et le sentiment
# positive ou bien negative.

import csv
# on stocke le resultat
# de la classification.
# (pour y venir et evaluer le resultat de la class.
# ificatopn)
with open('resultats_final.csv', 'a',encoding='utf-8', newline='') as file:

    writer = csv.writer(file)

    writer.writerow([tweet,classifier.classify(dict([token, True] for token in tweet['text']))])
```

Figure 4-10: Le code utilisé pour la classification des tweet collectés.

Etape 08 : Affichage du résultat

Le code suivant utilisé pour l'affichage du résultat obtenu

```
# on importe la bib matplotlib.
# ce module nous permet de visualiser les resultat.
# une circle ... contenant des pourcentages.
import matplotlib.pyplot as plt

# pourcentage des tweets pos. est calcule comme suit:
# number_of_postive_tweets / (number_of_postive_tweets + number_of_negative_tweets) * 100
# de meme pour le poucentage
# de tweets negatives.
final_result = [number_of_postive_tweets, number_of_negative_tweets]

# les etiquettes qui figurent dans le circle.
activities = ['Positive', 'Negative']

# les coleurs utilisées.
# g : green ou vert pour les tweets postives.
# r : red ou rouge pour les tweets negatives.
colors = ['g', 'r']

# on plot le resultat final .

plt.pie(final_result, labels=activities, colors=colors, startangle=90, autopct='%.1f%%')

# ici on va metre le titre
# c'est le terme recheché ex #Covid_19.
title = search_keyword
plt.title(title)

# on afficher la fenetre qu contient le resulat final.
plt.show()

