République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Saad Dahlab - Blida I -

N° D' ordre :.....

Faculté des sciences

Département d'informatique

Mémoire Présenté par :

BENYAHIA Hiba

En vue d'obtenir le diplôme de master

Domaine : Mathématique et informatique

Filière : Informatique

Spécialité : Informatique

Option : Génie des systèmes informatiques

Sujet : Analyse des sentiments dans les réseaux sociaux pour l'aide à l'amélioration des stratégies de Marketing

Soutenu le : 26 / 06/2016

Devant le jury :

Mr. S.Ferferra	Président
Mme. M.Arkam	Examinateur
Mme. MADANI Amina	Promotrice
M. BENYAHIA Zakaria	Encadrant

Promotion
2015/2016

Dédicace

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, de 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 chères frères Nacer, Zakaria et Aberraouf mes chères sœurs Kheira et Soumya, et à toute la famille qui n'ont cessé d'être pour moi des exemples de persévérance, de courage et de générosité.

Mes très chères belles sœurs Sara et Amina et mes chères neveux Wail, Samy, Tessnime et Wassime.
Que Dieu vous bénisse, trouvez ici l'expression de mon profond respect, mon amour,et ma gratitude.

Mon mon chère beau frère Abderrahmane et mon chouchou Yacine.
Je vous prie de trouver dans ce travail le témoignage de mon affection. Que Dieu vous comble de ses bienfaits.

Mes amis
Oussama :
J'espère que tu trouveras dans ce mémoire l'expression de mon profond respect, ma sympathie et ma grande gratitude.
Tu m'as beaucoup soutenu, et aidé, je te remercie infiniment.
Que Dieu te bénisse et te guide vers le meilleur inchaellah.

Amine :
J'espère que tu trouveras dans ce mémoire l'expression de mon profond respect, ma sympathie et ma grande gratitude.
Tu m'as beaucoup soutenu, et aidé, je te remercie infiniment.
Que Dieu te bénisse et te guide vers le meilleur inchaellah.

Mes très chère amis (Anis, Aycha, Yamina, Khansaa, Mouna, Kamélia) qui m'ont accompagné, aidé, soutenu et encouragé tout au long de la réalisation de ce mémoire.
Je vous dédie mon travail en témoignage de mon sincère attachement. Je prie Dieu pour vous donner santé, bonheur et prospérité.
Je n'oublierai jamais les moments agréables qu'on a vécus ensemble ...

Remerciements

On dit souvent que le trajet est aussi important que la destination, les cinq années d'apprentissage nous auront permis de bien comprendre la signification de cette phrase toute simple. Ce parcours en effet, ne s'est pas réalisé sans défis, et sans soulever de nombreuses questions dont les réponses ont nécessité de longues heures de recherche.

Tout d'abord, louange à " Allah " qui m'a guidée sur le droit chemin tout au long du travail et m'a inspiré les bons pas et les justes reflexes. Sans sa miséricorde, ce travail n'aura pas abouti.

Je tiens à remercier vivement ma promotrice, Mme MADANI Amina pour son aide précieu. Pour sa disponibilité et son soutien constant, tant matériel que moral et intellectuel.

Je tiens précisément à exprimer notre reconnaissance et nos remerciements à mon maitre de stage Mr BENYAHIA Zakaria , directeur manager au sein de l'entreprise MGPS , dont la disponibilité, le savoir-faire et le soutien ne m'a jamais fait défaut. Ses conseils, ses orientations ainsi que son soutien moral et scientifique ont contribué à mener ce projet à bon port.

Je remercie également toute l'équipe de MGPS pour leur accueil et particulièrement M. Christian TAORMINA le président de MGPS m'a accordé toute sa confiance et m'a consacré son temps précieux tout au long de cette période, sachant répondre à toutes mes interrogations.

Je tiens aussi à remercier les membres de l'équipe de MGPS avec qui j'ai travaillé, Teddy, Emma, Younes, Saad et Ali pour la bonne ambiance durant la période de mon stage, cela m'a permis de passer un stage très agréable.

Je désir exprimer ma reconnaissance envers mon collègue Oussama qui m'a apporté son support moral et intellectuel tout au long de ma démarche.

J'exprime également ma gratitude aux membres du jury, qui nous ont honorés en acceptant de juger ce modeste travail.

Enfin je tiens à remercier l'ensemble du corps enseignant Mr Mahdoum, Mr Chikhi, Mr cherif zahar, Mme zahra, Mme farah, Mme Arkam, Mme Rezoug et Mme Miloud.

Je garde un très bon souvenir de vous, J'apprécier toujours vos qualités humaines, votre savoir-faire et vos compétences scientifiques.

Mes remerciements vont également à nos parents, ainsi qu'à toutes les personnes qui nous ont aidées de près ou de loin par le fruit de leur connaissance pendant toute la durée de notre parcours éducatif.

Résumé

Les réseaux sociaux intègrent un volume et une variété de données textuelles.Ils ont changé la façon dont l'information est transmise aux clients, l'Analyse des Sentiments offre la possibilité de comprendre les status générés par l'internaute (le client) et explique comment un produit où une marque est perçus (l'avis public).

Dans ce contexte, ce mémoire présente les défis liés à la classification des sentiments afin d'aider les entreprises a mettre en place leurs panneaux publicitaires.

Dans notre travail nous allons utiliser une méthode d'apprentissage suppervisé et précisement le classifieur probabiliste Naïf Bayes (NB), en utilisant un corpus de données en anglais classifié manuellement et le testé soit avec un ensemble de tests où avec des tweets collectés à partir de Twitter par le mot clé "IPhone".

Nos résultats indiquent que le classifieur a abouti à une très bonne précision de 95.5% (le corpus sanders comme un ensemble de données d'apprentissage et tests) qui prouve une bonne implémentation de l'algorithme NB, et avec 70% avec les tests de sentiment140.

Mots clés : Analyse des sentiments, Classification des Sentiments, Panneaux Publicitaires, Apprentissage Suppervisé, Naïf Bayes.

Abstract

Social networks include quadrillions of textual data, they changed the way information is transmitted to the clients of enterprises. Sentiment analysis offers the possibility to understand the posts of social network users (clients), and also explains how a product is seen from the client's perspective.

In this subject, this memoir cites the challenges of sentiment classification, and explains the method that we used for sentiment analysis in order to help companies to choose best billboard locations for the ads for their products.

In our work, we will use the well-known supervised learning method the Naif Bayes Classifier (NB) which will learn from a manually classified English corpus, and then tested on both test sets and collected tweets on iPhone from Twitter.

The results we obtained indicate that the classifier we implemented could reach a precision of 95,5% when applied in tests on the same learning dataset (Sanders), and near 70% when applied on a different test set (Sentiment140) which proves that we succeeded to realise a good implementation for sentiment analysis on tweets.

Keywords : Sentiment Analysis, Sentiment classification, Billboards, Supervised learning, Naïf Bayes.

Abstract

الشبكات الاجتماعية تحتوي على حجم كبير من البيانات النصية المتنوّعة التي أحدثت تغيرا في طريقة إنتقال المعلومات إلى الزبائن، إن تحليل المشاعر يسمح بفهم تغيرات الزبائن، وإظهار الصورة الّتي ينظر بها الزبون للمنتج أو العلامة التجارية (الرأي العام).

في هذا السياق تظهر لنا هذه المذكرة مختلف التحديات والمشاكل المربوطة بتصنيف المشاعر لمساعدة الشركات في وضع اللوحات الإعلامية الخاصة بها. في عملنا هذا، سنقوم بإستعمال طريقة التعلم المشرَف عليه وعلى وجه التحديد المصنف الإحتمالي نايف بايز (الساذج بايز)، وذالك باستخدام مجموعة من البيانات باللغة الإنجليزية التي تم تصنيفها يدويا وإختبارها إما مع مجموعة من البيانات الخاصة بالإختبارات أو بالتغريدات والتي تم جمعها من تويتر والمتعلقة بكلمة ايفون.

تشير النتائج التي توصلنا إليها أن المصنف أدى إلى دقة عالية جدا بـ 95.5% كمجموعة من البيانات للتعلم والإختبار في آن واحد

هذا ما يثبت التطبيق الجيد لخوارزمية نايف بايز، مع 70% من البيانات الخاصة بالإختبار لـ Sentiment140

كلمات المفتاح: تحليل المشاعر، تصنيف المشاعر، اللوحات الإعلانية، التعليم المشرَف عليه، المصنف نايف بايز، تويتر.

Table des matières

Table des figures

Liste des tableaux

Introcution générale

De nos jours, l'internet est devenu un outil indispensable d'échange d'information. Il nous offre une quantité précieuse d'information et des données plus importantes, ses services s'adaptent de plus en plus aux besoins d'internautes, pendant les dernières années l'internet a connu une vaste portée grâce aux développements des médias sociaux qui favorisent les interactions sociales tels que Twitter et Facebook, ce sont des sites webs qui rassemblent des individus, des entreprises et des organisations.

Les réseaux sociaux s'intéressent au stockage d'informations afin de l'utiliser dans différents domaines d'applications tels que l'analyse de sentiment car il reflètent en temps réels l'internet public.
L'avis public se positionne sur le marché afin d'offrir aux entreprises une connaissance d'avantage poussée.

Notre travail s'interèsse plutôt à l'analyse de sentiments dans les plates formes sociales et plus précisément à la recherche dans les microblogs. Les microblogs sont des messages à faible taille environ 140 caractères à travers lesquels les individus consomment et produisent des informations intéressantes sous forme : d'opinions, status d'évenements ...etc.

Tous les récents développements dans le domaine de l'Analyse de Sentiments (AS) donnent un coup de vol aux applications informatiques conçues pour l'analyse et la détection de sentiment exprimé sur les réseaux sociaux.
Il est difficile de traiter un grand corpus de données et prédire la classe la plus probable de chaque document (dans notre cas le tweet), ce qui nous conduit a un problème de classification de sentiments.

A cet effet, une grande partie de cette étude sera consacrée à la classification automatique de sentiment avec un classifieur probabiliste pour prédire le sentiment publique afin d'aider les entreprises à mettre en place un plan de marketing pour leurs produits ou marques.

À travers ce travail, nous allons essayer de trouver une solution a la problématique d'utilisation de l'apprentissage automatique pour l'analyse de sentiment afin d'aider les entreprises à mettre en place un plan de marketing pour minimiser leurs dépenses publicitaires.

Objectifs

L'objectif de notre travail consiste à mettre en œuvre une application pour aider les entreprises à savoir l'avis public sur leurs produits afin de minimiser les coûts publicitaires pour savoir où mettre leurs panneaux publicitaires(générer la localisation de l'internaute).

Trois corpus comportant des tweets seront analysé lors de la présente étude. ces trois corpus sont : "sanders et sentiment140" destinés à être utilisé dans le marketing, et une autre à partir de Twitter collecté par mot clé "iPhone".

A partir des datasets nous allons entrainer notre algorithme probabiliste avec les données d'apprentissage, ensuite nous allons testé sa fiabilité et son exactitude avec les données de tests, afin de prédire le sentiment de chaque tweet avec le classifieur Naïf Bayes.

Le deuxième objectif est de faire rajouter à la lemmatisation (prétraitements) l'étiquetage morpho-syntaxique pour extraire seulement les adjectifs et les verbes car ils sont les porteurs d'opinions, après nous allons faire le construction des sacs de mots avec les n-grammes(unigrammes, bigrammes et trigrammes) por améliorer notre travail et aussi pour diminuer la taille de la base de données pour minimiser le temps de parcours et aussi le temps d' exécusion.

Présentation de L'entreprise "la structure d'accueil"

Etant en fin de cycle, Master Informatique spécialité Génie des Systèmes Informatiques à l'université de Saad Dahlab Blida, j'effectue actuellement un stage de six mois (du 01 Avril au 01 Juin) au sein de l'entreprise MGPS (Manutention Gérée Par Satellites). société a été créée en 2013 á partir de l'expertise développée par le service informatique de l'opérateur de terminal portuaire de EuroFOS.

Les projets de l'entreprise d'accueil s'inscrivent dans le cadre du développement des logiciels pour la logistique et la gestion des terminaux portuaires. Pour cela, MGPS travaille sur un projet nommé EGEE. Les projets en cours sont bien en avance et montrent un fonctionnement remarquable sur les terminaux FOS-Marseille. S'appuyant sur les bases du Système opérationnel actuel, EGEE doit créer une synergie inter opérateur qui n'existe pas sur les terminaux portuaires, d'où l'idée de réaliser dans un premier temps un démonstrateur EGEE sous forme d'une application web de gestion de flux d'activité de conteneurs qui doit montrer l'efficacité de la démarche EGEE à travers des statistiques, des graphes et des indices de performances qui inciteront les clients finaux à suivre la démarche EGEE.

MGPS est une société spécialisée dans le développement de solutions technologiques liées à l'ensemble des activité logiques. Elle dispose d'un personnel qui possède à son actif plusieurs décennies d'expérience dans la gestion des organisations portuaires et logistiques, et à ce titre cette dernière a décidé de s'imposer comme solution global inter opérateur, en fournissant ses compétences autour des axes suivants :

- Analyse des besoins en termes de sécurisation des marchandises et des conteneurs ;
- Analyse des flux de la Supply Chain et définition d'un processus " vertueux " capable de rationaliser efficacement cette chaine tout en apportant une véritables plus-value

en terme de développement durable, avec notamment ses effets sur les économies structurelles en CO2 sur l'ensemble des axes touchés.
— Par son niveau de compétences reconnu de l'ensemble des professionnels concernés, MGPS tentera de s'imposer sur le territoire national en utilisant les deux premiers axes majeurs que sont l'axe Rhône Saône et l'axe Seine, puis tentera ensuite d'exporter ce savoir-faire dans les autres pays du globe.
— Création, promotion, et commercialisation d'un démonstrateur d'efficacité qui associera l'ensemble des auteurs de ce projet.

Disposant idéalement de compétences organisationnelles ; MGPS se chargera de la conduite du changement opérée sur les différents intervenants, dans chacun de ses projets, tout en assurant, dans sa volonté promotionnelle l'adéquation entre les différentes corporations rencontrés. A ce stade, MGPS assure déjà un lien " privilégié " avec les métiers portuaires décriés à ce jour, et assumera l'efficacité de leur implication totale dans les différents projets afin d'ajouter aux innovations technologiques des améliorations sensibles dans la perception des places.

MGPS est le porteur du projet EGEE, et le but de cette entreprise reste de synthétiser les efforts de l'ensemble des opérateurs de la Supply Chain, afin de devenir le leader opérationnel (Europe surtout) comme leader de la solution d'hinterland. La solution, qui tournera essentiellement autour d'un démonstrateur vertueux, devra à la fois apporter cette synthèse sur des sites où les autres éléments du projet sont natifs ou présents, mais aussi, en assurant la cohérence de son rôle, sur les places où "tous les autres systèmes" sont différents de la solution globale.

MGPS et dans le cadre du projet EGEE, développera un réseau social pour les professionnels de la chaine logistique. Ce réseau social sera doté d'une API d'export des publications (géolocalisées).

La mission qui m'a été confiée durant ce stage concerne la mise en œuvre (comception) d'un logiciel pour la classification des sentiments des status échangés dans leur propre réseau social afin que les clients finaux suivent la chaine logistique des activités d'import et export maritime de leurs démonstrateurs EGEE (pour suivre la démarche EGEE).

Description des chapitres

Ce mémoire comprend six chapitres au total. Le deuxième chapitres présente une étude théorique sur les réseaux sociaux, les plateformes Twitter et Facebook, leurs fonctionnalités et leurs caractéristiques.

Le troisième chapitre état de l'art comprend les approches de classification de sentiments et les travaux éxistants aussi le domaine d'application, le chapitre quatre décrit la méthodologie et la conception détaillée du projet. Dans le cinquième chapitre les outils de développement, les tests et les résultats des datasets avec les méthodes d'évaluation (Rappel, Précision et F-mesure), le sixième chapitre renferme la conclusion et les perspectives de recherche qu'ouvre la présente étude.

Chapitre 1

Les Réseaux sociaux : Généralité

1 Introduction

Avec l'arrivé du Web 2.0 les réseaux sociaux sont devenus de plus en plus présents. Il apparait comme un outil de communication pour les entreprises à tous les niveaux (de la promotion de nouveaux produits à la recherche de nouveaux consommateurs). Un réseau social est un ensemble d'entités sociales (individu, groupe ou organisation) qui se caractérise par l'interaction sociales. Cette interaction peut être : familiale, sentimentale ou professionnelle avec des relations d'affaires et même du travail. Les réseaux sociaux ouvrent aux utilisateurs la possibilité de créer, partager, consulter et diffuser des informations, ils sont devenus l'assise sur laquelle dépendront les ressources humaines et les stratégies de marketing.
Dans ce cadre, il est intéressant de présenter les différents types de réseaux sociaux, les plateformes Twitter et Facebook ainsi que leurs fonctionnalités et leurs caractéristiques.

2 Le web 2.0

Web 2.0 est un terme qui décrit l'évolution des tendances dans l'utilisation des technologies du Word Wide Web. La conception Web vise à améliorer la créativité, le partage sécurisé de l'information afin d'accroitre la collaboration, et d'améliorer la fonctionnalité du Web [3] où l'utilisateur est lui-même un co-créateur de contenu. Cela signifie que Web 2.0 n'est pas un changement technique de la version du web, mais un changement de la manière d'interagir avec le web.

Ce nouvel aspect collaboratif du web, avec une accessibilité exponentielle à la technologie, a apporté d'une façon générale l'aspect "social" des communautés d'utilisateurs au web, qui a provoqué la naissance des médias sociaux prenant le rôle d'une boite à outils qui aide à la création et au partage d'information via le web.

3 Médias sociaux

Les médias sociaux qui sont une conséquence du Web 2.0, désignent des données envoyées par leurs utilisateurs, le partage d'information et la création et la mise en ligne de contenu et

l'interaction sociale. Les médias sociaux ont transformé la définition du média traditionnel "one to many " à "many to many " ou chacun peut devenir un producteur.
Un même média peut être accédé via plusieurs moyens de communications : un site web, un courriel ou une application mobile. Ils peuvent être vus comme un groupe d'application en ligne (blogs, réseaux sociaux, site de partage, etc) qui se fondent sur la philosophie et la technologie du Web 2.0.
Parmi les outils qui appartiennent aux médias sociaux, on trouve :

3.1 Les blogs (et même vlog)

Site Web personnel tenu par un ou plusieurs blogueurs qui s'expriment librement et selon une certaine périodicité, sous la forme de billets ou d'articles, informatifs ou intimistes, datés, à la manière d'un journal de bord, signés et classés par ordre chronologique [4] .

3.2 Les wikis

Un wiki est un type de site Web qui permet à plusieurs personnes d'écrire en collaboration ou modifier le contenu, même avec peu ou aucune connaissance de programmation ou de balisage langues du Web [5]. Exemple : Wikihow, Wikipédia.

3.3 Les portfolios

Un portfolio est un petit site web qui représente un dossier personnalisé, en partie photographique ou illustré, constitué pour qu'un professionnel pousse présenter ses travaux ou de promouvoir ses activités [6].

3.4 Les micro-blogs

Un microblog est un blog au contenu textuel court [7]. Ce terme est apparu en 2005, le contenu partagé (textes, liens, citations, vidéos, photos...), est comme pour les blogs, restent archivés sur le Web. L'usage de ce terme s'est généralisé en 2006 avec l'apparition de Twitter et renforcé par l'apparition de Tumblr.

3.5 Les réseaux sociaux

Les réseaux sociaux sont au cœur des médias sociaux, ils proposent un service de publication et de partage de contenu multimédia (article, vidéo, photo, musique, etc.) tout en reliant les personnes avec le contenu partagé ou avec autres personnes. Parmi les réseaux sociaux les plus utilisés à l'échelle international Facebook, Twitter, YouTube, Google+ [8] [9] , etc.
En france et dans le monde Twitter, Facebook et Google sont les principaux médias sociaux utilisés [10] [11] :

En 2015, 3,025 Milliards d'internautes à travers le monde 2,060 Milliards parmi eux sont

actifs et 68% des internautes et 28% de la population mondiale. L'europe est en 2 éme position dans l'utilisation des réseaux sociaux avec 44%, 42% en France.

— Facebook
C'est le plus large réseau social au monde, le plus idéal pour communiquer, chatter partager des vidéos, des images, des status et des opinions avec les amis et les proches. Il a été lancé le 4 février 2004 (bêta) et son ouverture public en 2006 le 26 septembre avec 1.65 milliards d'utilisateurs actifs mensuels (MAU) (Avril 2016) [12], 30 millions d'entre eux en France, les utilisateurs actifs mensuels sur mobile avec 1.314 milliards, 24 millions d'entre eux en France. Des utilisateurs actifs quotidiens (DAU) avec 968 millions, 20 millions en France.

— Twitter
il est centré sur le réseau d'amis proche, permet la liberté d'expression avec n'importe quel utilisateur(ami, marque ou personnalité), permet aussi le partage des messages qui aident a limiter l'opinion de l'internaute à 140 caractères seulement qui conduit à avoir une bonne analyse de sentiments. Il a été lancé le 21 mars 2006, il a 304 millions d'utilisateurs actifs mensuels (MAU) 2.3 millions en France. 500 millions de tweets envoyés chaque jour, 320 comptes créent chaque minute, d'utilisateurs actifs mensuels (MAU) avec 1.314 milliards, 24 millions d'entre eux en France.

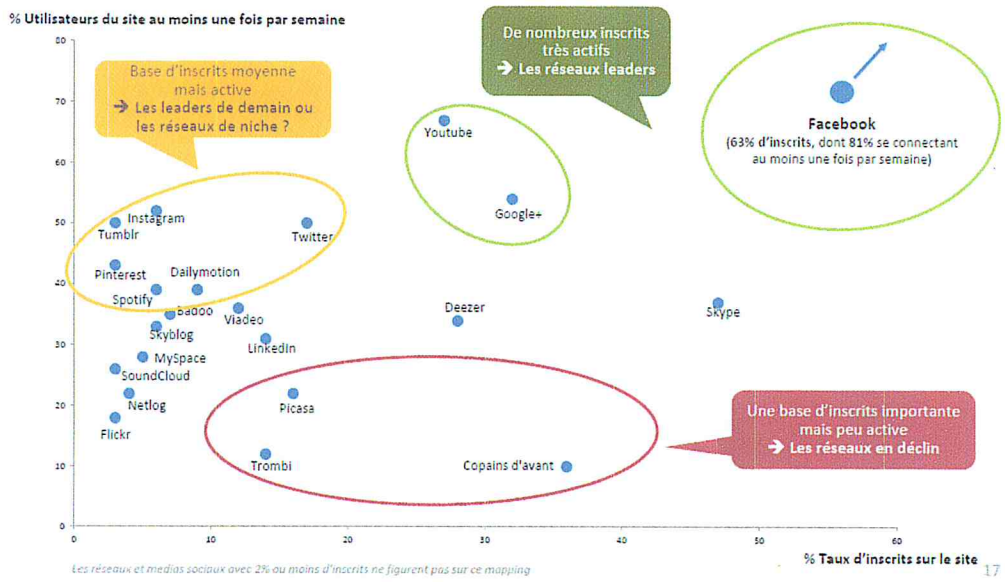


Figure 1.1 – Fréquence d'utilisation des réseaux sociaux [13]

4 L'impact et les conséquences des réseaux sociaux sur les entreprises

Les réseaux sociaux apparaissent comme le canal de communication le plus efficace dans le marketing qui affecte directement le rendement des entreprises. Ils représentent une vitrine virtuelle qui aide à présenter les biens et les services de l'entreprise, souvent sous forme de

publicité ciblée.

Facebook et twitter sont les réseaux sociaux les plus sollicités par les entreprises pour la publicité, car ils offrent des outils très avancés en termes de spécification de la clientèle ciblée par la publicité.

Ce ciblage de public offert par les réseaux sociaux n'est pas possible que par la collecte des informations sur les individus inscrits sur le réseau et qui ont été générées par leurs différentes interactions au sein de réseau (des " j'aimes", des favoris, des clicks, des relations etc). Cette collecte aide les réseaux sociaux à créer un profil pour chaque individu contenant ses envies, ses passions, ses préférences dans tous les côtés de la vie, ainsi que ses données plus personnelles comme l'âge, le sexe, intérêts et même ses habitudes de consommation.

