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 de diplôme de

Master en Informatique

Option : Ingénierie du logiciel

Thème

Détection d'opinions à partir de Twitter

Réalisé par :

- HASNI Bachir
- GOUDJIL Ayyoub

Encadré par :

Mr. MEGHATRIA Riadh

Soutenu devant le jury composé de :

S.HADJ SADOUK	Président
O.HARBOUCHE	Examinateur 1
N.AZZOUZA	Examinateur 2

Année Universitaire 2017/2018

Dédicace

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 mon frère amine.

A mon ami Chaouch Ibrahim.

A mon ami Benaziza Redhouane.

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

A mon père.

A tous mes amis (es).

BACHIR

Dédicace

Je dédie ce mémoire à :

Mes parents :

Ma mère, qui a œuvré pour ma réussite, de par son amour, son soutien, tous les sacrifices consentis et ses précieux conseils, pour toute son assistance et sa présence dans ma vie, reçois à travers ce travail aussi modeste soit-il, l'expression de mes sentiments et de mon éternelle gratitude.

Mon père, qui peut être fier et trouver ici le résultat de longues années de sacrifices et de privations pour m'aider à avancer dans la vie. Puisse Dieu faire en sorte que ce travail porte son fruit ; Merci pour les valeurs nobles, l'éducation et le soutient permanent venu de toi.

Mes frères qui n'ont cessé d'être pour moi des exemples de persévérance, de courage et de générosité.

AYOUB

Remerciements

Avant tout, nous remercions Dieu Tout-Puissant de nous avoir donné la force et le

Courage de faire ce travail avec altruisme.

Nos vifs sincères remerciements

S'adressent spécialement à,

MR.MEGHATRIA RIADH

Dont nous avons eu la chance de l'avoir comme Professeur, Encadreur et qui a bien voulu nous a confié ce riche travail d'expérience et nous a guidés à chaque étape de sa consécration. Vous m'avez toujours réservé un chaleureux accueil, malgré vos obligations et les contraintes professionnelles. Vos talents ainsi que vos compétences et votre sens du devoir m'ont marqué à jamais. Vos encouragements inlassables, votre amabilité, votre gentillesse et votre patience méritent toute notre attention. Veuillez trouver ici l'expression de notre estime et considération.

Nous remercions aussi,

Les membres du jury de nous avoir fait l'honneur de juger cette thèse. Veuillez accepter l'expression de notre profonde gratitude.

Enfin,

A toutes les personnes qui ont contribué de prés ou de loin, d'une manière directe ou indirecte à l'élaboration de ce travail de fin d'études.

Sommaire

Liste des abréviations

Cnet : Computer Network

Ebay : Electronic bidding all year

MSN : Microsofrt Network

CV : Curriculum Vitae

UMP : Union Mouvement Populaire

SMS : Short Message Service

Inc : Incorporated

Ei : le nom d'une entité

Aij : un aspect d'Ei

Hk : détenteur de l'opinion

Tl : le temps où l'opinion a été exprimée

NB : Naïve Bayes

SVM : Support Vector Machine

POS : Part Of Speech

MPQA : Corpus d'opinion

Net : network

Trec : Text Retrieval Conference

NIST : National Institute of Standard and Technology

DASA : Dissatisfaction-oriented Advertisingbased on Sentiment Analysis

K : K-moyennes

EM : Espérance-Maximisation

HTML : Hyper Text Markup Lanuage

URL : Uniform Resource Locator

CC : Conjonction de Coordination

RB : adverbes

IN : une préposition

NN : un nom

JJ : un adjectif

Pos : Positive

Neg : Negative

Obj : Objective

Ibm : International Business Machines

Json : Java Script Object Notation

NLTK : Natural Language Toolkit

API : Application Programming Interface

Càd : c'est-à-dire

Gpl : General Public License

Mac : Macintosh

Liste des tableaux

Liste des figures

ملخص :

تحليل المشاعر أو دراسة الرأي هو الدراسة الحسابية لآراء الناس, المشاعر, والمواقف التي يعبر عنها في النصوص المكتوبة.

تحليل المشاعر هو واحد من المجالات البحثية الأكثر نشاطًا في معالجة اللغات الطبيعية ودراسة النصوص في السنوات الأخيرة. يعود سبب شعبيتها بشكل رئيسي إلى مجموعة واسعة من التطبيقات لأن الآراء هي تقريبا محور جميع الأنشطة البشرية لاسيما في وسائل الإعلام الاجتماعية.

وبالتالي ليس من المستغرب أن يتزامن ذلك مع النمو السريع للنطاق خاصة بالشبكات الاجتماعية وعلى الويب. في الواقع انتشر البحث أيضًا خارج وسائل الإعلام الاجتماعية وعلوم الكمبيوتر إلى علوم الإدارة والعلوم الاجتماعية نظرًا لأهميتها بالنسبة إلى قطاع الأعمال والمجتمع.

في هذا التقرير ركزنا بشكل أساسي على تحليل التويت المختلفة باستخدام مجموعة بيانات عشوائية أو أي نص عشوائي. استخدمنا طريقة تعتمد على القاموس وهي طريقة بسيطة وفعالة حيث تم استخدامها منذ أجيال. وقد ثبت أنها موثوقة عندما يتعلق الأمر في تحليل المشاعر والآراء على أي تويت معينة وقد.قمنا أيضا بتنفيذ بعض التقنيات وبعض الخطوات الرئيسية للمعالجة المسبقة لنظامنا للحصول على دقة أفضل وأداء أعلى.

الكلمات المفتاحية : تحليل المشاعر ، الشبكات الاجتماعية ، الرأي.

Résumé :

L'analyse des sentiments ou l'analyse des opinions est l'étude computationnelle des opinions, des sentiments, des attitudes et des émotions des personnes exprimées dans le langage écrit.

L'analyse des sentiments est l'un des domaines de recherche les plus actifs dans le traitement du langage naturel et l'analyse des textes ces dernières années. Sa popularité est principalement due à la gamme large d'applications, car les opinions sont au cœur de presque toutes les activités humaines, en particulier dans les médias sociaux.

Il n'est donc pas surprenant que la création et la croissance rapide du domaine coïncident avec celles des médias sociaux sur le Web. En fait, la recherche s'est étendue aux sciences de gestion et aux sciences sociales en dehors des médias sociaux et de l'informatique en raison de son importance pour les entreprises et la société dans son ensemble.

Dans ce rapport, nous nous sommes concentrés principalement sur l'analyse de différents tweets en utilisant un ensemble de données aléatoires ou tout autre texte aléatoire. Nous avons utilisé une méthode basée sur un dictionnaire, qui est une méthode simple et efficace utilisée depuis des générations, pour analyser les sentiments et les opinions d'un certain tweet. Nous avons aussi implémenté quelques techniques et certaines des principales étapes de prétraitement de notre système pour une meilleure précision et une meilleure performance.

Mots clés : analyse des sentiments, réseaux social, opinion.

Abstract:

Sentiment analysis or opinion mining is the computational study of people's opinions, sentiments, attitudes, and emotions expressed in written languages.

Sentiment analysis is one of the most active research areas in natural language processing and text mining in recent years. Its popularity is mainly due to the wide range of applications because opinions are central to almost all human activities, especially in social media.

It is thus no surprise that the inception and the rapid growth of the field coincide with those of the social media on the Web. In fact, the research has also spread outside of social media and computer science to management sciences and social sciences due to its importance to business and society as a whole.

In this report, we focused mainly on analysing different tweets using random dataset or any random text for that matter. We used a dictionary based method which is a simple yet effective approach that has been used for generations and it has been proved to be quite reliable for analysing sentiments and opinions from a certain tweet. We also implemented some technics and some of the main preprocessing steps to our system for a better accuracy and a higher performance.

Keywords: Sentiment analysis, social media, opinion.

Introduction générale

Les opinions des autres ont une influence significative sur notre processus de prise de décision quotidien. Ces décisions vont de l'achat d'un produit comme un téléphone intelligent à des investissements dans le choix d'une école - toutes les décisions qui touchent divers aspects de notre vie quotidienne. Avant Internet, les gens cherchaient à obtenir des opinions sur des produits et services provenant de sources telles que des amis, des parents ou des rapports de consommateurs.

Cependant, à l'ère de l'Internet, il est beaucoup plus facile de recueillir des opinions diverses de différentes personnes à travers le monde. Les gens cherchent à passer en revue des sites (par exemple, CNET, Epinions.com), des sites de commerce électronique (Amazon, eBay), des sites d'opinion en ligne (TripAdvisor, Rotten Tomatoes, Yelp) et des médias sociaux (Facebook, Twitter). Obtenir des commentaires sur la façon dont un produit ou service particulier peut être perçu sur le marché.

De même, les organisations utilisent les enquêtes, les sondages d'opinion et les médias sociaux comme moyen d'obtenir des commentaires sur leurs produits et services. L'analyse des sentiments ou l'exploration de l'opinion est l'étude computationnelle des opinions, des sentiments et des émotions exprimés dans le texte, analyse des sentiments est utilisée dans plusieurs domaines tel que le marketing, la médecine, la politique ...

Actuellement, l'un des meilleurs exemples de réseaux sociaux permettant d'observer l'évolution de ces opinions est Twitter, Dans notre projet, on va analyser des différents tweets sur le réseau social Twitter.

1 Chapitre 1 : Etat de l'art

Introduction

Dans cette section nous allons aborder les différentes concepts et définitions comme les réseaux sociaux. En se basant particulièrement sur twitter, l'analyse des données, et les différentes méthodes et travaux qui existent en faisant une petite comparaison sur les concepts déjà abordés.

1.1 Les réseaux sociaux

Les réseaux sociaux sont une notion ambigüe car au sens strict, un réseau social désigne les différentes relations que les individus entretiennent entre eux et la façon dont celles-ci se structurent; ces différentes relations permettent de comprendre les comportements des individus [1].

1.1.1 Historique des réseaux sociaux

Le réseau social « Classmates » fut créer en 1995, il est catégorisé comme l'un de premiers réseaux sociaux puisqu'il permettait de retrouver des camarades de classes mais il n'apporte bien entendu pas du tout les mêmes possibilités des réseaux sociaux que l'on connait aujourd'hui. S'en est suivi en 1996 de l'apparition du site « SixDegrees » qui permettait de créer un profil et ses listes d'amis (un peu à la Facebook des temps anciens). A partir de 1998 il était possible de naviguer dans ces fameuses listes d'amis [4].

Entre 1997 et 2003, il y eut un affolement pour les réseaux sociaux. Il existait même des réseaux sociaux communautaires destinés à des communautés spécifiques comme par exemple « AsianAvenue » (communauté asiatique), « BlackPlanet » (communauté noire) et enfin « MiGente » (communauté latino) [4].

Le 22 juillet 1999 « MSN Messenger » débarque. Service de messagerie instantanée il connaitra un immense succès sur le web. Il était possible de discuter, de s'envoyer des photos ou d'interférer par webcam. En 2005, MSN devient Windows Live Messenger et c'est en 2013 que

Ce logiciel de messagerie instantanée tira enfin sa révérence [4] En mai 2003 le site « LinkedIn » est lancé sur internet. En clair, c'est un service en ligne qui catégorise tel un CV vos compétences, formations et expériences professionnels à la vu d'autres individus. Tout cela pour vous créer un réseau de connaissances qui peut vous faire avancer dans le monde professionnel [5]. La même année « **MySpace** » voit le jour. Les premiers utilisateurs de ce réseau social fut des groupes de rocks indépendants dans la région de Los Angeles. Par la suite, cela fut boule de neige et attira d'autres groupes de musiques populaires. Au fil du temps, MySpace permit le contacte entre les fans et artistes, jusqu'à devenir le réseau social atypique pour les adeptes d'artistes musicales. Aujourd'hui encore, MySpace est un réseau social qui compte plus 125 millions de visiteurs par mois [4].

Entre temps durant l'année 2004, Mark Zuckerberg était encore sur les bancs de l'université. Il fonde le plus gros réseau social à ce jour, Facebook. Au départ, c'était un réseau fermé réservé aux étudiants d'Harvard. Mais en 2005 Facebook élargira son audience jusqu'à que des inscriptions se fasse dans le monde entier. Aujourd'hui, Facebook est l'un des premiers sites de réseau social généraliste au monde avec ses plus d'un milliard d'utilisateurs dans le monde.

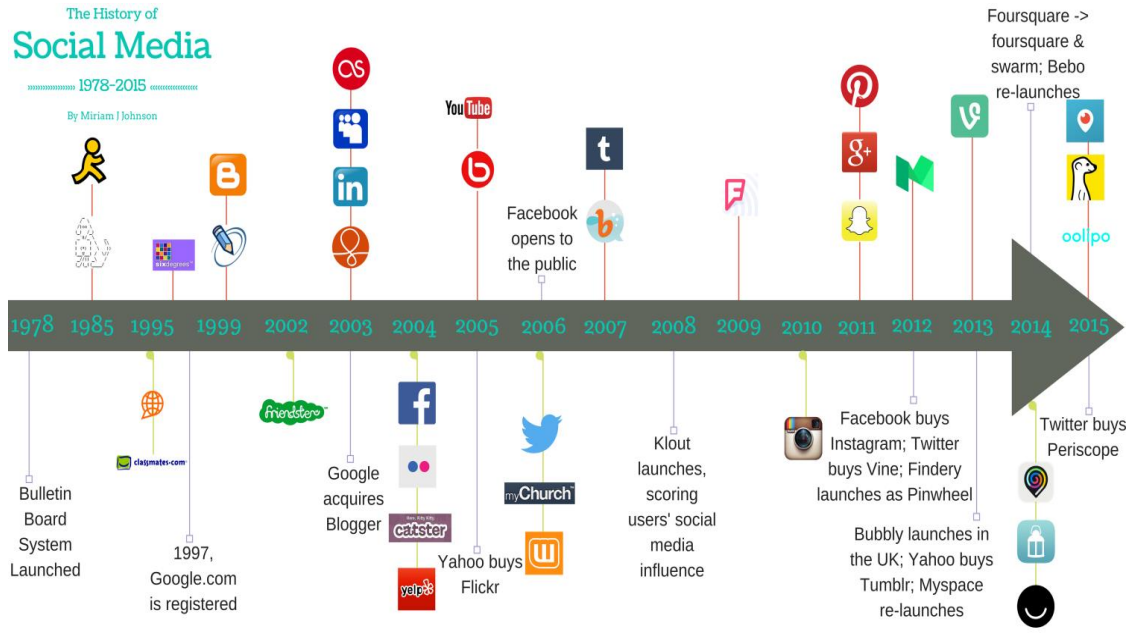

Figure 1-1: Enchainement des réseaux sociaux 1978-2015 [Web1].

1.1.2 Définition des réseaux sociaux

Dans le cadre d'une approche marketing, le terme de réseaux sociaux désigne généralement l'ensemble des sites internet permettant de se constituer un réseau d'amis ou de connaissances professionnelles et fournissant à leurs membres des outils et interfaces d'interactions, de présentation et de communication [2].