# fin de notre main.
quit_app = input('Vous voulez quitter l\'application ?  y/n')
```

Figure 4-11: Le code utilisé pour l'affichage de résultat.

4.4 Résultat obtenue

Après la terminaison de l'exécution du programme entièrement nous constatons l'apparition de 4 fichiers:

1. **negative_tweets.csv :** contient les tweets négatifs pour l'apprentissage.
2. **postive_tweets.csv :** contient les tweets positifs pour l'apprentissage (issue du fichier training.1600000.processed.noemoticon.csv).
3. **tweets.csv :** contient des tweets téléchargés depuis Twitter (les tweets contenant le terme à chercher par l'utilisateur de notre programme ex; #Covid_19).
4. **resultats_final.csv :** contient la classification final des tweets.

Après la phase d'apprentissage de notre modèle sur le corpus entier de 1600000 tweets, nous sauvegardons le modèle pour pouvoir l'utiliser pour prédire de futures données.

Les données à prédire sont des données en temps réels prises de twitter grâce à son API.

4.4.1 Interface

On a implémenté une interface graphique pour faciliter à l'utilisateur de chercher et télécharger les tweets correspond au mot qu'il doit le saisir .

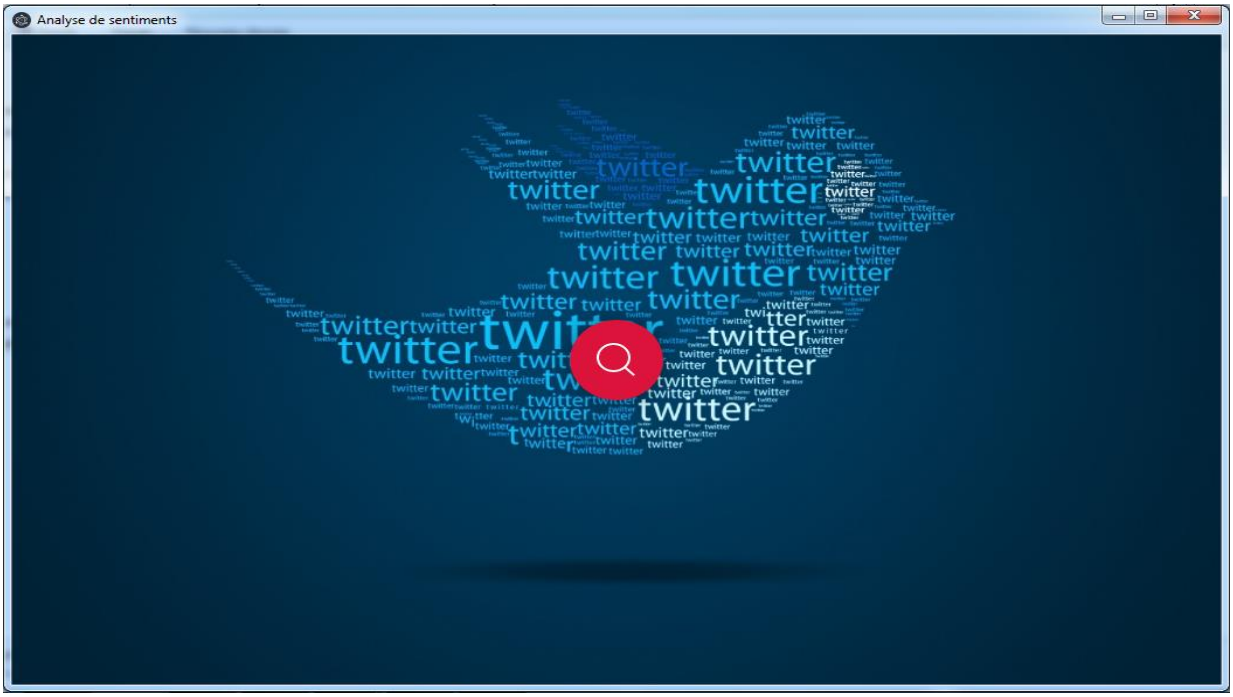

Figure 4-12: L'interface graphique pour l'utilisateur.

Remarque : Le mot que l'utilisateur souhaite saisir doit être suivi du nombre de tweets qu'il souhaite télécharger, comme exemple Corona 100.

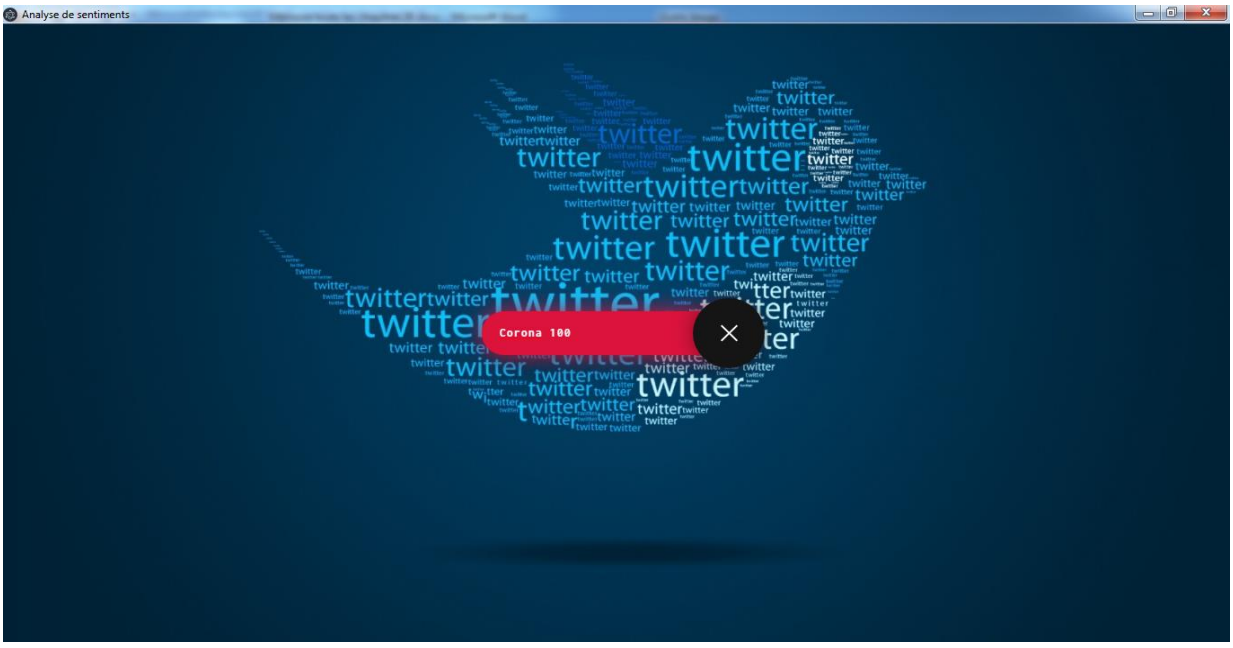

Figure 4-13: Un exemple de recherche par l'utilisateur.