Les réseaux sociaux sont un espace virtuel abstrait, mais ils peuvent affecter d'une manière significative les achats en ligne et en magasin physique. C'est le cas d'une startup de café nommée " Starbucks" qui a augmenté ses ventes jusqu'à 38% grâce au phénomène d'amplification du réseau social Facebook [9] .

Les réseaux sociaux et les entreprises vivent sur le principe de l'intérêt commun où les réseaux sociaux aident à augmenter le chiffre d'affaires des entreprises, et les entreprises représentent la ressource de revenu la plus importante aux réseaux sociaux. [9]

5 Twitter

Twitter est actuellement la plate-forme de microbloggage la plus populaire. Il peut répondre a cette question que faites-vous ? Cependant l'utilisation a pris une autre piste ou les utilisateurs échangent des avis et des informations, la question est devenu quoi de neuf ? Twitter est alors un site utile sur différents niveaux. Il peut être utilisé par des passionnés d'informatique pour effectuer de la veille technologique, par des entreprises pour communiquer, par des célébrités pour informer leurs fans.

Il est à la fois un réseau social qui permet aux utilisateurs de partager et de consulter un contenu varié (textuel et multimédia), et à la fois un microblog car le contenu partagé est limité à 140 caractères par message appelé " tweet" à l'aide de service SMS (short Message service) qui est limité a 160 caractères. Cet outil de réseau social permet d'améliorer la communication, cela devient en quelque sorte un média social.

Bien que ce service soit principalement connu aux états-Unis, il se popularise continuellement et commence à être utilisé en France et plus tard dans le monde. Lancé en Juillet 2006 par Jack Dorsey, Twitter est maintenant dans le top 10 des sites Internet les plus visités .

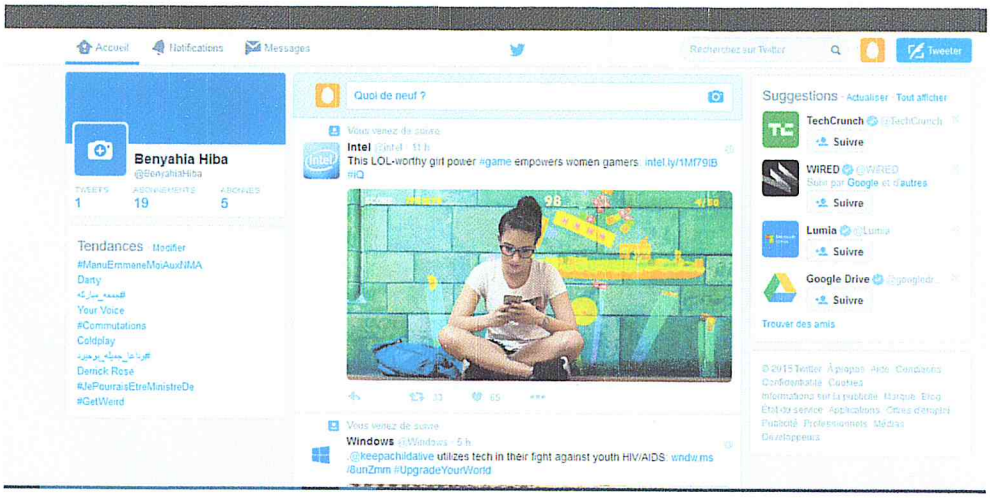


Figure 1.2 – Capture d'écran de la page d'accueil Twitter.

Fin 2011, twitter a lancé les pages entreprises qui ont des comptes spécifiques dédiés aux marques ou mêmes des entreprises. Vous pouvez consulter les premières entreprises partenaires de twitter (@AmericanExpress, @BestBuy, @bing, @CocaCola, @Hp, @intel, etc) pour avoir un aperçu de ces nouveaux types de profils.

5.1 Historique

L'histoire de Twitter a débuté autour de 2005 [14] chez un petit groupe de collaborateurs qui travaillait au sein de l'entreprise de démarrage "Odeo", fondée par Noah Glass à San Fransicso. Les trois autres comparses sont Jack Dorsey, Biz Stone et Evan Williams. Jack Dorsey, alors dans la fin vingtaine, entretient une passion pour la programmation. Inspiré par les communications radio des chauffeurs de taxi, le jeune homme rêve depuis quelques années déjà d'un système de communication par messages textes qu'on pourrait envoyer à un groupe d'amis par l'entremise de son téléphone cellulaire.

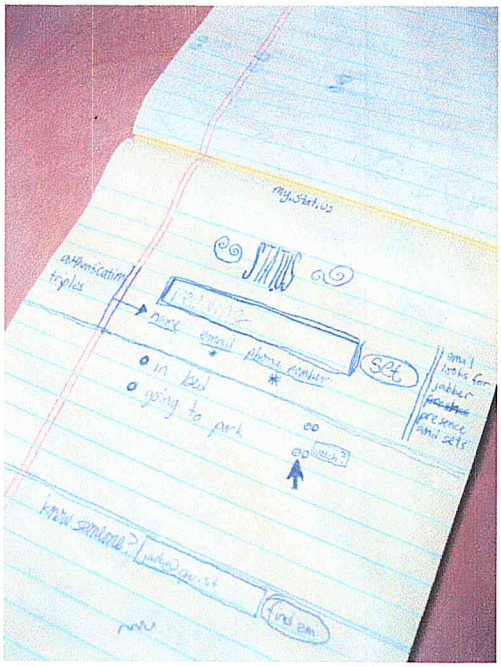


Figure 1.3 – Croquis préliminaire sur papier de Twitter, en 2006, par JackDorsey "premier essai" [15].

Twitter, qui à ses débuts se nomme Twttr en écho au site de photos Flickr, voit le jour en 2006 [14]. Le 21 mars2006, Jack Dorsey envoie le premier gazouillis ("Just setting up my twttr") [16] . L'été suivant, la plateforme est ouverte au public. à l'époque, il n'y avait pas de limite au nombre de caractères permis. Twitter compte alors une centaine d'abonnés. En avril 2007, Twitter devient une véritable entreprise et Jack Dorsey en prend les commandes.

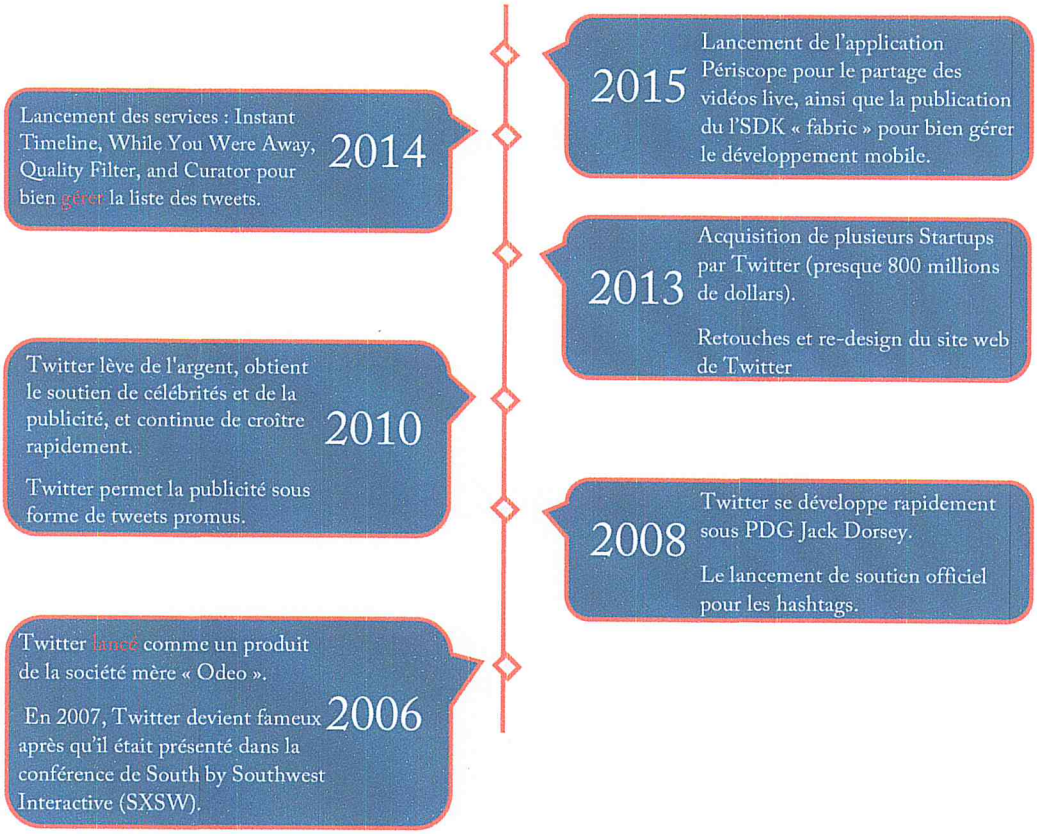


Figure 1.4 – Vue de l'ensemble d'historique de Twitter dans la timeline

Twitter connaîtra son véritable envol un an plus tard, au Festival South by South West (SXSW) à Austin, au Texas. Lieu de rencontre de l'avant-garde techno, SXSW récompense Twitter en lui accordant un Web Award. Les participants à la conférence adoptent Twitter sur-le-champ. C'est le coup de baguette magique qu'il fallait. En l'espace d'un week-end, le nombre de gazouillis envoyés passe de 20 000 à 60 000 [16].

Avec les années, les gens ont développé le réflexe de se tourner vers Twitter en temps de crise (séisme en Haïti, révolution en égypte) pour y trouver des informations en premier plan.

5.2 Tweet

Un tweet est un message posté sur Twitter de 140 caractères maximum, comme nous l'avons dit plus haut, il est utilisé par l'intermédiaire des SMS, (limité è 160 caractères), twitter prends 140 et conserve 20 pour son nom d'utilisateur.

5.3 Caractéristiques

Les utilisateurs de Twitter publient des messages " tweets " qui sont visibles par défaut pour tous (dites : public) et qui sont envoyés directement à leurs abonnés appelés " followers ". Twitter est utilisé pour la publication d'informations, et surtout pour exprimer des sentiments sur un sujet particulier qui sont subjectifs et limités à 140 caractères [17] :
Dans un tweet, les utilisateurs ont plusieurs possibilités, parmi eux :

- **Possibilité de republier :** Une personne " A " reçoit un tweet de la part d'une personne " B " alors tous les abonnées de " A " reçoivent le même tweet. Un tweet republié s'appelle un retweet et commence souvent par le texte " RT@ B ".
- **Possibilité d'ajouter des URLs :** Un tweet peut contenir des liens web (URL) mais vu que ces derniers peuvent dépasser 140 caractères, Twitter propose aux utilisateurs des services de " réduire " le nombre de caractères des URLs, en laissant toujours les pages accessibles par des liens très courts.
  Il existe plusieurs services de réduction comme "bit.ly" ou "tinyurl.co" et " t.co " ce dernier est crée par Twitter, il est utilisé seulement pour les URLs insérées dans les tweets. Par exemple l'URL : "http://www.iro.umontreal.ca/rubrique.php3?id_rubrique=13" deviendra : "http://t.co/NrUGjAtx" par le service "t.co".
- **Possibilité de contenir des hash-tags :** Un hash-tag est un mot précédé par le symbole "#". Il représente une étiquette attribuée par l'auteur de tweet pour caractériser brièvement son sujet. En cliquant sur le hash-tag (ex : #BMW), la liste des tweets qui contiennent ce hash-tag s'affiche grâce à une recherche par mots clés offerte par twitter, qui permet de retrouver tous les tweets parlant du même sujet pour que les autres puissent les suivent.
- **Possibilité d'être adressé à une personne spécifique :** Un tweet peut être adressé à une personne " A " en ajoutant le caractère " @ " avant son nom " @A ". Cette personne sera automatiquement informée, et le tweet peut être vu par les abonnés de l'auteur de tweet et les abonnés de la personne adressée aussi.

5.4 Fonctionnalités

- Le nom d'un utilisateur est un identifiant précédé par @.
- Dans un tweet on peut étiqueter les sujets les sujets dont on parle. Un sujet est précédé par un dièse # pour former un hashtag, mot-clic en français.
- Une réponse à un tweet de l'utilisateur x commence toujours par @x.
- Un retweet (réémission d'un tweet) commence par RT @Y tel que Y et le t'expéditeur du tweet original.
- Pour mentionner un utilisateur dans un tweet il suffit de taper son nom précédé par le caractère @.

5.5 Statistiques

5.5.1 L'usage de Twitter à travers le monde

Une étude, publiée par Beevolve [18], décrit le profil des utilisateurs de Twitter à travers le monde. Globalement, on apprend que 53% des membres sont des femmes et que la majeure

partie des usagers sont âgés de 15 à 25 ans (73,7%). Au niveau des pays, les états-Unis représenteraient plus de 50% de l'ensemble des membres, la France étant située à la septième place, avec seulement 1,76% des utilisateurs. Plus de 81% des comptes accueillent moins de 50 followers, tandis que seuls 3,4% sont suivis par plus de 500 personnes.

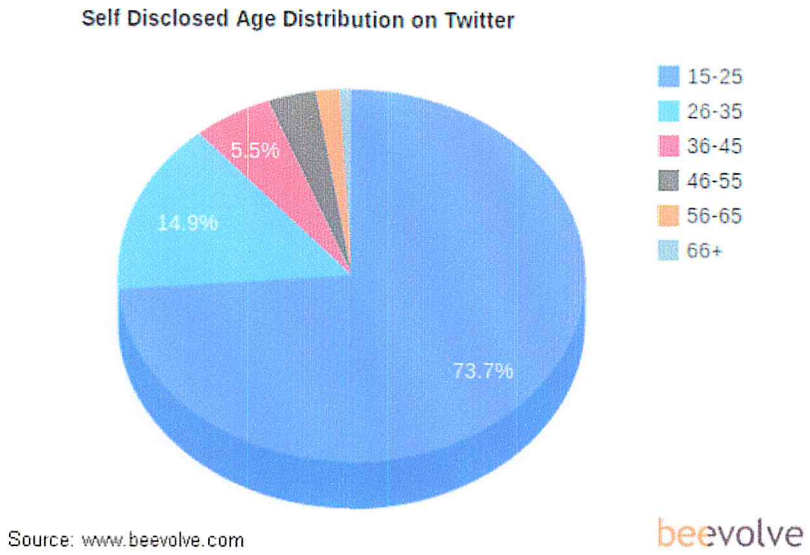


Figure 1.5 – Distribution de l'âge sur Twitter [18].

5.5.2 L'usage de Twitter en France selon conScore

Au début de l'année 2013 [10], comScore a publié une étude intéressante sur l'usage de Twitter en France. Au total, pas moins de 5,5 millions de français ont visité le site twitter.com en novembre 2012 (+53% en un an). Au niveau de la démographie, quelques surprises sont â signaler : les plus de 55 ans sont les plus nombreux. Cette tranche d'âge a doublé en un an, alors que les 25-34 ans sont passés dans le même temps de la première place â la troisième (comparaison novembre 2011-2012).

6 Facebook

Facebook est le plus grand réseau social, il compte presque 1 milliard et 400 millions utilisateurs actifs par moi , et le deuxième site web le plus visité au monde juste après Google . Il permet à ses utilisateurs de partager tout type d'information (texte, image, vidéo), et offre aussi une plateforme de communication via la messagerie instantanée et directe entre les utilisateurs, ainsi que des appels audio ou audio-visuel.

Facebook est un réseau social qui peut être utilisé de plusieurs manières : il peut être utilisé comme un outil pour suivre les actualités via des abonnements à des pages, ou bien

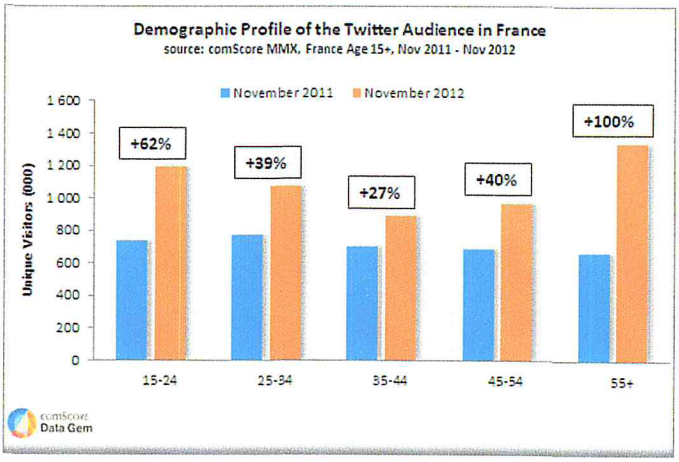


Figure 1.6 – Profile démographique des visiteurs de Twitter en France [19].

comme un outil de loisir grâce aux applications et les jeux au sein de Facebook, ou même comme un outil professionnel de marketing pour atteindre un grand nombre de clients.

Facebook, contrairement à Twitter, ne pose aucune contrainte sur la quantité d'information que les utilisateurs veulent partager, la chose qui donne plus de liberté aux utilisateurs pour s'exprimer et qui, par conséquence, donne plus d'informations à Facebook sur ses utilisateurs.

6.1 Historique

En 2004, Mark Zuckerberg un étudiant à l'université de Harvard a crée un site pour le partage de contenu entre étudiants appelé " The Facebook ". Cependant, ce site a connu une acceptation chaleureuse de la part des étudiants, par conséquence, le réseau a été offert aux étudiants d'autres universités et même aux écoles secondaires. En 2006, Facebook devient un réseau social ouvert à tout le monde, il permet aux utilisateurs d'etre " amis ", et lui permet aussi de publier des messsages appelés " statuts " ou de contenu multimédia. Facebook commence à offrir aussi la possibilité de créer des " Pages " pour des produits, des personalités, des groupes artistiques, et de promouvoir celles-ci à l'aide de publicités. Vers la fin de 2007, Facebook a eu 100.000 pages de produits et d'affaires.

Twitter : étude de l'entreprise	Données
Nombre total d'utilisateurs enregistrés Twitter	645,750,000
Nombre total d'utilisateurs actifs de Twitter	289,000,000
Nombre de nouveaux utilisateurs de Twitter inscrivant tous les jours	135,000
Nombre de visiteurs uniques du site Twitter chaque mois	190 million
Nombre moyen de tweets par jour	58 million
Nombre des requêtes de moteur de recherche Twitter tous les jours	2.1 milliards
Pourcentage des utilisateurs de Twitter qui utilisent leur téléphone pour tweeter	43%
Pourcentage des tweets qui proviennent de tiers candidats du parti	60%
Nombre de personnes qui sont employées par Twitter	2,500
Nombre d'utilisateurs actifs de Twitter chaque mois	115 million
Pourcentage de personnes qui ne tweetent pas mais qui regardent les tweets d'autres personnes	40%
Nombre de jours qu'il faut pour 1 milliard de tweets	5 jours
Nombre de tweets qui se produisent chaque seconde	9,100

Table 1.1 – Tableau des statistiques sur Twitter [1]

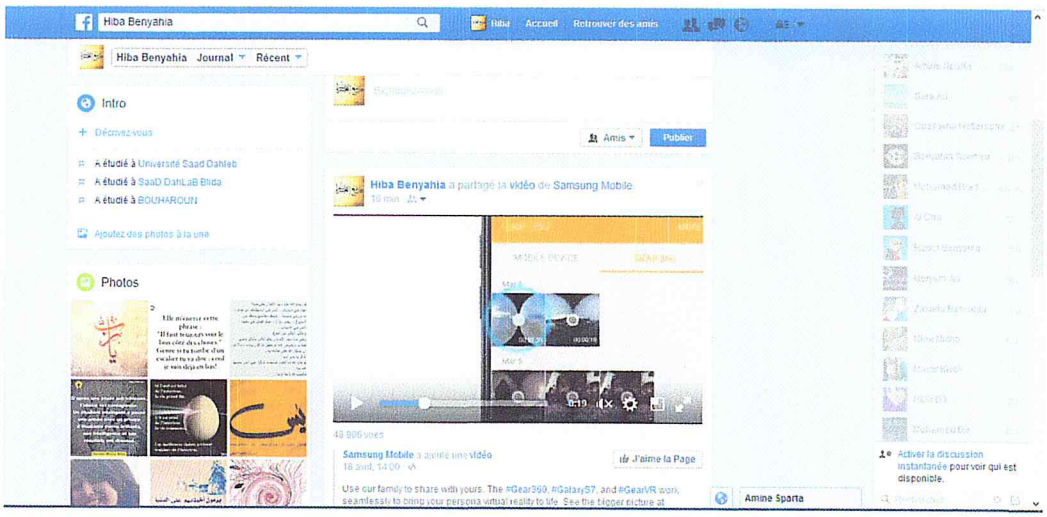


Figure 1.7 – Capture d'écran du profile Facebook.

6.2 Caractéristiques

6.2.1 Contenu Publié

Facebook offre à ses utilisateurs une variété de types de contenu qui peuvent être partagé :

- **Texte :** Les utilisateurs de Facebook peuvent partager des bouts de textes dits " Statuts ", ou même écrire des paragraphes longs sans aucune contrainte sur le nombre de caractères écrits.
- **Image :** Les images dans Facebook peuvent être partagées une par une ou par groupe de photos appelé " Album ". Facebook utilise des algorithmes de détection et d'identification de visage pour cadrer les visages trouvés dans une photo pour faciliter la tâche d'identification des amis.
- **Vidéo :** Dans le début de Facebook, les utilisateurs ne peuvent pas partager des vidéos sauf que des liens vers des vidéos dans des sites web externes (comme YouTube). Maintenant tous les utilisateurs peuvent partager des vidéos courtes ou longues d'une manière directe sur Facebook (depuis le fichier) sans utiliser des sites externes, des vidéos de différents formats et qualité sont acceptées et même des vidéos filmées en HD ou 360°. En Novembre 2015, Les vidéos de Facebook entaillent plus de 8 milliards de vues quotidiennes .
- **Hashtag :** Les hashtags transforment les sujets et les locutions en liens cliquables à l'intérieur des publications. Ils aident les gens à trouver des publications portant sur des sujets qui les intéressent. Les hashtags commencent par # (le symbole dièse) immédiatement suivi d'un sujet ou d'une locution.
- **Identification :** Les identifications peuvent pointer vers des amis ou n'importe qui d'autre sur Facebook. L'ajout d'une identification crée un lien que les personnes peuvent suivre pour en apprendre plus.
- **Sentiment :** Les utilisateurs de Facebook peuvent aussi partager leur humeur actuelle, ou bien leur sentiment vers un lieu ou une activité ou même leur état de santé.
- **Emplacement :** L'emplacement de l'utilisateur peut être partagé en utilisant deux manières, soit par l'activation de la géolocalisation qui va détecter automatiquement l'emplacement de l'utilisateur ou bien en faisant un lien vers un lieu enregistré déjà dans Facebook (Généralement une ville ou un monument ou site touristique... etc.)
- **Activité :** Les utilisateurs peuvent dire ce qu'ils mangent, ce qu'ils lisent, ce qu'ils regardent à la télévision ou bien ce qu'ils font comme sport.
- **Messagerie Instantanée :** Facebook offre à ses utilisateurs un service gratuit de messagerie instantanée qui donne aux utilisateurs la possibilité d'envoyer des messages texte, des émoticônes, des images et des vidéos, des fichiers de n'importe quel type à leurs amis ou abonnés.
- **Jeux et Applications :** Des applications et jeux sont développés pour améliorer l'expérience des utilisateurs sur Facebook, et aussi pour les faire passer plus de temps sur Facebook. Les applications et les jeux sur Facebook sont créés par des développeurs tiers, par conséquence les informations qui se trouvent dans ces applications et jeux sont enregistrées sur les serveurs des développeurs non hébergés par Facebook.

6.3 Statistiques

Les statistiques mentionnées dans le tableau précédent montrent d'une façon très claire que Facebook est devenu un monde digital entier avec presque 1.5 milliard d'utilisateurs dont la moitié se connecte chaque jour, le fait qui a attiré l'attention des grandes et petites entreprises pour profiter d'une publicité extrêmement ciblée. Le grand nombre des applications Facebook reflète la politique d'attraction des utilisateurs de Facebook, pour qu'ils puissent passer plus de temps dans le réseau, et par conséquence, voir plus de publicités.