L'ère numérique et l'apparition des réseaux sociaux n'apportent rien de nouveau aux besoins initiaux, ils viennent plutôt accélérer la transmission de l'information et réduire le temps de réponse [3].

Les réseaux sociaux les plus connus sont Facebook, Twitter, Linkedin, Viadeo, Pinterest, etc. Youtube peut également être considéré partiellement comme un réseau social dans la mesure où le service a développé des outils d'interactions entre ses membres [2].

Voici quelques statistiques sur le réseau social:

- Répartition par âge :

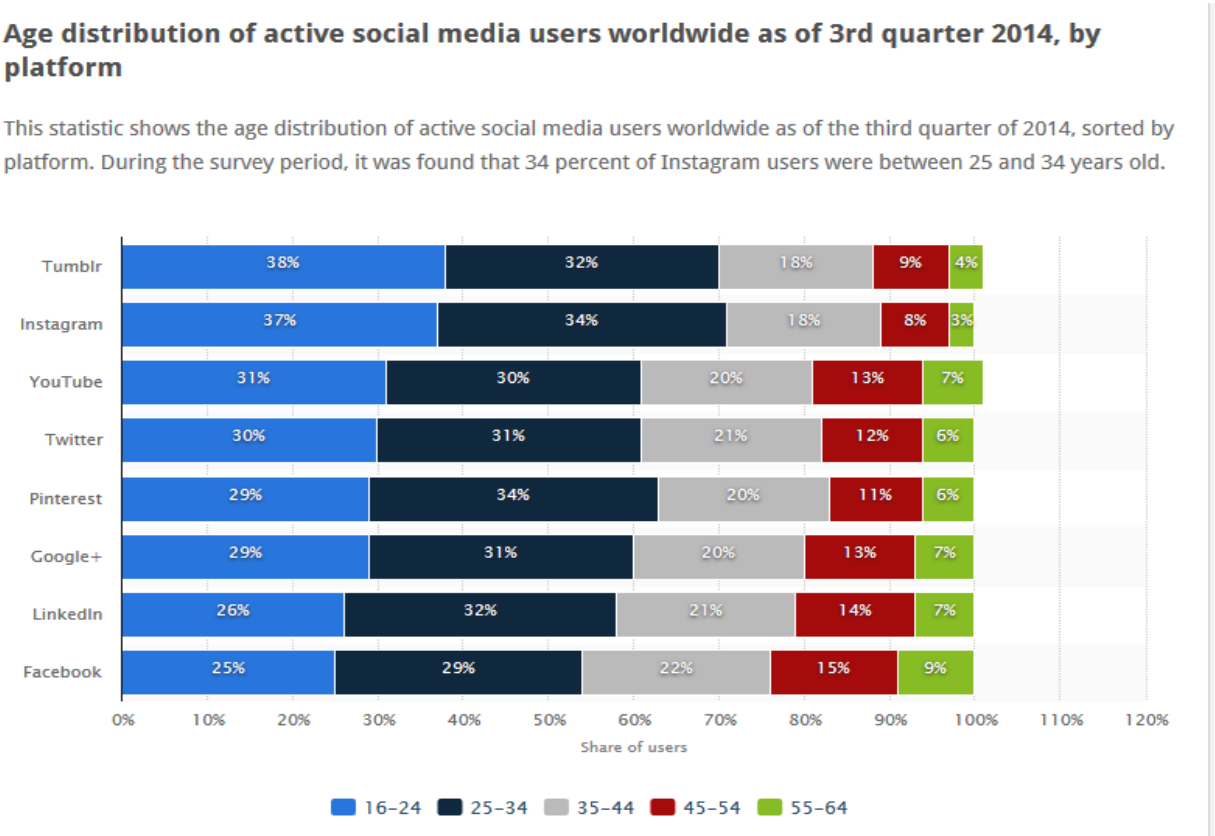

Figure 1-2:Répartition des réseaux sociaux par âge dans le monde [Web2].

- Repartions par régions :

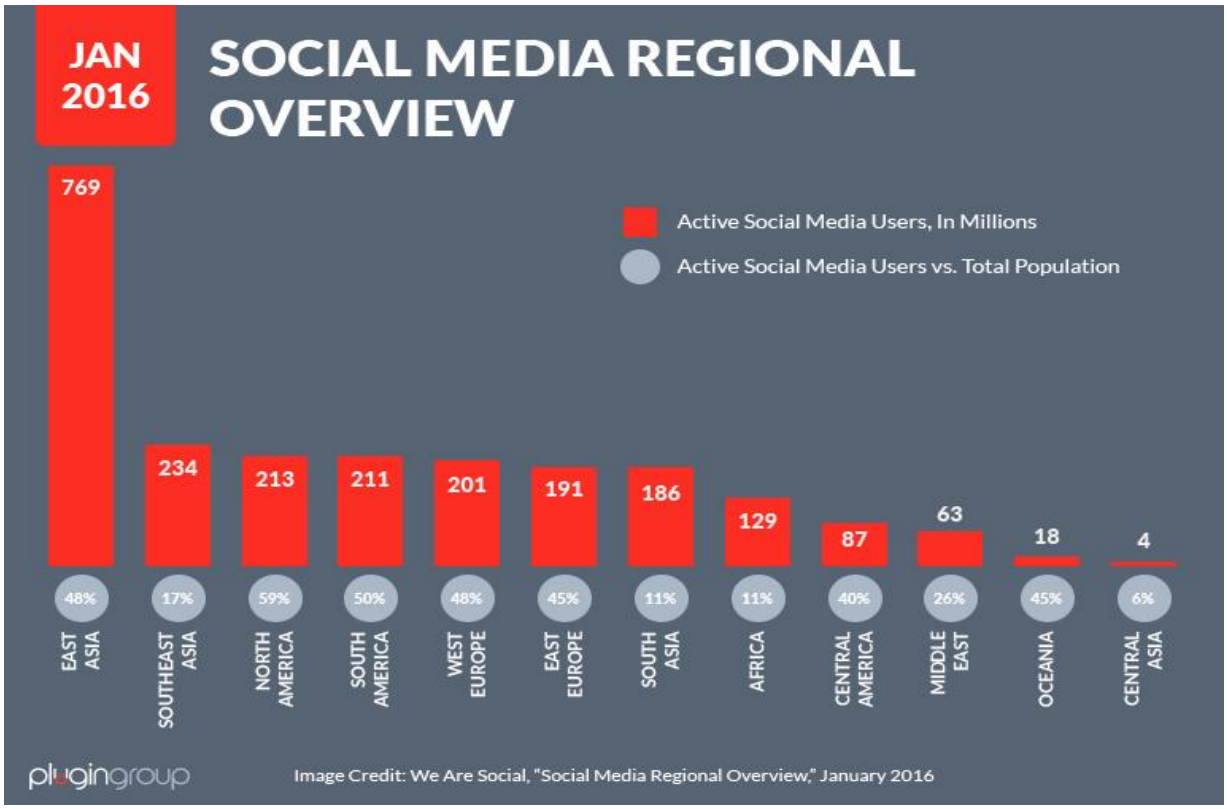

Figure 1-3:Réparation des réseaux sociaux par régions [Web3].

- Statistiques des utilisateurs :

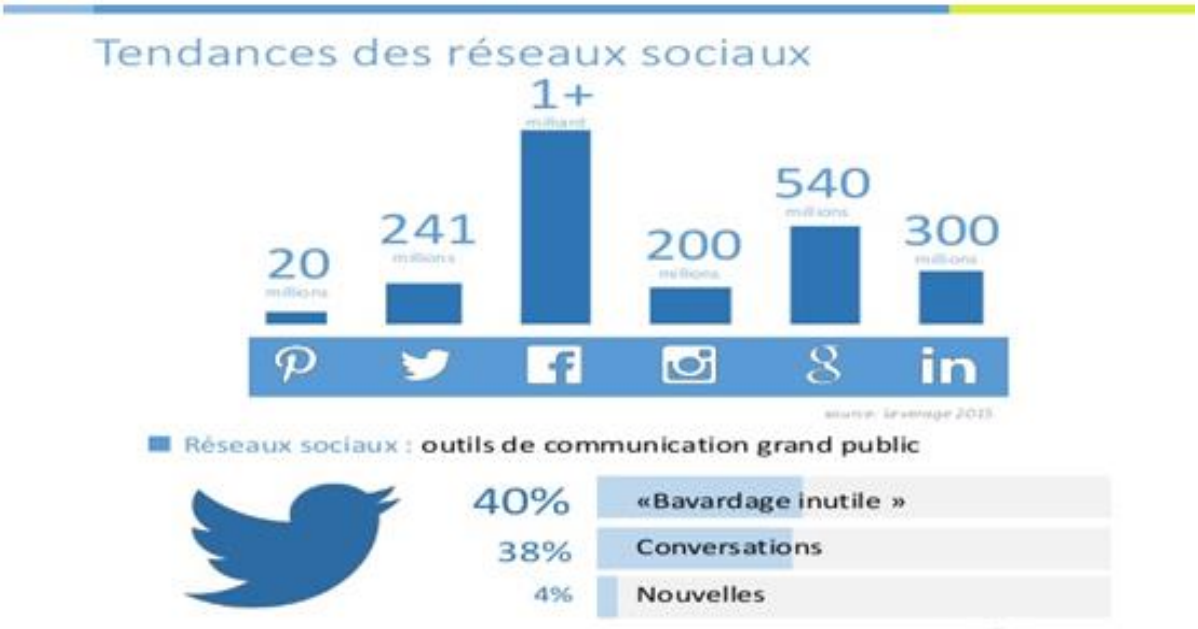

Figure 1-4:Statistiques d'utilisateurs des réseaux sociaux [Web4].

1.1.3 Différent types de réseaux sociaux

Le monde des réseaux sont très diversifiés, il existe de ce fait plusieurs plateformes de réseaux sociaux, parmi ces plateformes, il faut distinguer deux catégories, ceux à usage exclusivement professionnel, orienté sur la mise en valeur et les échanges professionnels de ses membres, et ceux à usage privé, ceux qui sont devenus grand-public comme MySpaces (construit au départ pour favoriser la mise en relation d'artistes) ou Facebook (conçu à l'origine par et pour des universitaires) [6].

1.1.3.1 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.
- Twitter : 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 : [7]

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

1.1.3.2 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. [7]
- InterFrench : réseau francophone mondial pour les projets à l'étranger.
- Piwie : le premier chat business.

1.1.3.3 Les réseaux sociaux dits de service

- Ma-residence : 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.
- RéseauxLycée 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. [7].

1.1.3.4 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 évenements, débats, partage d'idées, de liens, de vidéos…
- Créateurs de possible : c'est les réseaux social de l'UMP (Union pour un mouvement populaire) lancé en janvier 2010. Propose des fonctionnalités similaires à Coolpol [7].

1.2 La plate-forme de microblogging Twitter

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 [56].

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 » [6].

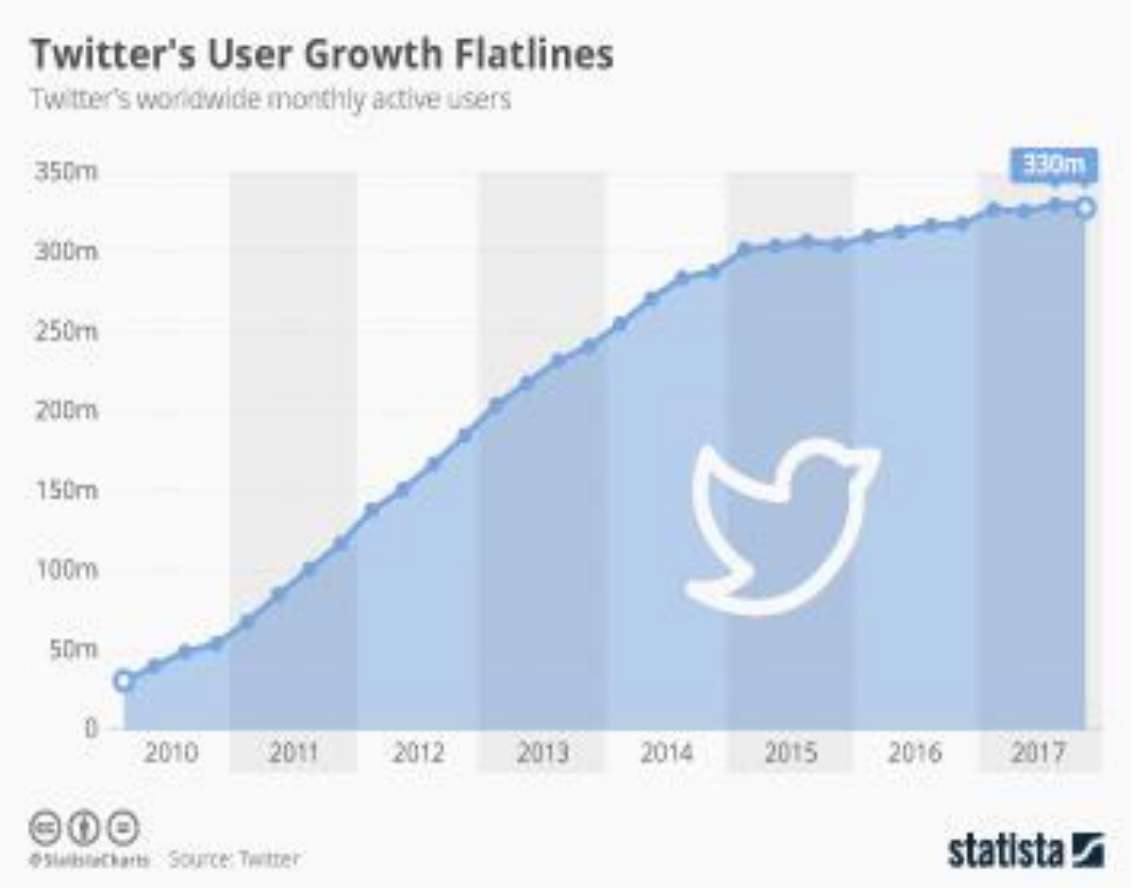

Figure 1-5:Statistique d'utilisateurs twitter[Web5].

1.2.1 Historique

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 Southwest à Austin, au Texas, en mars 2007. Le mois suivant, Twitter Inc. a été créé en tant qu'entité corporative de capital-risque [26].

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 [27].

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.

1.2.2 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 [22].

- **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 [23].

- **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.

1.3 Définition

Dans le domaine de l'analyse des sentiments, les gens confondent souvent une opinion avec un sentiment, c'est pour cela que nous devrions définir brièvement chaque terme.

1.3.1 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 [42].

Le dictionnaire Larousse de Poche 2017 définit le sentiment comme étant un état affectif complexe et durable lié à certaines émotions ou représentations.