Après que l'utilisateur a appuyé sur le bouton Entrée, notre application affiche la précision de notre classifieur et les vingt mots plus pertinents et sa classification en mots positifs ou négatifs pour que l'utilisateur sache si notre système performant et donne des bons résultats ou non.

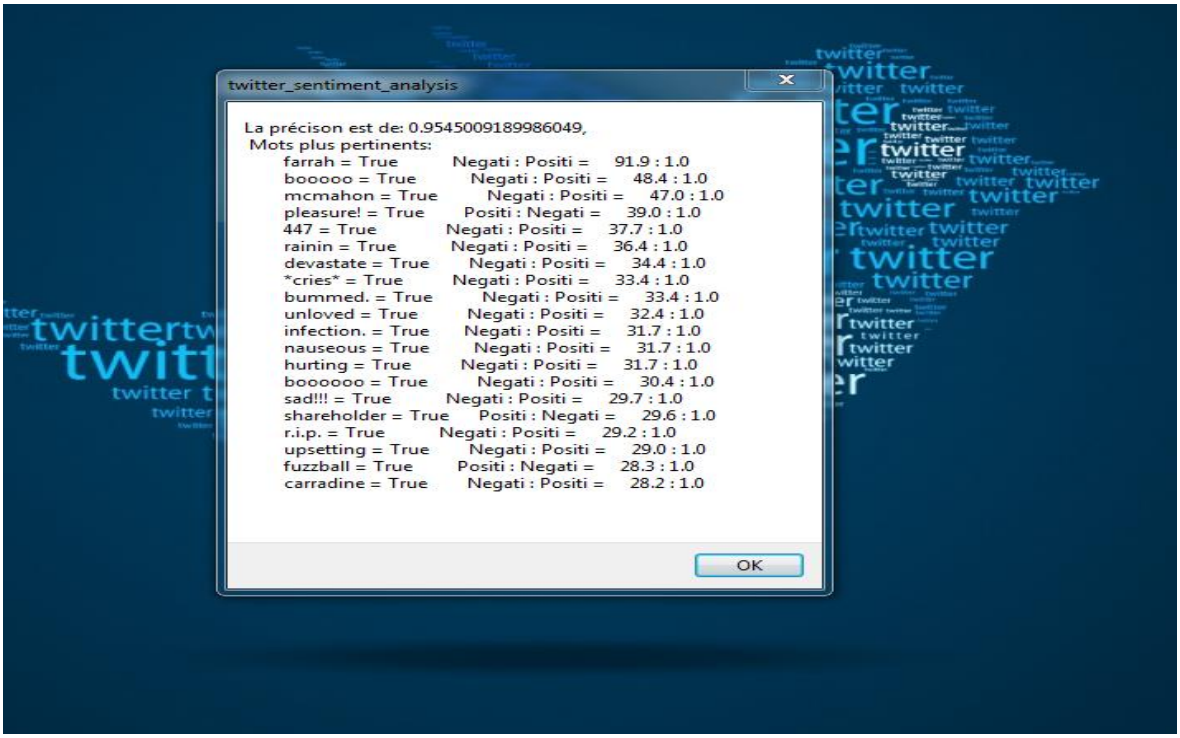

Figure 4-14: La précision de classification de notre modèle.

finalement notre système classe l'ensemble des tweets collectés en tweets positifs et négatifs .