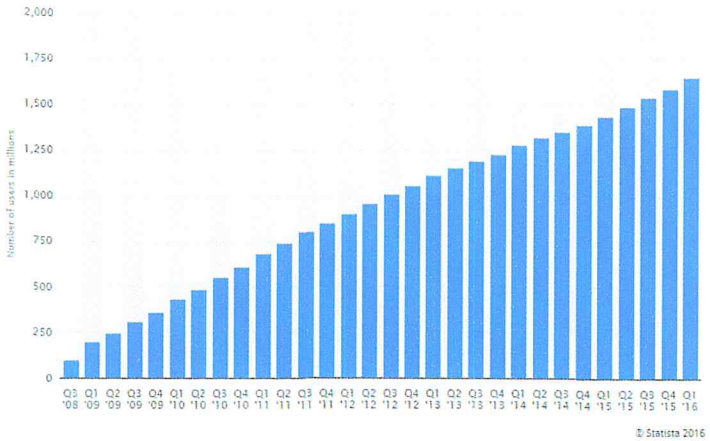


Figure 1.8 – Nombre d'utilisateurs enregistré sur Facebook [20].

Facebook : étude de l'entreprise	Données
Nombre total d'utilisateurs actifs	1,440,000,000
Pourcentage d'utilisateurs qui se connectent tous les jours	48%
Nombre total de page Facebook	74,200,000
Le nombre total des applications et des sites Web a intégré en Facebook	7,000,000

Table 1.2 – Tableau des statistiques sur Facebook. [2]

7 Conclusion

Dans ce chapitre, nous avons exploré les idées principales sur le premier axe de notre travail : les réseaux sociaux, spécifiquement, Twitter en tant que microblog le plus utilisé en monde, et Facebook en tant que réseau social le plus connu du monde.

Nous pouvons conclure que l'utilisation exponentielle des millions et même milliard d'internautes des réseaux et médias sociaux comme Facebook et Twitter enjendreraient une forte réponse des entreprises pour promouvoir leurs produits et services.

Dans le prochain chapitre nous allons présenter les approches de classifications de sentiments en se basant sur les méthodes d'apprentissage supervisé, ainsi que sur les travaux récents et les domaines d'application

Chapitre 2

Etat de l'art : Les approches de classification de sentiments et les comportements humains

1 Introduction

Avec l'essor des multimédias et des réseaux sociaux, les internautes et les entreprises, génèrent des données précieuses et tès importantes qui contiennent un réservoir volumineux de connaissances, qui couvrent le domaine de recherche d'informations.
Ils stockent ces informations afin de les utiliser dans différents domaines tels que l'analyse des sentiments.

L'analyse des sentiments s'utilise entre autres pour la détection d'opinion, sur des sites web et des réseaux sociaux. Elle consiste à rechercher des textes évaluatifs sur internet tels que des critiques, et à analyser de manière manuelle ou automatique, les sentiments qui y sont exprimés afin de comprendre l'opinion publique, que ce soit pour les raisons personnelles, commerciales ou politiques. Ansi de nombreux systèmes autonomes ont dèjá été développés pour l'analyse automatique des sentiments. Généralement, ces systèmes étaient entraînés aux textes non traditionnels tels que les messages envoyés via les réseaux sociaux, ceux-ci constituent une source précieuse d'opinions échangées entre internautes.

2 Généralité

Lorsque les consommateurs doivent prendre une décision ou un choix concernant un produit, une information importante est la réputation de ce produit, qui est dérivé de l'opinion des autres. L'analyse des sentiments peut révéler ce que les autres pensent d'un produit.

La première application de l'analyse des sentiments donne ainsi une indication et des recommandations dans le choix des produits en fonction des opinions des utilisateurs. Là, quand le client choisie un produit, il est généralement attiré par certains aspects spécifiques du produit. Une note globale unique pourrait être trompeuse. L'analyse des sentiments peut

regrouper les opinions des inernautes et les évaluer sur certains aspects du produit.

Une autre utilité de l'analyse des sentiments est pour les entreprises qui veulent connaître l'opinion des clients (l'opinion publique) sur leurs produits. Ils peuvent alors améliorer les caractéristiques que les clients ont trouvé insatisfaisantes. L'analyse des sentiments peut également déterminer les caractéristiques qui sont importantes pour les clients.

Liu et al. [21] ont dit que l'opinion est l'expression des sentiments d'une personne envers une entité étant subjective et opposée, ainsi le terme fouille d'opinion est indiqué pour évoquer le traitement automatique d'opinion et de sentiment et aussi la subjectivité dans le texte.

Pour Weibe et al. [22] la subjectivité est l'expression linguistique de l'opinion d'une personne, des sentiments, des émotions, des évaluations, croyances et des états privés (la spéculation) qui est d'après Quirck et al. [23] "private state" est un état qui n'est pas ouvert à l'observation objective ou de vérification.

3 Littérature

3.1 Opinion

Un opinion est un jugement, avis, sentiment qu'un individu ou un groupe émet sur un sujet, des faits, ce qu'il en pense [24].

3.2 Sentiment

Un sentiment est le processus affectif durable en l'absence des objets déclencheurs et du contexte, par exemple l'amour et la haine, s'associent aux humeurs, pour entretenir un tonus de base qui colore la vie psychique, ainsi la bonne humeur (l'euphorie) et la mauvaise humeur (dysphorie, dépression). [25]

3.3 L'analyse de sentiments

L'analyse de sentiment (SA), également appelée l'exploitation ou l'analyse d'opinion (Opinion Mining OM), est le domaine d'études qui analyse les avis, les sentiments, les évaluations, et les émotions des personnes envers des entités telles que des produits, des services, des organismes, des individus, des issues, des événements, des sujets, et leurs attributs [26].

4 Formulation de la problématique

L'analyse des sentiments consiste à identifier et extraire le sentiment ainsi que plusieurs d'autres informations nécessaires pour faire une bonne analyse. L'exemple suivant présente un cas réel à étudier. [27] " (1) J'ai acheté un iPhone 6 il y a six mois. (2) Je l'aime simplement. (3) La qualité des photos est étonnante. (4) Son poids est bien léger. (5) Cependant, mon frère pense qu'il est trop cher pour lui."

— Un sentiment doit avoir une cible qui peut être n'importe quelle entité (phrase 2) ou aspect de l'entité (phrase 3 et 4) au sujet de laquelle un sentiment a été exprimé. Une entité (e) est un produit, un service, un sujet, une issue, une personne, une organisation, ou un événement. Elle est décrite avec une paire : e (T, W), où T est une sous-partie ou dite aspect de l'entité et W est un ensemble de ses attributs [26].
— (s) est un sentiment positif, négatif, ou neutre, ou un score numérique exprimant l'intensité sentiment (par exemple, 1 à 5).
— Le positif, le négatif et le neutre s'appellent des orientations de sentiment ou les polarités.
— Il faut définir la source de sentiment (dite : Opinion holder h ou sentiment holders) certaines études [28] considérent que la PERSONNE et l'ORGANISATION sont les seuls supports possibles de sentiment. (Dans l'exemple, le support de sentiment de la phrase 5 n'est pas le même des autres phrases)
— La date de publication de sentiment est importante pour les études statistiques.

On peut représenter un sentiment par le quintuple :

$$(e_i, a_{ij}, s_{ijkl}, h_k, t_l) \tag{2.1}$$

e_i : est le nom d'une entité
a_{ij} : est un aspect de e_i
s_{ijkl} : est le sentiment sur l'aspect a_{ij} de l'entité e_i
h_k : est le support de sentiment
t_l : est le moment où l'opinion est exprimée par h_k

5 La subjectivité du texte

L'étude des textes subjectifs s'est développée en relation avec leurs accessibilité sur le web, donc ils favorisent l'expression des internautes afin de générer des textes non formatés, non signés et non indexés qui posent des problèmes d'analyse qu'il faut les formaliser pour pouvoir être traités. L'analyse des sentiments oblige les chercheures de faire la distinction au niveau d'un document entre textes subjectifs et textes objectifs. Ce qui rend la décidabilité de la classification du texte crucial (problème de classification). Alors deux démarches sont applicables :

— Soit le texte comporte de l'opinion, reste juste à le classer selon sa polarité positive ou négative.
— Soit il faut repérer tout d'abord la phrase porteuse de l'opinion (classement objectif/subjectif) pour ensuite leur attribuer une polarité (positive ou négative). [29]

Trancher entre le caractère objectif et le caractère subjectif revient à détecter dans le texte les indices de la présence de la subjectivité. La présence d'adjectif est un très bon indicateur de caractère subjectif de phrases [29]. D'autres proposent d'identifier des indices co-occurents de subjectivité pour en déduire le caractère subjectif des phrases. On peut aussi tenter de détecter la subjectivité d'un texte en utilisant différents indices et caractéristiques (enmplacements des mots .. etc). Enfin, certains ont tenté de séparer les opinions des faits, au niveau du document et au niveau de la phrase, en utilisant un classificateur Naïf Bayes.

6 Les approches de classification de sentiments

Les travaux existants sur l'analyse des sentiments peuvent être classés à partir de différents points de vue : la technique utilisée, vue du texte, le niveau de détail de l'analyse de texte, niveau de notation, etc. D'un point de vue technique, nous avons identifié l'apprentissage par machine, basé lexique, basée sur la règle statistique.

6.1 L'approche à base de régles

Dans cette approche on cherche les mots d'opinion dans un texte, puis les classer en fonction du nombre de mots positifs et négatifs. Elle considère différentes règles de classification, tels que le dictionnaire de polarité, mots de négation, mots de rappel, les expressions idiomatiques, des émoticônes, des opinions mixtes, etc.

6.2 Les approches lexicales

Ces approches utilisent des dictionnaires de mots subjectifs. Ces dictionnaires peuvent être généraux comme le General Inquirer [30], Sentiwordnet Figure 2.1, Opinion Finder [31], NTU Sentiment Dictionary (NTUSD) [32]. Pour chaque dictionnaire une polarité est associée a priori à chaque mot afin de donner un score d'opinion pour chaque document en fonction le l'apparition du mot issus de ces dictionnaires dans le texte. L'approche basée sur le lexique consiste à calculer le sentiment de polarité pour un review en utilisant l'orientation sémantique des mots ou des phrases dans la revue ."l'orientation sémantique" est une mesure de la subjectivité et de l'opinion dans le texte.

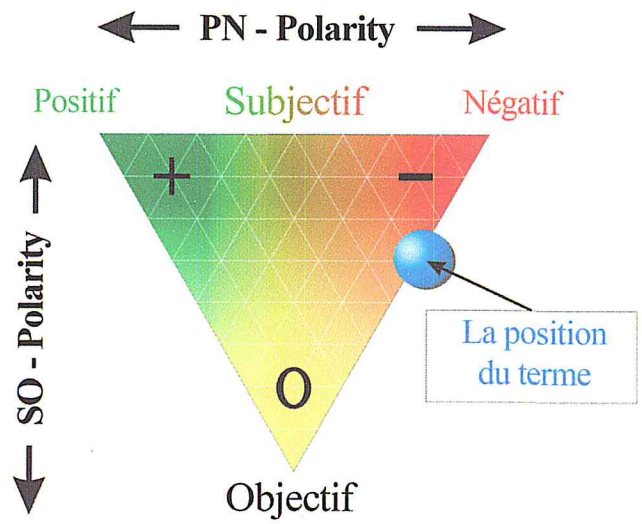


Figure 2.1 – SentiWordNet : Analyse des sentiments á l'échelle des mots entre subjectivité et objectivité et positivité et négativité [33].

6.3 Méthodes statistiques

Les méthodes statistiques les plus populaires de l'analyse de sentiment sont "Le Classification Naïve Bayésienne" et les "Séparateurs à Vaste varge SVM". En alimentant un algorithme d'apprentissage automatique d'un large corpus d'entraînement de textes affectivement annotés, il est possible pour le système d'apprendre non seulement la valence affective de l'affect mots-clés (comme dans l'approche de repérage de mot-clé (spotting approach)), mais aussi de prendre en compte la valence de l'autre arbitraire mots-clés (comme affinité lexicale), les fréquences de ponctuation, et le mot de co-occurrence.

6.4 Sélection et classification d'un matériau subjectif

Certains travaux choisissent du premier coup des textes subjectifs. C'est exactement les avis des consommateurs ou des critiques de produits. Comme les critiques télévisuelles et vidéoludiques de Sébastien Gillot dans fouille d'opinions. Les sites d'avis de consommateurs sont la source de la subjectivité sur le produit/service. La classification devient plus compliquée dans le cas des blogs, ou l'avis peut être noyé dans un contenu qui ne traite pas seulement du produit/service qui nous intéresse.
D'autres recherches utilisent des indice de subjectivité comme l'émoticons au sein d'une phrase donc il sera classé comme un matériau subjectif.

6.5 Les Méthodes d'apprentissage automatique

Dans lesquelles on utilise plusieurs algorithmes d'apprentissage pour déterminer le sentiment qui s'entraînent sur un ensemble de données (datasets) connu.

L'apprentissage automatique est la science qui donne aux ordinateurs (ou machine en générale) la capacité d'apprendre sans être explicitement programmée [34]. L'apprentissage automatique est utilisé dans le domaine de l'analyse de sentiment dans l'étapes de classification de sentiment où on doit décider si un sentiment appartient à la classe des sentiments positifs ou négatifs ou neutres.

6.5.1 Apprentissage supervisé

Le processus d'apprentissage automatique supervisé se passe en deux phases. Lors de la première phase (dite phase hors ligne ou phase d'apprentissage), il s'agit de déterminer un modèle de données étiquetées. Là, on donne des phrases dont les sentiments ont été reconnus et étiquetés par un ou plusieurs spécialistes de domaine (ces phrases étiquetées sont appelées : Training set).

La seconde phase (dite en ligne, ou phase de test) consiste à prédire l'étiquette d'une nouvelle donnée, connaissant le modèle préalablement appris. Là on donne aux classifieur des phrases dont les sentiments ne sont pas reconnus (Test set) et on obtient comme résultat l'orientation du sentiment de cette phrase [35]. Parmi les approches d'apprentissage supervisé les plus utilisées :

— Classification probabiliste : utilisant les classifieurs suivant : classifieur Naïve Bayésien, Réseau Bayésien, principe d'Entropie Maximale.

— Classification linéaire : parmi ces méthodes : Séparateurs à Vaste Marge (Support Vector Machine ou SVM), Réseau de Neurones.
— Classification par les arbres de décision : utilisant les méthodes ID3, C4.5 ou C5, et Maximum Spanning Tree MST (Arbre couvrant de poids minimal).

1. Naïve Bayes est un algorithme de classification simple mais efficace. L'algorithme de NB est très utilisé dans la classification des documents [36], [37]. L'idée de base est d'estimer la probabilités de catégories qui reçoivent un document de test à l'aide des probabilités conjointes de mots et de catégories. la partie naïve d'un tel modèle est dans l'indépendance des mots. La simplicité de cette hypothèse rend le calcul de classification naïve bayésienne beaucoup plus efficace [38].
2. Machines à Vecteurs de Support (SVM), un classificateur discriminatif est considéré comme la meilleure méthode de classification de texte [37], [39], [40] et [41]. La méthode SVM est une méthode de classification statistique proposée par Vapnik [42], Elle est basée sur le principe structurel de minimisation des risques à partir de la théorie de l'apprentissage informatique, SVM vise à trouver une surface de décision pour séparer les points de données d'entraînement en deux classes et prend des décisions sur la base des vecteurs de support qui sont sélectionnés comme les seuls éléments efficaces dans l'ensemble de la formation. De multiples variantes de SVM ont été développés dans lesquels la classe multi SVM est utilisée pour la classification du sentiment [43].
3. L'idée derrière l'algorithme de classification barycentre (centre de gravité) est extrêmement simple et directe (Songho tan, 2008). Initialement, le vecteur prototype ou vecteur centroïde (CG) pour chaque classe de formation est calculé, puis la similitude entre un document d'essai à tous centroïde est calculée, enfin, sur la base de ces similitudes, le document est attribué à la classe correspondante au centre de gravité le plus semblable.
4. Le voisin de K-plus proche (KNN) est un exemple typique de classificateur à base qui ne construit pas, une représentation déclarative explicite de la catégorie, mais repose sur la catégorieétiquettes apposées sur les documents de formation similaires au document de test.étant donné un document d'essai, le système trouve les k voisins les plus proches parmi les documents de formation. Le score de similitude de chaque document voisin le plus proche du document de test est utilisé comme le poids des classes du document voisin (Songho tan, 2008).

6.5.2 Apprentissage non supervisé

Dans la phase hors-ligne de l'apprentissage supervisé, une grande quantité de texte étiqueté par la polarité de sentiment est incontournable pour faire la classification. Cependant, dans la classification de sentiment, il est parfois difficile de créer ou trouver une telle ressource. Les méthodes d'apprentissage non supervisé surmontent ces difficultés en essayant de remplacer complétement la phase hors-ligne par la collecte et l'étiquetage automatique de texte [44], ou bien parfois par un apprentissage supervisé faible [45].

7 Les Travaux éxistans sur Twitter

Dans la littérature relative à l'analyse de sentiments dans les textes longs, on peut distin-

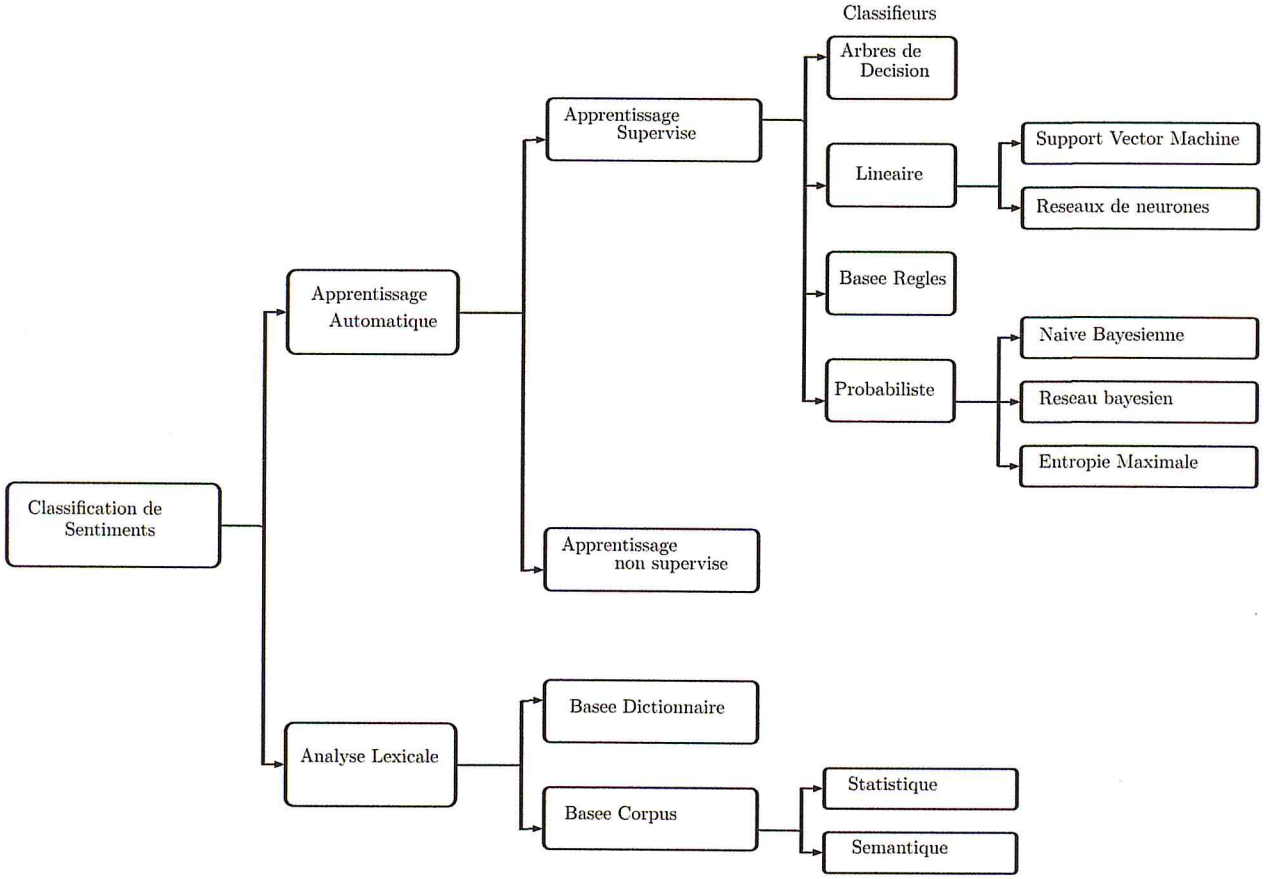


Figure 2.2 – Méthodes de Classification de Sentiments

guer deux approches différentes. Dans la première, nous supposons que le texte en générale est une opinion, et donc nous avons besoin uniquement de calculer sa polarité (classification de polarité). Dans la deuxième approche, et avant de mesurer la polarité, il est nécessaire de déterminer si le texte est subjectif ou objectif (classification de la subjectivité). En ce qui concerne l'étude de la polarité sur Twitter, la plupart des expériences supposent que les tweets sont subjectifs. L'une des premières études sur la classification de la polarité des tweets a été réalisée par Go et al. [46]. Ils ont mené une étude de classification supervisée sur les tweets en anglais.

7.1 Le corpus d'étude

Twitter est caractérisé par la grande quantité d'informations publiées et la grande variété de sujets sur lesquels les utilisateurs écrivent. Cela rend très difficile et coûteux de construire et étiqueter manuellement un corpus pour la classification supervisée de la polarité. Des chercheurs ont utilisé (Read 2005 [47]) les émoticônes qui apparaissent habituellement dans les tweets pour différencier les tweets positifs et négatifs. Grâce à la recherche de Twitter API, les auteurs ont généré un corpus de tweets positifs et négatifs selon les émoticônes positives :) , ou négatives :(. Le corpus est utilisé pour spécifier la meilleure caractéristique ainsi que les

meilleurs algorithmes pour la classification de la polarité sur Twitter.

En conclusion, les auteurs ont trouvé les mêmes résultats que Pang et al [48]. pour les trois méthodes(SVM, NB et ME), cela pose l'hypothèse que l'analyse des sentiments depuis les revues ne se diffère pas de celle des tweets.

L'importance croissante de Twitter dans la société a été montrée à la mort de chanteur "Michael Jackson". Selon Kim et al. (2009), suite à l'annonce de la mort de Jackson, de 21h à 22h le 25 Juin 2009, environ 279 000 tweets ont été publiés, environ 78 par seconde. Cetteétude présentait une analyse de l'humeur des utilisateurs qui ont posté des tweets à propos de la mort du célèbre chanteur. Les auteurs de n'utilisent pas un algorithme d'apprentissage automatique, mais plutôt les expressions de tristesse dans les tweets en fonction du score que le lexique (ANEW - Affective Norms for English Words) a attribué à chaque terme figurant dans les tweets.Le lexique ANEW (Bradley et Lang 1999) fournit un ensemble de 1.034 mot anglais, qui sont notés sur uneéchelle de 1-9 concernant trois humeurs différentes : valence (plaisir/déplaisir), excitation (excitation/calme) et la domination (force/faiblesse).

7.2 La classification de texte sur Twitter

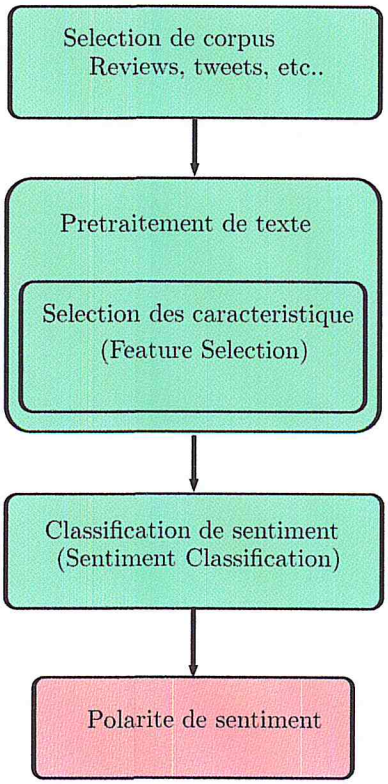


Figure 2.3 – Les étapes de l'analyse de Sentiments généralement suivies sur Twitter

Dans le travail de Pang et al. [48] (2002) Les auteurs ont voulu savoir la différence entre la catégorisation des textes par topique et la classification des textes par sentiment en utilisant les mêmes méthodes utilisées pour la catégorisation par topique pour l'analyse de sentiment : naïve bayes, entropie maximale, et les SVM. Pour l'analyse de sentiment, les auteurs ont obtenu des résultats meilleurs dans le cas de l'utilisation des unigrammes pour la sélection des caractéristiques et surtout avec la classification par SVM. Cependant, l'utilisation de l'étiquetage morphosyntaxique ne fournit pas une amélioration significative à la précision de la classification.
Les auteurs ont tiré la conclusion suivante : dans le cas de classification de sentiment, les méthodes ne fournissent pas un résultat assez bon que dans la catégorisation des textes par topique.