1.3.2 Opinion

Une opinion c'est un Jugement, avis, sentiment qu'un individu ou un groupe émet sur un sujet, des faits, ce qu'il en pense [43].

Une opinion comme un quintuple, (Ei, Aij, Sijkl, Hk, Tl), où Ei est le nom d'une entité, Aij est un aspect de Ei, Sijkl est le sentiment sur l'aspect Aij de l'entité Ei, Hk est détenteur de l'opinion, et Tl le temps où l'opinion a été exprimée par Hk. Sijkl est positive, négative ou neutre, ou peut être exprimé par des niveaux d'intensité [44].

1.3.2.1 Types d'opinion

Pouvant être positive, négative ou neutre, l'opinion en tant qu'information de base est retrouvée en deux catégories principales [12] :

a) Opinion usuelle ou comparative

Opinion usuelle : Une opinion usuelle est une expression simple d'un avis, pouvant viser de manière directe ou indirecte le sujet principal.

Les opinions directes sont actuellement les plus exploitées dans la plupart des études de recherches en *opinion mining* pour leur simplicité et facilité à déterminer les différentes parties de l'opinion elle-même.

Opinion comparative : L'opinion comparative représente l'expression d'un avis sur un sujet en le comparant à un autre. Cette relation de comparaison aide à déterminer la valeur de l'opinion émise à l'égard du premier sujet.

b) Opinion explicite ou implicite

Opinion explicite : Une opinion explicite est souvent un avis subjectif, exprimé de manière simple ou à travers une comparaison. Opinion implicite Une opinion implicite est un avis généralement objectif qui sous-entend l'expression d'une opinion usuelle ou comparative.

Cette catégorie d'opinion étant plus difficile à déterminer, et clairement moins explorée de par le nombre d'études à son compte.

1.4 Analyse des sentiments

En informatique, sentiment analysis aussi appelé (l'opinion mining) est l'analyse des sentiments à partir de sources textuelles dématérialisées sur de grandes quantités de données (big data).

Ce procédé apparait au début des années 2000 et connait un succès grandissant dû à l'abondance de données provenant de réseaux sociaux, notamment celles fournies par Twitter.

L'objectif de l'opinion mining est d'analyser une grande quantité de données afin d'en déduire les différents sentiments qui y sont exprimés. Les sentiments extraits peuvent ensuite faire l'objet de statistiques sur le ressenti général d'une communauté [17].

1.4.1 Définition de 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 du sentiment est également connue sous le nom d'extraction d'opinion (opinion mining).

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[46].

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. [8].

Les sentiments sont généralement classés en trois types : négatifs, neutres ou positifs.

1.4.2 Historique de l'analyse des sentiments

Dans cette partie nous présentons un bref historique de l'analyse des sentiments. Nous nous somme basés sur l'article [28].

Hatzivassiloglou et McKeown en 1997, travaillaient au niveau de document (corpus critiques de films) et utilisaient " World Street Journal " comme source de données. Leurs travaux se basent sur les conjonctions et les adjectives et créent un modèle de Log Linear Regression. Dans le même niveau document Pang et al. En 2002 effectuaient une analyse avec des modèles d'apprentissage Naïve Bayes (NB), Support Vector Machine (SVM), Maximum Entropy (ME). Ils ont utilisée Unigram, bigram, effet contextuel de la négation, et les fréquences. Ils ont appliqué de tels modèles sur les critiques des films. Nous citons aussi autres travaux sur l'analyse des sentiments au niveau documents : Das et Chen en 2001, Turney 2002, Morinaga et al 2002, Turney et Littman 2003 et Pang et Lee 2004.

Nigam et Hurst en 2004, travaillaient au niveau des expressions en se basant sur le lexique des phrases polaires et leurs parties du discours (POS Part Of Speech) avec un modèle basé sur des règles syntaxiques en utilisant Usenet[1] message board et autres sources en ligne comme source de donnes.

Au niveau des phrases Wilson et al en 2005, développent un modèle nommé BoosTexter qui s'intéresse à la subjectivité lexique avec un corpus de MPQA[2].

Dans les niveaux de phrase et expression nous pouvons citer aussi Aue et Gamon en 2005, les descripteurs (méthodes combinées) utilisés sont les termes lemmatisés, leur fréquences et poids, Popescu et Etzioni en 2005, proposent un modèle de dépendance syntaxique qui se bases sur les conjonctions et disjonctions de WordNet[3], Cesarano en 2006, mis en jeux les modèles sur POS et N − grams.

Hu et Liu en 2005, d'enveloppaient un modèle d'extraction d'opinions à base de mots et extension d'agrégation avec WordNet. Les descripteurs utilisés sont les mots d'opinion et les phrases d'opinions. Ils utilisaient Cnn.net de Amazon comme source de donnés. En 2007 Godbole et al, ont développée un modèle lexicale à base de WordNet.

Ils ont utilisé des descripteurs qui se basent sur des mesures de distance dans le graphe entre les mots en tenant compte les relations de synonymie, anonyme, commun des mots.

[1] Usenet est un système de discussion distribué dans le monde entier disponible sur les ordinateurs
[2] Le Corpus d'Opinion MPQA contient des articles de presse provenant d'une grande variété de sources d'information annotées manuellement pour des opinions et d'autres états privés
[3] WordNet est une grande base de données lexicale de l'anglais. Les noms, les verbes, les adjectifs et les adverbes sont regroupés dans des ensembles de synonymes cognitifs (synsets)

Les sources de données utilisées sont les journaux et les publications. Ferguson et al, à l'année 2009, utilisaient les articles de blog financiers comme source de données pour créer un modèle Multinomial Naïve Bayes (MNB) en manipulant des vecteurs binaires des descripteurs au niveau des phrases.

Au niveau des mots, Melville et al en 2009, effectuaient une classification bayésienne avec les lexiques et documents d'apprentissage en utilisant les post des blogs, site des avis, des blogs politiques et des critiques des filmes.

Concernant l'analyse des sentiments sur Twitter, Pak et Paroubek en 2010, Barbosa et Feng en 2010, ont travaillé au niveau des phrases des messages Twitter. Les premiers auteurs utilisent les N−gram[4] et POS−tags[5] comme des descripteurs avec le classificateur MNB, tandis que les derniers utilisent les retweets, hashtags, liens et le point ponctuations en conjonction avec les descripteurs comme les polarités à priori des mots ainsi que les POS de mots pour créer le modèle SVM.

1.4.3 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 [20] :

1.4.3.1 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") [18].

1.4.3.2 Ambiguïté de sentiment

Une phrase avec un mot positif ou négatif n'exprime pas nécessairement un sentiment.

[4] N−gram est une chaîne d'éléments (tels que des lettres, des mots ou des phonèmes) qui apparaît dans une séquence plus longue
[5] POS−tags est le processus qui consiste à associer aux mots d'un texte les informations grammaticales correspondantes

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.

1.4.3.3 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).

1.4.3.4 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 Coke"? 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.

1.4.3.5 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!» Vs «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 américain 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.

1.4.4 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 [9] :

- 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.

Voici quelques travaux de quelques universités et laboratoires de recherche qui se sont basés sur le lexique pour déterminer l'opinion

L'université de Glasgow a été exposée et présentée un article en 2007, Deux approches pour la recherche d'opinion, l'un d'eux est L'approche basée sur le lexique[6].

Cette université a participé à la campagne de TREC[7] 2007 dans plusieurs tâches notamment celle de la recherche d'opinion. Les auteurs utilisent dans ce contexte là une approche basée sur le lexique.

1.4.5 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.

L'infographie ci-dessous [25], présente quelques domaines d'application dans le domaine d'analyse des sentiments.

[6] L'approche basée sur le lexique consiste à calculer l'orientation d'un document à partir de l'orientation sémantique de mots ou d'expressions dans le document.

[7] Text Retrieval Conference désigne l'ensemble des conférences organisées par le NIST «National Institute of Standard and Technology» sur la recherche d'informatio

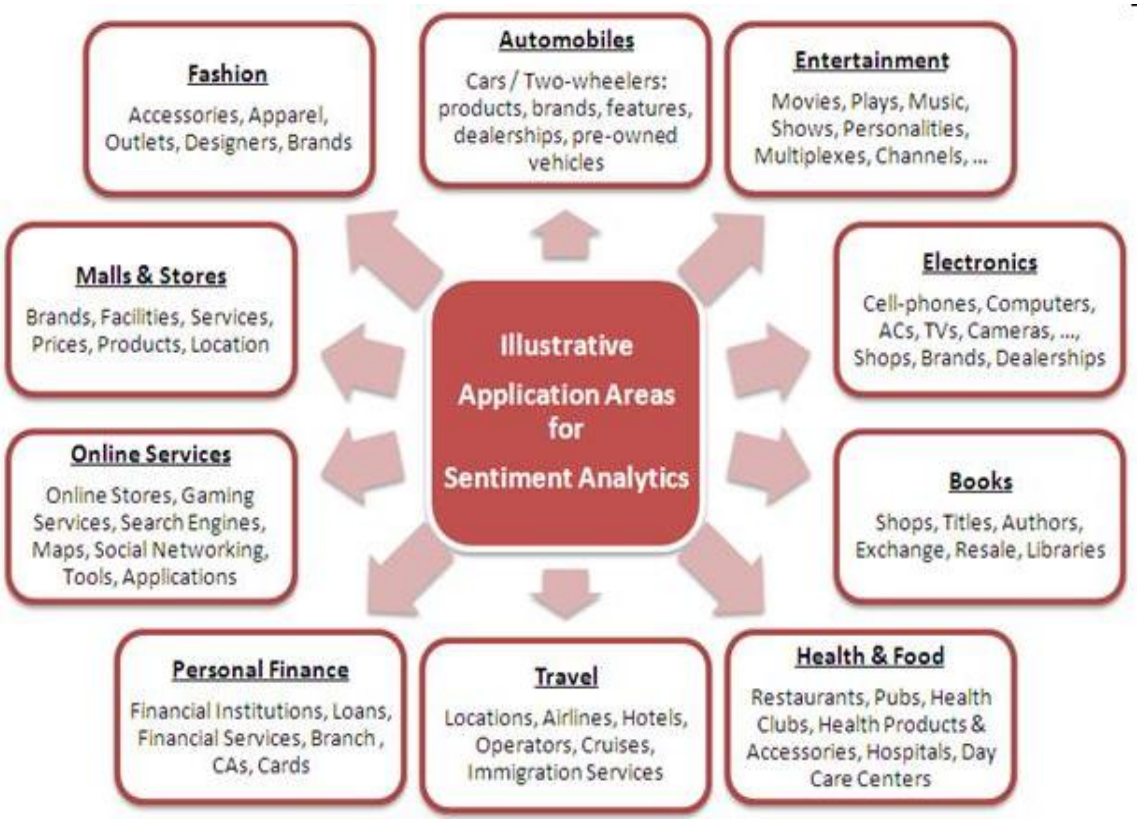

Figure 1-6: Domaines d'application d'analyse des sentiments [Web6].

Les principales applications d'analyse des sentiments sont les suivantes [24] :

1.4.5.1 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 structurée et compréhensible.

1.4.5.2 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 avec leurs goûts.

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 [47].

1.4.5.3 Dans le domaine 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*, qui a pointé vers une victoire de Trump avant même que les votes aient été jetés.

1.4.5.4 Domaine de la publicité en ligne

Si la publicité en ligne, ciblée et contextuelle, s'est considérablement développée ces dernières années, elle 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.

1.4.5.5 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.

1.5 **Outils d'analyse**

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 de recherche nous nous intéressons aux travaux traitant les opinions ordinaires, qui sont classées par polarité à savoir positive, négative ou neutre Dans la suite, nous détaillerons celles-ci dans la suite du rapport. Les trois approches qui existent sont:

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

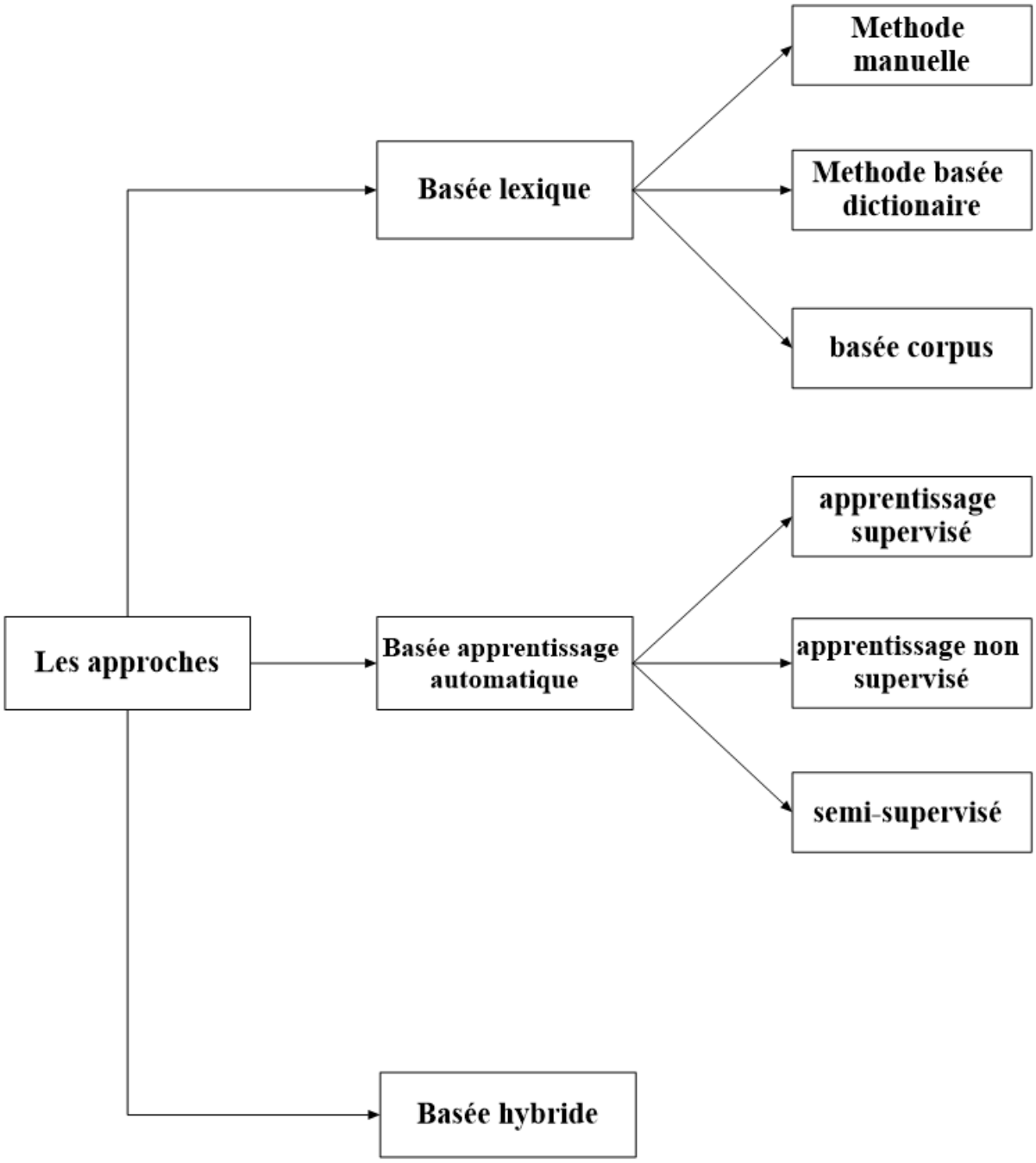

Figure 1-7: Hiérarchie des approches d'analyse.