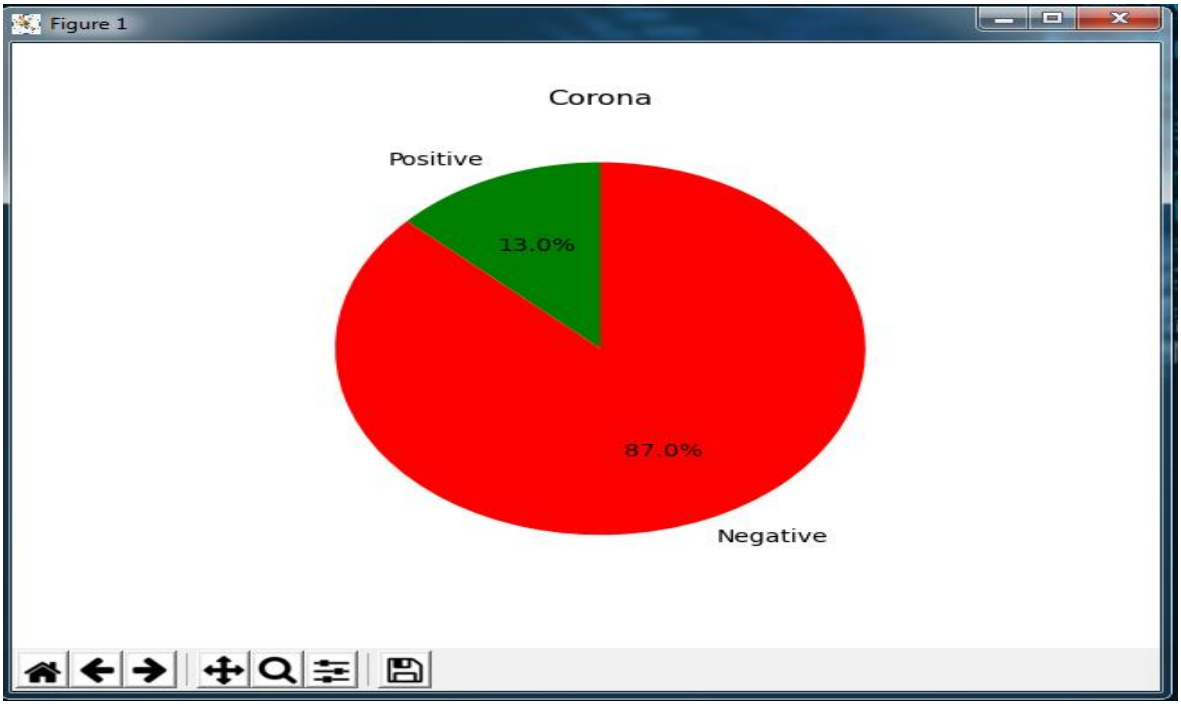

Figure 4-15: Le résultat de classification.

4.4.2 Interprétation des résultats

- **Point positifs de notre système:**

➢ premièrement notre système est basé sur un apprentissage supervisé qui est le très connu dans ce style de recherche et de la classification et d'analyse des sentiments.

➢ Vue qu' une majorité très large des études et travaux fait auparavant dans ce domaine fait par le choix d'un algorithme qui est intégré dans l'approche supervisé comme en a expliquer dans l'état de l'art.

➢ Le jeu de données (dataset) que nous avons utilisé pour faire l'apprentissage de notre classifieur contient 1.6 million de tweets qui augmente le taux de réussite et de précision et plus fiable dans la classification par rapport a des datasets de petite taille.

➢ L'utilisation d'un ensemble de données pour tester la fiabilité et la précision de notre classifieur.

➢ La simplicité des calculs, l'efficacité des résultats et la facilité de l'implémentation.

➢ la complexité de temps d'exécution de la classification par rapport au nombre de tweets téléchargés est O(n) donc notre application ou système de classification est optimal .

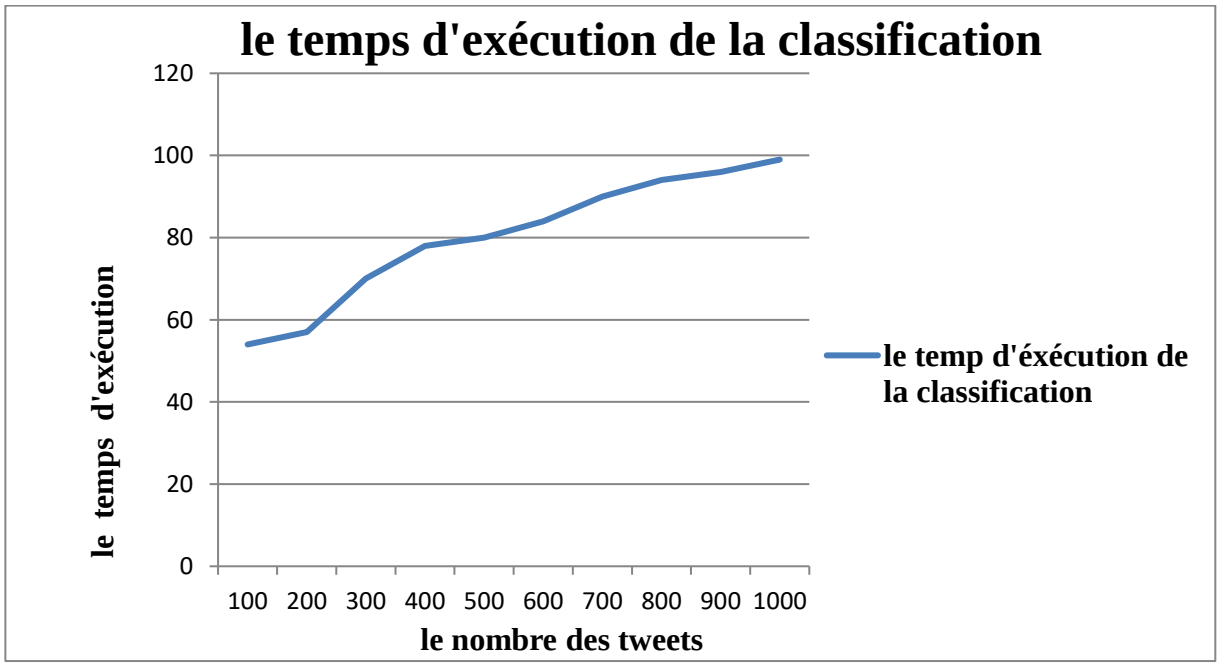

Figure 4-16: Le temps d'exécution de la classification.

- **Limites**

➢ la non-existence du sentiment neutre dans la classification parce que le dataset ou jeu de données utilisé pour l'apprentissage ne contient pas la classe neutre mais d'après les recherches qu'on a fait précédemment on a récolté une multitudes de dataset mais celles qui contient la classe neutre ou l'étiquette neutre n'été pas adéquate a notre projet parce qu'elles contiennent qu'un ensemble minime de tweets et vue que nous cherchons la fiabilité et que notre classifieur détecte autant de possible donc on à choisis un dataset très puissant et contient 1.6 millions de tweets malgré qu'il contient pas l'étiquette neutre.

➢ Le choix de la langue de tweet n'est pas disponible dans notre système alors que la langue anglaise est régler comme la langue de recherche par défaut.