Une étude très intéressante est réalisée par Ley Zhang et al. [21], dans lequel un procédé hybride pour la classification de la polarité est appliqué à Twitter. Comme cela est bien décrit par les auteurs, dans l'analyse de sentiments il existe deux paradigmes, l'un basé sur l'utilisation des ressources lexicales telles que lexiques et un autre basé sur l'utilisation de techniques d'apprentissage machine. Ceux qui sont fondés sur les ressources lexicales ont souvent le problème de l'obtention de faibles valeurs de rappel, car ils dépendent de la présence des mots comprenant le lexique dans le message pour déterminer l'orientation de l'opinion.

Les méthodes basées sur l'apprentissage par machine dépendent de la disponibilité des ensembles de données étiquetées. En ce qui concerne l'analyse de sentiments sur Twitter, la première stratégie vise à surmonter le problème de la nature variée et changeante de la langue utilisée sur Twitter, la seconde difficulté est d'obtenir un vaste corpus de tweets étiquetés. Pour surmonter ces problèmes, les auteurs proposent un système hybride pour l'analyse des opinions au niveau de la phrase sur Twitter. Pour leurs expériences, ils ont utilisé un corpus de tweets en anglais sur cinq entités distinctes (Obama, Harry Potter, Tangled, iPad et Packers), auquel a été appliqué un prétraitement qui commence par la suppression des retweets, la traduction des abréviations en termes d'origine et la suppression des liens, en ajoutant un processus de tokenization et l'étiquetage morphosyntaxique (POS). Une fois que le corpus a été nettoyé, une méthode basée sur le lexique pour classer les tweets en fonction de leur polarité a été appliquée. Les auteurs ont choisi un ensemble de mots subjectifs de tout ce qui est disponible en anglais et ils ont ajouté des hashtags avec une signification subjective.

Notez que des règles spéciales pour le traitement des jugements comparatifs, le traitement de la négation, et le traitement des expressions qui peuvent changer l'orientation d'une phrase. Pour résoudre le problème du rappel (recall) souvent inhérent à ces méthodes, les auteurs ont tenté d'identifier un plus grand nombre de mots indiquant le contenu subjectif. Ainsi, ils ont appliqué le test $\chi 2$, avec l'idée que si un terme est plus susceptible d'apparaître dans un jugement positif ou négatif, il est plus susceptible d'être un identifiant de contenu subjectif. De cette façon, et automatiquement, ils ont réussi à augmenter le nombre de tweets étiquetés. L'étape suivante consiste à appliquer une méthode d'apprentissage automatique pour la classification des nouveaux tweets, dans ce cas, en utilisant l'algorithme SVM.

Bien qu'il existe des études controversées sur l'utilisation des émoticônes en tant qu'un cor-

pus valide de Twitter, récemment Davidov et al. [49] ont utilisé 50 hashtags et 15 émoticônes en tant que des étiquettes de sentiment pour entraîner un classificateur de sentiment supervisé en utilisant l'algorithme des K voisins plus proches (KNN). Les expériences sont validées par des juges humains et les résultats obtenus sont très prometteurs. Au contraire, la plupart des travaux sur Twitter utilisent première représentation de mot (n-gramme) comme une caractéristique pour construire un modèle pour la détection de sentiment. Toutefois, l'inclusion de certaines méta-informations (ex. : Les tags POS) et l'utilisation des fonctions syntaxiques des tweets (par exemple hashtags, retweets et liens) peuvent améliorer le résultat obtenu par Barbosa et al. [50].

Sur la base de ce travail précédent, Jiang et al. [51] étudient la classification de sentiment dépendante de la cible de tweets en utilisant SVM et General Inquirer. Ils classent les sentiments des tweets comme positif, négatif ou neutre selon une requête donnée. Ainsi, la requête sert à une cible de sentiments. En outre, ils appliquent également une approche sensible au contexte pour incorporer le contexte de tweets dans la classification.

Dans Agarwal et al. [52], une étude a été réalisée sur les différentes caractéristiques à prendre en compte dans l'analyse de sentiments sur Twitter. L'étude est menée sur un corpus réduit de tweets étiqueté manuellement. L'expérimentation teste les différentes méthodes de classification de polarité commençant par un cas basic, qui est pour les auteurs l'utilisation des unigrammes, puis un modèle fondé sur les arbres, et le troisième modèle est l'utilisation de diverses caractéristiques linguistiques et, enfin, une combinaison des différents modèles proposés. Une caractéristique commune utilisée à la fois dans le modèle fondé sur un arbre et dans le troisième modèle basé sur les caractéristiques est la polarité des mots apparaissant dans chaque tweet. Pour calculer cette polarité, les auteurs utilisent le dictionnaire DAL (Whissell 1989).À noter également l'ensemble complet de fonctionnalités qu'ils utilisent et l'étude qu'ils effectuent sur lequel on donne plus d'informations. Après une expérimentation extensive, ils ont conclu que les deux méthodes : basée arbre et basée caractéristiques ont donné un bon résultat pour le cas basic. La dernière conclusion, les auteurs parviennent contredit les recherches menées jusqu'à présent, puisque la plupart des auteurs ont tendance à indiquer que les caractéristiques spécifiques à Twitter impliquent l'utilisation d'autres techniques ou une adaptation particulière des techniques de l'analyse de sentiments sur de longs textes. Les auteurs, cependant, ont trouvé une forte indication dans ce cas que l'analyse de sentiment sur Twitter ne diffère pas de l'analyse sur les longs textes.

Bifet et al. [53] optent pour l'utilisation d'algorithmes de flux de données pour la classification de polarité sur Twitter. En outre, ils proposent l'utilisation de la mesure d'évaluation "Kappa" [54] pourévaluer de grandes quantités de données déséquilibrées. Dans leur expérimentation, ils ont utilisé un framework d'analyse de flux de données appelé : Massive Online Analysis (MOA) [55].Les données utilisées ont été le corpus généré par Go et al. [46] et Petrovic et al. [56], à partir de laquelle, après un processus de tokenization, les mots d'arrêt de la liste des WEKA [57] (Waikato Environnement pour l'analyse des connaissances) ont été éliminés et représentés par des vecteurs binaires, où la présence de chaque unigramme a été indiquée. Les algorithmes testés dans l'expérimentation étaient le Naive Bayes Multinomial, Stochastic Gradient Descent (SGD) et l'arbre de Hoeffding. Les auteurs concluent que pour la classifi-

cation de polarité sur Twitter l'algorithme qui retourne les meilleurs résultats est SGD. Les résultats obtenus avec le corpus de Go et al. [46] pour Naive Bayes sont comparables à celles de Go et al. [46] (2009).

Hernández et al. [58] proposent une méthode non supervisée pour réduire les caractéristiques pour l'analyse de sentiments. Leur méthode est basée sur l'allocation Dirichlet (LDA), qui est résumée dans l'article. La méthode est évaluée avec un corpus de 10 000 tweets en anglais sur la tablette iPad, qui ont été téléchargés au cours des mois de Mars et Avril 2011. Après avoir nettoyé le corpus, les tweets sont représentés suivant le modèle d'espace vectoriel et en utilisant la métrique TF-IDF pour pondérer les termes. Une fois que tous les tweets sont représentés, les auteurs appliquent leur proposition pour la réduction des caractéristiques. Ils ne procèdent pas à un processus de classification de polarité pour comparer les performances de l'ensemble de données complet et l'ensemble de données réduit. Ils ont conclu que le modèle réduit est mieux parce que sa valeur d'entropie est inférieure à (meilleure) celle du modèle complet.

8 Travaux récents

Jansen et al. [59] ont fait une études assez détaillée et complète sur Twitter en terme de sa possibilité d'être un outil de marketing eWOM valide.

Le "Word Of Mouth WOM" ou "de bouche à oreille" est le faite qu'une personne recommande un produit à une autre pesonne en se baseant sur son connaissance ou experience avec ce produit. Ce type de communication est connu dans le marketing comme un outil plus efficace que la publicité pour pousser les gens à choisir un produit.

Dans leur travail, les auteurs ont analysé 150 000 tweets contenants des noms de marques. ils ont trouvé que presque 19% des tweets montionnent une marque quelconque, ceci est un bon pourcentage et indique que le moyen de micro-blogging est une zone viable pour les organisations pour les campagnes de marketing viral, la gestion des relations clients, et pour influer sur leurs efforts de marketing utilisant eWOM. Parmis ces 19% on trouve 20% contient une expression d'un sentiment sur un produit, un service de la marque ou la marque elle même, d'après eux, Il est évident que les entreprises peuvent recevoir une exposition positive ou négative de marque via les followers et d'autres qui postent des tweets sur la société et ses produits Figure2.4.

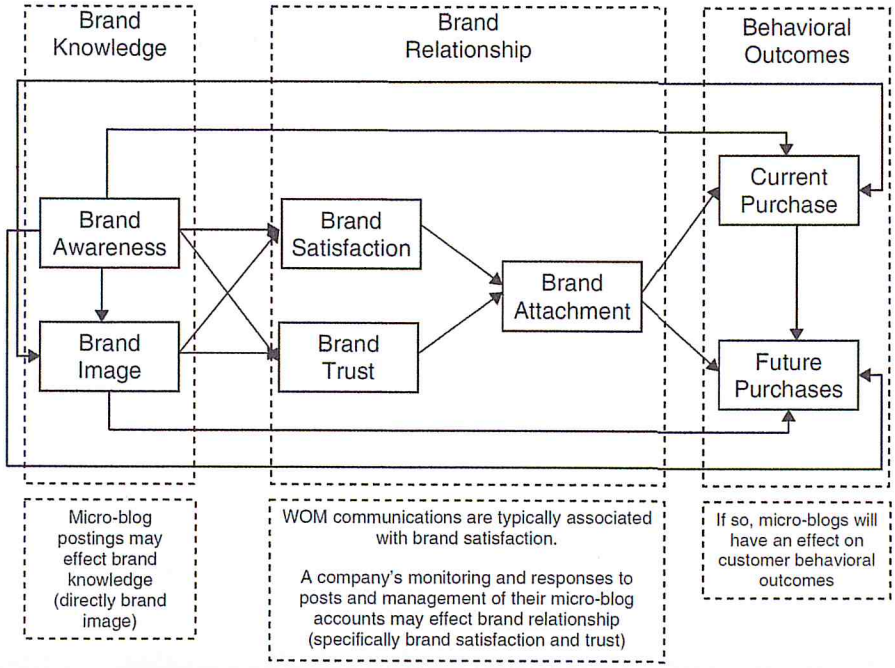


Figure 2.4 – Modèle général de la relation entre une marque et le microblogging [59]

Les auteurs ont trouvé que le microblogging en général, et spécifiquement les tweets ont un impact direct sur la communication eWOM (Bouche à Oreille) car ils permetent aux gens de partager ces pensées et ses sentiments sur une marque, presque partout (en voiture, aux café, dans leur bureau) et par n'importe quel média (Web, smartphone, messagerie instantanée, e-mail) et à une échelle colosale qui n'a pas été vue dans le passé.

Aisopos et al. [60] notent certaines difficultés rencontrées dans l'analyse de sentiments sur Twitter : la rareté des données, l'utilisation d'un vocabulaire non-standard, la faible qualité de la grammaire des tweets et de la nature multilingue des textes publiés sur Twitter. Pour surmonter ces problèmes, les auteurs étudient différents modèles pour représenter les tweets. Les modèles sont divisés en deux groupes, ceux qui sont basés sur le contenu de chaque tweet et celles qui utilisent le contexte du tweet. Le modèle de contenu qui obtient les meilleurs résultats consiste à représenter chaque tweet comme un graphe de caractères n-grammes. En raison de cette utilisation de caractères n-grammes, la solution devient indépendante de la langue, et les erreurs de syntaxe et de grammaire rendue hors de propos, de sorte que certains des problèmes soulignés sont surmontés. Le modèle de contexte était simple. La classification supervisée a trouvé des preuves de la bonne performance du modèle de contenu et pas si bonne du modèle de contexte.

En 2014, dans le travail de Montejoàez et al. [61], le problème de classification de sentiment d'un tweet est résolu en combinant les scores SentiWordNet avec une analyse de la marche aléatoire (Random Walk Analysis) des concepts trouvés dans le texte sur le graphe WordNet. Les auteurs ont construit un corpus de tweets publiés en anglais suivant la procédure décrite dans Go et al. (2009). Ils ont réuni un ensemble de 376 296 tweets (181 492 étiquetés comme tweets positifs et 194 804 étiquetés comme tweets négatifs selon la liste des émoticônes posi-

tives et négatives), puis ils ont fait un pré traitement et nettoyage des tweets collectés afin de ne pas affecter négativement le résultat de l'analyse (la figure).

Après le nettoyage, et pour chaque tweet, chaque mot de ce tweet sera étiqueté par le code de son synset approprié dans WordNet, ce code sera utilisé par la méthode de PageRank qui utilise la méthode de Randow Walk pour trouver des concepts reliés ou proches au concept de ce mot dans WordNet, le résultat de cette étape est un vecteur personalisé de PageRank (Personilized PageRank Vector PPV), dans lequel chaque mot est étiqueté par deux valeurs : code de son synset de WordNet et son score PageRank. Puis pour chaque mot on l'attribue son score de polarité dans SentiWordNet. Ce score sera multiplié par le score de PageRank, le résultat sera sommé avec les résultats de tous les autres mots de tweet pour obtenir un seul score final qui représente la polarité du sentiment de tweet. Si le score final est positif, le tweet et positif, sinon c'est le contraire.

Leur méthode a obtenu un résultat très proche aux SVM mais pas mieux, c'est pour cela ils ont considéré que leur solution non supervisée est une alternative intéressante à celui supervisé.

Dans le travail de Go et al. [46] les auteurs ont décidé d'utiliser les mêmes méthodes d'apprentissage par machine utilisées par Pang et al. [48] mais leur butétait de trouver un moyen automatique pour avoir une large base de données de tweets étiquetés positifs et négatifs pour l'utiliser dans la phase d'apprentissage avant la classification. Ils ont constitue une collection (dataset) de 1 600 000 tweets, dont ils considèrent que les tweets avec le sentiment positif contient des émoticônes qui représentent la joie comme :D, et les tweets avec le sentiment négatif contient des émoticônes qui représentent la tristesse comme :) .

Avant de faire l'analyse de sentiments, les auteurs voient que leur prétraitement des tweets est indispensable pour avoir une bonne classification de sentiments, ainsi il offre une réduction qui arrive jusqu'à 45.85 % de sa taille originale.
Les prétraitements qu'ils ont faits sont :

- Élimination de toutes les émoticônes avant la classification.
- Élimination des Retweets et les tweets répétés.
- Remplacer tous les mots qui commencent par @ par le mot USERNAME.
- Remplacer tous les liens (Ex. http ://tinyurl.com/glay3z8) par le mot URL.
- Toute lettre qui surviennent plus de deux fois successives sera remplacée par deux occurrences (Ex. I'm haaaaappyyyy sera remplacée par I'm haappyy).

9 Tableau résumant les approches utilisées

Il est tentant de comparer cettes approches pour avoir une vision claire de ces similitudes et de ces différences et il est utile de construire un tableau basé sur un choix de critères de comparaison des crétiques.
Ce tableau présente à la fois les différentes approches mentionné dans l'état de l'art, ainsi les ressources de données, la langue, le modèle et l'algorithme utilisé, le résultat trouvé, les critiques et enfin le but de chaque approche afin de comparer les approches de classification de sentiments.

- L'élément Approche : considérer comme les noms des auteurs des approches utilisées dans l'état de l'art.
- L'élément Ressources de données : soit des collections manuelles (datasets) ou des tweets à partir de Twitter.
- L'élément Langue : la langue utilisée dans le corpus collecté.
- L'élément Algorithme : l'algorithme adapté pour la classification de sentiment et pour le traitement de données.
- L'élément Résultat : représenté par un indice (+) comme : (+) : un bon résultat. (++) : un très bon résultat. (+++) : le meilleur résultat.
- L'élément Critiques : représente les points forts où les points faibles de chaque approche.
- l'élément But : dans quel domaine est appliqué l'approche.

Table 2.1 – Comparaison des approches de classification de sentiments - Partie 1

Approche	**Ressource de données**	**Langue**	**Modéle utilisée**	**Algorithme**	**Résultat**	**Crétiques (Points faibles ou points forts**	**But**
Pang et al 2002	movie reviews, Internet Movie Database (IMDb) archive of the rec.arts.movies.reviews	Anglais	basé sur les méthodes statistiques(vecteurs, réseaux bayésiens..)	Naive Bayes, Maximum Entropy, Support Vector Machines	(+), (++), (+++)	thwarted expectations et la présence d'un corpus pour entrainer les classifieurs	savoir la capabilité des méthodes de classification de texte par topic dans la classification des textes par sentiments
Lei Zhang et al.	Twitter⇒ Tweets	Anglais	hybride, une basé sur les ressources lexicales et l'autre sur les techniques d'apprentissage supervisé	étiquetage morpho-syntaxique, et l'algorithme SVM	(+), (++), (+++)	réussir à augementer automatiquement le nombre de tweets étiquetés	classification de polarité est appliqé à Twitter
Davidov et al.	Twitter⇒ Tweets	Anglais	basé sur une méthode supervisé	L'algorithme des K voisins plus proches (KNN), inclusion de certaines méta-informations(les tags et les fonctions syntaxiques des tweets)	ils ont amélioré le résultat obtenu par Barbosa et al. (+++)	l'utilisation de méta-informations et aussi les fonctions syntaxiques comme les hashtags, les retweets et les liens.	construire un modèle pour la détection de sentiment
Agarwal et al.	Twitter⇒ Tweets	Anglais	méthode hybride, basé sur le lexique et apprentissage suppervisé	trois modèles : le premier modèle fondé sur les arbres, le 2éme sur les unigrammes et le 3 éme sur les caractèristiques et la polarité des mots (l'utilisation du dictionnaire DAL)	(+++) Bon resultat	basé arbre et basé caractéristiques ont donné de bon résultat + ils ont trouvé une forte indication l'analyse de sentiment sur Twitter ne diffère pas de l'analyse sur les longs textes	construire un modèle pour la détection de sentiment et de polarité sur Twitter.

Table 2.2 – Comparaison des approches de classification de sentiments - Partie 2

Approche	Ressource de données	Langue	Modéle utilisée	Algorithme	Résultat	Crétiques (Points faibles ou points forts	But
Hernández et al. 2009	Twitter⇒ Tweets	Anglais	méthode non supervisée	basée sur l'allocation Dirichlet (LDA), la métrique TF-IDF (pondérer les termes)	(+++)	le modèle réduit et mieux car sa valeur d'entropie est inférieure à celle du modèle complet	réduire les caractéristiques pour l'analyse de sentiments(résumé)
Alec Go 2009	Twitter⇒ Tweets	Anglais	basé sur les méthodes statistiques(vesteurs, réseaux bayésiens..)	Naive Bayes, Maximum Entropy, Support Vector Machines	même résultat que pang et al., (+), (++), (+++)	-	la possibilité de créer automatiquement une dataset de tweets pour entrainer les méthodes d'apprentissage
Bifet et al. 2011	Twitter⇒ Tweets	Anglais	utilisation d'algorithme de flux de données	Naive Bayes Multinomial, stochastic Gradient Descent et l'arbre de Hoeffding	SGD (+++), les autres (+) comparable à celle de Geo et al. pour le NB	concluent que pour la classification de polarité l'algorithme SGD retourne les meilleurs résultats	construire un modèle pour la classification de polarité sur Twitter.
Jansen et al. 2012	Microblogging "Twitter" ⇒ Tweets	Anglais	Outils Marketing	ewow (bouche à l'oreille)	+	-	-
Rachel Bugeja 2013	Twitter ⇒ tweets	Anglais	NER Reconnaissance des Entités Nommées"	GATE "pour éffectuer l'analyse de sentiments"	75%	-	Fournir à l'utilisateur un outil de recherche de marché
Montejoàez et al.	Twitter ⇒ tweets	Anglais	méthodes non supervisée, Random Walk (sur WordNet) + les scores de SentiWordNet (RW-SWN)		Proche à SVM	n'arrive pas a avoir un résultat comme la méthode supervisée	proppser une nouvelle méthode statique non suppervisée de classification de sentiment

10 Les Domaines d'application

Les analysites d'affaires ont accès aux réseaux sociaux où les opinions et les sentiments sur les entreprises, les produits et les politiques sont exprimées sous forme non structurée. l'exploitation des informations provenant de source publiques est d'une grande importance pour de nombreuses applications de business intelligence.

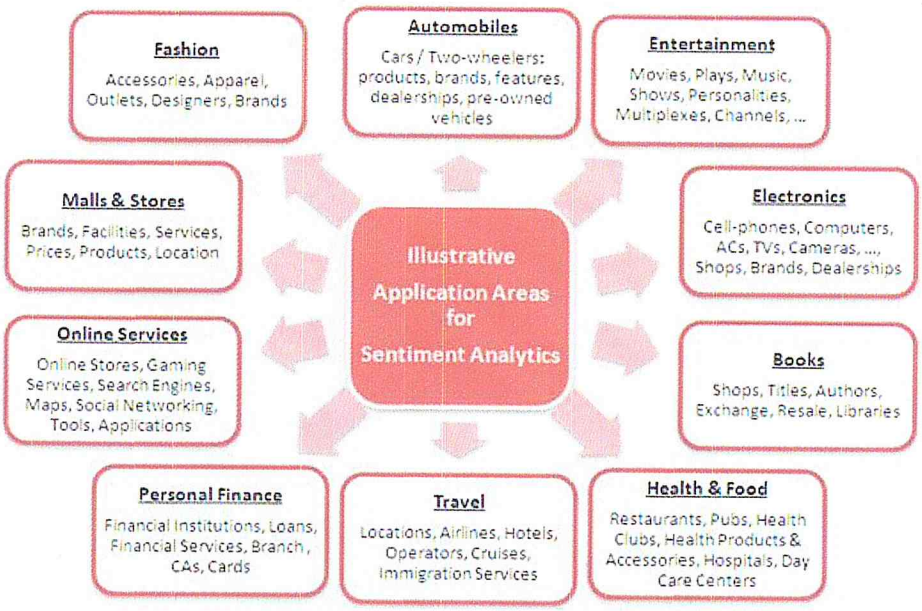


Figure 2.5 – Domaines d'application de l'analyse de sentiment

10.1 Concernant les entreprises

Opinion mining est régulièrement convoqué dans les processus décisionnel, que ce soit en vue d'un achat, dans le contexte d'une élection, ou pour évaluer la réputation de son entreprise ou de sa marque. La collecte et l'analyse des opinions des individus sont devenues des sources d'informations précieuses pour les entreprises. Le Marketing a rapidement compris l'intérêt de l'analyse de sentiments. Des agences vendent aux entreprises des informations sur leur image et leurs produits.

Parmis les domaines d'application de l'analyse de sentiment :

10.1.1 Domaine du product review mining

A partir des sites d'avis de consommateurs qui donnent la main aux consommateurs d'échanger des avis et retrouver leurs décisions d'achat (produits), Alors l'analyse de sentiment permet de catégoriser les opinions sur le sujet d'un produit, et d'en proposer des résumés des revues ou des avis on parle alors sur les " features ", qui aident à détecter les spam commerciaux et les faux avis.

— **Pour les entreprises :** Acquérir des connaissances sur des consommateurs ou anticiper leurs attentes est possible à partir de la collecte des avis de consommateurs sur un produit, un service ou une marque. Contrôle de qualité des produits et améliore la relation client/fournisseur , eBay à proposer une application d'analyse de sentiment qui permet de rechercher la présence de mot-clés sur Twitter pour détecter les pannes du service signalées par les utilisateurs, avant même que l'alerte ne soit donnée par le système d'enchère lui-même.