1.5.1 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: l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage semi-supervisé.

1.5.1.1 Apprentissage supervisé(Classification)

L'apprentissage supervisé (ou classification) consiste à construire un modèle basé sur un jeu d'apprentissage et des labels (nom des catégories ou des classes) et à l'utiliser pour classer des données nouvelles [29].

Cette technique est utilisée dans plusieurs 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 utilisés pour la classification supervisée telles 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 Sparns) pour séparer les bons courriels des mauvais [29]. 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) [29].

Diverses études ont montré que Naive Bayes et SVM sont les meilleurs pour les microblogs.

Par exemple l'étude de Fang en 2015 [50], 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 [29].

Essam Al-Mansouri Sean Amos [51] 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 [29].

En particulier dans la fouille d'opinions, ils ont l'avantage d'être lisibles et très faciles à comprendre et à interpréter ce qui est une des raisons de leurs succès [49].

- 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 [29].

1.5.1.2 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 [29] :

- 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.
- Fuzzy KMeans : 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 [29].

1.5.1.3 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 [29].

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

1.5.2 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[52] 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 :

1.5.2.1 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.

1.5.2.2 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 [19].

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. [13].
- 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 [15] : il s'agit également d'un travail basé sur WordNet permettant de polariser les mots de façon plus précise.
- Subjectivity Lexicon [16] : 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 [17] 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.

1.5.2.3 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 [19].
Voici la différence principale entre Methode basée corpus et methode basée lexique [48]:

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...)

1.5.3 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 [20] 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 [21], 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%.

1.6 Conclusion

Nous avons vu plusieurs travaux sur la fouille d'opinions et énuméré toutes les approches existantes. Il est à noter que certains travaux dans l'approche basée lexique travaillent avec une méthode dictionnaire et d'autres avec les mots du corpus. L'avantage de la méthode dictionnaire est d'englober un grand ensemble de mots mais elle a l'inconvénient d'être indépendante du domaine et du contexte.

Globalement, nous avons remarqué dans les travaux existants que dans l'approche basée lexique, les meilleurs résultats sont obtenus quand une force de polarité est donnée aux mots du vocabulaire et quand la négation[8] est considérée.

L'approche basée apprentissage automatique donne d'excellents résultats grâce aux algorithmes sophistiqués employés pour construire le modèle de classification. Cela fait que cette approche s'adapte aux mots employés dans le corpus. Cependant, son inconvénient majeur est la nécessité de l'annotation manuelle, ce qui est difficile à réaliser dans de grand corpus et ce qui est très problématique étant donné que théoriquement, tout dépend de ce que vous travaillez avec.

Dans le chapitre suivant, nous allons décrire, processus par processus, la conception de notre projet sur la plate-forme de microblogging Twitter.

[8] Négation est une construction linguistique très commune qui affecte la polarité des mots, donc la négation inverser fondamentalement la valeur d'un mot. Exemple: good=+1, not good= -1.

2 Chapitre 2 : conception

Introduction

Pour la conception de ce système nous avons choisi l'utilisation d'une approche basée dictionnaire et étudier les opinions sur le réseau social Twitter. En d'autres termes, nous utilisons un dictionnaire pour déterminer la polarité d'une opinion dans un tweet. Nous utilisons également des dictionnaires pour les prétraitements que nous devons effectuer sur les textes.

Le processus général de la méthodologie du système à été utilisé pour faciliter la compréhension de la conception.

2.1 Description des données

Cette section est consacrée à la description des données utilisées en entrée et à celles des résultats produits.

2.1.1 Données en entrée

Nous avons choisi, comme mentionné précédemment, de nous concentrer sur le réseau social Twitter. Donc l'entrée est un texte court (tweet).

2.1.1.1 Tweet

Les données en entrée que nous exploitons sont des textes qui représentent des statuts postés sur le réseau social Twitter.

2.1.1.2 Caractéristiques d'un Tweet

Un tweet est accompagné de plusieurs informations :

- Le nom de l'utilisateur, son pseudonyme et son image de profil.
- La date et l'heure d'émission du tweet.

• Le nombre de "j'aime" et de retweet.

• Le lieu où a été posté le tweet.

2.1.2 Données en sortie

La sortie est un texte et sa polarité (positive, négative ou neutre). L'algorithme traite l'entrée et nettoie le bruit en utilisant les étapes de prétraitement.

On attribue à chaque tweet une opinion sous forme de valeur numérique, mesurant la polarité de cet avis.

2.2 Architecture du système

Les processus composant ce système sont présentés dans la figure 2-1, représentant le processus général de la méthodologie du système.

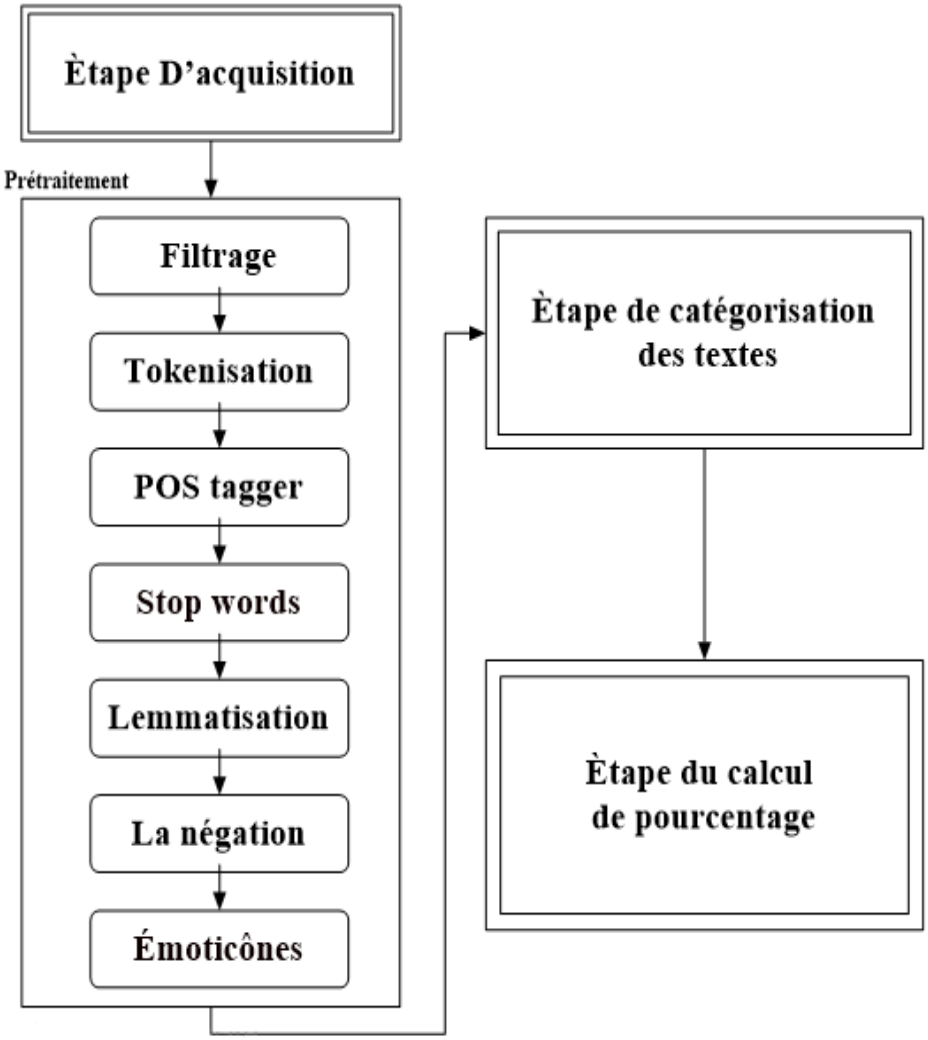

Figure 2-1: Processus général de la méthodologie du système.

2.2.1 Étape d'acquisition

Ce processus est chargé d'insérer des données (tweets) soit manuellement ou à travers une base de données. Notre système fonctionne uniquement avec des tweets en anglais.

2.2.2 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].

- Objectif de prétraitement : 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]

- Les phases de prétraitement : 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.

Figure 2-2: Les processus de prétraitement.

2.2.2.1 Filtrage

L'opération de filtrage a pour but la suppression des métadonnées contenues dans les messages, nous pouvons citer comme exemples [35] :

- **Liens URL** : si le message contient un lien URL (exemple : http://bit.ly/KCairo) l'opération de filtrage le supprime, étant donné qu'il ne contient aucune information qui influe sur l'opinion exprimée dans le message.

Figure 2-3: Exemple d'un tweet contenant lien url [Web7].

- **Les symboles** : dans le cas des tweets et des posts, souvent les utilisateurs utilisent des symboles (exemple : le hashtag « # ») pour mettre en évidence un mot précis, ces symboles doivent être supprimés de façon à pouvoir utiliser les mots par la suite.

Figure 2-4: Exemple de Tweet avec des hashtags [Web8].

Après analyse du tweet, le processus de filtrage supprime le hashtag

Tweet devient : (FIFAarrests FIFAgate) a la place de #FIFAarrests #FIFAgate

- **Les noms d'utilisateurs** : nous pouvons les détecter à travers les liens hypertexte, sur les tweets comme pour les posts. Les utilisateurs se servent du symbole « @ » afin de mentionner d'autres utilisateurs ou pages.

Pour le cas de twitter, le symbole apparait dans le tweet donc il est facile de supprimer le nom d'utilisateurs. Par contre pour Facebook, l'opération doit tenir compte des liens hypertextes qui indiquent la page pour effectuer la suppression.

Le filtrage peut comprendre d'autres tâches qui dépendent de la nature du corpus et de l'objectif du travail.

Figure 2-5: Exemple de tweet portant des noms d'utilisateurs [Web9].

2.2.2.2 Tokenisation

Dans cette partie, l'opération de Tokenization [34] est l'acte de décomposer une séquence de chaînes en morceaux tels que des mots, des mots-clés, des phrases, des symboles et d'autres éléments appelés jetons (tokens).

Les jetons peuvent être des mots individuels, des phrases ou même des phrases entières. Dans le processus de tokenisation, certains caractères comme les signes de ponctuation sont ignorés. Les jetons deviennent l'entrée d'un autre processus comme l'analyse syntaxique et l'exploration de texte.

La tokénisation est utilisée en informatique, où elle joue un rôle important dans le processus d'analyse lexicale, d'où l'utilisation de cette fonction dans notre projet.

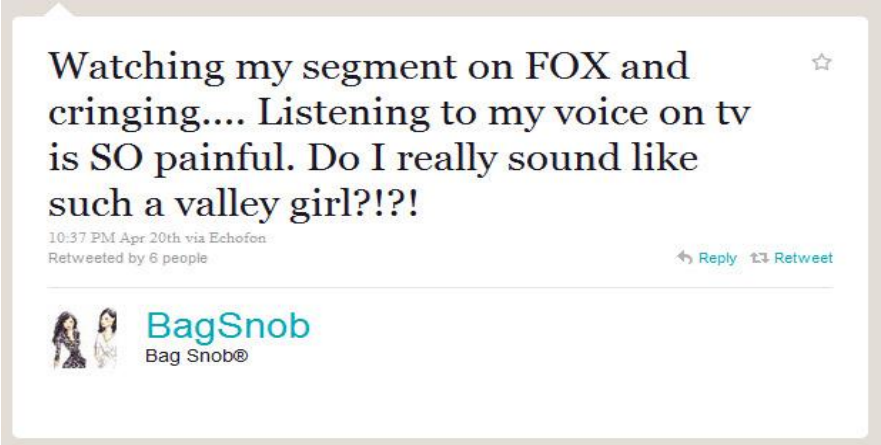

Figure 2-6: Tweet compsé du texte et des signes de ponctuation [Web10].

Après analyse du tweet, le processus de tokenisation transformer le texte en liste de tokens (mot, ponctuation,..) donc le tweet de la figure 2-6 devient : (‘watching’ , ‘my’ , ‘segment’ , ‘on’ , ‘FOX’ , ‘and’ , ‘cringing’ , ‘.’ , ‘.’ , ‘.’ , ‘.’ , ‘Listening’, ‘to’ , ‘my’ , ‘voice’ , ‘on’ , ‘tv’ , ‘is’ , ‘SO’ , ‘painful’ , ‘.’ , ‘Do’ , ‘I’ , ‘really’, ‘sound’ , ‘like’ , ‘such’ , ‘a’ , ‘valley’ , ‘girl’ , ‘ ?’ , ‘!’ , ‘ ?’ , ‘ !’).

2.2.2.3 POS tagger

Chaque mot de la phrase a son rôle syntaxique qui définit comment le mot est utilisé. Ces rôles syntaxiques font partie du discours lié à ce mot. 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 [33].

Le processus consistant à classer les mots dans leurs parties du discours et à les étiqueter en conséquence est connu sous le nom de marquage de partie de discours, POS-tagging, ou simplement marquage [32].On a utilisée des bibliothèques externes pour la segmentation des mots. Voici un exemple qui montre comment fonctionne le processus du tokenisation des mots

```
>>> text = word_tokenize("And now for something completely different")
>>> nltk.pos_tag(text)
[('And', 'CC'), ('now', 'RB'), ('for', 'IN'), ('something', 'NN'),
('completely', 'RB'), ('different', 'JJ')]
```

Figure 2-7: POS tagging exemples [Web11].

text = And now For something completely diferent.

Tokens= (And, now, for, something, comleptely, different).

Après POS_TAG la phrase devient : [(and, CC), (now, RB), (for,IN), (something,NN), (completely,RB), (different,JJ).

Ici, nous voyons cela ‘and ‘ est CC (une conjonction de coordination), ‘now’ et ‘completely’ sont RB (adverbes), ‘for’ est IN (une préposition), ’something’ est NN (un nom), ‘différent’ est JJ (un adjectif).

2.2.2.4 Suppression des mots vides

Dans ce cas, nous effectuons une suppression des mots qui n'influent pas sur l'opinion 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).

Figure 2-8: Exemple de Tweet contentant des mots vides (stopwords) [Web12].

Pour cibler et éliminer les mots vides, nous avons utilisé le dictionnaire Wordnet Comme tout dictionnaire, il a sa propre liste de mots vides.