➢ le manque de correcteur d'orhographe des mots mal écrits dans les tweets que ce soit l'ors de l'apprentissage ou dans la phase de téléchargement des tweets a cause de manque de ressources nécessaire (un processeur très puissant et performant et une grande taille de mémoire vive) .

➢ Nous n'avons pas pu comparer notre modèle avec d'autres modèle existent dans l'apprentissage supervisé à cause de la limite de temps.

- **Perspectives**

Les premiers résultats de cette application semblent prometteurs mais les études dans le domaine doivent être poursuivies, à cet égard et dans la même optique de recherche, on aperçoit de nombreuses pistes qui restent à explorer qui déclarent plusieurs chantiers ouverts:

➢ Tous les étapes de prétraitement faites dans la phase d'apprentissage sont manquantes d'une étape de correction d'orthographe puisque nous avons entrain de traiter des tweets qui peut être avoir des fautes d'orthographe malgré, donc nous devons ajouter ce facteur de prétraitement en considération au futur.

➢ Nous avons des perspectives de développer notre application de sorte qu'on introduit d'autre langues pour la recherche des tweets .

➢ En général, Le modèle Naïve Bayes classent bien dans le domaine de microblogging , et en particulier notre classifieur a amené à des résultats très encourageants avec une précision de 90% mais insuffisants à l'égard des méthodes connues dans la littérature par la qualité de leurs résultats, donc on a besoin dans le futur de exploiter les autres modèle de classification.

4.5 Conclusion

Dans ce chapitre, nous avons présenté l'essentiel de notre travail qui consiste à créer un système de detection de polarité des points de vue dans le réseau social twitter.

Pour l'implémentation, nous avons utilisé l'une des méthodes de classification les plus connue méthode de naïve bayes. en examinant les tweets pour les classifier en: positifs, négatifs .

Notre système s'intègre dans le domaine d'intelligence artificielle précisément " Machine Learning" Car la précision de la classification augmente à chaque fois quand exécute l'algorithme de classification.

Conclusion générale

L’analyse des sentiments et l'opinion Mining est un domaine émergent, dans ces dernières années plusieurs recherches s'intéressent à la tâche de l'analyse du sentiment, en particulier dans le domaine du micro-blogging.