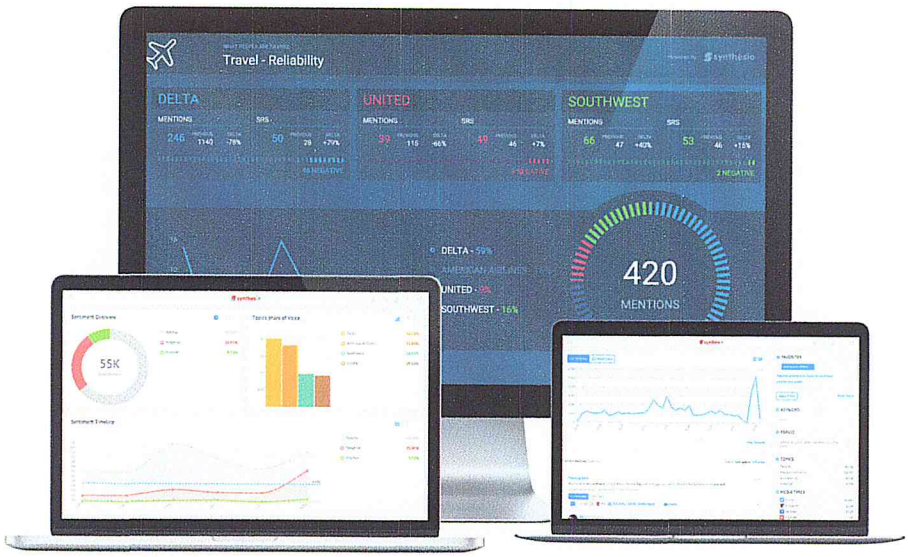


Figure 2.6 – "Synthsio" : est une plateforme d'analyse des sentiments des clients pour les entreprises, Twitter est parmis ces ressources d'informations les plus importantes. Elle était parmis les TOP outils analytiques et de management des médias sociaux dans les choix des éditeurs du PC MAG en 2015 [62].

— **Pour les clients :** selon le baromètre sur les comportements d'achat des internautes, et d'après la Fédération du E-commerce et de la Vente à Distance (FEDAV) disent que prè de 9 internautes sur 10 (86%) déclarent avoir consulté un site internet avant d'acheter un produit sur internet ou en magasin [63].

10.1.2 Domaine financier

Prédiction de tendances de marché, Gilad et al. [64] utilisent des techniques d'analyse de sentiment pour améliorer la prédiction du succès commercial d'un film à partir des blogs.

— Construire de meilleures prédictions que la mesure du simple buzz, et surtout si elle est associée à d'autre types de données comme le genre du film et le moment de sa sortie.

— La classification des dépêches financières afin d'observer l'impact éventuel de ces dernières sur le prix des actions cotées en bourse. C'est ce type d'application que présentent M.Généreux et al. [65] en faisant l'hypothése que "la création du marché

suite à la publication d'une dépêche reliée à une action particulière est un bon indicateur de la polarité de la nouvelle et qu'un algorithme d'apprentissage à partir de ces dépêches permet de construire un système qui donne à l'investisseur une source d'information supplémentaire qui peut être exploitée de façon avantageuse dans une stratégie d'investissement".

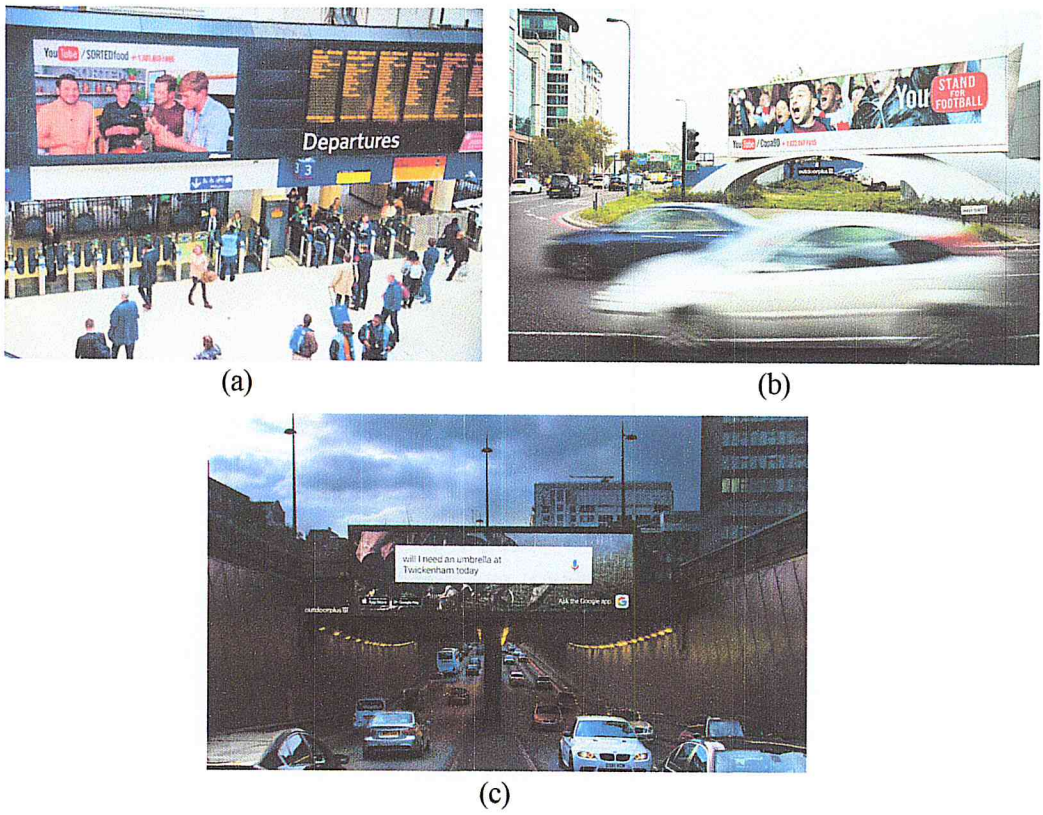
10.1.3 Domaine de la veille

- **Journalisation :** L'analyse de sentiment permet de classer de grandes quantités de textes, rapports, conversations informelles sur des produits ou des dirigeants d'entreprises etc. peuvent être utilisées dans le domaine de la veille, qu'elle soit éconimique, technologique, stratégique ou institutionnelle. Ainsi espère-t-on par exemple mettre en place des systèmes d'évaluation de la réputation des entreprises en rassemblant dans des bases de données des faits et opinions trouvés sur le web et permettant de tracer le profil de telle ou telle entreprise ?.
- **Secteur industriel :** L'intérêt pour la surveillance de données extraites des médias sociaux est considérable dans le secteur industriel. En effet, ces données sont susceptibles d'aider en optimisant de manière importante l'efficacité de la veille stratégique. L'intégration de telles données aux systèmes de veille stratégique déjà en place permet aux entreprises d'atteindre différents objectifs, notamment concernant la stratégie de marque et la notoriété, la gestion des clients actuels et potentiels et l'amélioration du service à la clientèle. Le marketing en ligne, la recommandation de produits et la gestion de la réputation ne sont que quelques exemples d'applications concrètes de l'ASMS.

10.1.4 Domaine de la publicité en ligne

Si une annonce publicitaire est efficace quand elle apparaitre au bon endroit et au bon moment elle sera plus efficace si elle s'adapte aux prédictions des avis des consommateurs vis-à-vis d'un produit dans les réseaux sociaux. Comme la stratégie 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 [66].

Un nouveau concept de publicité était prposé dans mon projet de fin d'études de licence est maintenant le sujet d'un test fait par Google pour utiliser le même principe des publicités en ligne dans le monde réel. La publicité "*DoubleClick*" de Google Figure 2.7 vise à mettre en place des panneaux publicitaires dynamiques, leurs contenu se change selon le sentiment public, la météo, les informations, les événements sportifs, et les scores, etc. Selon ces informations il décide quels messages créatifs à afficher et quels panneaux d'affichage pour les afficher ainsi que le meilleur moment pour les montrer.



(a) (b) (c)

Figure 2.7 – La publicité "DoubleClick" de Google : (a) Une annonce YouTube servie par un programme sur un panneau d'affichage numérique de JCDecaux dans la gare de *Waterloo* à *Londre*, (b) Une annonce YouTube au rond-point *Vauxhall*, (c) Une publicité pur Google App suit la météo sur *Euston Road*, *Londres* [67].

10.2 Autres domaines

10.2.1 Domaine politique

La publication croissante sur internet de textes à teneur politique (lois, rapports, billets de blogs politiques, etc.) et le constat politique ne se fait plus seulement dans les hémicycles mais aussi dans les débats en ligne, a conduit certain chercheurs à utiliser l'analyse de sentiments pour déterminer l'accord ou le désaccord des internautes avec telle ou telle proposition de loi. Dans leur article [68] M. Thomas et al. espèrent faciliter la reconnaissance du positionnement d'un orateur (speaker) dans un débat politique grâce au sentiment analysis. D'autres recherches tentent par exemple d'analyser en masse les commentaires et opinions des citoyens américains lors de l'élaboration des réglementations proposées par les agences indépendantes du gouvernement [29].

10.2.2 Soin de santé

Les forums de discussion qui sont des espaces d'échanges asynchrones de messages textuels sont très prisés par certains patients. En effet, ils sont associés à un véritable espace de

liberté du discours. Ainsi, l'utilisation de Twitter ou des forums comme des plates-formes de discussion sur des sujets tels que les maladies, les traitements, les médicaments ou les recommandations à l'intention des professionnels et des bénéficiaires (patients, familles et aidants) illustre bien la pertinence des médias sociaux dans ce domaine. Par ailleurs, dans ce contexte éminemment subjectif, la caractérisation et la compréhension des perceptions que les patients ont de leur maladie et du suivi médical représentent un enjeu sociétal particulièrement intéressant pour les professionnels de santé [69].

11 Conclusion

Dans ce chapitre nous avons présenté les méthodes de classification de texte dans l'analyse de sentiment sur Twitter, Ainsi que les travaux existants dans le même domaine.
Dans le prochain chapitre nous montrons l'utilisation des techniques présentées ici, et la conception détailler de notre approche.

Chapitre 3

Méthodologie et Conception

1 Introduction

L'analyse de sentiments est la partie du text mining qui essaye de définir les opinions, elle est particulièrement utilisée en marketing pour analyser les commentaires des internautes dans les réseaux sociaux, et même les textes des blogueurs, aussi pour sonder l'opinion publique.

Dans ce chapitre nous allons présenter les étapes de l'analyse de sentiments ainsi que la conception suivie dans notre projet.

2 Conception générale

Dans la partie antérieure, nous avons expliqué l'ensemble des traitements et prétraitements nécessaires pour mettre en oeuvre la classification des sentiments utilisant l'apprentissage automatique supervisé par Naïf Bayes.

Dans la partie suivante, nous présentons la conception dans laquelle tous les traitements et prétraitements ont été organisés, utilisant un haut niveau d'abstraction appelé "4C".

La méthode "4C" (abriviation pour : context, containers, components and classes) peut donner une image constituée de plusieurs calques pour bien comprendre la complexité et les relations entre les entités de traitement. Le point fort de cette méthode est qu'elle peut servir à la fois à l'architecture logicielle et à la conception de l'idée [70].

Les calques de la conception utilisant "4C" sont les suivants :

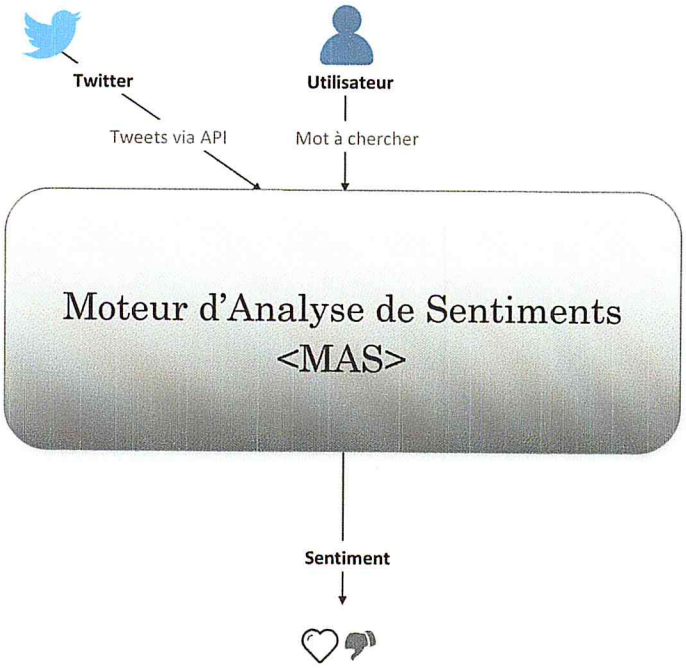


Figure 3.1 – Premier calque de la conception (Vue globale)

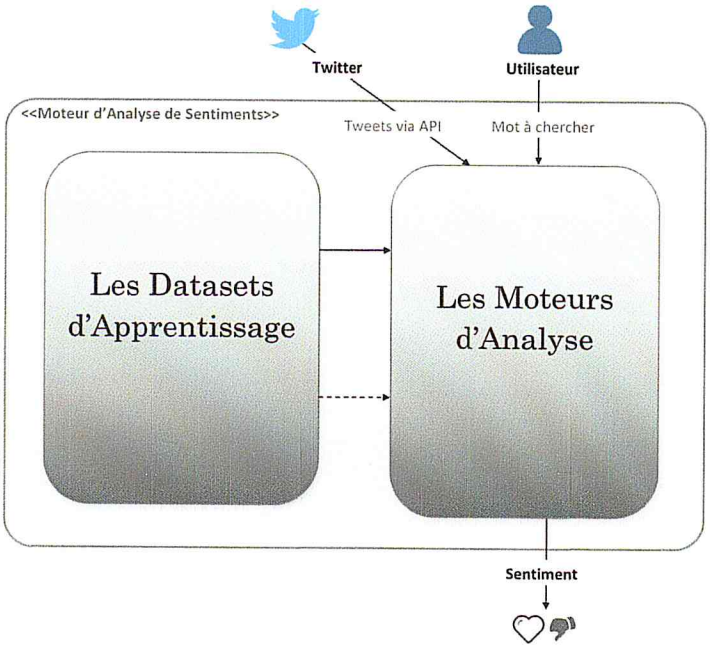


Figure 3.2 – Deuxième calque de la conception (Vue globale plus détaillée)

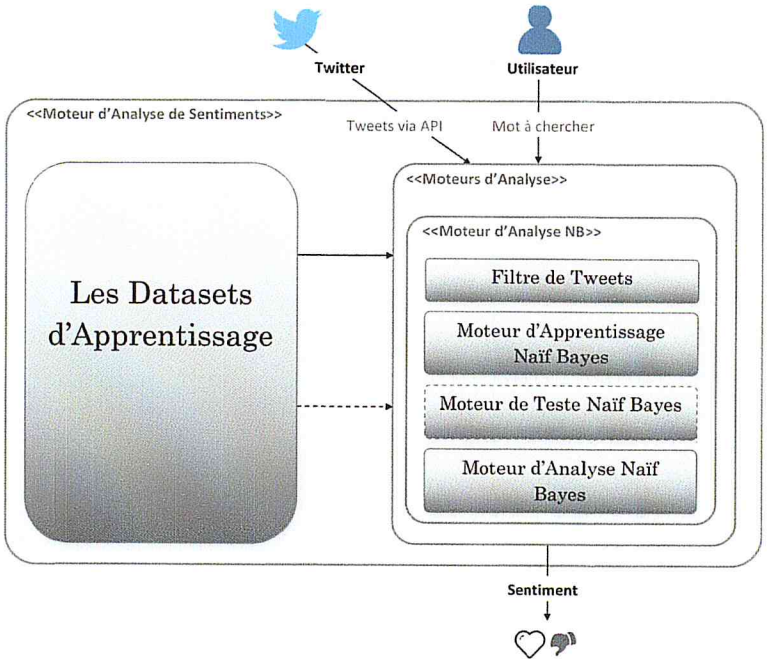


Figure 3.3 – Troisième calque de la conception (Composants du moteur d'analyse NB)

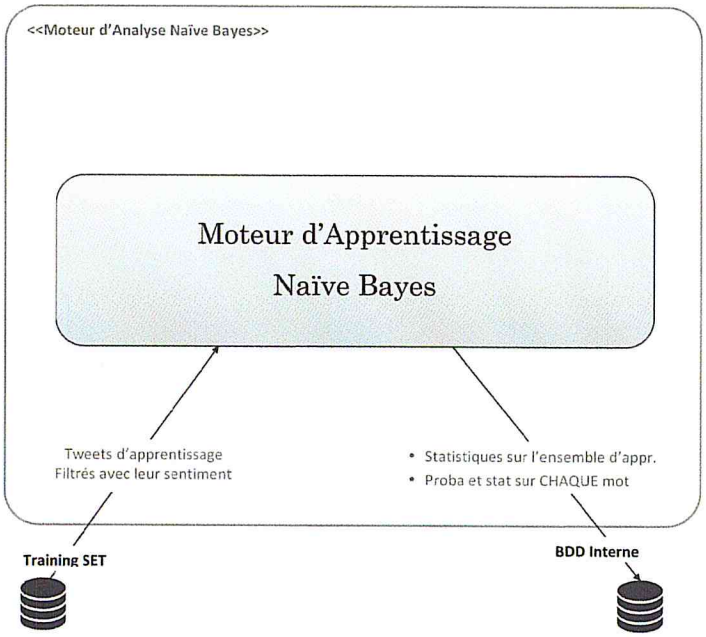


Figure 3.4 – Quatriéme calque de la conception du composant : Moteur d'apprentissage NB (Vue globale)

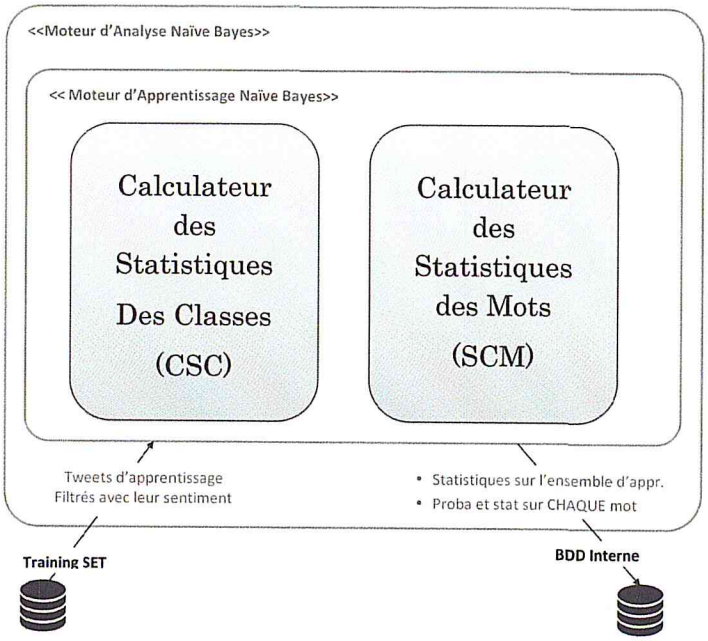


Figure 3.5 – Cinquième calque de la conception du composant : Moteur d'apprentissage NB (Vue générale plus détaillée)

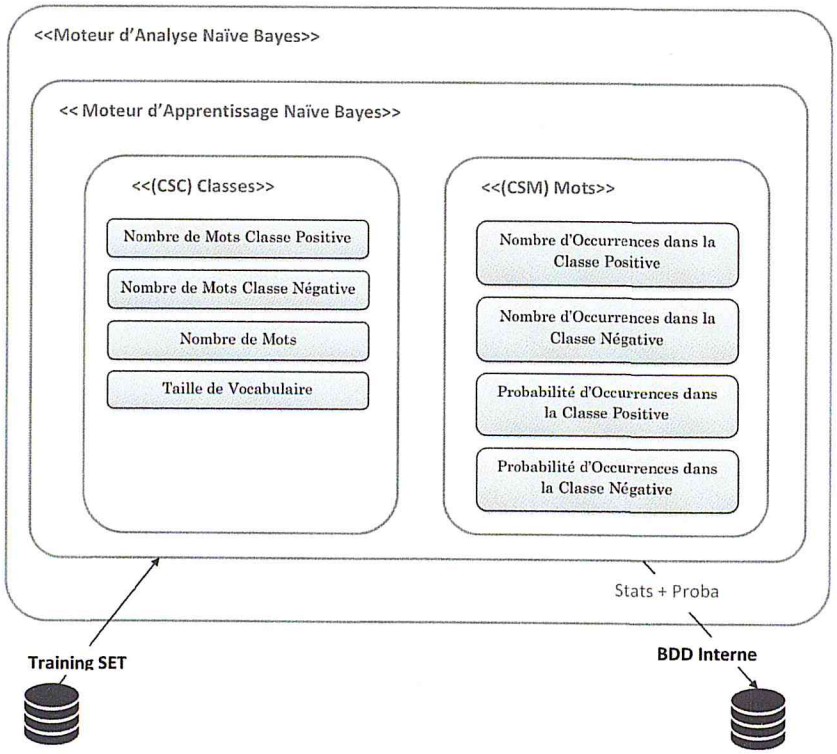


Figure 3.6 – Sixième calque de la conception du composant : Moteur d'apprentissage NB (Vue détaillée)

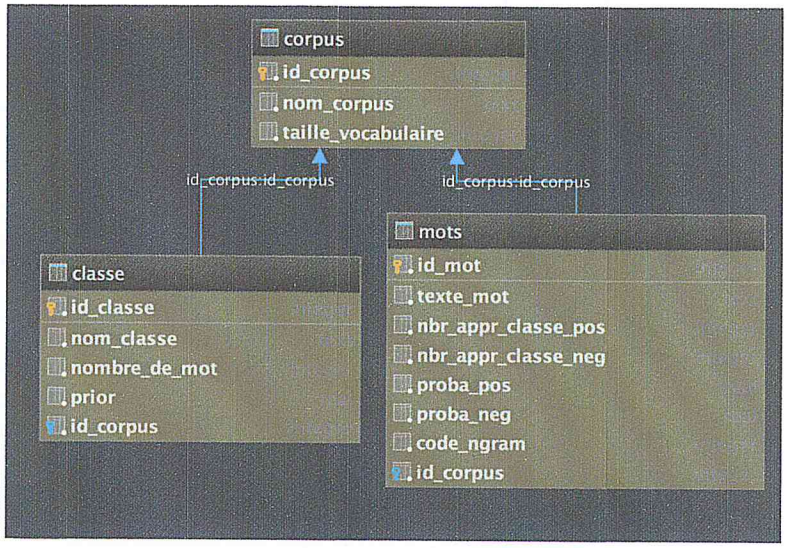


Figure 3.7 – Conception de la base de donnée interne utilisée pour enregistrer les résultats de l'apprentissage

3 Étapes d'analyse de sentiments

Selon pang et lee (2008) l'analyse de sentiments consiste à plusieurs étapes, la première consiste à distinguer entre les textes subjectifs, objectifs et neutre. Ensuite le système doit détecter les phrases porteuses de l'opinion, finalement toutes les informations doivent être présentées dans une analyse globale du sentiment exprimé dans le texte.

Cela va concevoir un système autonome d'analyse automatique de sentiment proche de celui du cerveau humain.

L'analyse peut s'effectuer à différents niveaux :

— Niveau document :
Vise à classifier les documents, chaque document est considéré comme une unité entière d'information de base.

— Niveau de phrase :
Vise à classifier le sentiment exprimé en chaque phrase (positif, négatif où neutre). Wilson et al. ont souligné que les expressions de sentiments ne sont pas nécessairement de nature subjective. Cependant, il n'y a pas de différence fondamentale entre le niveau document et le niveau de phrase car les phrases ne sont que de courts documents. Classer le texte au niveau du document ou au niveau de la phrase ne fournit pas les détails nécessaires sur tous les aspects de l'entité qui est nécessaire dans de nombreuses applications. Pour obtenir ces détails ; nous avons besoin de passer au niveau aspect.

— Niveau aspect :
Vise à classer le sentiment à l'égard des aspects spécifiques d'entités. La première étape consiste à identifier les entités et leurs aspects. Les émeteurs d'opinion peuvent donner des avis différents pour différents aspects de la même entité.

3.1 Selection de corpus(tweets)

Les critéres les plus importants pour sélectionner une source de données sont l'API où l'accessibilité et droits aux données, et la taille des textes.