Après analyse du tweet de la figure 2-8, la phrase devient : (willing stand , vote , organize , will finish started)

2.2.2.5 Lemmatisation

La lemmatisation représente un procédé plus avancée que la racinisation[9], elle réduit un mot à sa forme canonique appelée lemme qui est toujours un mot correcte contrairement à la racine [30], c'est à dire pas de pluriels, pas des verbes conjugués. La lemmatisation recherche généralement dans un dictionnaire pour trouver le lemme d'un mot, car il est difficile dans certains cas d'avoir le lemme d'un mot en se basant uniquement sur des règles morphologique et syntaxiques.

La lemmatisation est étroitement liée au la racinisation[54] La différence est qu'un stemmer fonctionne sur un seul mot sans connaissance du contexte, et ne peut donc pas faire la distinction entre des mots qui ont des significations différentes selon la partie du discours. Cependant, les stemmers sont généralement plus faciles à implémenter et à exécuter plus rapidement, et la précision réduite peut ne pas avoir d'importance pour certaines applications.

[9] La racinisation est le procédé de transformation d'un mot en gardant que sa racine, donc la partie du mot restante une fois que l'on a supprimé son préfixe et son suffixe

La normalisation des verbes est une forme de la lemmatisation :

a) Normalisation des verbes

Ce sous-module se charge de remettre à l'infinitif tout verbe conjugué au passé pour les mêmes raisons que nous transformons tout nom ou adjectif pluriel en singulier.

En anglais, les verbes réguliers sont conjugués au passé en ajoutant un "ed" à leur infinitif mais il existe des exceptions à cette règle. De plus, il existe également une longue liste de verbes irréguliers. Pour pallier à toutes ces exceptions, un dictionnaire de conjugaison est utilisé lors de cette étape, comme il est montré dans la figure 2-9.

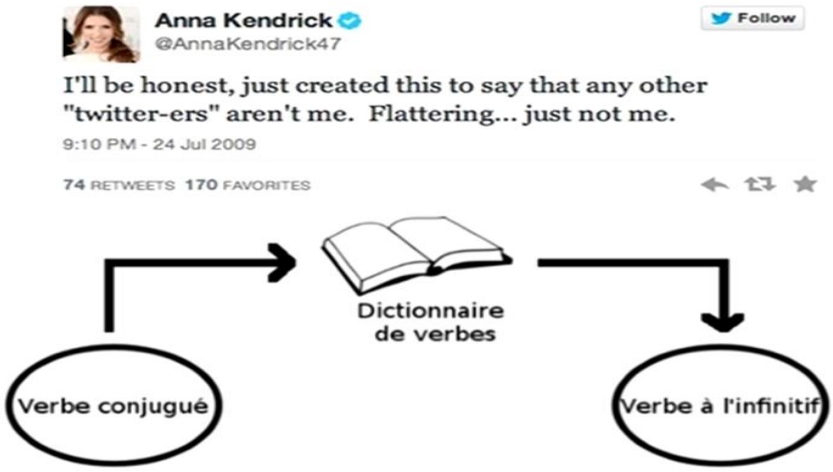

Figure 2-9: Normalisation des verbes [Web13].

Après analyse du tweet, le processus de lemmatisation remplace le texte du tweet de la figure 2-9 par :

(i be honest , just create this to say any other ' twitter-ers ' aren't me , flater … just not me)

2.2.2.6 La négation

Comme son nom l'indique, l'opération consiste à construire un ensemble des mots dans une liste de négation qui est un ensemble de mots consécutifs. Si nous prenons l'exemple de la négation « Not bad at all », quand elle est attachée à un mot qui est suivi d'un mot négatif dans le sens, alors le sens de la phrase devient positif et vice versa Mais l'algorithme qui calcul le score ne peut voir cette différence si l'on ne lui indique pas cela.

La solution est de mettre une condition pour tester si un mot de la liste de négation existe dans le texte.

Après avoir trouvé un mot de négation, le mot suivant juste après la négation, sa polarité serait affectée (inversée) par exemple : (Beautiful= +3, not beautiful= -3)

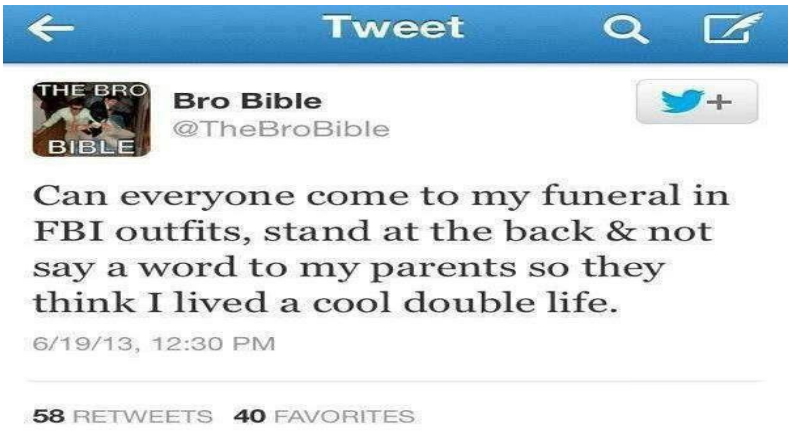

Figure 2-10: Exemple de Tweet contenant une négation [Web14].

2.2.2.7 Émoticônes

Pour autant que nous sachions, les emojis n'a pas été utilisé aussi souvent que les émoticônes dans les études d'analyse des sentiments. Cependant, les émoticônes ont été beaucoup utilisés en association avec les corpus de Twitter qui représentent la grande majorité des études de messages texte. Il y a beaucoup de méthodes pour détecter les émoticônes. Nous avons choisi le plus simple qui est une technique basée sur un dictionnaire [37].

Figure 2-11: Exemple de tweet contenant des émoticônes [Web15].

2.2.3 Étape de catégorisation des textes

Après avoir fini de prétraiter notre tweet, l'étape suivante consiste à évaluer et à analyser chaque token (que ce soit un mot ou une émoticône) dans le tweet. Cette section est divisée en deux parties majeures, à savoir la description des données et celle de la catégorisation des textes.

Avant de parler de la méthode de catégorisation, nous mentionnons quelques ressources importantes qui ont été utilisées pour développer le programme telles que les dictionnaires de mots, dictionnaire d'émoticône et aussi des ensembles de données.

2.2.3.1 Description des données

Nous avons utilisé dans notre conception quelques dictionnaires pour un taux plus précis. Et nous allons définir chaque dictionnaire, et expliquer le concept principal de leur fonctionnement.

a) Dictionnaire des mots

Voici les deux dictionnaires qui ont été utilisés dans la conception de notre programme

- **Hu et Liu**

Une liste de mots d'opinion positifs et négatifs en anglais ou de mots de sentiment (environ 6800 mots) faite par Hu et Liu [45].

Cette liste a été compilée sur plusieurs années à partir du premier article de Hu et Liu en 2004 Si vous ouvrez le dictionnaire, vous remarquerez beaucoup des fautes d'orthographes dans la liste. Ce ne sont pas des erreurs. Ils sont inclus car ces mots mal orthographiés apparaissent fréquemment dans le contenu des médias sociaux parce que c'est tout l'objectif de l'étude, ils ont utilisé une méthode d'apprentissage en profondeur pour voir quels mots ont plus d'influence et d'effet dans les médias sociaux.

Dictionnaire	Négatif	Neutre	Positif
Hu et Liu	4782	0	2004

Tableau 2-1: Description du dictionaire Hu et Liu.

- **Sentiwordnet**

SentiWordNet est une ressource lexicale pour l'extraction d'opinion. SentiWordNet attribue à chaque synset[10] de WordNet trois scores de sentiment: positivité, négativité, objectivité [10].

SentiWordNet est constitué de dizaines de milliers de mots, de significations, de parties de discours représentées et de degré de positivité et de négativité du mot compris entre 0 et 1. Ces mots proviennent tous de la base de données WordNet 2.0, qui est une base de données des mots anglais et leurs significations où les termes sont organisés selon des relations ou des significations sémitiques. Ces mots sont tous regroupés par leurs synonymes dans ce qu'on appelle synsets [11].

Chaque synset de WORDNET (version 2.0) est associé à trois scores numériques Obj (s), Pos (s) et Neg (s), décrivant comment Objective, Positive et Negative les termes sont contenus dans le synset. L'hypothèse qui sous-tend notre passage des termes aux synsets est que différents sens du même terme peuvent avoir différentes propriétés liées à l'opinion. Chacun des trois scores varie de 0,0 à 1,0, et leur somme est de 1,0 pour chaque synset. [55].

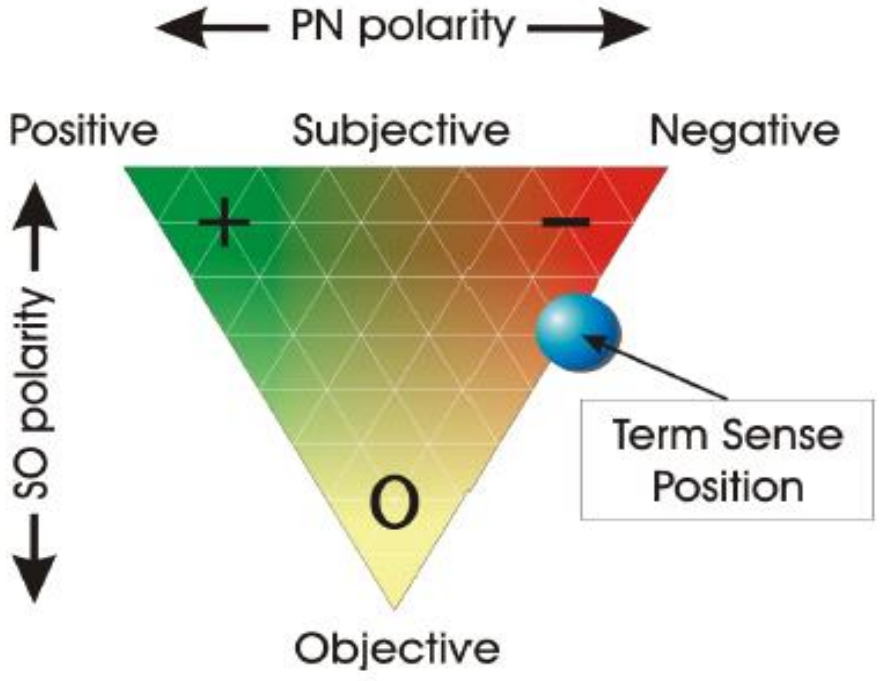

Figure 2-12: Représentation des propriétés liées à l'opinion d'un terme [Web16].

[10] Synset C'est une fonction dans la bibliothèque NLTK sous la branche wordnet et sentiwordnet. Elle renvoie une liste de tous les synonymes d'un mot plus un positive_score, negative_score et objective_score de ce mot

b) Dictionnaire des émoticônes

L'explosion des emojis et des émoticônes [14] dans la parole écrite est un des phénomènes récents les plus intéressants sur internet, l'analyse des emojis et des émoticônes est très important pour la compréhension de sentiment d'une certaine conversation ou texte, c'est pour ça nous avons décidé d'utiliser une liste d'émoticônes donc nous pouvons avoir une analyse plus précise.

```
# Happy and sad emoticons

HAPPY = set([
    ':-)', ':)', ';)', ':o)', ':]', ':3', ':c)', ':>', '=]', '8)', '=)', ':}',
    ':^)', ':-D', ':D', '8-D', '8D', 'x-D', 'xD', 'X-D', 'XD', '=-D', '=D',
    '=-3', '=3', ':-))', ":'-)", ":')", ':*', ':^*', '>:P', ':-P', ':P', 'X-P',
    'x-p', 'xp', 'XP', ':-p', ':p', '=p', ':-b', ':b', '>:)', '>;)', '>:-)',
    '<3'
    ])

SAD = set([
    ':L', ':-/', '>:/', ':S', '>:[', ':@', ':-(', ':[', ':-||', '=L', ':<',
    ':-[', ':-<', '=\\', '=/', '>:(', ':(', '>.<', ":'-(", ":'(", ':\\', ':-c',
    ':c', ':{', '>:\\', ';('
    ])
```

Figure 2-13: La liste des émoticônes [Web17].

c) Ensemble des données (dataset)

Pour tester le taux de réussite nous avons travaillé avec deux différentes sources des données contenant les tweets et leurs polarités.

- **Première ensemble des données**

Tout d'abord Nous avons utilisé l'ensemble des données SemEval2017-task4-test.subtask-A.english, il est présenté sous forme de fichier text (.txt) contenant 12284 tweets avec leurs tags (polarité) traité par Noura Farra de l'Université Columbia et Preslav Nakov du Qatar Computing Research Institute et Sara Rosenthal d'IBM Watson Health Research, pour la classification ils ont utilisé des méthodes d'apprentissage en profondeur (deep learning) pour obtenir un bon résultat.

Cet ensemble de données comporte 3 class des sentiments : positive, négative ou neutre, il traite plusieurs et différents sujets tel que (produits, célébrités, film, sports) chaque entrée de notre ensemble de données est structurée comme suit :

- Tweet id : un identifiant d'un tweet
- Sentiment : tag qui peut être : positive, negative ou neutre
- Tweet texte : le contenu publié par les utilisateurs

- **Deuxième source des données**

Après avoir utilisé le premier ensemble de données, nous avons essayé un autre dataset de NLTK il est présenté sous forme de fichier JSON (.JSON).

Le corpus Twitter de NLTK[11] contient actuellement un échantillon de tweets de 20k (nommés 'twitter_samples') récupérés à partir de l'API Twitter Streaming, ainsi qu'un autre 10k qui sont divisés selon le sentiment en négatif et positif.

Nous avons travaillé avec les 10K qui sont divisés donc en 5000 tweets positives et 5000 tweets de polarité négatives, cet ensemble des données a été traité manuellement et les tweets ont été séparés en deux fichiers (negative_tweets) et (positive_tweets) contrairement à l'ensemble de données précédent qui a été traité avec une méthode d'apprentissage automatique.

2.2.3.2 La catégorisation

Après avoir mentionné toutes les ressources qui ont été utilisées pour le développement de notre programme. Nous allons ci-après présenter la phase de classification, qui est fondamentalement l'objectif principal de ce projet.

Chaque mot du vocabulaire lui est associé un poids. Donc tout ce que nous avons à faire est de sommer ces poids pour déterminer la polarité de ce tweet. Nous allons détailler chaque étape brièvement :

- Pour trouver le sentiment d'un mot symbolique, tout d'abord nous vérifions s'il s'agit d'une émoticône, par exemple " ☹ ☺ ".

[11] Natural Language Toolkit (NLTK) est une bibliothèque logicielle en Python permettant un traitement automatique des langues

On lui donne une valeur de +0.5 si l'émoticône est dans l'ensemble HAPPY_SET ou une valeur de -0.5 s'il appartient à l'ensemble *SAD_SET*.