En effet, jusqu’à présent, aucun outil n’est encore arrivé au point de faire une analyse de sentiments parfaite, même pas les êtres humains à cause de la subjectivité et transparence de certains discours. Par raison de difficulté de l’analyse des opinions (humour, ironie, sens caché…. etc..).

Notre travail s'intègre dans ce même axe de recherche, nous proposons un système de Classification subjective des points de vue des utilisateurs du réseau social twitter sur n'importe quel sujet, cette classification faite en deux catégories : positif et négatif on a négligé la catégorie neutre parce que le corpus utilisé dans notre système et très large et annotée en polarité positif et négatif.

Nous avons suivi une méthodologie basée sur l'approche d'apprentissage supervisé, cette méthodologie basé trois axe principaux le premier axe c'est de récupérer ou télécharger un corpus d'entrainements ou d'apprentissage dans notre cas c'est un dataset de 1.6 million de tweets en langue anglaise annotée en deux catégorie positives et négatives et de faire préparer et nettoyer ce dataset de tout types de bruits qu'on a déjà décortiquer dans le chapitre de la conception , le deuxième axe c'est de faire le modèle d'apprentissage dans notre cas nous avons implémenter la méthode de naïve bayes pour classifier les données prétraités en données ou tweets positifs et négatifs ,le troisième axe c'est l'implémentation de notre modèle sur des tweets téléchargées et contenant un terme chercher par l'utilisateur de notre application ou sytéme et le résultat final c'est le nombre de tweets positifs nettoyés et tweets négatifs nettoyés contenant dans l'ensemble des tweets téléchargés .

Références bibliographiques

[1] https://www.cadre21.org/ressources/historique-des-reseaux-sociaux/

[2]: ZEMMAR Nisrine, ″Réseaux Sociaux numériques : essai de catégorisation et Cartographie des controverses, ″Université Rennes 2, Thèse de doctorat 2012.

[3]: HEER Jeffrey, BOYD Danah, ″Vizster: Visualizing online social networks, ″Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on. IEEE, pp. 32-39, 2005.

[4]:LES Z'ED. [En ligne]. http://les-zed.com/ qu'-est-ce-que-les-réseaux-sociaux.

[5]: PASCAL Choteau - Formateur TIC Réseautage social et internet ,mars 2011 [En ligne]. https://reseaufoad.files.wordpress.com/2011/09/le_reseautage_social_sur_internet_sur_lefam_pascal_choteau-mars2011.pdf.

[6]: Les différents types de réseaux sociaux. (2013, 03). Consulté le 04 10, 2020, sur Le web 2.0 et les réseaux sociaux: https://megganeangellotti.wordpress.com/titre-5/

[7] FRAISIER Ophélie, ″Détection de points de vue sur les médias sociaux numériques" EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse, Mémoire présenté En vue de l'obtention du doctorat de l'université de toulouse, le 7 décembre 2018.

[8] FILLIETTAZ Francois, GREGORI Macro, ″Comprendre les réseaux sociaux numériques, Direction des systèmes d'information et service écoles- médias, 2011.

[9] BOYD danah, ″Social NetworkSites: Public, Private, or What?″Knowledge Tree May 2007.

[10] Rosen, A. (0117, 11 tuessdady). Tweeting Made Easier. Consulté le 15/04/2020, sur blog.twitter: https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html

[11] BELBACHIR Faiza, Expérimentation de fonctions pour la d´etection d'opinions dans les blogs, Université de Paul Sabatier, Institut de Recherche en Informatique de Toulouse 2010.

[12] Britannica, T. E. (2009). The History Of Twitter. ENCYCLOPÆDIA BRITANNICA

[13] MacArthur, A. (2020, 04 20). life wire. Consulté le 18/04/2020, sur life wire: https://www.lifewire.com/history-of-twitter-3288854.

[14] Cha, P. G. (2013). Comparing and combining sentiment analysis methods. Boston, Massachusetts: COSN '13 Proceedings of the first ACM conference on Online social networks.