Pour La méthode basée sur des corpus, il est important de préciser la différence entre deux méthodes d'apprentissage différentes : l'apprentissage supervisé et l'apprentissage non supervisé. L'apprentissage supervisé implique l'élaboration de deux corpus : un corpus d'apprentissage et un corpus de test.

La méthode s'appelle " supervisée" parce que le système automatique est entraîné à traiter une base de données en se basant sur un corpus d'apprentissage qui comporte des modèles déjà traités (dans la présente étude, ces modèles sont des tweets). En revanche, la méthode non supervisée ne requiert qu'un corpus. Cette méthode implique que le système autonome doit lui-même structurer les informations au sein du corpus en les divisant en groupes. Ainsi, il doit organiser la base de données d'une telle manière que les données les plus similaires soient associées dans un groupe et les données différentes dans un autre.

3.2 Prétraitement de texte

Pour réduire certaines ambiguités, il est souvent nécessaire d'orienter les traitements des tweets, ce qui nous oblige à faire un enchaînement des traitements proposés dans différentes chaînes de prétraitement des tweets. Les étapes sont définies comme suit :

1. **Elimination des tweets non anglais**
   Dans cette étape, touts les tweets qui ne sont pas écrits en anglais serons supprimés. La langue des tweets est obtenue depuis le tweet comme attribut fourni par Twitter.
2. **Elimination des tweets répétés**
   Dans la phase d'apprentissage, il faut avoir une seule occurance pour chaque tweet, pour ne pas affecter les probabilités d'occurance des mots, qui est la caractéristique la plus importante dans la phase d'apprentissage.
3. **Suppression des URLs**
   Les liens du web dans un tweet, certainement, ne contiennent pas des sentiments, donc il est préférable de les supprimer pour obtenir plus de précision.
4. **Elimination des noms d'utilisateurs**
   Les noms d'utilisateur dans les tweets sont toujour précédés par un "@", ils ne fournissent pas une information sur le sentiment.
5. **Elimination des émoticônes**
   Les émoticônes dans les tweets peuvent porter plusieurs sentiments, mais leur forte présence peut affecter négativement la précision de l'algorithme, pour cette raison tous les émoticônes seront supprimés des les tweets, cela inclu une liste de plus de 1300 émoticônes.
6. **Elimination de ponctuation**

Pour l'élimination des mots vides, tokenization, stemming, lemmatisation et construction des n-grammes ils seront comptabilisés dans la partie construction des bags of words.

3.2.1 Sélection des caractéristiques (Feature selection)

L'une des sous-tâches les plus importantes de classification de motifs sont l'extraction et la sélection de caractéristiques ; les trois principaux critères de bonnes caractéristiques sont énumérées ci-dessous :

— Saillant
Les caractéristiques sont importantes et significatives par rapport au domaine du problème.

— Invariant
Invariance est souvent décrite dans le contexte de la classification d'images : Les caractéristiques sont insensibles à la déformation, mise à l'échelle, l'orientation, etc [71]
.

— Discriminatoire
Les entités sélectionnées portent suffisamment d'informations pour bien distinguer entre les modèles lorsqu'ils sont utilisés pour former le classificateur.

Avant le montage du modèle et l'utilisation des algorithmes d'apprentissage automatique pour l'entrainement, nous devons réfléchir à la façon de représenter au mieux un document texte en tant que vecteur de caractéristiques. Un modèle couramment utilisé dans Natural Language Processing est le sac de mots.

L'idée derrière ce modèle est vraiment aussi simple que cela puisse aparaître. Vient d'abord la création du vocabulaire - la collection de tous les différents mots qui se produisent dans l'ensemble de la formation et chaque mot est associé à un compte de la façon dont il se produit. Ce vocabulaire peut être compris comme un ensemble d'éléments non redondants où l'ordre n'a pas d'importance. Soient D1 et D2 deux documents dans un ensemble de documents dans un dataset d'entrainement :

— D1 : "Each state has its own laws."
— D2 : "Every country has its own culture."

Sur la base de ces deux documents, le vocabulaire peut être écrit comme
$V = each : 1, state : 1, has : 2, its : 2, own : 2, laws : 1, every : 1, country : 1, culture : 1$
Le vocabulaire peut ensuite être utilisé pour construire les vecteurs caractéristiques de la dimension d des documents individuels dans lesquels la dimension est égale au nombre des différents mots du vocabulaire (d = V). Ce processus est appelé vectorisation.

	each	state	has	its	own	laws	every	country	culture
X_{D1}	1	1	1	1	1	1	0	0	0
X_{D2}	0	0	1	1	1	0	1	1	1
Σ	1	1	2	2	2	1	1	1	1

Table 3.1 – Représentation d'un sac de mots de deux simple documents D1 et D2

Étant donné l'exemple dans le tableau 3.1 une question est de savoir si les 1 et les 0 des vecteurs caractéristiques sont les chiffres binaires (1 si le mot se produit dans un document particulier, sinon 0) ou chiffres absolus (combien de fois le mot se produit dans chaque document). La réponse dépend du modèle probabiliste est utilisé pour le classificateur Naïf bayes :

1. **Tokenization**
   Décrit le processus général de décomposer un corpus de textes en éléments individuels qui servent d'entrée pour différents algorithmes de traitement du langage naturel. Habituellement, tokenization est accompagné par d'autres étapes de traitement optionnelles, telles que la suppression des mots vides et des caractères de ponctuation, issus ou lemmatisation, et la construction de n-grammes. Voici un exemple d'une étape simple mais typique tokenization qui divise une phrase en mots individuels, supprime la ponctuation, et convertit toutes les lettres en minuscules.

A swimmer likes swimming, thus he swims.						
a	swimmer	likes	swimming	thus	he	swims

Table 3.2 – Exemple de tokenization

2. **Stop Words**
   Les mots vides sont des mots qui sont particulièrement fréquents dans un corpus de texte et donc considérés comme plutôt non informatifs (par exemple, des mots tels que oui, et, ou, le ...etc.). Une approche alternative est de créer une liste des mots vides en triant tous les mots dans l'ensemble du corpus de texte par la fréquence. La liste des mots vides - après conversion en un ensemble de mots non redondants - est ensuite utilisée pour enlever tous ces mots à partir des documents d'entrée qui sont classés parmi les n premiers mots dans cette liste.

A swimmer likes swimming, thus he swims.					
swimmer	likes	swimming	,	swims	.

Table 3.3 – Exemple d'élimination des mots vides

3. **Stemming and Lemmatisation**
   Stemming (stemmatisation), c'est de trouver la racine de chaque mot,ce traitement est basé sur un dictionnaire de suffixes qui permet d'extraire le radical du mot grâce à l'étude morphologique du mots [72] [73] .
   l'analyse la plus complexe c'est la lemmatisation fondée sur un lexique.

A swimmer likes swimming, thus he swims.							
a	swimmer	like	swim	,	thu	he	swim	.

Table 3.4 – Exemple de stemming

A swimmer likes swimming, thus he swims.							
A	swimmer	like	swimming	,	thus	he	swim	.

Table 3.5 – Exemple de lemmatisation

4. **n-grammes**
   Nous avons essayé de déterminer les meilleurs réglages pour l'obtention des n-grammes Pang et al. (Pang et al., 2002), et les bigrammes ont mieux fonctionné pour la classification des sentiments d'aprés Dave et al., ils ont confirmé que les bigrammes et les trigrammes ont mieux fonctionné pour la classification de polarité sur les produits et les reviews, La procédure d'obtention de n-grammes à partir d'un tweet est comme suit :

— Filtrage
  nous supprimons les liens d'URL (par exemple http ://example.com), les noms d'utilisateur Twitter (par exemple @ Alex - avec le symbole @ indiquant un nom d'utilisateur), aussi, les Mots spéciaux (tels que "RT" 6) et émoticônes (RT signifie la citation ou republication d'un message).
— Tokenization
  nous segmentons le texte en le divisant par des espaces et des signes de ponctuation, et de former un sac de mots.
— Suppression des mots vides : Retrait des mots vides comme ("am", "all", "any", "but", "by" ...) du sac de mots.
— Construction des n-grammes : nous construirons des unigrammes, des bigrammes et des trigrammes, exemple [74] :
  1. Unigrammes :
     From companies experience best customer service aside { from, companies, experience, best, customer, service, aside}
  2. Bigrammes :
     { from-companies, companies-experience,experience-best, best-customer, customer-service, service-aside}
  3. Trigrammes :
     { from-companies-experience,companies-experience-best, experience-best-customer, best-customer-service, customer-service-aside}

3.3 La technique de classification de sentiments

3.3.1 Techniques de classification

Il y a plus d'un demi-siècle, les scientifiques sont devenus très sérieux au sujet de répondre à la question : "Peut-on construire un modèle qui apprend à partir des données disponibles et automatiquement rend les bonnes décisions et les prévisions ?" En regardant en arrière, la réponse peut être trouvée dans de nombreuses applications qui émergent dans les domaines de la classification des formes, l'apprentissage de la machine, et l'intelligence artificielle.

L'un des sous-domaines de la modélisation prédictive est la classification de motif supervisé ; la classification de motifs supervisée est la tâche de l'entrainement d'un modèle basé sur des données d'entrainement marquées qui peut ensuite être utilisée pour attribuer une étiquette pré-classe définie à de nouveaux objets.

Les classificateurs Naive Bayes sont des classificateurs linéaires qui sont connus pour être simple mais trés efficace. Le modèle probabiliste de classificateurs de Bayes Naïf est basé

sur le théorème de Bayes, et l'adjectif Naïf vient de l'hypothèse que les caractéristiques d'un ensemble de données sont mutuellement indépendants.

La technique de classification utilisée dans l'analyse de sentiment est le classifieur Naïf Bayes. Cette méthode est une méthode de classification par apprentissage automatique supervisé. cette derniére utilise un ensemble d'apprentissage où les données sont déjà classifiées pour construire une base de connaissances sur laquelle les données non classiées le seront. L'approche sac de mot ou "bag of words" est un des premiers modèles de représentation textuelle, qui est souvent utilisé pour l'analyse de sentiment. Peut être représenté comme un ensemble de n-grammes.

Plusieurs approches de classification utilisent cette représentation pour construire des systèmes de classification en sentiment de texte. l'exactitude de cette approche peut être très elevée.Dans ce chapitre nous allons nous baser sur le classifieur Naïf bayes.

La construction d'un classifieur consiste à donner en entrée le ducument "représente un tweet" qui est composé d'un ensemble de mots "bag of words" et appartient à un ensemble de classe "P ou N".

3.3.2 Classificateur de Bayes

Étant donnée une fonction SCS la fonction de classification de sentiment (Sentiment Classification Status) prends des valeurs entre 0 et 1, $c_i \in C$ et C = {0, 1}.

Les Tweets seront ensuite classés en fonction de leur valeurs SCS_i, cette fonction est définie en terme de probabilité et la construction d'un classifieur peut consister la définition d'une fonction $SCS_i : D \rightarrow \{P, N\}$ où d'une fonction $SCS_i : D \rightarrow [0, 1]$ un paramètre de seuil est défini tel que $SCS_i(d_j) \geq s_i$ est interprété comme positif-P et $SCS_i(d_j) \leq s_i$ est interprété comme Négatif-N.

Les classificateurs probabilistes interprètent la fonction $SCS_i(d_j)$ en terme de $P(c_i/d_j)$ c'est un classifieur linéaire probabiliste basé sur le théorème de bayes, simple et avec une forte indépendance(naïve) des hypothèses. Malgré sa simplicité, Naïf Bayes peut surpasser les méthodes de classification plus sophistiquées.

Soient $T_1...T_k$, $p(c_i/d_j)$ représente la probabilité qu'un document ou un tweet Tous les attributs sont des tweets représentés par un vecteur(ensemble) de terme $d_j =< t_1, t_2, t_3...t_k >$ qui appartient à la classe c_i. On peut calculer cette probabilité en utilisant le théorème de Bayes, définie par [75] :

$$\underbrace{P(c_i/d_j)}_{probanilite\ a\ posteriori} = \underbrace{P(d_j/c_i)}_{maximum\ de\ vraissemblance} \times \underbrace{P(c_i)}_{probabilite\ a\ priori} /P(d_j) \tag{3.1}$$

L'objectif principale du théorème NB est de maximiser la probabilité a posteriori, la classification consiste à calculer la probabilité a postériori suivante :

$$Pr = (C = c_j \mid T_1 = t_1....T_k = t_k) \tag{3.2}$$

est maximale.
En appliquant la règle de Bayes :

$$Pr = (C = c_j \mid T_1 = t_1 .... T_k = t_k) \tag{3.3}$$

$$= \frac{(T_1 = t_1, ..., T_k = t_k \mid C = c_j) Pr(C = c_j)}{Pr(T_1 = t_1, T2 = t_2, ... T_k = t_k)} \tag{3.4}$$

$$= \frac{(T_1 = t_1, ..., T_k = t_k \mid C = c_j) Pr(C = c_j)}{\sum_{i=0}^{T} Pr(T_1 = t_1, T2 = t_2, ... T_k = t_k) Pr(C = c_j)} \tag{3.5}$$

le dénominateur $P(T_1 = t_1, ..., T_k = t_k)$ est sans importance pour la prise de décision puisque c'est le même pour toutes les classes.

On a besoin uniquement de $P(T_1 = t_1, ..., T_k = t_k \mid C = c_j)$.
Maintenant une hypothèse est nécessaire : **Tous les attributs sont conditionnellement indépendants étant donné une classe** $C = c_j$
donc :

$$Pr(T_1 = t_1 \mid T = 2 = t_2, ..., T_k = t_k, C = c_j) = Pr(T_1 = t_1 \mid C = c_j) \tag{3.6}$$

Et ainsi de suite pour T_2 à T_k

$$Pr(T_1 = t_1, T = 2 = t_2, ..., T_k = t_k \mid C = c_j) = \prod_{i=1}^{T} Pr(T_i = t_i \mid C = c_j) \tag{3.7}$$

Étant donné un exemple de test, on fait le calcul suivant afin de déterminer la classe la plus probable pour l'instance de test.

$$SCS_i = c_{NB} = argmax Pr(c_j)_{c_j \in C} \prod_{i=0}^{T} Pr(T_i = t_i \mid C = c_j) \tag{3.8}$$

$$c_{MAP} = argmax Pr(c_j) Pr(x_1, x_2, ..., x_n \mid c_j) \tag{3.9}$$

Dans cette formule, f représente une fonction et n i (d) représente le nombre de fonction f_i trouvée dans le tweet d. Il y a un total de m fonctions. Les paramètres $P(c)$ et $P(f \mid c)$, sont obtenus par le biais des estimations du maximum de vraisemblance, et ajoutez 1 comme lissage est utilisé pour les fonctions invisibles.

3.3.3 Apprentissage par Naïf Bayes

— Estimation du maximum de vraissemblance
il suffit d'utiliser les fréquences dans les données
Nombre de fois le mot w_i apparaît parmi tous les mots dans les documents de topic c_j

$$p(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V}^{n} cout(w, c_j)} \tag{3.10}$$

Créer méga-documents pour le sujet j par concaténation de tous les documents dans ce sujet.

La fréquence de l'utilisation de w dans méga-documents.

Pour bien lisser Naïf Bayes et éviter que le produit soit nul on rajoute un 1 alors :

$$p(w_i \mid c_j) = \frac{count(w_i, c_j) + 1}{\sum_{w \in V}^{n} cout(w, c_j) + \mid V \mid} \tag{3.11}$$

Donc on va affecter chaque mot a une classe $P(mot \mid classe)$ après avoir la probabilité de chaque mot sachant qu'on affectera à la classe tout le tweet :

$$P(classe \mid tweet) = P(classe) \times \prod P(mot \mid classe) \tag{3.12}$$

	tweets	bags of words(mots)	classes
Training	1	Chinese Beijing Chinese	P
	2	Chinese Chinese Shanghai	P
	3	Chinese Macao	P
	4	Tokyo Japan Chinese	P
Test	5	Chinese Chinese Chinese Tokyo Japan	N

En utilisant les lois de la probabilité conditionelle à priori :

Probabilité à priori :

$$P(c) = \frac{N_c}{N} \tag{3.13}$$

$P(c) = \frac{3}{4}$ et $P(c) = \frac{1}{4}$

Probabilité conditionelle :

$$p(w_i \mid c_j) = \frac{count(w_i, c_j) + 1}{\sum_{w \in V}^{n} cout(w, c_j) + \mid V \mid} \tag{3.14}$$

$$\begin{aligned} P(Chinese \mid P) &= (5+1)/(8+6) = 6/14 = 3/7 \\ P(Tokyo \mid P) &= (0+1)/(8+6) = 1/14 \\ P(Japan \mid P) &= (0+1)/(8+6) = 1/14 \\ P(Chinese \mid N) &= (1+1)/(3+6) = 2/9 \\ P(Tokyo \mid N) &= (1+1)/(3+6) = 2/9 \\ P(Japan \mid N) &= (1+1)/(3+6) = 2/9 \end{aligned} \tag{3.15}$$

Choisir la classe la plus probable :

$$\begin{aligned} P(P \mid T5) &\approx 3/4 \times (3/7)^3 \times 1/14 \times 1/14 \approx 0.0003 \\ P(N \mid T5) &\approx 1/4 \times (2/9)^3 \times 2/9 \times 2/9 \approx 0.0001 \end{aligned} \tag{3.16}$$

Alors la classe la plus probable est la classe positive.

3.3.4 Implémentation de Naïf Bayes

L'algorithme qui représente l'idée de la classification des sentiments avec le classifieur NB est le suivant :

Algorithm 1: Algorithme Classifieur de Naïf de bayes

Input :
$C = \{c1, c2\}$; // ensemble de classes P ou N
$V = w_1, w_2, ..., w_n$; // ensemble de mots d'apprentissage (Vocabulaire)
$V_t = w_1, w_2, ..., w_m$; // ensemble de mots de tests (vocabulaire de tests)
Output : $CSC \Longrightarrow C$; //classifieur entraîner
Description :
1: // calculer les termes P(c_j)
Pour chaque c_j dans C do {
$count(w_i) \leftarrow$ tout l'ensemble de mots avec la $classe = c_j$;
P(c_j) $=\leftarrow \frac{|count(w_i)|}{|V|}$;
// calculer $P(w_i \mid c_j)$
Pour chaque nombre d'occurences de w_k dans le vocabulaire w_t (de tests)
$P(w_k \mid c_j) \leftarrow \frac{count(w_i,c)+1}{count(c_j)+|V|}$;
$SCS_{NB} = argmax_{c_j \in C} P(c_j) \prod_{i \in positions} P(w_i \mid c_j)$;
}
return $classe = c_j$;

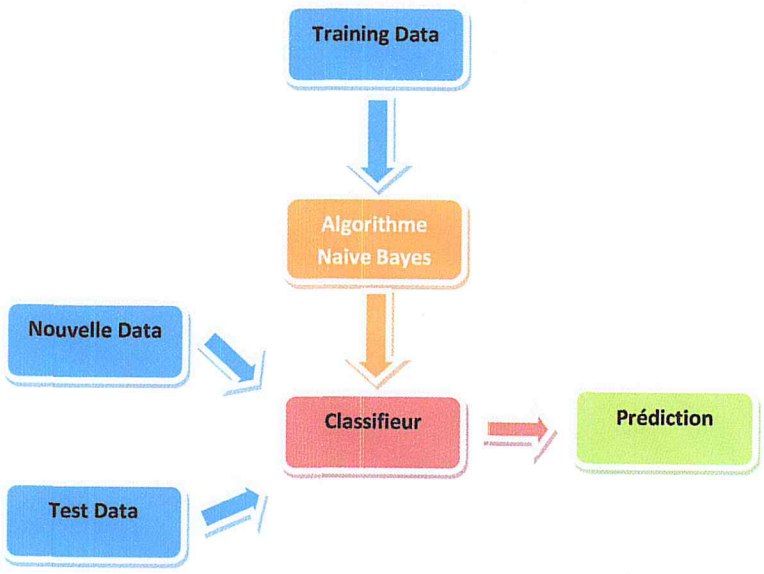


Figure 3.8 – Schéma explicatif pour le modèle générale de la classification avec le Naïf bayes

3.4 Polarité de sentiment

Une étape plus avancée dans la classification de sentiment est le calcul de de la polarité de sentiment qui représente le degrée de sentiment. Cela peut être fait en utilisant le calcul de la polarité des mots individuel de tweet.

4 Conclusion

Dans ce chapitre nous avons présenté les techniques de domaine de SA et les étapes de la classification des sentiments.

Nous avons présenté les techniques d'apprentissage en précisant les étapes d'analyse de sentiments, aussi les différents moyens de représentátion de corpus de données.

Ainsi quel'algorithme du classifieur NB et la conception détaillée de l'approche $EGEE_{status}$. Dans le chapitre qui suit nous allons montrer l'utilisation des techniques présentées dans ce chapitre pour l'AS " opinion mining", ainsi que l'implémentation et les résultats trouvé.

Chapitre 4

Implémentation et Résultats

1 Introduction

Après avoir présenté les diférentes étapes pour appliquer l'analyse des sentiments d'une manière appropriée sur les tweets dans le chapitre précédent, il est temps de présenter l'implémentation de ce processus, et discuter les résultats obtenus en prennant en considération l'utilisation des datasets connus dans ce domaine pour juger d'une façon rationnelle l'efficacité de notre approche $Intelligent_{business}$ d'analyse des sentiments.

2 Conception

2.1 Partie Logique

Afin de maintenir une approche de programmation efficace pour l'implémentation des méthodes d'analyses des sentiments par apprentissage automatique, nous avons utilisé le modèle de conception orientée objet.

L'utilisation de la programmation orientée objet permet de respecter les principes de la bonne conception des logiciels comme l'encapsulation, la cohésion, couplage faible, etc. La POO utilise le concept des classes et des objets, et rend le code source plus lisible et compréhensible, et aussi plus maintenable.