Le pseudo-code montré dans l'algorithme 1 est appliqué pour tester est ce que token courant est un mot ou une émoticône, et dans le cas où c'est un émoticône, nous verrons à quel ensemble appartient-t-il (HAPPY) ou (SAD).

Algorithme 1 : calcul des émoticônes

```
Score =0
if tag is "emote" :
    if word in sentiment.util.HAPPY :
       return +0.5
    elif word in sentiment.util.SAD :
       return -0.5
```

- Dans la deuxième étape, dans le cas où le mot parcourus n'est pas une émoticône. Nous obtenons à partir d'un dictionnaire (WordNet) une liste de tous les senti_synsets de ce mot. Lorsque la fonction senti_synsets retourne automatiquement un positive_score, negative_score et objective_score avec une liste des synonymes. Nous avons seulement besoin des positive_score et negative_score, et puisque les scores de chaque mot sont des valeurs absolues, nous calculons le score du mot avec formule suivante :

$$Word\_Sentiment = score\_positive - score\_negative.$$

Après avoir récupéré une liste de synonymes d'un mot spécifique de la fonction senti_synsets, on calcule d'abord le poids de chaque mot avec la formule précédente puis on sommes les scores de tous les synonyme et on divise le poids total sur le nombre total de synonymes pour obtenir une moyenne du score d'un mot.

Il y a des conditions sur le synonyme. Nous l'avons seulement ajouté si le mot à la même étiquette que le synonyme, par exemple le mot "good" peut avoir l'étiquette « adjectif » ou « nom ».

Le calcul de poids (scores) des mots et leur moyenne dans senti_wordnet peut être résumé par l'algorithme suivant :

Algorithme 2 : calcule la moyenne du poids d'un mot

```
Get_sentiment_NLTK_sentiword
Results_list=list ( sentiwordnet.senti_synsets (word) )
List_length = len (Results_list)
Total_outer = 0
If List_length > 0 :
    Sum_Score = 0
    Total = 0
    For result in Results_list :
        If result.synsets._pos = = tag or ( result.synsets._pos = = "s" and tag = = "a" ) :
            Sum_score + = result.pos_score () – result.pos_score ()
            Total + = 1
    If total > 0 :
        Score = Sum_Score / total
        Total_outer + = 1
```

- Dans la troisième étape, si le mot existe Dans le dictionnaire Hu et Liu, nous l'ajoutons aussi au calcul afin d'obtenir une moyenne entre celui-ci et le score SentiWordNet pour avoir un résultat statistiquement plus précis.

La moyenne totale du poids d'un mot après avoir vérifié le mot dans les deux dictionnaires est déterminé comme le décrit le pseudo-code 3 :

Algorithm 3 : Calcul de la moyenne totale du poids d'un mot

```
If word in self.File_positive_words_set :
  Score + =  1
  Total_outer + = 1
elif wrd in self.File_negative_words_set :
  Score + = - 1
  Total_outer + = 1
If Total_outer > 0
  Score = score / Total_outer
```

Après avoir réussi d'obtenir le sentiment du mot, nous l'ajoutons au sentiment de tweet actuel. L'algorithme continue à analyser d'autres mots dans la liste en utilisant le même concept (les mêmes étapes) que nous avons mentionné, donc l'analyse continue jusqu'à ce qu'à la fin du tweet.

La négation est prise en considération si nous travaillons avec les listes de négation. Dans ce cas, la valeur du prochain mot est affectée. Voici un exemple pour démontrer l'importance de détecter la négation dans un texte et comment cela peut affecter le sens d'une phrase quand il est ignoré. Exemple de tweet : (i don't like this weather).

Disons que nous avons ignoré la négation et notre programme n'a pas réussi à le reconnaître. Le programme ainsi, après avoir analysé ce tweet, renvoi un sentiment positif pour cet example. Or que c'est clairement un tweet négatif. Donc pour une analyse meilleure et plus précise, l'algorithme devrait détecter la négation. L'algorithme 4 suivant, décrit comment calculer le sentiment d'un tweet en tenant compte de la négation.

Algorithme 4 : Calcul de la valeur d'opinion totale du tweet

```
Tweet_sentiment = 0
Negation_flag = false
Lemmed_words_list = []
For word_and_tag_tuple in lemmedWords_plus_wordtag_list :
    Word_sentiment = 0
    Word = word_and_tag_tuple [0]
    Tag = word_and_tag_tuple [1]
    Lemmed_words_list.append (word )
  If word not in negation_words_list :
    Word_sentiment = self.Get_sentiment_NLTK_sentiword (word, tag)
  If negation_flag = = true :
    Word_sentiment * = -1
    Negation_flag = false
  If word in negation_words_list :
    Negation_flag = true
    Tweet_sentiment + = word_sentiment
Return ( list(Lemmed_words_list), Tweet_sentiment )
```

La polarité des messages est déterminée en appliquant une fonction (la somme) sur les mots d'opinions présentes dans le message.

Quand l'algorithme finit de parcourir tous les mots du text ou tweet en entrée, le sentiment du message sera :

- Si (la polarité est < 0) : le message exprime une opinion négative.
- Si (la polarité est > 0) : le message exprime une opinion positive.
- Si (la polarité est = 0) : le message n'exprime pas d'opinions (neutre).

Le pseudo-code montré dans l'algorithme 5 est appliqué pour classifier les tweets :

Algorithmme 5 : Mesure de polarité

```
If   Sentiment_number = = 0
  Return  "Neutral"
If   Sentiment_number  >  0
  Return  "Possitive"
If   Sentiment_number  <  0
  Return  "Negative"
```

2.2.4 Étape du calcul de pourcentage

L'objectif de ce processus est de déterminer l'exactitude et la précision du processus de catégorisation du système. Concernant le fonctionnement de ce processus, d'abord, nous commençons par choisir un ensemble de données qui soient étiquetés (c.à.d polarité réelle connue de chaque tweet), pour pouvoir les comparer avec les polarités trouvés par notre système. Puis nous traitons et analysons chaque tweet dans cet ensemble de données avec notre algorithme et cela génère un sentiment soit positif, négative ou neutre. Ensuite nous comparons ce résultat avec le résultat de l'ensemble de données. La formule qui été choisi pour le calcul du pourcentage est simple et efficace.

La formule est : nombre de tweet correctement classés divisé par le nombre total de tous les tweets qui ont été traités

2.3 **Conclusion**

Nous nous sommes concentrés principalement dans ce chapitre sur la méthodologie générale de notre 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 ou neutre.

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

3 Chapitre 3 : Réalisation

Introduction

Dans ce chapitre, nous allons parler des concepts de base de tous les dictionnaires et les ensembles de données qui ont été utilisés dans notre implémentation. Nous allons également mentionner tous les tests majeurs qui ont été faits depuis le début avec leur évaluation et interprétation. Avant d'entrer dans les détails, nous décrivons d'abord notre environnement de travail.

3.1 Environnement de Travail

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

3.1.1 Environnement matériel

Quand il s'agit de l'exigence de matériel, vous n'avez pas vraiment besoin de quelque chose de fantaisie. N'importe quel 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 Lenovo équipé d'un processeur I3.

3.1.2 Environnement logiciel

Nous avons utilisé le langage de programmation Python la version 3.6.4. Python est un langage de programmation général, interprété, interactif, orienté objet et de haut niveau [38]. Il a été créé par Guido van Rossum entre 1985 et 1990. Comme Perl, le code source Python est également disponible sous licence GNU ou General Public License (GPL). Cette définition donne assez de compréhension sur le langage de programmation Python.

Python est un langage de script de haut niveau, interprété, interactif et orienté objet. Python est conçu pour être très lisible. Il utilise fréquemment des mots-clés anglais alors que d'autres langues utilisent la ponctuation, et il a moins de constructions syntaxiques que d'autres langues.

Figure 3-1: Logo de Python [Web18].

- 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.

3.1.2.1 Modules externe utilisé

Nous avons utilisé des différentes packages et bibliothèques comme :

a) Bibliothèque Nltk

Natural Language ToolKit (NLTK) est une plate-forme leader pour la construction de programmes Python pour travailler avec des données de langage humain [39].

Il fournit des interfaces faciles à utiliser pour plus de 50 ressources corporelles et lexicales telles que WordNet. Ainsi qu'une suite de bibliothèques de traitement de texte pour la classification, tokenization, stemming, étiquetage, analyse et raisonnement sémantique, *wrappers* pour les bibliothèques NLP de puissance industrielle, et un forum de discussion actif.

Grâce à un guide pratique présentant les fondamentaux de la programmation ainsi que des sujets en linguistique informatique. Ainsi qu'une documentation API complète. NLTK convient aux linguistes, ingénieurs, étudiants, enseignants, chercheurs et utilisateurs de l'industrie.

NLTK est disponible pour Windows, Mac OS X et Linux. Le meilleur de tous, NLTK est un projet libre, open source, axé sur la communauté. NLTK a été appelé "un outil merveilleux pour enseigner, et travailler dans, la linguistique computationnelle en utilisant Python". Et " une bibliothèque incroyable pour jouer avec le langage naturel." Natural Language Processing avec Python fournit une introduction pratique à la programmation pour le traitement du langage. Écrit par les créateurs de NLTK. Il guide le lecteur à travers les bases de l'écriture de programmes Python, en travaillant avec des corpus, en catégorisant le texte, en analysant la structure linguistique, etc.

b) Package re (Regular expressions)

Ce module fournit des opérations correspondant aux expressions régulières. Les expressions régulières utilisent le caractère barre oblique inverse ('\') pour indiquer des formes spéciales ou pour permettre l'utilisation de caractères spéciaux sans invoquer leur signification particulière [40].

c) Package Tkinter

("interface Tk") est l'interface standard de Python. Tk et tkinter sont tous deux disponibles sur la plupart des plates-formes Unix, ainsi que sur les systèmes Windows. Le paquet tkinter est une fine couche orientée objet au-dessus de Tcl / Tk [41]. Pour utiliser tkinter, vous n'avez pas besoin d'écrire le code Tcl, mais vous devrez consulter la documentation de Tk, et occasionnellement la documentation de Tcl. tkinter est un ensemble de *wrappers* qui implémentent les *widgets* Tk sous forme de classes Python. De plus, le module interne tkinter fournit un mécanisme *threadsafe* qui permet à Python et Tcl d'interagir.

d) Package word_tokenize

Chaque mot est un jeton (*token*) lorsqu'une phrase est "tokenisée" en mots, il divise une phrase à une liste des mots séparés. On va essayer la segmentation mot par mot avec cet exemple [53].

```
from nltk.tokenize import sent_tokenize, word_tokenize

data = "All work and no play makes jack a dull boy, all work and no play"
print(word_tokenize(data))
```

Figure 3-2: Tokenisation des mots [Web19].

La sortie est: [All, 'work' , 'and' , 'no' , 'play' , 'makes' , 'jack' , 'a' , 'dull' , 'boy' , ',' , 'all' , 'work' , 'and' , 'no' , 'play'].

e) Package TweetTokenizer

Compatible avec Twitter, développée pour être flexible et facile à adapter à de nouveaux domaines et tâches. La logique de base est la suivante:

- Le tuple regex_strings définit une liste de chaînes d'expressions régulières.
- Les chaînes regex_strings sont placées dans l'ordre, dans un objet d'expression régulière compilé appelé word_re.
- La tokenisation est effectuée par word_re.findall (s), où s est la chaîne fournie par l'utilisateur, à l'intérieur de la méthode tokenize () de la classe Tokenizer.

3.2 Test et évaluation

Cette partie représente toutes les phases des tests qui ont été effectués depuis le début. Nous avons implémenté dans chaque test des différentes méthodes et techniques, On essayant d'explorer différentes options. Donc nous avons utilisé deux dictionnaires différents et deux ensembles de données pour une analyse plus précise.

Nous allons expliquer brièvement chaque test dans le suivant :

- Le premier test était avec le 1ere ensemble de données et avec le Hu et Liu dictionnaire, Cependant, nous avons implémenté seulement cinq étapes de prétraitement qui a été mentionné dans le chapitre précédent. Nous avons obtenu une précision de 57%. Comme montre la figure3-3.

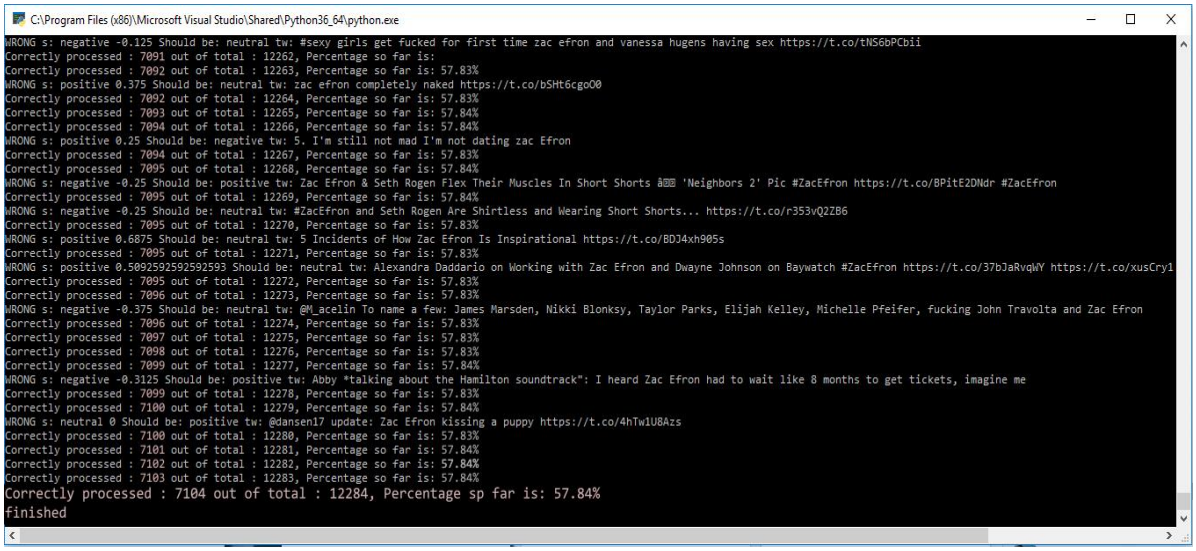

Figure 3-3 : Premier test

- Le deuxième test, à ce moment-là, nous n'avions pas pris en considération la liste de négation comme une étape supplémentaire pour les étapes de prétraitement. typiquement après que nous avons essayé de l'implémenter à l'algorithme pour voir si cela affectait le taux de précision et apparemment il n'a pas, au moins pas pour cet ensemble de données parce que le pourcentage enregistré après était 53%.