[15] wikihow. (s.d.). Récupéré sur Comment envoyer un message privé sur Twitter: https://fr.wikihow.com/envoyer-un-message-priv%C3%A9-sur-Twitter

[16] Wikipédia. (s.d.). Consulté le 20/04/2020, sur https://fr.m.wikipedia.org/wiki/Sentiment

[17] Techopedia Sentiment Analysis. (s.d.). Consulté le 24/04/2020, sur Techopedia: https://www.techopedia.com/definition/29695/sentiment-analysis

[18] B.Bathelot. Définition : Analyse des sentiments. Consulté le 28/04/2020, sur définition marketing:https://www.definitions-marketing.com/definition/analyse-des-sentiments/

[19] Choudhary, R. N. (2009). Sentiment analysis of conditional sentences. Singapore: Association for Computational Linguistics Stroudsburg, PA, USA ©2009.

[20] Korashy, W. M. (2014). Sentiment analysis algorithms and applications:. Ain Shams.

[21] Levene, A. M. (2012). Combining lexicon and learning based approaches for concept-level sentiment analysis. Beijing, China: WISDOM '12 Proceedings of the Fir st International Workshop on Issues of Sentiment Discovery and Opinion Mining.

[22] Liu, B. (2012). Sentiments analysis and opinig mining. ISBN-10: 1608458849: Morgan & Claypool.

[23] Ziqiong Zhang, Q. Y. (2011). Sentiment classification of Internet restaurant reviews written in Cantonese. chine.

[24] InsightsAtlas. (s.d.). Consulté le 05 01, 2020, sur blog.insightsatlas: https://blog.insightsatlas.com/7 -benefits-of-sentiment-analysis-you-cant-overlook

[25] BELAININE Billal, ″ Classification supervisée de textes courts et bruités : Application au domaine des médias sociaux" , ″ Université du Québec à Montréal, Mémoire présenté comme exigence partielle de la maitrise en informatique ,Avril 2017.

[26] Zhan, X. F. (2015). Sentiment analysis using product review data. Journal of big data.

[27] Amos, E. A.-M. (2016). Using Artificial Neural Networks and. In partial fulfillment of the requirements for the.

[28] El-Beltagy, S. R. (2006). A Phrase and Word Level Sentiment Lexicon for Egyptian and Modern Standard Arabic. Egypt: Nile University Juhayna Square.

[29] Vaithyanathan, B. P. (2002). sentiment classification using machine learning techniques. Stroudsburg: EMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10.

[30] SenticNet : http://sentic.net/about/, Consulté le 06/05/2020.

[31] Gervas, J. C. (2014). An easily scalable concept-based affective lexicon.

[32] Hoffmann, T. W. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. Vancouver, British Columbia, Canada.

[33] Lohard, D. B. (2012). OPINION MINING ET SENTIMENT ANALYSIS. Marseille

[34] Techopedia Sentiment Analysis. (s.d.). Consulté le 09/05/2020, sur Techopedia : https://www.techopedia.com/definition/29695/sentiment-analysis

[35] In NLP, what is the difference between a Lexicon and a Corpus? (Hady Elsahar answer). Consulté le 12/05/2020, sur Quora:https://www.quora.com/In-NLP-what-is-the-difference-between-a-Lexicon-and-a-Corpus.

[36] Choudhary, R. N. (2009). Sentiment analysis of conditional sentences. Singapore: Association for Computational Linguistics Stroudsburg, PA, USA ©2009.

[37] Patrick Paroubek, Twitter as a Corpus for Sentiment Analysis and Opinion Mining Document de conference, janvier 2010.

[38] D.LEWIS David, Naive bayes at Forty : The independence Assumption in Information Retrieval ,Avril 2004.

[39] Hassane HILALI, ″ Application de la classification textuelle pour l'extraction des règles d'association maximal " , mémoire présenter à l'université du Québec à Trois-Rivières comme exigence partielle de la maitrise en mathématiques et informatique appliquées, Avril 2009.

Référence web

:

[Web1] : http://www.booksaresocial.com/history-of-social-media-part-i/

[Web2]: http://www.barriblog.com/2017/10/lo-siempre-quiso-saber-del-api-twitter-nunca-se-atrevio-preguntar-actualizado-2017/, Consulté le 15/04/2020 .

[Web3] : http://books.openedition.org/oep/204, Consulté le 04/05/ 2020.

[Web4] :https://code.tutsplus.com/fr/tutorials/introducing-the-natural-language-toolkit-nltk--cms-28620

[Web5] : https://www.nltk.org/book/ch05.html