La structure du projet est la suivante :

```
Projet
├── java ..... Package contient que les classes Java
│   └── com
│       ├── app ..... Package de l'application
│       │   ├── helper ..... Contenant des classes d'aide
│       │   ├── main ..... Contenant la classe principale
│       │   ├── model ..... Contenant les classes modèle des objets
│       │   ├── temp ..... Contenant des classes pour traitement temporaire
│       │   └── ui ..... Contenant les classes de l'interface graphique
│       │       └── controllers ..... Contenant les controlleurs de JavaFx
│       └── apphelper ..... Package des applications d'aide
│           ├── sanders ..... Préparation de Sanders
│           └── sentiment140 ..... Préparation de Sentiment140
├── ressources ..... Package contient autres fichiers
│   ├── datasets ..... Contients les fichiers csv des datasets
│   ├── fxml ..... Contient les fichiers fxml de l'interface JavaFx
│   ├── images ..... Contient les images utilisées
│   ├── émoticônes ..... Contient le fichier csv de tous les émoticônes
│   └── motsvides ..... Contenant les fichiers des mots vides
└── tests ..... Package contenant les classes des tests en JUnit
```

2.2 Partie Interface Utilisateur

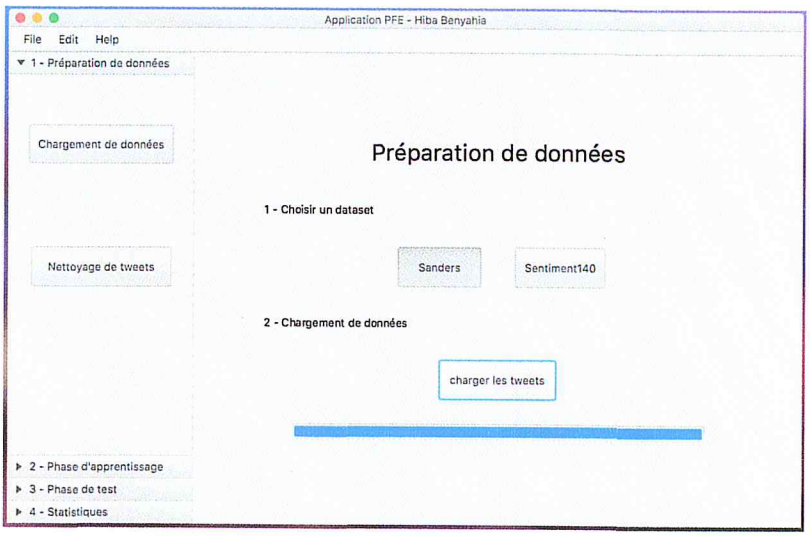


Figure 4.1 – L'interface utilisateur principale : le premier traitement (chargement d'un dataset)

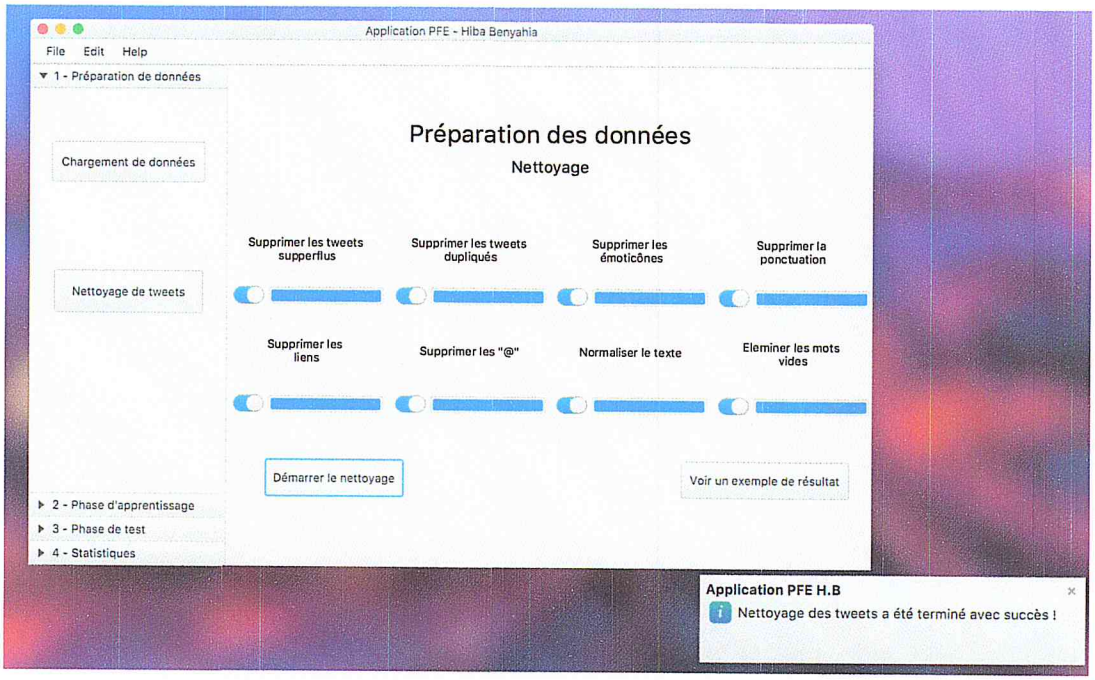


Figure 4.2 – Interface utilisateur de nettoayge de tweets

Figure 4.3 – Interface utilisateur de comparaison entre les tweets avant/après nettoyage

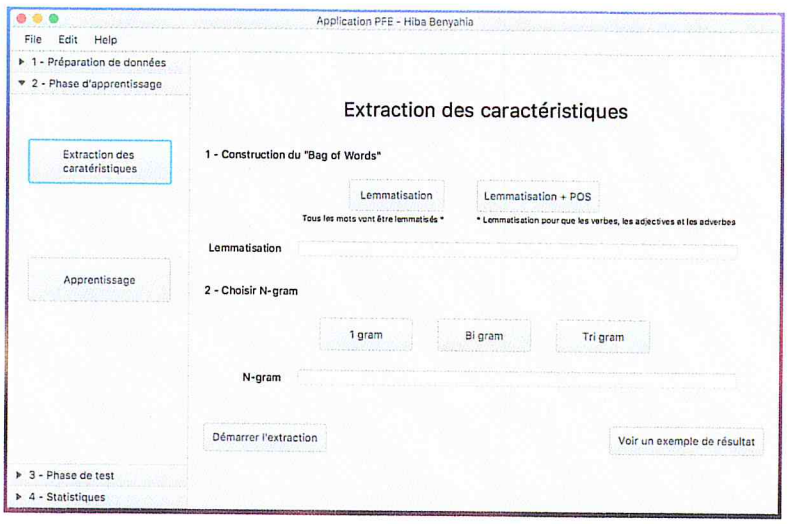


Figure 4.4 – Interface utilisateur d'extraction des caractéristiques

2.3 Pipeline de traitements

2.3.1 Notre approche $Intelligent_{business}$

Notre approche est constituée éssentiellement de six étapes dont chacune est composée de plusieurs phases 4.5 :

— Chargement de dataset 4.1

1. Téléchargement de tweets.
2. Organisation des données.

— Nettoyage de tweets 4.2

1. Élimination des tweets non anglais.
2. Élimination des tweets répétés.
3. Suppression des URLs.
4. Élimination des noms d'utilisateurs.
5. Élimination des émoticônes.
6. Normalisation.
7. Élimination des mots vides.
8. Élimination de ponctuation.

— Extraction des caractéristiques. 4.4

1. Lemmatisation.
2. Étiquetage morpho-syntaxique. 4.4
3. Extraction des N-grammes.

— phase d'apprentissage

1. Création du vocabulaire.

2. Calcul des probabilités.

— Phase de tests. 4.7

1. Préparation des tweets de tests.
2. Éxécution de classification.

— Statistiques. 4.10 et 4.11

1. Benchmark des résultats.
2. Comparaison des résultats

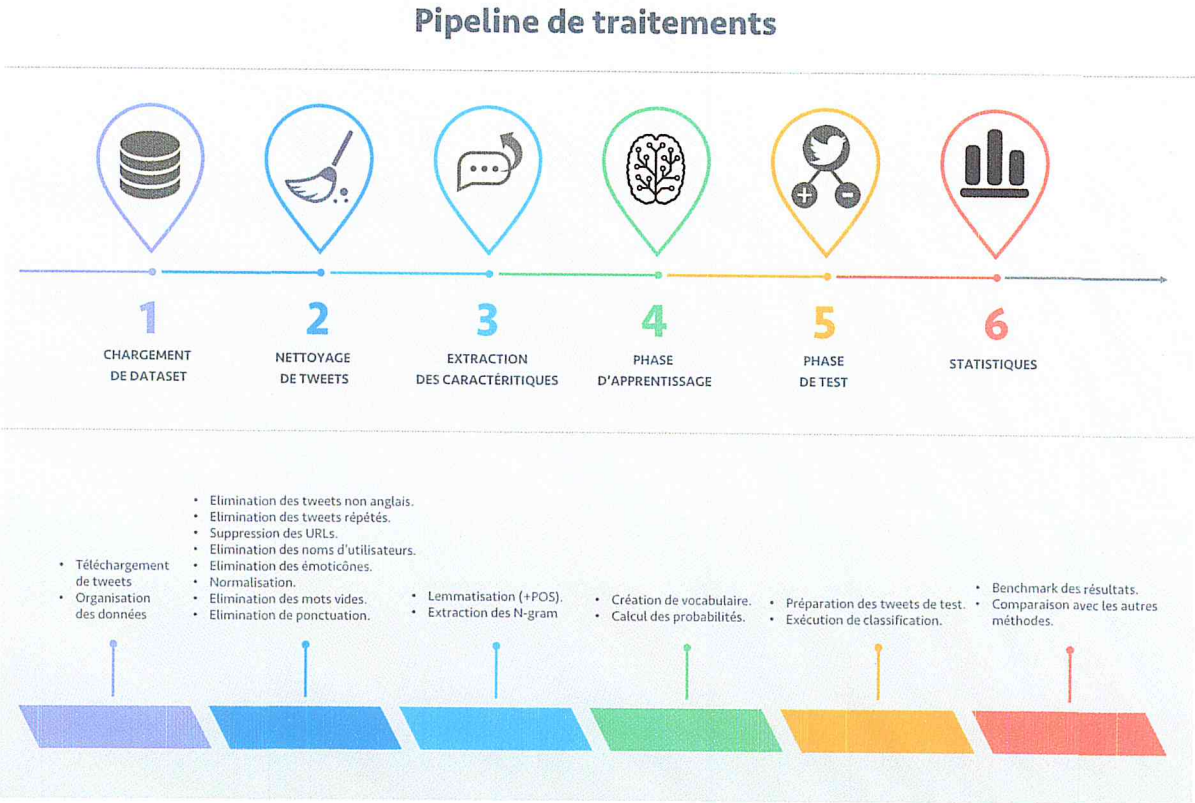


Figure 4.5 – Conception générale de notre approche

3 Imlémentation

3.1 Software

3.1.1 Langage de programmation

Le langage de programmation choisi pour l'implémentation de l'application est Java. Ce choix viens de la présence de plusieurs motivations : la première est d'aquérir une bonne expérience de programmation avec ce langage et la deuxième viens du faite que Java est considéré comme étant un des meilleurs langages de programmation multi-plateforme.

Développé par James Gosling chez Sun Microsystems en 1990, et maintenant son développement est maintenu est pris en charge et développé par Oracle depuis 2010 Java est un langage de programmation orienté objet qui répond à une variété de besoins de développement : développement des applications mobiles, applications Web et d'entreprise, services Web, applications et logiciels de bureau et qui s'exécutent sur tous les systèmes d'exploitation utilisant le même code source (cette caractéristique est appelée "WORM : Write Once and Run Anywhere"). Actuellement, Java reste le langage de programmation le plus populaire [76].

3.1.2 Environnement de développement

L'environnement de développement utilisé pour l'application est : IntelliJ IDEA ultimate (2016.1). Nous l'avons utilisé pour écrire et gérer le code source avec plus de flexibilité et efficacité. IntelliJ IDEA Ultimate est un IDE payant, mais actuellement, il est parmi les environnements de développement les plus intelligents de Java. Il est utilisé par Google, et les grandes entreprises de développement de logiciels très complexes [77].

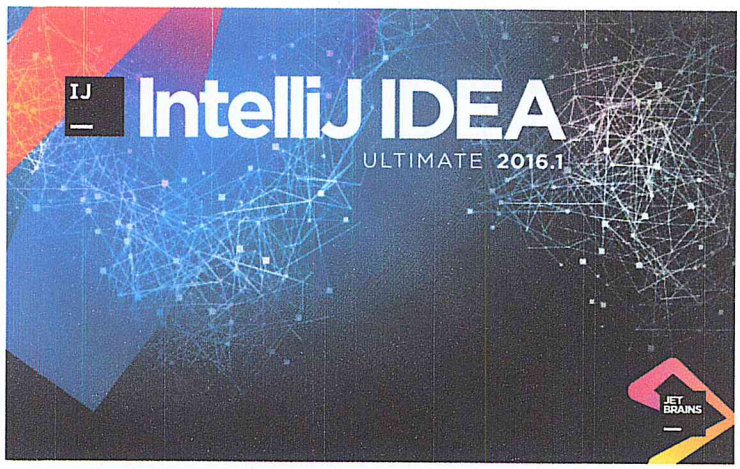


Figure 4.6 – IntelliJ 2016.1 Ultimate Intro

3.1.3 Bibliothèques tières

— **uniVocity Parsers (Version 2.1.0) :** est une collection de parseurs extrêmement rapides et fiables pour Java [78]. Il fournit une interface cohérente pour gérer différents formats de fichiers. Leur parseur CSV est le plus rapide parmis tous les parseurs des fichiers CSV pour Java [79], il a été utilisé pour la lecture et l'ecriture des fichiers CSV des datasets.

— **Twitter4j (Version 4.0.3) :** Twitter4J est une bibliothèque Java pour l'API de Twitter. Avec Twitter4J, on peut facilement intégrer une application Java avec le service Twitter. Il existe d'autres bibliothèques Java comme : Java-Twitter et JTwitter. JTwitter est considérée comme la bibliothèque la plus faible des trois. Elle ne semble pas avoir été mis à jour, en plus elle ne couvre pas beaucoup de l'API de Twitter, et ne semble pas avoir d'autres versions majeurs en dehors de la version initiale. Les deux autres, Java-Twitter et Twitter4J elles sont beaucoup plus proches en termes

de qualité. Elles couvrent l'API de base et le nouveau API 1.1, les fichiers sources sont libres et disponibles ainsi que la documentation en ligne. Un point de plus pour Twitter4J et qu'elle est utilisée généralement dans les logiciels professionnels tel que "Mathematica".

— **Stanford CoreNLP (Version 3.6.0) :** La bibliothèque Stanford CoreNLP [80] fournit un ensemble d'outils d'analyse du langage naturel. Elle peut donner les formes de base de mots, leurs parties du discours (POS Tags), et même la reconnaissance des entités nommées tels que les organisations, le lieu, etc. Cette bibliothèque peut même faire la lemmatisation (que pour la langue anglaise), et elle peut marquer la structure des phrases et les dépendances de mots, etc.
Nous avons utilisé cette bibliothèque pour faire l'étiquetage morphosyntaxique et la lemmatisation.

— **JavaFX (Version 8) :** Une bibliothèque native de Java [81]. Elle offre une interface utilisateur graphique très personnalisable, belle, riche et intéractive. Les interfaces graphiques conçues utilisant JavaFx sont connues pour leurs fluidité, stabilité et rapidité, même avec la présence des traitements lourds dans l'application. C'est pour cela les ingénieurs de la NASA utilisent Java et JavaFx pour le développement logiciel [82]. Nous avons utilisé JavaFx à cause de sa rapidité qui ne peut pas être affectée par la complexité de nos traitements de "big data".

3.1.4 Environnement d'exécution

— **Machine Virtuelle :** Java JVM (SE Version 8 mise à jour 91).
— **Systèmes d'exploitation :** Windows 10 Pro x64, MacOS X El Capitan 10.11.5 x64.

3.2 Hardware

— **PC Portable Lenovo** : CPU : Intel i3 - 3210 @ 3.20 GHz, RAM : 4 Go DDR3, GPU : Intel Graphics, , Disque Dur : 1 Tb HDD.
— **MacBook Pro** : CPU : Intel i7 @ 2.5 Ghz, RAM : 16 Go DDR3, GPU : NVIDIA GeForce GT 750M 2048 Mo, Disque Dur 512 Gb SSD.

4 Résultat d'exécution

4.1 Phase d'apprentissage

4.1.1 Sanders dataset

L'ensemble de données Sanders se compose de 5512 tweets sur quatre sujets différents (Apple, Google, Microsoft, Twitter). Chaque tweet a été étiqueté manuellement par un annotateur comme positif, négatif, neutre, ou non pertinentes par rapport au sujet. Le processus d'annotation a donné lieu é 654 négative, 2.503 neutre, 570 positif et 1.786 tweets pertinents. L'ensemble de données a été utilisé dans [83], [84], [85] pour la polarité et la classification de la subjectivité des tweets. L'ensemble de données Sanders est disponible à " http ://www.sananalytics.com/lab".

— **Statistiques sur le dataset :** Figure 4.7

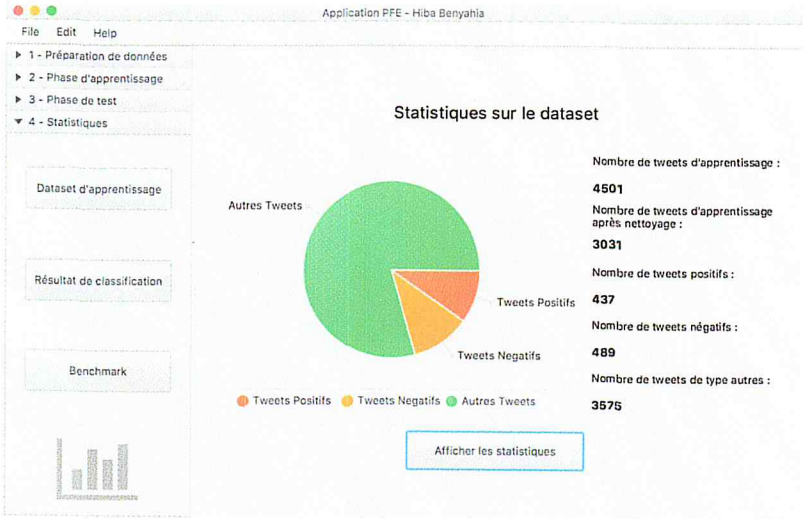


Figure 4.7 – Statistiques sur le dataset d'apprentissage de Sanders

— **Statistiques sur l'opération d'apprentissage :**
- Temps de chargement de tweets $< 1sec$
- Temps de nettoyage $< 4sec$, Résultat : **Figure 4.8**
- Temps d'extraction des caractéristiques $< 8sec$, Résultat : **Figure 4.9**
- Temps d'apprentissage $< 20sec$, Résultat : **Tableau 4.1**

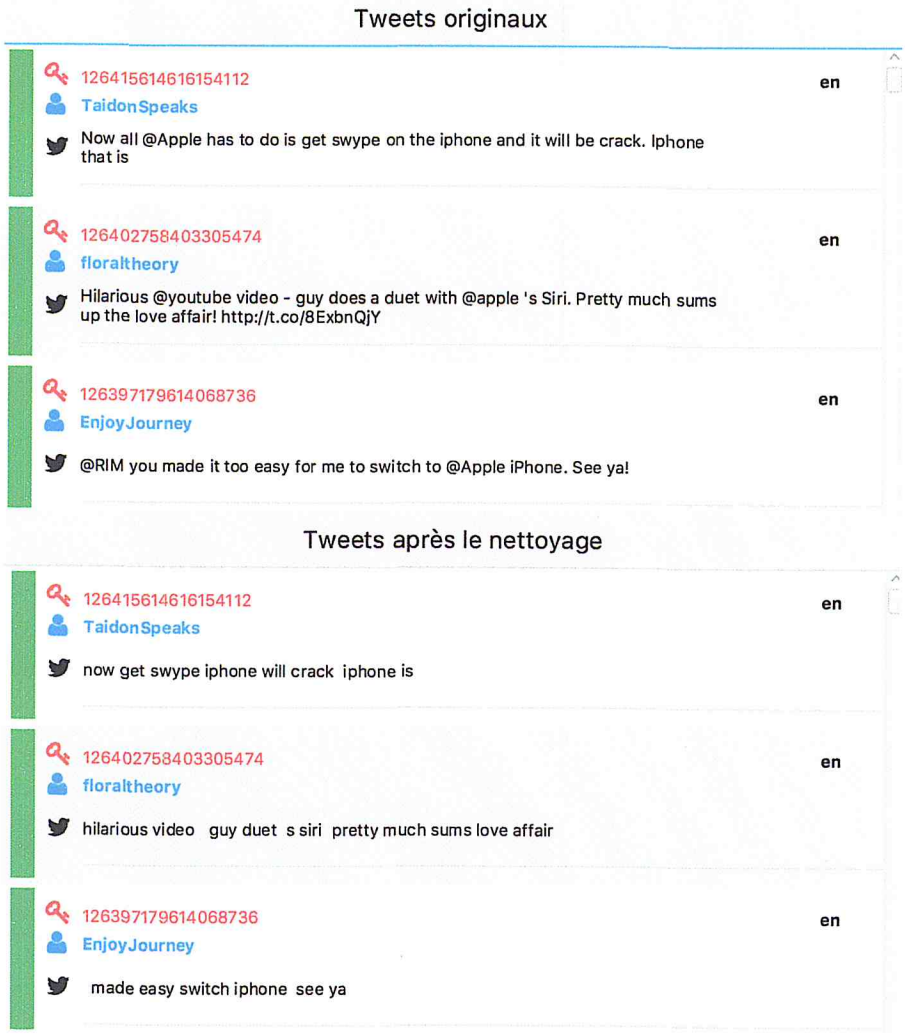


Figure 4.8 – Résultat de nettoyage des tweets de dataset d'apprentissage de Sanders

Figure 4.9 – Résultat de l'extraction des caractéristique des tweets

Taille de vocabulaire d'apprentissage	2499
Nombre de mot dans la classe positive	4275
Nombre de mot dans la classe négative	5131

Table 4.1 – Résultat d'apprentissage sur Sanders

4.2 Phase de test

Dans la phase de test, la classification a été appliquée sur 3 ensembles de données, à base de l'apprentissage sur le dataset de Sanders.

4.2.1 Qu'est ce qu'un bon classifieur

On à 4 cas :
Vrai positif (true positive) : ex : positif classé positif.
Vrai négatif (true negatif) : ex : négatif classé positif.
Faux négatif (false negatif) : ex : positif classé négatif.
Faux positif (false positive) : ex : négatif classé positif.

Classé / Vrai classe	Pos	Nég
Pos	VP	FP
Nég	FN	VN
total	P	N

Table 4.2 – Mesures de performance d'un algorithme de classification

Evaluation de résultats :

— Le Rappel
Le rappel est défini par le nombre de documents pertinents retrouvés au regard du nombre de documents pertinents que possède la base de données. Cela signifie que lorsque l'utilisateur interroge la base il souhaite voir apparaître tous les documents qui pourraient répondre à son besoin d'information. Si cette adéquation entre le questionnement de l'utilisateur et le nombre de documents présentés est importante alors le taux de rappel est élevé. À l'inverse si le système possède de nombreux documents intéressants mais que ceux-ci n'apparaissent pas dans la liste des réponses, on parle de silence. Le silence s'oppose au rappel.

— La Précision
La précision est le nombre de documents pertinents retrouvés rapporté au nombre de documents total proposé par le moteur de recherche pour une requête donnée.
Le principe est le suivant : quand un utilisateur interroge une base de données, il souhaite que les documents proposées en réponse à son interrogation correspondent à son attente. Tous les documents retournés superflus ou non pertinents constituent du bruit. La précision s'oppose à ce bruit documentaire. Si elle est élevée, cela signifie que peu de documents inutiles sont proposés par le système et que ce dernier peut être considéré comme "précis".

— Le F-mesure
Une mesure populaire qui combine la précision et le rappel est leur moyenne harmonique, nommée F-mesure (soit F-measure en anglais) ou F-score.

Précision :

$$\frac{TP}{TP+FP} \tag{4.1}$$

Recall :

$$\frac{TP}{TP+FN} \tag{4.2}$$

F-score :

$$\frac{2\times P\times R}{P+R} \tag{4.3}$$

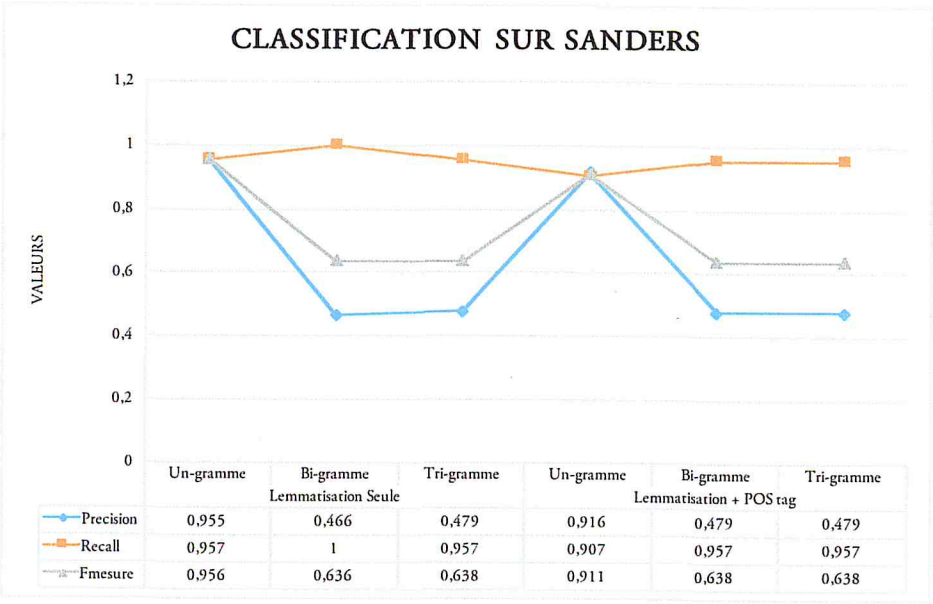
4.2.2 Ensemble de tests Sanders

Pour tester l'algorithme, la classification a été appliquée sur l'ensemble d'apprentissage lui même (les tweets de Sanders).

D'après les statistiques obtenues Figure 4.10 l'algorithme est arrivé à une très bonne précision de 95,5%, même cas pour le Recall et le F-mesure utilisant le résultat de l'apprentissage avec la lemmatisation seule des 1-grammes. Cela signifie que notre implémentation du Naïf Bayes est juste.

Cependant, la précision de classification utilisant la lemmatisation et l'étiquetage mrophosyntaxique a dépassé 91% utilisant seulement les lemmes des verbes et des adjectifs et des adverbes, qui est un vocabulaire de taille très réduite par rapport à celui de lemmatisation seule, cela signifie qu'on peut atteindre un très bon taux de précision avec l'ajout du POS-tagging, et avec une réduction de la taille du vocabulaire. Cela signifie aussi que les verbes, les adjectifs et les adverbes porte la majorité des informations sur le sentiment des tweets.

On peut remarquer aussi que l'utilisation des bi-grammes et tri-grammes n'améliore pas le resultat de la classification.



	Un-gramme	Bi-gramme	Tri-gramme	Un-gramme	Bi-gramme	Tri-gramme
		Lemmatisation Seule			Lemmatisation + POS tag	
Precision	0,955	0,466	0,479	0,916	0,479	0,479
Recall	0,957	1	0,957	0,907	0,957	0,957
Fmesure	0,956	0,636	0,638	0,911	0,638	0,638

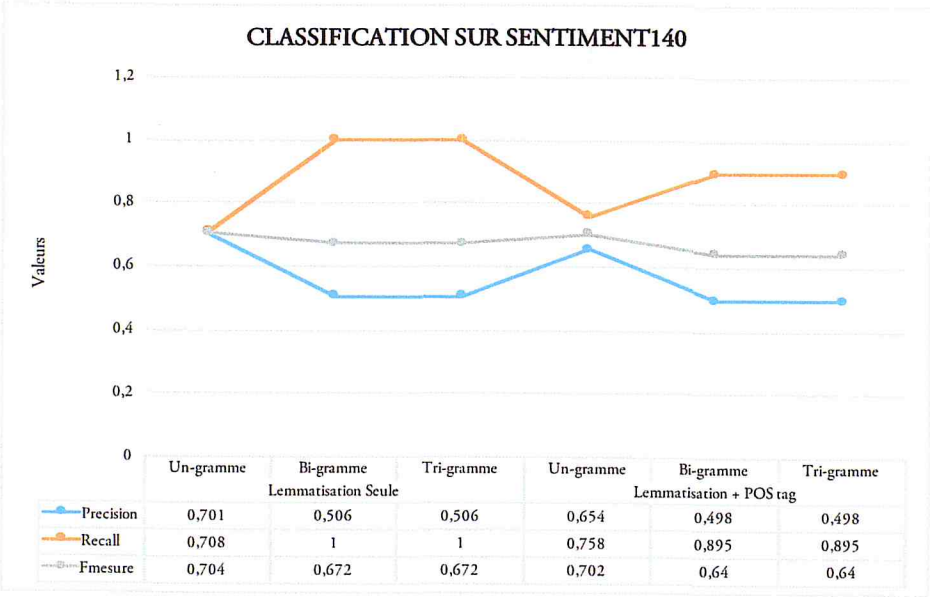
Figure 4.10 – Benchmark des résultats de classification des tweets de Sanders utilisant des étapes d'apprentissage différentes

4.2.3 Ensemble de tests Sentiment140

Dans ce test, la classification est faite sur un ensemble de tests fournis avec le dataset Sentiment140, qui est constitué de 497 tweets manuellement annotés par leurs sentiments.

D'après les statistiques obtenues Figure 4.11 La précision que nous avons obtenus dépace 70%, un bon résultat vu que nous avons fait l'apprentissage sur seulement 3000 tweets.

Les même remarques sur les résultats précédents concernant l'utilisation de POS-tagging sont appliquées sur ce cas.



	Un-gramme	Bi-gramme	Tri-gramme	Un-gramme	Bi-gramme	Tri-gramme
		Lemmatisation Seule			Lemmatisation + POS tag	
Precision	0,701	0,506	0,506	0,654	0,498	0,498
Recall	0,708	1	1	0,758	0,895	0,895
Fmesure	0,704	0,672	0,672	0,702	0,64	0,64

Figure 4.11 – Benchmark des résultat de de classification des tweets de test de Sentiment140 utilisant des étapes d'apprentissage différentes

4.2.4 Ensemble de tweets téléchargés

Dans ce cas, la classification des tweets est faite sur un ensemble de 2470 tweets que nous avons téléchargé sur le produit "iPhone" dont le sentiment n'est pas connu pour vérifier la validité de l'analyse des sentiments par notre méthode dans le marketing.

La vérification manuelle partielle des tweets classés via une interface dédiée Figure 4.12 montre que l'algorithme fait une bonne classification des sentiments, cela implique que cette méthode peut être utilisée pour classer les sentiments sur les produit dans les tweets.

Un petit ensemble des tweets téléchargés et classés contient des informations sur la géolocalisation des tweets qui sera très importante pour les entreprises. La figure 4.13 montre l'interface dédiée à l'affichage des tweets classés.

Figure 4.12 – Résultat de classification de tweets téléchargés sur iPhone

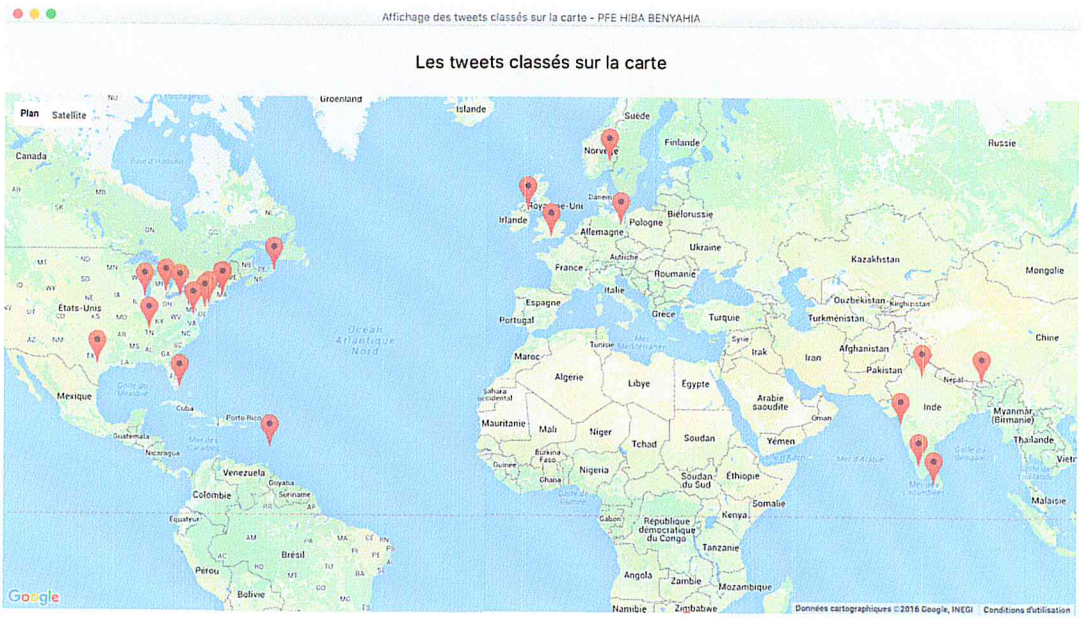


Figure 4.13 – Affichage des tweets classés sur la carte

5 Les limites des données recueillies sur les réseaux sociaux

Malgré toutes ces améliorations et touts ces bienfaits dans le domaine commerciales et publicitaires, il est nécessaire de cité certaines limites qui sont lié à la nature informelle des données extraites et à leurs enjeux stratégiques : Les opinions sont-elles sincères : Les défis liés au traitement du contenu des réseaux sociaux :

— Problèmes liés à la nature de traitement de text
  - La langue.
  - La présence de vocabulaire (soit pour le nétoyage, soit pour l'extraction des caratéristiques ou pour l'extraction de sentiment).
  - Un sens explicite est integré dans la phrase.

— Problemes liés à Twitter
  - Limite d'extraction des tweets.
  - Beaucoup de tweets repetés.
  - Extraction des caractéristiques dificile à cause de grande "sparcity" de Twitter.
  - Les abriviations.

— Problème specifique à l'analyse de sentiment
  - Le sentiment est une chose morale.
  - Il existe des caractéritiques qui peuvent inverser le sentiment.
  - Bien selectionner les paramètres de sentiment.

— Problemes liés aux algorithmes de classification
  - Choisir les bonnes caractéritiques pour le bon algorithme.
  - Trouver un dataset adéquat dans le cas de l'utilisation d'un algorithme d'apprentissage automatique supervisé.

6 Conclusion

Dans ce chapitre nous avons présenté en détail notre approche implémentée et l'interface de notre application ainsi quelques résultats de tests. Nous avons collecté et récupéré les tweets contenant le terme "IPhone", les résultats de la classification ont était présenté et bien analysé en utilisant les mesures de performance des algorithmes de classification (Rappel, Précision et F-mesure), ces derniers ont trouvé de bon résultats de classification de sentiments. Nous avons illustré les sentiments (positifs et négatifs) sur une Map intéractive du monde entier afin d'aider les entreprises à afficher les panneaux publicitaires dans les régions de sentiments négatifs. Ce qui permet de minimiser le coût de publicité. Notre approche $Intelligent_{business}$ peut être utilisée pour afficher les panneaux publicitaires de n'importe quel produit de n'importe quelle entreprise.

Conclusion générale

Synthèse

D'après cette étude nous avons trouvé notre intérêt dans la classification des sentiments des tweets collectés, nous avons développé un système de classification d'opinions exprimées dans les critiques des produits sur Twitter.

le principal objectif était d'affecter :

- une analyse de sentiments dans la problématique est de classifier les sentiments à partir d'un clasifieur probabiliste qui s'appelle le classifieur Naïf Bayes.
- d'appliquer les prétraitements sur les datasets avec le nettoyage de données.
- de composer le résultat de nettoyage en ngrammes afin de créer les sacs de mots(notre vocabulaire). et faire un étiquetage morpho-syntaxique pour extraire seulement les verbes et les adjectifs afin de minimiser la taille du vocabulaire qui implique la minimisation du temps de traitement et aussi le temps d'exécution.
- d'entrainer notre algorithme avec un corpus de données (datasets sanders) classifier manulellement et appliquer les tests du dataset (sentiment140) et les tweets collecté à partir de Twitter par le mot clé "iPhone".
- de mesurer la classification et évaluer l'apprentissage par les mesures d'évaluations Rappel, Précision et f-mesure.

Le deuxième objectif était de proposer une méthode aux entreprises à mettre en place un plan de marketing pour leurs produits où marques, cela permet aux firmes d'identifier l'avis public qui compose les marché auxquels elles s'adressent. l'étape importante ici est de récupérer la géolocalisation pour les tweets qui contiennent la localisation géographique.

Nous avons illustré les sentiments (positifs et négatifs) sur une Map du monde entier afin d'aider l'entreprise à mettre en place les panneaux publicitaires dans les régions de sentiments négatifs. Ce qui permet de minimiser le coût de publicités.

Après les tests effectués, nous pouvons constater que nous avons réussi à implanter une méthode innovante basée sur un classificateur probabiliste qui est le classifieur naïf bayes. Les résultats obtenus après cette classification donnent une plus grande satisfaction. Nous pouvons donc conclure que la classification de sentiment avec les méthodes d'apprentissage suppervisé profonde est une voie importante de recherche dans le domaine de l'Analyse de Sentiments.

Perspectives

le système présenté dans ce travail est une proposition logicielle pour l'aide à l'amélioration des stratégies de marketing afin d'aider les entreprise a minimiser leurs dépenses publicitaires. Mais aussi il est utile de préciser que cette approche donne de nombreuses perspectives de recherche.

Dans le cadre de la recherche , il serait intéressant d'optimiser le temps d'éxécution et surtout avec la manipulation de grands corpus de données comme sentiment140.

Lors de l'aboutissement de ce travail, une deuxième perspective apparaît est de changer en temps la publicité tout dépend de l'emplacement du client c'est-à-dire suivre le déplacement du client (l'internaute).

Bibliographie

[1] Staticbrain. Twitter statistics. http://www.statisticbrain.com/twitter-statistics/, 2015. Accessed : 23-03-2016.

[2] Staticbrain. Facebook statistics. http://www.statisticbrain.com/facebook-statistics/, 2015. Accessed : 23-03-2016.

[3] Per Andersen. *What is Web 2.0 ? : ideas, technologies and implications for education*, volume 1. JISC Bristol, 2007.

[4] le Grand Dictionnaire Office québécois de la langue française. Blogue. http://www.granddictionnaire.com/ficheOqlf.aspx?Id_Fiche=8362053, 2012.

[5] Wikis. In *Internet Cool Tools for Physicians*, pages 113–116. Springer Berlin Heidelberg, 2009.

[6] Maria Teresa Zanola. Les anglicismes et le français du xxie siècle : La fin du franglais ? *Synergies Italie*, 4 :87–96, 2008.

[7] Larousse. Microblog. http://www.larousse.fr/dictionnaires/francais/microblog/186517, 2000.

[8] Mr RAHILA Abdelkader. Thème.

[9] Hugo Lauras. L'impact des réseaux sociaux sur les entreprises a-t-il un rôle essentiel sur leur image. http://www.andlil.com/limpact-des-reseaux-sociaux-sur-les-entreprises-a-t-il-un-role-essentiel-sur-leur-i html, 2013. Accessed : 08-05-2016.

[10] Blog du modérateur. Les chiffres clés des réseaux sociaux en 2015. http://www.blogdumoderateur.com/chiffres-reseaux-sociaux/, 2015. Accessed : 12-04-2016.

[11] Blog econcepto. Les chiffres clés des réseaux sociaux en 2015. http://www.econcepto.com/chiffres-cles-reseaux-sociaux-2015/, 2015. Accessed : 12-04-2016.

[12] Facebook Investor Relations. Latest community stats. https://investor.fb.com/home/default.aspx, 2016. Accessed : 21-04-2016.

[13] IFOP. Sondage. http://www.ifop.com/?option=com_publication&type=poll&id=2436, 2013. Accessed : 23-03-2016.

[14] Hugo Lauras. L'histoire de twitter en (un peu plus de) 14 caractére. *La presse*, 4 :18–20, 2011.

[15] Jack Dorsey. twttr sketch. `https://www.flickr.com/photos/jackdorsey/182613360`, 2016. Accessed : 13-03-2016.

[16] Médias sociaux. le français est 7 éme position des langues utilisées sur twitter. *Semlocast*, 4 :18–20, 2013.

[17] Alina STOICA, Philippe SUIGNARD, and Lambert PEPIN. Twitter : Extraction, regroupement et visualisation pour la veille strategique. page 3, 2011.

[18] Beevolve. An exhaustive study of twitter users across the world. `http://www.beevolve.com/twitter-statistics/`, 2012. Accessed : 13-03-2016.

[19] Blog du modérateur. Profil démographique des visiteurs de twitter en france. `http://www.blogdumoderateur.com/profil-demographique-des-visiteurs-de-twitter-en-france/`, 2013. Accessed : 21-04-2016.

[20] Statista. Number of monthly active facebook users worldwide as of 1st quarter 2016 (in millions). `http://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/`, 2016. Accessed : 23-03-2016.

[21] Lei Zhang, Bing Liu, Suk Hwan Lim, and Eamonn O'Brien-Strain. Extracting and ranking product features in opinion documents. In *Proceedings of the 23rd international conference on computational linguistics : Posters*, pages 1462–1470. Association for Computational Linguistics, 2010.

[22] Theresa Wilson and Janyce Wiebe. Annotating attributions and private states. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II : Pie in the Sky*, pages 53–60. Association for Computational Linguistics, 2005.

[23] Randolph Quirk, David Crystal, and Pearson Education. *A comprehensive grammar of the English language*, volume 397. Cambridge Univ Press, 1985.

[24] Dictionnaire de français Larousse. Définitions : opinion. `http://www.larousse.fr/dictionnaires/francais/opinion/56197`, 2016. Accessed : 08-04-2016.

[25] Jacques Cosnier. *Psychologie des émotions et des sentiments*. Retz, 1994.

[26] B. Liu. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool, 2012.

[27] Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1) :1–167, 2012.

[28] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics, 2004.

[29] Dominique Boullier and Audrey Lohard. *Opinion mining et Sentiment analysis : Méthodes et outils*. Openedition Press, 2012.

[30] Philip J Stone, Dexter C Dunphy, and Marshall S Smith. The general inquirer : A computer approach to content analysis. 1966.

[31] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Opinionfinder : A system for subjectivity analysis. In *Proceedings of hlt/emnlp on interactive demonstrations*, pages 34–35. Association for Computational Linguistics, 2005.

[32] Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI spring symposium : Computational approaches to analyzing weblogs*, volume 100107, 2006.

[33] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet : A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer, 2006.

[34] Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3) :210–229, 1959.

[35] G Vinodhini and RM Chandrasekaran. Sentiment analysis and opinion mining : a survey. *International Journal*, 2(6), 2012.

[36] Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. Data quality from crowdsourcing : a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*, pages 27–35. Association for Computational Linguistics, 2009.

[37] Rui Xia, Chengqing Zong, and Shoushan Li. Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, 181(6) :1138–1152, 2011.

[38] Qiang Ye, Rob Law, and Bin Gu. The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, 28(1) :180–182, 2009.

[39] Ziqiong Zhang, Qiang Ye, Zili Zhang, and Yijun Li. Sentiment classification of internet restaurant reviews written in cantonese. *Expert Systems with Applications*, 38(6) :7674–7682, 2011.

[40] Songbo Tan and Jin Zhang. An empirical study of sentiment analysis for chinese documents. *Expert Systems with Applications*, 34(4) :2622–2629, 2008.

[41] Rudy Prabowo and Mike Thelwall. Sentiment analysis : A combined approach. *Journal of Informetrics*, 3(2) :143–157, 2009.

[42] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3) :293–300, 1999.

[43] Kaiquan Xu, Stephen Shaoyi Liao, Jiexun Li, and Yuxia Song. Mining comparative opinions from customer reviews for competitive intelligence. *Decision support systems*, 50(4) :743–754, 2011.

[44] Youngjoong Ko and Jungyun Seo. Automatic text categorization by unsupervised learning. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 453–459. Association for Computational Linguistics, 2000.

[45] Yulan He and Deyu Zhou. Self-training from labeled features for sentiment analysis. *Information Processing & Management*, 47(4) :606–616, 2011.

[46] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1 :12, 2009.

[47] Jonathon Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL student research workshop*, pages 43–48. Association for Computational Linguistics, 2005.

[48] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up ? : sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[49] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd international conference on computational linguistics : posters*, pages 241–249. Association for Computational Linguistics, 2010.

[50] Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, pages 36–44. Association for Computational Linguistics, 2010.

[51] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*, pages 151–160. Association for Computational Linguistics, 2011.

[52] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*, pages 30–38. Association for Computational Linguistics, 2011.

[53] Albert Bifet and Eibe Frank. Sentiment knowledge discovery in twitter streaming data. In *Discovery Science*, pages 1–15. Springer, 2010.

[54] Jacob Cohen. Weighted kappa : Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4) :213, 1968.

[55] Albert Bifet, Geoff Holmes, Richard Kirkby, and Bernhard Pfahringer. Moa : Massive online analysis. *The Journal of Machine Learning Research*, 11 :1601–1604, 2010.

[56] Saša Petrović, Miles Osborne, and Victor Lavrenko. Streaming first story detection with application to twitter. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189. Association for Computational Linguistics, 2010.

[57] Stephen R Garner et al. Weka : The waikato environment for knowledge analysis. In *Proceedings of the New Zealand computer science research students conference*, pages 57–64. Citeseer, 1995.

[58] Sergio Hernández and Philip Sallis. Sentiment-preserving reduction for social media analysis. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 409–416. Springer, 2011.

[59] Bernard J Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power : Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11) :2169–2188, 2009.

[60] Fotis Aisopos, George Papadakis, Konstantinos Tserpes, and Theodora Varvarigou. Content vs. context for sentiment analysis : a comparative analysis over microblogs. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, pages 187–196. ACM, 2012.

[61] Arturo Montejo-Ráez, Eugenio Martínez-Cámara, M Teresa Martín-Valdivia, and L Alfonso Ureña-López. Ranked wordnet graph for sentiment polarity classification in twitter. *Computer Speech & Language*, 28(1) :93–107, 2014.

[62] Synthesio. Social media intelligence platform. `http://www.synthesio.com/our-platform/`, 2016. Accessed : 02-04-2016.

[63] FEVAD Fédération du E-commerce et de la Vente à Distance. Les consommateurs de plus en plus connectés via les réseaux sociaux et l'internet mobile. `http://www.fevad.com/espace-presse/7eme-barometre-sur-les-comportements-d-achats-des-internautes`, Juin 2011. Accessed : 01-04-2016.

[64] Gilad Mishne, Natalie S Glance, et al. Predicting movie sales from blogger sentiment. In *AAAI Spring Symposium : Computational Approaches to Analyzing Weblogs*, pages 155–158, 2006.

[65] Michel Généreux, Thierry Poibeau, and Moshe Koppel. Sentiment analysis using automatically labelled financial news items. In *Affective Computing and Sentiment Analysis*, pages 101–114. Springer, 2011.

[66] Guang Qiu, Xiaofei He, Feng Zhang, Yuan Shi, Jiajun Bu, and Chun Chen. Dasa : dissatisfaction-oriented advertising based on sentiment analysis. *Expert Systems with Applications*, 37(9) :6182–6191, 2010.

[67] Business Insider. Google is bringing doubleclick to billboard ads for the first time which could be huge for outdoor advertising. `http://www.businessinsider.com/google-brings-doubleclick-to-ooh-billboards-2015-10?op=1`, Octobre 2015. Accessed : 03-04-2016.

[68] Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote : Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 327–335. Association for Computational Linguistics, 2006.

[69] Soumia Melzi, Amine Abdaoui, Jérôme Azé, Sandra Bringay, Pascal Poncelet, and Florence Galtier. Que ressentent les patients ? In *EGC'2014 : 14èmes Journées Francophones" Extraction et Gestion des Connaissances"*, 2014.

[70] Simon Brown. Software architecture for developers. *Coding the Architecture*, 2013.

[71] Cong Yao, Xin Zhang, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. Rotation-invariant features for multi-oriented text detection in natural images. *PloS one*, 8(8) :e70173, 2013.

[72] Martin F Porter. An algorithm for suffix stripping. *Program*, 14(3) :130–137, 1980.

[73] Michal Toman, Roman Tesar, and Karel Jezek. Influence of word normalization on text classification. *Proceedings of InSciT*, 4 :354–358, 2006.

[74] Ioannis Kanaris, Konstantinos Kanaris, Ioannis Houvardas, and Efstathios Stamatatos. Words versus character n-grams for anti-spam filtering. *International Journal on Artificial Intelligence Tools*, 16(06) :1047–1067, 2007.

[75] Stanford NLP Dan Jurafsky. Naïve bayes. `https://web.stanford.edu/class/cs124/lec/naivebayes.pdf`, —-.

[76] IEEE Spectrum ranking. Interactive : The top programming languages 2015. `http://spectrum.ieee.org/static/interactive-the-top-programming-languages-2015`, 2015. Accessed : 07-05-2016.

[77] JetBrains. Intellij idea the java ide. `https://www.jetbrains.com/idea/`, 2016. Accessed : 07-05-2016.

[78] uniVocity. univocity - etl, data integration and data synchronization for java. `http://www.univocity.com`, 2016. Accessed : 07-05-2016.

[79] uniVocity GitHub. csv-parsers-comparison : Comparaison entre tous les parseur csv pour java. `https://github.com/uniVocity/csv-parsers-comparison`, 2016. Accessed : 07-05-2016.

[80] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

[81] Oracle. Javafx - java platform, standard edition (java se) 8. `http://docs.oracle.com/javase/8/javase-clienttechnologies.htm`, 2016. Accessed : 07-05-2016.

[82] Geertjan Wielenga. Developing nasa's mission software with java. `https://jaxenter.com/netbeans/developing-nasas-mission-software-with-java`, 2014. Accessed : 07-05-2016.

[83] Felipe Bravo-Marquez, Marcelo Mendoza, and Barbara Poblete. Combining strengths, emotions and polarities for boosting twitter sentiment analysis. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 2. ACM, 2013.

[84] William Deitrick and Wei Hu. Mutually enhancing community detection and sentiment analysis on twitter networks. 2013.

[85] Kun-Lin Liu, Wu-Jun Li, and Minyi Guo. Emoticon smoothed language models for twitter sentiment analysis. In *AAAI*, 2012.