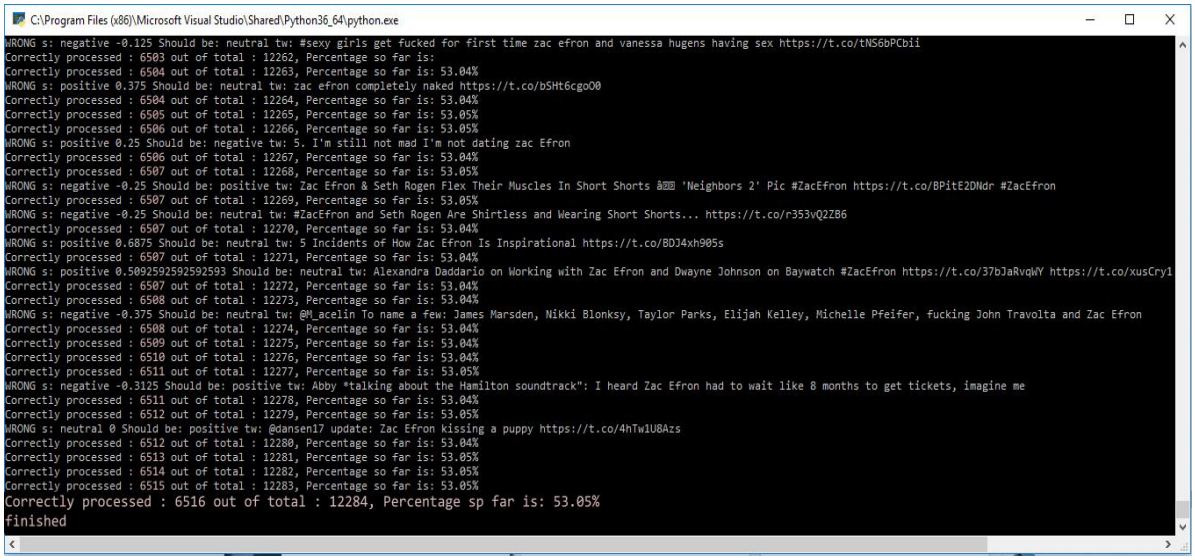

Figure 3-4 : Deuxième test

- Le troisième test, nous avons essayé la technique de combiner différents dictionnaires pour réduire la marge d'erreur et pour une meilleure optimisation. Alors on a utilisé Hu et Liu avec SentiWordNet en essayant d'arriver à un résultat statistiquement plus précis.

Nous avons essayé cette technique toujours avec le premier ensemble de données malheureusement la précision a chuté encore plus bas à 51%.

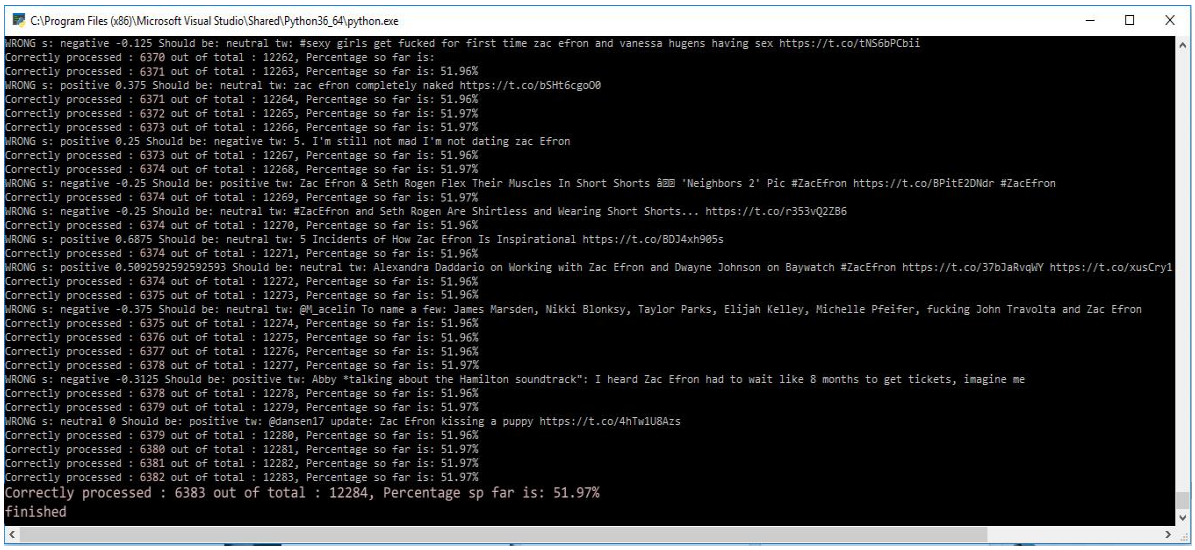

Figure 3-5 : Troisième test

- Après cela pour le quatrième test, nous avons pensé à passer à un autre ensemble de données en essayant d'obtenir un meilleur pourcentage de précision, par suite nous avons utilisé twitter_samples de la deuxième source des données qui a donné un meilleur résultat de 62%.

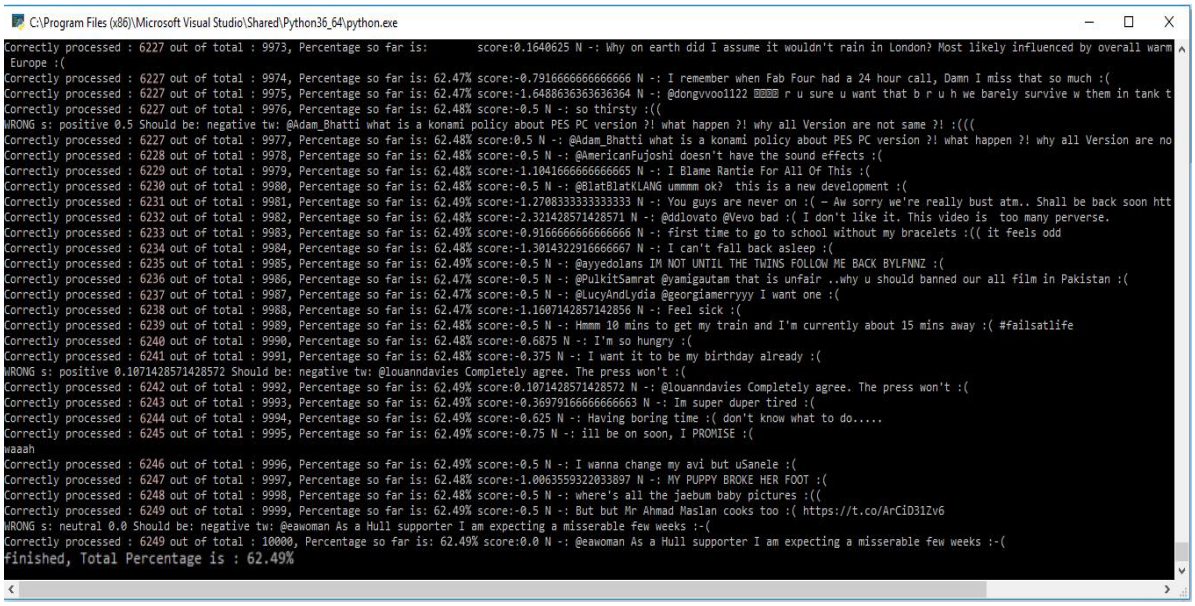

Figure 3-6 : Quatrième test

- Après avoir passé à un autre ensemble, il n'y avait vraiment plus rien à faire sauf pour améliorer notre algorithme pour une meilleure performance et une meilleure détection des sentiments.

Donc après avoir lu quelques recherches et quelques méthodes proposées dans ce domaine, nous avons décidé d'ajouter une liste d'émoticônes. Ainsi pour le cinquième test, nous avons pris les émoticônes en considération et les prenons en compte dans le score total d'une certaine phrase par opposition à avant (car ils étaient indétectables par notre programme antérieur). Après avoir réussi à le faire (traiter les émoticônes) étonnamment le pourcentage est allé jusqu'à 84.03%.

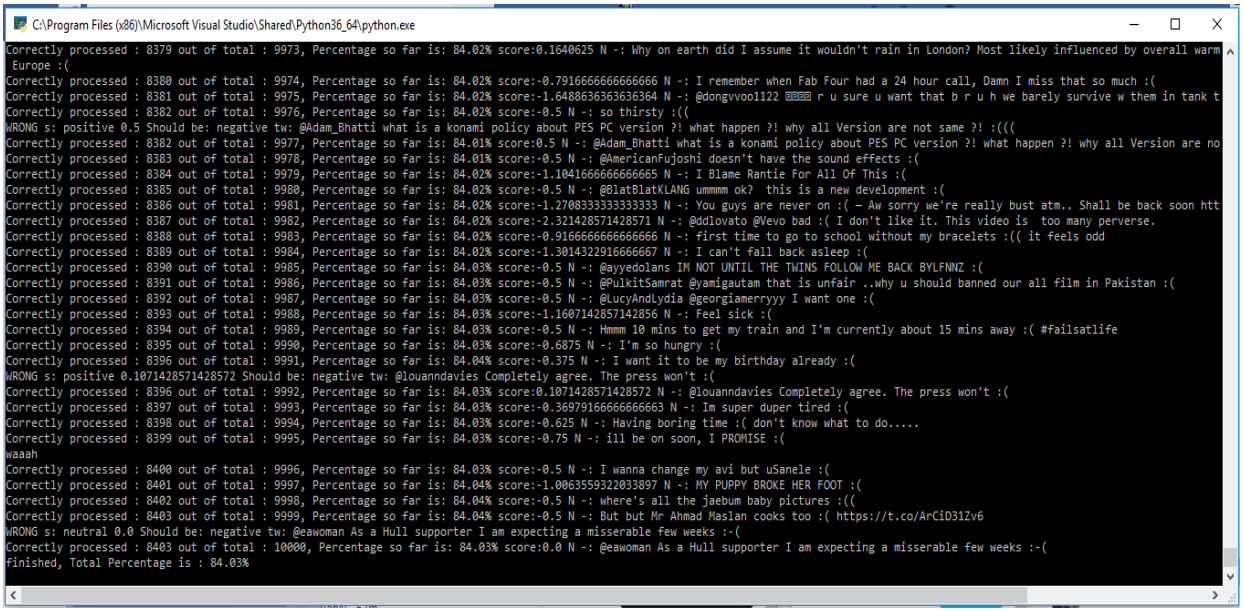

Figure 3-7 : Cinquième test

- Après avoir obtenu un bon taux de précision du dernier test, par curiosité, nous voulions donner un autre test au premier ensemble de données pour tester si le problème est l'ensemble de données lui-même ou quelque chose d'autre.

En utilisant toutes les méthodes et techniques mentionnées précédemment, nous avons également utilisé les deux dictionnaires mais le pourcentage est descendu à 47%.

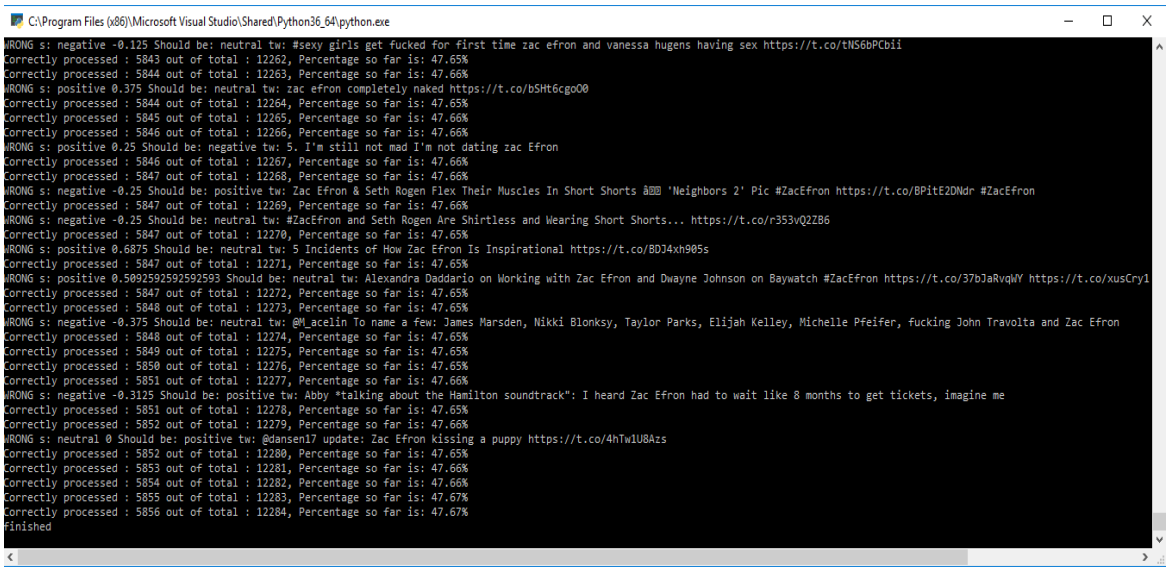

Figure 3-8: Sixieme test

Le tableau 3-1 ci-dessous résumera toutes les phases de tests que nous avons effectuées jusqu'à présent. Chaque test est personnalisé avec des données spécifiques et dépend d'un ensemble des données et d'un dictionnaire particuliers.

Tests	**Dataset1**	**Dataset2**	**Caractéristique de test**	
			Dictionnaires utilisé	**Le prétraitement étapes utilisé**
Première test	57%		Hu et Liu dictionnaire	les cinq premières étapes du prétraitement processus
Deuxième test	53%		Hu et Liu dictionnaire	les cinq premières étapes du prétraitement processus + implémentation de a négation
Troisième test	51%		Hu et Liu + Senti_wordnet	les six premières étapes du prétraitement processus
quatrième test		62%	Hu et Liu + Senti_wordnet	les six premières étapes du prétraitement processus
cinquième test		84%	Hu et Liu + Senti_wordnet	les six premières étapes du prétraitement processus + liste des émoticônes
Sixième test	47%		Hu et Liu dictionnaire + Senti_wordnet	Les sept premières étapes du prétraitement processus

Tableau 3-1 : évaluations des testes

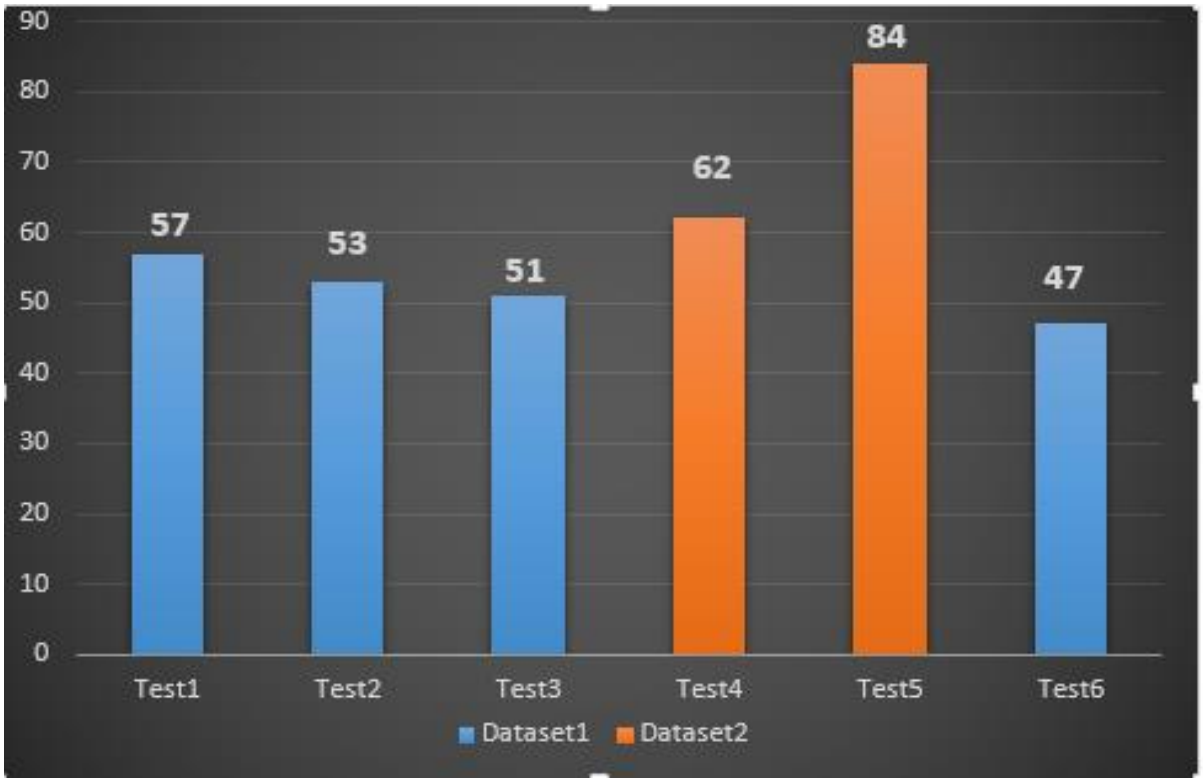

Figure 3-9 : Représentation graphique des phases des tests

3.3 **Conclusion**

Dans ce chapitre nous avons mentionné les principaux outils et paquets qui ont été implémentés dans notre programme, afin de savoir quel niveau d'exactitude le programme donne. Comme dans tous les systèmes d'analyse de sentiments, notre programme prend en entrée un ensemble de tweets à analyser. Pour notre cas, nous avons testé avec deux datasets différents.

La fiabilité du programme dépend de la précision de votre analyse .mais ce n'est pas toujours le cas, car les données à analyser jouent un grand rôle sur la performance de résultat obtenu. Car si ces données contiennent trop d'ambiguïté ou trop de sarcasme, C'est vraiment difficile de les analyser et détecter leurs polarités.

Donc soit votre programme ne parviendra pas à le détecter ou vous finissez par prendre des prédictions.

Conclusion générale

L'analyse de sentiment est de différencier et de classer les points de vue ou des sentiments ou des évaluations dans le contenu composé. Les sentiments des gens peuvent être exprimés en positif, façons négatives ou neutres.

La tâche de l'analyse des sentiments, en particulier dans le domaine du micro-blogging, est Encore en développement et loin d'être complète, en fait il s'agit d'un important domaine de recherche actuel. Et ce n'est pas seulement le microblogging, Il peut être utilisé avec une variété des domaines d'applications comme le domaine politique, le marketing ou prédiction de stock.

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..). Plusieurs recherches en ingénierie linguistique montrent qu'il est possible de le faire avec la méthode des dictionnaires de sentiments (comme wordnet), la méthode de filtrage et autres étapes de prétraitement à effectuées pour optimiser l'analyse des opinions. Nous avons spécifiquement implémenté les étapes qui ont le plus impact sur le traitement. Nous nous appuyons également sur ces technologies pour construire notre propre méthode.

La mise en œuvre et la conception de notre projet visent à déterminer la polarité des tweets rédigés en langue anglaise.

Le concept fondamental de notre programme est de traiter un tweet en utilisant les principales étapes de prétraitement Pour une meilleure analyse. Pour que le traitement fonctionne correctement, il doit être basé sur un bon dictionnaire. Dans notre cas, nous avons choisi deux dictionnaires. Nous l'avons utilisé pour réduire la marge d'erreur. On vérifie d'abord le mot dans le dictionnaire wordnet. On utilise une fonction (senti_synsets) qui est implémentée en NLTK sous la branche wordnet, elle nous donne une liste des synonymes du mot ainsi que la polarité de chacun. On calculer donc la moyenne sentimental du mot à partir de ses synonyme. Si le mot existe Dans d'autre dictionnaire par exemple (Hu et Liu). Nous l'ajoutons aussi, puis nous obtenons une moyenne entre celui-ci et le score SentiWordNet pour un meilleur résultat. Après cela nous appliquant la fonction de la somme pour déterminer la polarité du texte.

Reference web

[01] *Les Réseaux Sociaux en général*. (s.d.). Consulté le 05 26, 2018, sur OverBlog: http://socialonline.over-blog.com/les-reseaux-sociaux-en-general.html

[02] B.Bathelot. (2016, 01 27). *L'encyclopédie illustrée du marketing*. Consulté le 05 26, 2018, sur définition marketing: https://www.definitions-marketing.com/definition/reseaux-sociaux/

[03] *Historique des réseaux sociaux*. (s.d.). Consulté le 05 26, 2018, sur cadre21: https://www.cadre21.org/ressources/historique-des-reseaux-sociaux/

[04] Olivier. (2017, 04 14). *Histoire des reseaux sociaux*. Consulté le 05 26, 2018, sur GLABS: http://glabs-consulting.fr/histoire-des-reseaux-sociaux/

[06] *Les réseaux sociaux - BTS SIO*. (2012, 04 01). Consulté le 05 26, 2018, sur canalblog: http://reseauxlapie.canalblog.com/archives/2012/04/01/24096259.html

[07] *Les différents types de réseaux sociaux*. (2013, 03). Consulté le 05 26, 2018, sur Le web 2.0 et les réseaux sociaux: https://megganeangellotti.wordpress.com/titre-5/

[08] B.Bathelot. (2018, 02 02). *Définition : Analyse des sentiments*. Consulté le 05 26, 2018, sur définition marketing: https://www.definitions-marketing.com/definition/analyse-des-sentiments/

[10] *SentiWordNet*. (s.d.). Consulté le 06 07, 2018, sur http://sentiwordnet.isti.cnr.it/

[11] *Using SentiWordNet for Sentiment Classification What is SentiWordNet*. (s.d.). Consulté le 06 07, 2018, sur http://www.academia.edu/4062253/Using_SentiWordNet_for_Sentiment_Classification_What_is_SentiWordNet

[13] (2012). Consulté le 06 10, 2018, sur SenticNet: http://sentic.net/about/

[14] *Source code for nltk.sentiment.util*. (2018, 05 06). Consulté le 06 07, 2018, sur NLTK Project: https://www.nltk.org/_modules/nltk/sentiment/util.html

[27] MacArthur, A. (2018, 04 27). *lifewire*. Consulté le 06 01, 2018, sur lifewire: https://www.lifewire.com/history-of-twitter-3288854

[31] Patrick, P. &. (2010, 05 17-23). *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*. (DPLB, Éditeur, L. L.-C. Universit´e de Paris-Sud, Producteur, & Universit´e de Paris-Sud, Laboratoire LIMSI-CNRS, Bˆatiment 508) Consulté le 06 02, 2018, sur Researchgate: https://www.researchgate.net/publication/220746311_Twitter_as_a_Corpus_for_Sentiment_Analysis_and_Opinion_Mining

[32] *Categorizing and Tagging Words*. (2015, july 01). Consulté le 06 03, 2018, sur nltk book: https://www.nltk.org/book/ch05.html

[34] *Tokenization*. (s.d.). Consulté le 06 04, 2018, sur techopedia: https://www.techopedia.com

[38] *Python - Overview*. (s.d.). Consulté le 06 06, 2018, sur Tutorials Point Simply Easy Learning: https://www.tutorialspoint.com/python/python_overview.htm

[39] Project, N. (Éd.). (2018, 05 06). *Natural Language Toolkit*. Consulté le 06 06, 2018, sur NLTK 3.3 documentation: https://www.nltk.org/

[40] *Regular expression operations*. (2018, 05 16). Consulté le 06 06, 2018, sur Documentation « The Python Standard Library » 7. String Services »: https://docs.python.org/2/library/re.html

[41] *docs.python Graphical User Interfaces with Tk*. (s.d.). Consulté le 06 08, 2018, sur docs.python 3.6.5: https://docs.python.org/3/library/tk.html

[42] *Wikipédia*. (s.d.). Consulté le 06 06, 2018, sur https://fr.m.wikipedia.org/wiki/Sentiment

[43] *dictionaire de francais larousse*. (s.d.). Consulté le 06 07, 2018, sur https://www.larousse.fr/dictionnaires/francais/opinion/56197

[44] Liu, B. (s.d.). Consulté le 06 07, 2018, sur Morgan & claypool publishers: https://www.morganclaypool.com/doi/abs/10.2200/s00416ed1v01y201204hlt016

[45] Liu, M. H. (2004, 05 15). *Opinion Mining, Sentiment Analysis, and Opinion Spam Detection*. Consulté le 06 07, 2018, sur https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[46]*Techopedia Sentiment Analysis*. (s.d.). Consulté le 06 08, 2018, sur Techopedia : https://www.techopedia.com/definition/29695/sentiment-analysis

[47] *InsightsAtlas*. (s.d.). Consulté le 06 10, 2018, sur blog.insightsatlas: https://blog.insightsatlas.com/7-benefits-of-sentiment-analysis-you-cant-overlook

[48] *In NLP, what is the difference between a Lexicon and a Corpus?* (s.d.). Consulté le 06 10, 2018, sur Quora: https://www.quora.com/In-NLP-what-is-the-difference-between-a-Lexicon-and-a-Corpus

[53] *Tokenizing Words and Sentences with NLTK*. (s.d.). Consulté le 06 10, 2018, sur python tutorials: https://pythonspot.com/tokenizing-words-and-sentences-with-nltk/

[54] *Dive Into NLTK, Part IV: Stemming and Lemmatization*. (s.d.). Consulté le 06 10, 2018, sur Text Mining Online: https://textminingonline.com/dive-into-nltk-part-iv-stemming-and-lemmatization

[56] Rosen, A. (0117, 11 tuessdady). *Tweeting Made Easier*. Consulté le 06 13, 2018, sur blog.twitter: https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html

---

[Web1] http://www.booksaresocial.com/history-of-social-media-part-i/, Consulté le 06 05, 2018

[Web2] https://www.pinterest.com/pin/513128951274150409/, Consulté le 06 05, 2018

[Web3] https://plugingroup.com/top-15-popular-social-media-sites-november-2016/, Consulté le 06 05, 2018

[Web4] https://fr.slideshare.net/amjedbj/accs-linformation-dans-les-rseaux-sociaux-quelles-formes-de-collaboration, Consulté le 06 06, 2018

[Web5] http://www.barriblog.com/2017/10/lo-siempre-quiso-saber-del-api-twitter-nunca-se-atrevio-preguntar-actualizado-2017/, Consulté le 06 06, 2018

[Web6] http://books.openedition.org/oep/204, Consulté le 06 07, 2018

[Web7] http://azhari.typepad.com/just_another_serendipity/2013/05/hashtags-the-key-to-reaching-a-wider-audience-on-twitter.html, Consulté le 06 07, 2018

[Web8] https://www.business2community.com/twitter/behold-the-amazing-power-of-twitter-hashtags-01265767, Consulté le 06 07, 2018

[Web9] https://twitter.com/search?q=SB%20Projects%20%23problem%20&src=typd, Consulté le 06 07, 2018

[Web10]http://fr.newtabmedia.com/?pub_id=3521&sub_id=152890736606950526180053567844364 46&srcid=2038307, Consulté le 06 07, 2018

[Web11] http://www.nltk.org/book/ch05.html, Consulté le 06 07, 2018

[Web12] https://sproutsocial.com/insights/great-twitter-promoted-tweets/, Consulté le 06 07, 2018

[Web13] https://www.quora.com/What-are-some-examples-of-really-savvy-tweets-by-celebrities-on-Twitter, Consulté le 06 07, 2018

[Web14] https://failblog.cheezburger.com/failbook/tag/bro-bible, Consulté le 06 08, 2018

[Web15] https://instituteforpr.org/crisis-ready-top-tips-social-media-managers-crisis/, Consulté le 06 08, 2018

[Web16] https://ileriseviye.wordpress.com/2010/06/16/sentiwordnet-sentiment-analysis-and-opinion-mining-with-wordnet/, Consulté le 06 08, 2018

[Web17] https://www.nltk.org/_modules/nltk/sentiment/util.html, Consulté le 06 08, 2018

[Web18] https://www.python.org/community/logos/, Consulté le 06 09, 2018

[Web19] https://pythonspot.com/tokenizing-words-and-sentences-with-nltk/, Consulté le 06 09, 2018

Références bibliographiques

[05] RICHAUD, N. (2015, 10 30). les echos.fr. *LinkedIn passe le cap des 400 millions de membres* .

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

[12] Amira, A. B. (2016). *Diffusion des opinions dans les réseaux.* alger.

[15] Gerv´as, J. C. (2014). *An easily scalable concept-based affective lexicon.*

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

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

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

[19] 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.

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

[21] Lei Zhang, R. G. (2011). *Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis* . HP Laboratories.

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

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

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

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

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

[28] Prem Melville, W. G. (Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification). 2009, Watson Research Ctr. P.O. Box 218 Yorktown Heights.

[29] Ribeiro, C. S. (2010). *Inductive infeTence foT large scale text classification.* berlin heidelberg: springer-verlag.

[30] zakaria, B. m. (2017). *fouille d'opinions dans les réseaux sociaux.* ALG: ESI.

[33] Garje, A. R. (2017). Sentiment Analysis Text POS Tagging on Movie reviews using NLTK. *Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad - 431004, Maharashtra, India* , 5.

[35] Mehdi, M. M. (2016-2017). *Fouille d'opinion dans les reseaux sociaux.* alger.

[36] Shi, X. L. (2013). *The Role of Text Pre-processing in Sentiment Analysis.* London: Department of Information System and Computing, Brunel University, London,UB8 3PH, UK b CAS Research Centre of Fictitious Economy & Data Science, Chinese Academy of Sciences, Beijing, 100080, PR China.

[37] Gaël Guibon, M. O. (2016). *From Emojis to Sentiment Analysis.* Marseille , france: Université de toulon , lsis UMR7296,13397,

[49] Leo Breiman, J. F. (1984). *Classification and Regression Trees.* Chapman and Hall/CRC.

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

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

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

[55] FabrizioSebastiani, A. a. *SENTIWORDNET: APubliclyAvailableLexicalResource forOpinionMining.* Pisa, Italy : Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche .