Ordre...../F.S.S.A/UAMOB/2018

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE AKLI MOHAND OULHADJ-BOUIRA

Faculté des Sciences et des Sciences Appliquées
Département d'Informatique

Mémoire de fin d'étude

Présenté par :
Mlle MERZOUK Kahina
Mlle MERZOUK Sabrina

En vue de l'obtention du diplôme de**Master 02**en **:**

Filière :INFORMATIQUE
Option : Ingénierie des systèmes d'information et du logiciel

Thème :

Détection automatique des sentiments dans les réseaux sociaux

Devant le jury composé de :

Dr.MESSAOUDI Oussama		UAMOB	Président
Dr.AMAD Mourad	MCA	UAMOB	Encadreur
Mme.ALIOUATWahiba		UAMOB	Examinateur 1
Mr.HAYIMohamedY		UAMOB	Examinateur 2

Année Universitaire 2018/2019

Remerciements

En tout premier lieu, nous remercions le bon Dieu, tout puissant, de nous avoir donné la force pour survivre, ainsi que l'audace pour dépasser toutes les difficultés.

Nous tenons à exprimer notre reconnaissance à notre encadreur monsieur *AMAD Mourad*, pour avoir accepté de nous encadrer dans cette étude. Nous le remercions pour son implication, son soutien, et ses encouragements tout au long du travail.

Un remerciement spécial à notre chère enseignante *AID Aicha*, pour son temps, sa patience et ses remarques qu'elle trouve ici le témoignage de ses conseils.

Nous souhaitons adresser nos remerciements les plus sincères au corps professoral et administratif de l'Université *Akli Mouhand Oulhadj Bouira*, pour la richesse et la qualité de leur enseignement et qui déploient de grands efforts pour assurer à leurs étudiants une formation actualisée.

Nous tenons à remercier les membres du jury d'avoir accepté très aimablement de juger ce travail. Veuillez trouver ici l'expression d'une infinie reconnaissance.

Nous adressons nos plus sincères remerciements à nos familles : nos parents, tous nos proches et amis, qui nous ont accompagné, aidé, soutenu et encouragé tout au long de la réalisation de ce mémoire.

Afin de n'oublier personne, nos vifs remerciements s'adressent à tous ceux qui nous ont aidé à la réalisation de ce modeste travail.

Dédicaces

Je dédie ce modeste travail :

À *ma grande mère maternelle* que dieu garde son âme dans son vaste paradis.

À mes très chers parents *Bouaziz et Nadia*, aucune dédicace ne saurait être assez éloquente pour exprimer la profondeur des sentiments d'affection, d'estime et de respect que je vous porte, pour l'amour dont vous m'avez toujours comblé, l'éducation et le bien être que vous m'assurez, pour votre soutien, vos sacrifices et vos prières.

À mes frères et sœurs je ne peux trouver les mots justes et sincères pour vous exprimer mon affection et mes pensées, vous êtes les ami(e)s sur qui je peux compter.

À tous les membres de ma deuxième famille qui font partie de ces personnes rares par leur gentillesse, leur tendresse et leurs grands cœurs. Qu'ils trouvent ici, le témoignage de tout mon amour et toute ma reconnaissance pour leur inlassable soutien et pour tous les merveilleux moments que nous avons passé ensemble dans un environnement familier.

À tous mes professeurs qui nous ont prodigués de conseils et de sagesse pour réussir notre parcours et consolider notre formation. Je vous remercie pour vos encouragements et votre entière disponibilité.

À *Sabrina* chère amie avant d'être binôme qui a contribué à la réalisation de ce travail, et son aimable famille je leur souhaite le bonheur de la vie.

À tous ceux qui m'aiment et que j'aime.

MERZOUK Kahina

Dédicaces

Au nom du dieu clément et miséricordieux

Je dédie ce modeste travail :

À *mon père* qui a récolté les épines pour me paver le chemin de la science et de la connaissance, que dieu prolonger sa vie, procure sa santé et du bien-être. À la personne la plus précieuse de ma vie, *ma mère*, qui a illuminé mon chemin et orné ma vie. Merci pour votre bonne éducation et vos sacrifices, j'espère être à la hauteur de vos souhaits.

À celle qui m'inspire la patience et la force, ma meilleure amie *Ziriya*, je la prends non seulement pour amie mais pour la sœur d'âme. Toujours su me faire le courage dans mes bons et durs moments. Que dieu la garde en bonne santé, l'entourer des gens géniaux et fidèles dans sa nouvelle vie.

À mon adorable frère et mes aimables deux sœurs, je leurs souhaite la réussite dans leurs vies professionnelle et quotidienne.

À toutes ma famille, mes chers grand pères, grande mères, oncles, tantes, cousins et cousines que dieu les accorder une grande vie pleine de paix et de bonheur.

Á mon binôme *Kahina* et sa belle famille, je leur souhaite d'être toujours solide et tous le bonheur de la vie.

À mes enseignants dont les conseils précieux m'ont guidée.

Á tous mes amies et collègues qui m'ont encourager et soutenue à chaque fois que j'en avais besoin.

MERZOUK Sabrina

Résumé

L'analyse des sentiments ou l'opinion mining est l'étude informatique des opinions, sentiments, attitudes et émotions exprimés dans un langage écrit. C'est l'un des domaines de recherche les plus actifs dans le traitement du langage naturel et l'extraction de texte au cours des dernières années. Ce problème a beaucoup de solutions proposées avec différentes techniques de classification de l'apprentissage machine. Nous avons utilisé l'une des technique de l'apprentissage supervisé et une autre méthode basée dictionnaire et nous avons appliqué certaines modifications afin d'avoir des meilleurs résultats.

Mots clés : analyse des sentiments, opinion mining, apprentissage machine, apprentissage supervisé, dictionnaire.

Abstract

Sentiment analysis or opinion mining is the computational study of people's opinions, sentiments, attitudes, and emotions expressed in written language. It is one of the most active research areas in natural language processing and text mining in recent years. This problem has many proposed solutions with different machine learning classification techniques. We used one of the supervised learning technique and another dictionary-based method and we applied some modifications in order to have better results.

Key words : Sentiment analysis, opinion mining, machine learning, supervised learning, dictionary.

تلخيص

تحليل المشاعر أو التنقيب عن الرأي هو الدراسة الحسابية لآراء الناس, والمشاعر, والمواقف, والعواطف المعبر عنها بلغة مكتوبة. إنها واحدة من أكثر مجالات البحث نشاطا في معالجة اللغات الطبيعية واستخراج النصوص في السنوات الأخيرة. هذه المشكلة لها العديد من الحلول المقترحة مع تقنيات تعلم الآلة المختلفة. استخدمنا أحد أساليب التلقين الذاتي وطريقة أخرى تعتمد على القاموس وطبقنا بعض التعديلات من أجل الحصول على نتائج أفضل.

الكلمات المفتاحية : تحليل المشاعر, التنقيب عن الرأي, تعلم الآلة, التلقين الذاتي, القاموس.

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

OM	**O**pinion **M**ining
TALN	**T**raitement **A**utomatique du **L**angage **N**aturel
RI	**R**echerche **I**nformation
CV	**C**urriculum **V**itæ
ML	**M**achine **L**earning
NB	**N**aïf **B**ayes
NLTK	**N**atural **L**anguage **T**oolkit
BN	**B**ayésien **N**etwork
ME	**M**aximum **E**ntropy
SVM	**S**upport **V**ector **M**achine
NN	**N**eural **N**etwork
EM	**E**spérance **M**aximisation
POS	**P**art **O**f **S**peech
RST	**R**hetorical **S**tructure **T**heory
HTML	**H**ypertext **M**ark-up **L**anguage
BOM	**B**yte **O**rder **M**ark
URL	**U**niform **R**esource **L**ocator
HTTP	**H**yper**T**ext **T**ransfer **P**rotocol
WWW	**W**orld **W**ide **W**eb
XML	**E**xtensible **M**ark-up **L**anguage

Introduction générale

Dans ces dernières années, les plateformes des réseaux sociaux sont devenues de plus en plus populaires, considérées comme un moyen de communication entre plusieurs humains où ils partagent leurs pensées et leurs idées sur des sujets ou des problèmes. Ce qui donne beaucoup d'informations diffusées au tour de monde. Ces informations peuvent devenir des données importantes pour nous, elles peuvent s'agir des avis sur une réservation d'un ticket d'avion, lire des commentaires sur un candidat pour les élections, ou sur des nouvelles technologies qui nous intéressent, ..etc. Par conséquent, la circulation des informations sur le web hérite l'augmentation de la quantité de données volumineuses. Le problème qui se pose ici n'est pas de trouver ces données mais plutôt de les classer en considération sociologique et significative, ce qui à pousser à la naissance d'un nouveau domaine.

L'analyse des sentiments *(en anglais Sentiment Analysis)* ou l'opinion mining *(OM)*, appelé aussi analyse de subjectivité *(en anglais Subjectivity Analysis)*, est une discipline réunit plusieurs domaines : traitement automatique des langages naturels *(TALN)*, la recherche d'information *(RI)*, et la linguistique. Elle a comme différentes tâches : l'extraction d'opinions, l'analyse d'émotions, les avis mining, ...etc. Ces analyses consistent à rechercher des données textuelles à caractère évaluatif sur Internet, telles que les tweets, qui nous intéresse particulièrement dans ce travail.

Dans ce projet, l'objectif principal est l'exploration de domaine d'analyse de sentiment pour extraire des opinions à partir des tweets en langue anglaise et de les classifier selon leurs orientations sémantiques en trois catégories principales : positive, négative et neutre. Donner une vue sur les différentes techniques, approches utilisées et proposées avec leurs résultats. Nous avons un but de présenter l'approche basée sur un dictionnaire et de développer notre propre modèle en tant qu'assistant à la problématique d'analyse de

sentiment.

Le manuscrit de notre travail est structuré comme suit :

— **Chapitre 1** présente des généralités sur les réseaux sociaux qui introduit les caractéristiques principales et typologies des réseaux sociaux, leurs avantages et inconvénients, ainsi leurs problèmes courant qui influence sur le coté personnel et social.

— **Chapitre 2** est spécifié pour le domaine d'analyse des sentiments et opinion mining. Nous présenterons l'analyse des sentiments, les disciplines en relation avec l'analyse de sentiments, leurs méthodes de classification et ses outils permettant de réaliser l'analyse et les travaux connexes réalisés sur l'analyse des sentiments dans chacune des classes *(approche basée lexique, basée apprentissage automatique, basée hybride)*.

— **Chapitre 3** est consacré à la conception, afin d'expliquer les étapes nécessaires et les démarches détaillées pour entraîner et tester le classificateur naïve bayes et aussi notre modèle proposé sous forme de plusieurs améliorations.

— **Chapitre 4** représente notre réalisation avec une brève présentation des outils de programmation utilisés, l'implémentation, l'évaluation de classificateur naïve bayes, la méthode basée sur dictionnaire et l'utilisation d'un outil d'analyse des sentiments "Sentiment Analyzer" pour finir avec les résultats d'exécution.

Finalement, nous clôturons ce travail par un rappel résumé sur les points importants abordés sur ce mémoire avec la présentation de quelques perspectives.

Chapitre 1

Généralités sur les réseaux sociaux

1.1 Introduction

De nos jours, les réseaux sociaux sur Internet apparaissent comme un nouveau moyen important de communication. Internet a consacré la montée en puissance des réseaux sociaux qui permettent aux internautes et aux professionnels de créer une page profil et de partager des informations, photos et vidéos avec leurs réseaux. Des espaces de partage qui se distinguent par leurs utilités (*personnel, professionnel, rencontres...*), leurs logos et leurs audiences. Le réseau le plus connu est évidemment Facebook [1].
La communication est évidemment un élément central des réseaux sociaux qui proposent tous les outils de communication synchrones (*ex : chat ou vidéoconférence*) et asynchrones (*ex : commentaires, forum*). Ce chapitre est consacré pour présenter les principaux réseaux sociaux connus, leurs caractéristiques, en mettant l'accent sur leurs impacts et leurs problèmes courants.

1.2 Réseaux sociaux

La définition des réseaux sociaux ne se résume pas à une seule précise, dans ce qui suit nous présentons quelques unes :

1. Selon *le grand dictionnaire terminologique* [2], le réseau social est une communauté d'internautes reliés entre eux par des liens amicaux ou professionnels, regroupés

1. https ://www.facebook.com
2. http ://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx ?Id_Fiche=26503112

ou non par secteurs d'activité, qui favorise l'interaction sociale, la création et le partage d'informations.

2. Selon *Wikipédia*[3], l'expression réseau social désigne un agencement de liens entre des individus et/ou des organisations, constituant un groupement qui a un sens : famille, collègues, groupe d'amis, communauté, ...etc.

3. Un réseau en général, est un terme polysémique [1] qui représente des nœuds liés entre eux par des liens. Un réseau social en particulier, est un ensemble d'acteurs *(individus, groupes ou organisations)* reliés par des interactions sociales de différentes natures *(familiales, sentimentales, relation d'affaire, de travail)* voir la FIGURE1.1.

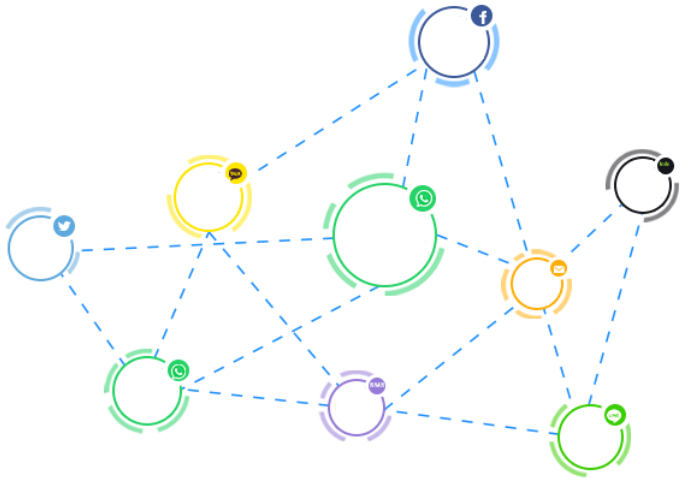


FIGURE 1.1 – Réseau sous forme de nœuds.

1.3 Typologie des réseaux sociaux

Les deux auteurs **Christophe Dubois** et **Catherine Chatet** [2] proposent de classer les réseaux sociaux en 5 *(voir la* FIGURE1.2*)* types qui sont illustrés comme suit :

1. **Réseaux sociaux de masse :** réseaux de personnes connectés par des systèmes d'amis, de fans. Exemple : Facebook, MySpace[4].

2. **Agrégateurs sociaux :** sites dont les contenus importants sont choisis par la

3. https ://fr.wikipedia.org/wiki/R%C3%A9seau_social
4. https ://www.myspace.com

communauté. Exemple : Wikio [5], Reddit [6].

3. **Marque page social :** sites qui stockent, organisent, identifient, gèrent et cherchent les marques pages/favoris/signets. Exemple : Delicious [7], Diigo [8].
4. **Media sociaux et partage de contenus :** sites qui permettent la publication de contenus générés (*ex : vidéos, photos, etc*) par les utilisateurs. Exemple : Youtube [9], Slideshare [10].

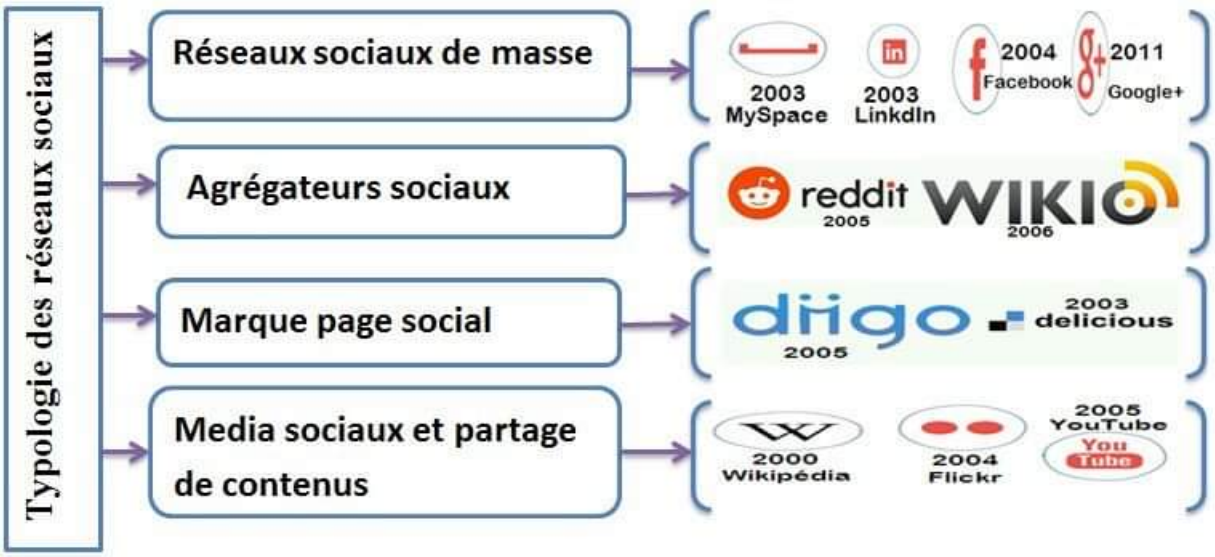


FIGURE 1.2 – Typologie des réseaux sociaux.

1.4 Caractéristiques des réseaux sociaux

Les réseaux sociaux sont généralement divisés en deux catégories [3] : professionnel (*ex : Viadeo* [11], *Linkedin* [12]) ou privé (*ex : Facebook, Twitter* [13]) mais ils partagent les mêmes caractéristiques qui se présentent comme suit :

5. http ://www-wikio.over-blog.com/
6. https ://www.reddit.com/
7. https ://del.icio.us
8. https ://www.diigo.com/
9. https ://www.youtube.com/
10. https ://www.slideshare.net/
11. http ://dz.viadeo.com/en/
12. https ://www.linkedin.com/
13. https ://twitter.com/

— Multiplicité des plateformes, généralistes ou spécialisées (*multiplication des applications*).
— Plateformes adaptées et appropriées aux propres usages des utilisateurs.
— Les utilisateurs sont liés de façon bilatérale ou via des groupes : profils individuels, constitution de communautés, interaction avec le cercle des relations.
— Principe de cooptation et de recommandation.
— Gratuité et ouverture *(pour la plupart des plateformes).*

1.5 Principaux réseaux sociaux

On distingue les réseaux sociaux grands publics et les réseaux sociaux professionnels.

1.5.1 Exemple de réseaux sociaux grands publics

1. **Facebook :** c'est un site web de réseau social gratuit et populaire [4] qui permet aux utilisateurs enregistrés de créer des profils, de télécharger des photos et des vidéos, d'envoyer des messages et de rester en contact avec leurs amis, leurs familles et leurs collègues. Le site, disponible en 37 langues différentes, permet de réagir sur les commentaires et news postés par ses amis via le "Like" ou "J'aime". C'est un moyen pour dire que l'on a trouvé un commentaire ou un post à son goût.
2. **Twitter :** c'est un site web de réseau social [4], qui permet aux utilisateurs de publier des messages courts qui sont visibles par les autres utilisateurs. Ces messages sont appelés tweets et peuvent contenir 280 caractères. Les utilisateurs ont trouvé de nombreuses utilisations différentes de Twitter, notamment la communication de base entre amis et la famille, un moyen de faire connaître un événement ou un outil de relation client permettant aux entreprises de communiquer avec leurs clients.
3. **Google+** [14] **:** est un réseau social qui fonctionne sur le principe de ≪cercles≫ de contacts, permettant de choisir facilement avec quel ≪cercle≫ de contacts on souhaite partager le contenus (*amis, famille, collègues, clients, prospects,... etc.*) [5]. Les cercles permettent de segmenter les contacts et donc d'adapter les messages en fonction de la cible. Ce réseau social a fermé ses fonctionnalités *avril 2019*[15].

14. https ://accounts.google.com
15. https ://support.google.com/plus/answer/9195133 ?hl=fr

4. **Tumblr** [16] **:** est une plateforme permettant de publier des textes, citations, liens, photos, sons et vidéos de manière ultra-simple sans passer par la création fastidieuse d'un blog [5].

1.5.2 Exemple de réseaux sociaux professionnels

1. **LinkedIn :** c'est un réseau professionnel international [6] permet de partager des informations relatives au travail avec d'autres utilisateurs et de conserver une liste en ligne de contacts professionnels. Toutefois, les profils créés dans LinkedIn sont davantage axés sur les entreprises que sur les particuliers. Par exemple, un profil LinkedIn met en évidence les études et l'expérience professionnelle antérieure, ce qui le fait ressembler à un CV *(curriculum vitæ)*. Les profils répertorient également les connexions avec d'autres utilisateurs de LinkedIn, ainsi que les recommandations faites ou reçus de la part des autres utilisateurs.
2. **Viadeo :** c'est un réseau social professionnel Web 2.0 dont les membres comprennent des propriétaires d'entreprise, des entrepreneurs et des gestionnaires. En 2014, le site comptait 65 millions de membres [6].
3. **Wizbii** [17] **:** est une plateforme professionnelle pour l'emploi et l'entrepreneuriat [7]. Elle est entièrement dédiée aux étudiants et aux jeunes diplômés. En 2017, Wizbii a permis à environ 40000 jeunes de trouver un emploi.

1.6 Avantages et inconvénients des réseaux sociaux

Les réseaux sociaux présentent des avantages et des inconvénients dont on doit tenir compte lors de leur utilisation.
Nous présentons les avantages et les inconvénients [8] incontournables des réseaux sociaux afin de savoir comment les utiliser d'une manière plus sûre et plus précieuse possible :

16. https ://www.tumblr.com/
17. https ://www.wizbii.com/

1. **Avantages**
   - *Connectivité mondiale :* on peut se connecter avec n'importe qui, n'importe où dans le monde. Peu importe qu'on l'utilise pour des raisons professionnelles, de divertissement, de romance, de conseil, de recherche d'emploi, de clubs religieux / scolaires ou pour apprendre, c'est facile et rapide.
   - *Sécurité :* se connecter avec tous les amis du monde entier dans le confort du foyer est plus sûr et plus réconfortant de savoir que personne ne peut faire de mal, car on se connecte numériquement au lieu de devoir assister physiquement à des réunions, on se réunissent virtuellement. On peut rencontrer nos amis chaque fois qu'on a une connexion Internet et chaque fois qu'on les trouve en ligne.
   - *Publicité et marketing gratuits :* les réseaux sociaux pourraient être facilement utilisés pour commercialiser des produits ou des services, ainsi que des stratégies et des campagnes rentables pouvant générer des résultats viraux.
   - *Partage d'informations :* c'est tellement efficace pour quiconque de partager des nouvelles, des informations utiles, des potins, ce qui se passe, surtout pour ceux qui étudient, cela peut être la plate-forme pour partager des copies d'examens antérieurs, entraider pour les travaux à domicile et préparer ensemble aux tests et des examens.
2. **Inconvénients**
   - *Contrecoup :* une blague entre amis est une chose, mais une blague avec le monde en général est bien différente. Lorsqu'un contenu potentiellement offensant est mis en ligne, le nombre de réactions peut être excessif et souvent brutal. Cela est particulièrement vrai avec des sujets très controversés comme la politique et la religion.
   - *Cyberintimidation et crimes contre les enfants :* l'utilisation des réseaux sociaux peut exposer les individus à d'autres formes de harcèlement ou même à des contacts inappropriés. Cela peut être particulièrement vrai pour les adolescents et les jeunes enfants. À moins que les parents ne filtrent avec diligence le contenu web de leurs points de vue familiaux, les enfants pourraient être exposés à la pornographie ou à tout autre contenu inapproprié.
   - *Invasion de la vie privée par les entreprises :* les réseaux sociaux invitent les

grandes entreprises à envahir la vie privée et à vendre les informations personnelles.

— *Perte de temps :* afin d'obtenir le plein effet du réseau social, on doit comprendre son fonctionnement, quand et comment l'utiliser, et les canaux sur lesquels se concentrer, en fonction de l'objectif final d'utilisation des médias sociaux.

Les réseaux sociaux possèdent des avantages mais aussi des inconvénients, c'est pour cela qu'il faut prendre conscience des risques lors de leurs utilisation, malgré leurs nombreux dangers, ils restent aujourd'hui un moyen indispensable pour faciliter le quotidien de leurs utilisateurs.

1.7 Problèmes courants dans les réseaux sociaux

Le développement des réseaux sociaux a notamment un impact sur le respect de la vie privée, qui découle du fait que tout y est partagé, à savoir aussi bien les activités des personnes inscrites que leurs *"heures de présence"* sur le réseau, notamment à travers l'affichage de l'heure de la dernière connexion ou encore par le biais du logo *"Vu"* affiché par la messagerie instantanée. Mais ces nouveaux outils de communication génèrent autant de risques pour les personnes morales (*ex : entreprises, établissements publics, associations, etc.*) qu'il appartient aux directeurs et responsables des systèmes d'information de gérer cette nouvelle situation résultante de l'émergence des réseaux sociaux et leur insertion aussi bien dans la sphère privée que professionnelle risque de poser de graves problèmes juridiques. Il s'agit notamment pour les utilisateurs de savoir comment protéger leur vie privée et leur liberté d'expression.
Dans ce qui suit nous présentons quelques problèmes courants [9] au niveau des réseaux sociaux dans la vie privée et professionnelle :

1.7.1 Droit au respect de la vie privée *(anonymat)*

— *L'accès à des données personnelles :* les informations transmises à l'occasion de l'inscription à un réseau social sont stockées dans la base de données de ces fournisseurs de service. Le problème posé, les informations transmises sont toujours accessibles par d'autres acteurs intervenant sur le web. En outre, on remarque l'absence fréquente de traduction en français des conditions d'utilisation de ces réseaux,

ne permettant pas aux internautes et notamment aux plus jeunes d'appréhender réellement les conséquences de la communication de données personnelles sur ces sites. Il existe des sociétés spécialisées dans le traitement des informations données sur les réseaux sociaux. L'activité de ces sociétés consiste à collecter les informations recueillies sur les internautes pour permettre aux publicitaires d'opérer un ciblage précis des consommateurs. Les moyens mis en place pour recueillir ces informations peuvent être les cookies, scripts et spywares.

— *Une frontière floue entre publication privée et publique sur Internet :* l'hypothèse où publier une photographie privée sur Internet peut poser certains problèmes juridiques, pour diffuser une photo sur Internet on doit obtenir une autorisation préalable de la part de la personne concernée. Ainsi même si la photo a un caractère public par sa visibilité, elle conserve un statut de publication privée et donc elle donne droit au respect de la vie privée.

1.7.2 Présence des réseaux sociaux sur le lieu de travail

— *Une redéfinition des contours de la liberté d'expression du salarié :* le propre des réseaux sociaux est de permettre aux internautes de s'exprimer et communiquer avec une foule d'autres personnes. Or cette possibilité peut mener à des dérives lorsque les informations échangées par des salariés concernent les produits ou services d'une entreprise, la société elle-même ou encore ses dirigeants. Les risques résultants de ces outils par des salariés sont l'atteinte à la réputation de l'entreprise et ses marques, à celle de ses dirigeants, la diffusion de fausses rumeurs,... etc. L'enjeu de l'utilisation des réseaux sociaux est ici primordial, car on peut alors être condamné à verser des dommages et intérêts pour réparer le préjudice causé à l'entreprise par laquelle on est employés.

— *L'utilisation des réseaux sociaux :* quoi qu'il en soit, l'utilisation des réseaux sociaux implique une grande vigilance à la fois au regard de la vie privée mais aussi de la vie professionnelle. Les conséquences pouvant en effet être très fâcheuses comparées à un bénéfice plutôt négligeable.

1.8 Conclusion

Dans ce chapitre, nous avons abordés les notions de base des réseaux sociaux ainsi que leurs effets et difficultés essentielles.
Les avis et les points de vue des utilisateurs est un moyen pour mesurer l'importance de tel service vis-à-vis un autre.
Pour évaluer l'importance d'un tel service on utilise les avis et les points de vue des internautes. L'objectif de l'analyse des sentiments et l'opinion mining est collecter et analyser ces opinions, cette tâche sera détaillée dans le prochain chapitre.

Chapitre 2

État de l'art sur l'analyse des sentiments

2.1 Introduction

En informatique, sentiment analysis aussi appelé *opinion mining* est l'analyse des sentiments à partir de sources textuelles dématérialisées sur de grandes quantités de données. Cette analyse est utilisée pour mieux comprendre la perception, les opinions et les émotions exprimées dans une mention en ligne. Donc, c'est une tâche de traitement automatique des langues et d'extraction d'information pour les déterminer généralement sur trois niveaux *(positive, neutre et négative)*.

2.2 Définition

2.2.1 Sentiment

Le sentiment est la composante de l'émotion qui implique les fonctions cognitives de l'organisme, la manière d'apprécier [10]. Le sentiment est à l'origine d'une connaissance immédiate ou d'une simple impression. Il renvoie à la perception de l'état physiologique du moment. Le sens psychologique de sentiment qui comprend un état affectif est à distinguer du sens propre de la sensibilité.
Le dictionnaire Larousse[1] définit le sentiment comme étant un état affectif complexe et durable lié à certaines émotions ou représentations.

1. https ://www.larousse.fr/dictionnaires/francais/sentiment/72138 ?q=sentiment#71335

2.2.2 Analyse des sentiments

L'analyse des sentiments *(également appelée exploration d'opinion)* fait référence à l'application du traitement du langage naturel, de la linguistique informatique et de l'analyse de texte pour identifier et classer les opinions subjectives dans des sources [11]. Les données analysées quantifient les sentiments ou réactions du grand public envers certains produits, personnes ou idées et révèlent la polarité contextuelle de l'information. En général, elle a été étudié principalement à trois niveaux : niveau de document qui détermine l'opinion général du l'ensemble de document, niveau de la phrase qui consiste à classer chaque phrase selon l'opinion qu'elle s'exprime et le niveau d'aspect *(niveau de caractéristiques)* effectue une analyse plus poussée et de meilleure qualité.

2.3 Besoin de connaître les opinions des autres

Connaitre l'opinion des autres personnes a toujours été un élément d'information important durant le processus de décision [12]. Les gens souvent demandent à d'autres de leurs recommander un mécanicien d'automobiles ou d'expliquer leurs choix de votes aux élections par exemple.
Avant de prendre des décisions, les gens s'intéressent énormément aux avis des autres personnes dans différents domaines. Ils consultent les avis des autres consommateurs avant d'effectuer un achat, ou avant de voir un film au cinéma.
Aujourd'hui, plusieurs personnes donnent leurs avis sur différents sujets, ces avis sont à la disposition de tout le monde sur Internet.

2.4 Disciplines en relation avec l'analyse des sentiments

L'analyse des sentiments a une relation avec plusieurs disciplines à savoir :

2.4.1 Fouille de texte

La fouille de texte ou text mining est l'ensemble des méthodes scientifiques destinées à l'exploration et l'analyse de grande quantité de données informatiques [13] en vue de

détecter des profils-type, des comportements récurrents, des règles, des liens, des tendances inconnues, des structures particulières restituant de façon concise l'essentiel de l'information qu'elle est utile pour l'aide à la décision.
Cette technique permet d'extraire depuis les réseaux sociaux les données puis les analyser et les classer. Elle permet d'accomplir les quatre types d'analyse suivant : classification, estimation, segmentation et prévision. Ces types d'analyse se répartissent dans deux catégories descriptives et prédictives :

— Techniques descriptives (*ex : classification*) : ces techniques essayent de mettre en évidence des informations présentes mais cachées par le volume des données.
— Techniques prédictives (*ex : estimation, segmentation et prévision*) : consiste à estimer une valeur future d'un champ à partir des données réelles possédées. Cette technique est très utilisée dans le domaine d'intelligence artificielle, spécialement dans les algorithmes machine learning.

2.4.2 Traitement automatique du langage naturel (TALN)

Le traitement automatique du langage naturel est le domaine de l'intelligence artificielle qui concerne le traitement et la compréhension du langage humain [14]. Depuis sa création dans les années 50, la compréhension du langage par la machine a joué un rôle essentiel dans la traduction, la modélisation de sujets, l'indexation de documents, la récupération d'informations et l'extraction. De nos jours, il est utilisé pour alimenter les moteurs de recherche, filtrer le spam et obtenir des analyses de manière rapide et évolutive. Les chercheurs peuvent même se vanter d'avoir atteint la perfection au niveau humain dans un bon nombre de ces tâches, la plus importante étant la traduction automatique. L'objectif ultime de TALN est de lire, déchiffrer, comprendre et donner un sens aux langages humains d'une manière précieuse. La plupart des techniques de TALN reposent sur l'apprentissage automatique pour dériver le sens des langages humains.

2.4.3 Apprentissage automatique (*en anglais Machine Learning ML*)

L'apprentissage automatique [15], branche de l'intelligence artificielle, concerne la construction et l'étude de systèmes pouvant apprendre à partir de données.

Un algorithme qui implémente la classification, en particulier dans une implémentation concrète, est appelé classificateur. Le terme classificateur fait aussi parfois référence à la fonction mathématique, mise en œuvre par un algorithme de classification, qui trie des données inutiles à une catégorie. En l'entrainant, cela signifie les former sur des intrants particuliers afin de pouvoir les tester pour des intrants connus pour lesquels ils peuvent classer ou prédire en fonction de leur apprentissage. La classification des données est une tâche courante dans l'apprentissage automatique.

2.4.4 Deep learning

Deep learning *(ou apprentissage profond)* fait partie d'une famille de méthodes d'apprentissage automatique fondées sur l'apprentissage de modèles de données, c'est un ensemble varié d'algorithmes qui tente d'imiter le fonctionnement du cerveau humain en utilisant des réseaux de neurones artificiels pour traiter les données. Ces algorithmes ont récemment contribué à faire progresser les performances des systèmes d'analyse des sentiments, qu'il s'agisse d'analyser des textes écrits [16].

2.5 Techniques de classification des sentiments

Les techniques de classification des sentiments [17] peuvent être grossièrement divisées en une approche basée apprentissage automatique, une approche basée lexique et une approche hybride *(voir la* FIGURE2.1*)*.

— **L'approche basée apprentissage automatique :** applique les célèbres algorithmes ML et utilise des fonctionnalités linguistiques.

— **L'approche basée lexique :** s'appuie sur un lexique de sentiments, un ensemble de termes de sentiments connus et précompilés.

— **L'approche hybride :** combine les deux approches et est très courante, les lexiques de sentiment jouant un rôle clé dans la majorité des méthodes.

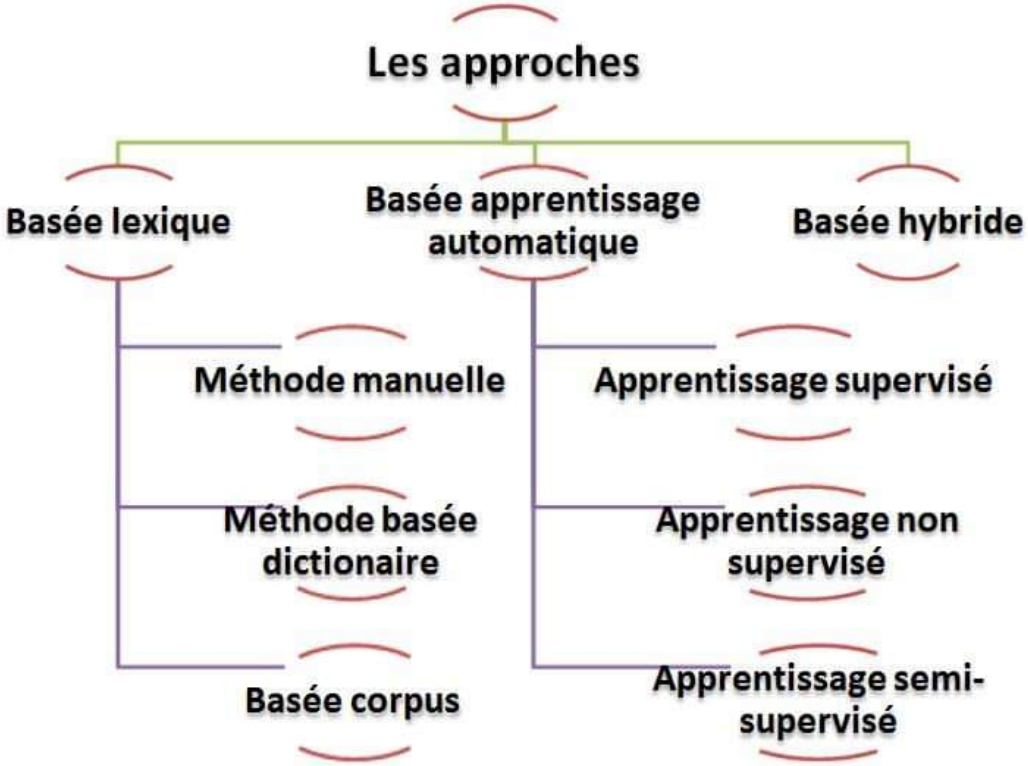


FIGURE 2.1 – Méthodes de classification des sentiments.

2.5.1 Approche basée apprentissage automatique

L'apprentissage automatique est une tentative de comprendre et reproduire la faculté de l'apprentissage humain dans des systèmes artificiels [17]. Il s'agit de concevoir des algorithmes capables, à partir d'un nombre important d'exemples, d'en assimiler la nature afin de pouvoir appliquer ce qu'ils ont ainsi appris aux cas futurs. Ainsi, le but essentiel de l'apprentissage automatique est de déterminer la relation entre les objets et leurs catégories pour la prédiction et la découverte des connaissances.
On distingue ainsi trois types d'apprentissage : l'apprentissage supervisé, non supervisé et semi-supervisé.

Apprentissage supervisé

L'apprentissage supervisé *(ou classification)* [17] consiste à construire un modèle basé sur un jeu d'apprentissage et des labels *(nom des catégories ou des classes)* et à l'utiliser pour classer des données nouvelles. Il existe de nombreux types de classificateurs supervisés. Dans ce qui suit, nous présentons brièvement certains des classificateurs les plus

fréquemment utilisés dans l'analyse des sentiments.

1. **Classificateurs probabilistes :** utilisent des modèles de mélange que chacun suppose que chaque classe est un composant du mélange [17]. Chaque composant est un modèle génératif qui fournit la probabilité d'échantillonnage d'un terme particulier pour ce composant. Ces types de classificateurs sont également appelés classificateurs génératifs. Trois des plus célèbres classificateurs probabilistes sont discutés dans ce qui suit.
   — *Classificateur Naïve Bayes (NB) :* le classificateur NB est l'une des techniques de classification supervisée permettant de classer un texte *(phrase)* appartenant à une classe donnée. C'est un algorithme probabiliste qui calcule la probabilité de chaque mot du texte *(phrase)*, le mot ayant la probabilité la plus élevée est considéré comme une sortie.

   En Juin 2019, *Kavya Suppala* et *Narasinga Rao* [18] ont développé un modèle pour l'analyse de sentiments sur des données twitter à l'aide de technique machine learning, ce modèle a été construit à l'aide de la boite à outils en langage naturel *(Natural Language Toolkit NLTK)* sur l'ensemble de données contenant des tweets. Son concept est basé sur le sac de mot qui contient des mots positifs et des mots négatifs séparément.

   La classification de sentiments a été faite à l'aide du classificateur naïve bayes en calculant la probabilité de nouvelles données d'entrée où le tweet avec la valeur la plus élevée est considéré comme positif ou négatif. Ils ont choisi le jeu de donnée d'entrée twitter pour améliorer l'efficacité et la précision du classificateur.

   Ce classificateur est introduit par un algorithme, dans ce qui suit, nous indiquons ses étapes.

Algorithm 1 Classificateur NB

1: Considérer un ensemble de données d'apprentissage D constitué de documents appartenant aux différentes classes, telles que les classes A et B.

2: Calculer la probabilité préalable des classes A et B comme suit :
classe A = nombre d'objets de la classe A / nombre total d'objets.
classe B = nombre d'objets de la classe B / nombre total d'objets.

3: Calculer le nombre total de fréquences de mots des classes A et B, c'est-à-dire ni :
na = nombre total de fréquence de mots de classe A.
nb = nombre total de mots de fréquence de la classe B.

4: Calculer la probabilité conditionnelle d'occurrence de mot-clé pour une classe donnée :
$P(mot1|classeA)$= nombre de mots / ni (A), $P(mot1|classeB)$ = nombre de mots / ni (B).
$P(mot2|classeA)$ = nombre de mots / ni (A), $P(mot2|classeB)$ = nombre de mots / ni (B).
...
$P(motn|classeA)$ = nombre de mots / ni (A), $P(motn|classeB)$ = nombre de mots / ni (B).

5: Effectuer des distributions uniformes afin d'éviter le problème de fréquence nulle.

6: Classer un nouveau document M sur la base du calcul de la probabilité pour les classes A et B, $P(M|phrase)$:
$P(A|phrase)$ = P(A) * $P(mot1|classeA)$ * $P(mot2|classeA)$..... * $P(motn|classeA)$.
$P(B|phrase)$ = P(B) * $P(mot1|classeB)$ * $P(mot2|classeB)$..... * $P(motn|classeB)$.

7: Attribuer la classe ayant la probabilité la plus élevée au nouveau document M, après le calcul de la probabilité pour les classes A et B.

— *Réseau Bayésien (Bayesian Network BN) :* est un graphe acyclique dirigé dont les nœuds représentent des variables aléatoires et les arêtes les dépendances conditionnelles [17]. BN est considéré comme un modèle complet pour les variables et leurs relations. Par conséquent, une distribution de probabilité conjointe complète sur toutes les variables est spécifiée pour un modèle. Dans le text mining, la complexité de calcul du BN est très coûteuse, c'est pourquoi,

il n'est pas fréquemment utilisé.

— *Classificateur d'entropie maximale (Maximum Entropy ME) :* le classificateur Maxent *(connu sous le nom de classificateur exponentiel conditionnel)* [17] convertit les ensembles de caractéristiques étiquetés en vecteurs à l'aide du codage. Ce vecteur codé est ensuite utilisé pour calculer des pondérations pour chaque entité, qui peuvent ensuite être combinées pour déterminer l'étiquette la plus probable pour un ensemble d'entités. Ce classificateur est paramétré par un ensemble de X poids, qui permet de combiner les entités jointes générées à partir d'un ensemble de fonctionnalités par un X encodage. En particulier, l'encodage mappe chaque paire C *(fonctionnalités, étiquette)* sur un vecteur.

2. **Classificateurs linéaires :** le terme de classificateur linéaire [19] représente une famille d'algorithmes de classement statistique. Son rôle est de classer dans des classes les échantillons qui ont des propriétés similaires, mesurées sur des observations. Ce classificateur calcule la décision par combinaison linéaire des échantillons ce qui le rend particulier.

   — *Classificateurs de machines à vecteurs de support (Support Vector Machine SVM) :* son principe de base est déterminer dans l'espace de recherche des séparateurs linéaires capables de séparer au mieux les différentes classes [17]. Sur la FIGURE 2.2, il existe 2 classes +, o et 3 hyperplans A, B et C. L'hyperplan A fournit la meilleure séparation entre les classes, car la distance normale de n'importe quel point de données est la plus grande marge maximale de séparation. Les données de texte conviennent parfaitement à la classification SVM en raison de la nature dispersée du texte, dans lequel peu de caractéristiques sont sans importance, mais elles tendent à être corrélées les unes aux autres et généralement organisées en catégories pouvant être séparées linéairement.

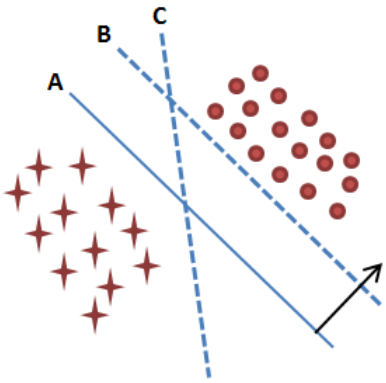


FIGURE 2.2 – Exemple d'utilisation de l'algorithme SVM.

— *Réseaux de neurones (Neural Network NN) :* les réseaux de neurones [20] *(voir la* FIGURE2.3*)* sont généralement optimisés par des méthodes d'apprentissage de type probabiliste, en particulier bayésien. Ils sont placés d'une part dans la famille des applications statistiques, qu'ils enrichissent avec un ensemble de paradigmes permettant de créer des classifications rapides, et d'autre part dans la famille des méthodes de l'intelligence artificielle auxquelles ils fournissent un mécanisme perceptif indépendant des idées propres de l'implémenteur, et fournissant des informations d'entrée au raisonnement logique formel.

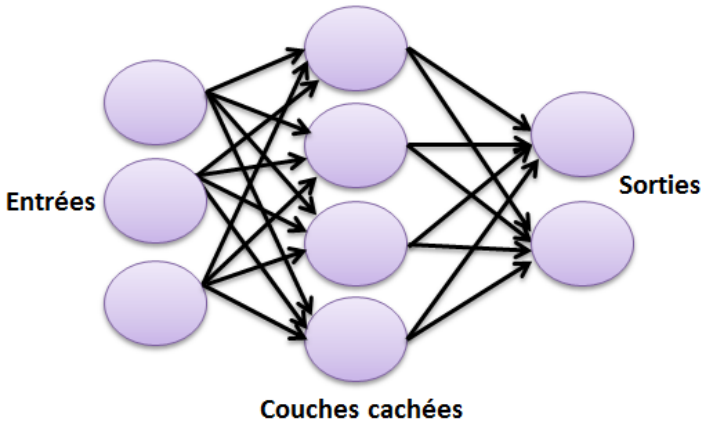


FIGURE 2.3 – Structure d'un réseau de neurone.

3. **Classificateur d'arbre de décision :** le classificateur d'arbre de décision [17] *(voir la* FIGURE2.4*)* fournit une décomposition hiérarchique de l'espace de données d'apprentissage dans lequel une condition sur la valeur d'attribut est utilisée pour

diviser les données. La condition *ou prédicat* est la présence/l'absence d'un ou plusieurs mots. La division de l'espace de données est effectuée de manière récursive jusqu'à ce que les nœuds feuilles contiennent un nombre minimum d'enregistrements utilisés aux fins de la classification.

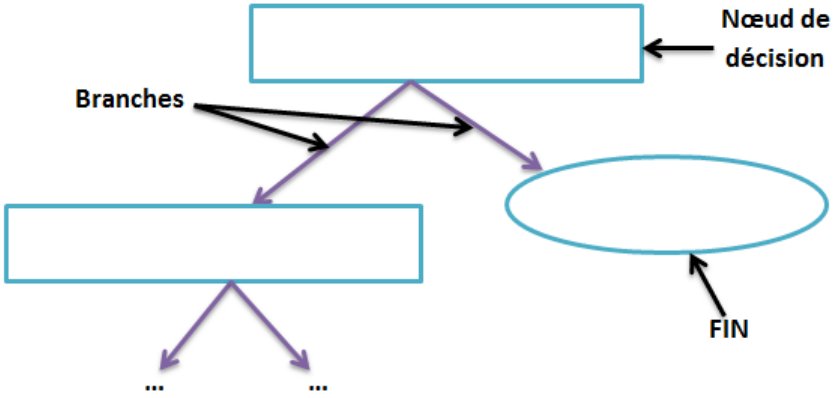


FIGURE 2.4 – Exemple d'arbre de décision.

4. **Classificateurs basés sur des règles :** dans ces classificateurs l'espace de données est modélisé avec un ensemble de règles [17]. Le côté gauche représente une condition sur le jeu de caractéristiques exprimé sous forme normale disjonctive, tandis que le côté droite est le libellé de la classe. Les conditions sont sur le terme présence. L'absence temporaire est rarement utilisée, car elle n'est pas informative dans les données rares.

5. **K-NN *(K-Nearest Neighbors)* :** l'algorithme K-NN qui signifie k-voisins les plus proches utilise l'intégralité du dataset en tant qu'entraînement, au lieu de diviser ce dernier en un *training* et *testing set* [21].

   Quand un résultat est requis pour une nouvelle instance de données, l'algorithme K-NN parcourt l'intégralité du dataset pour rechercher les k-instances les plus proches de la nouvelle instance ou le nombre k d'instances les plus similaires au nouvel enregistrement, puis renvoie la moyenne des résultats ou la classe à laquelle appartienne cette instance si c'est un problème de classification l'utilisateur spécifie lui même la valeur de k *(voir la* FIGURE2.5*)*.

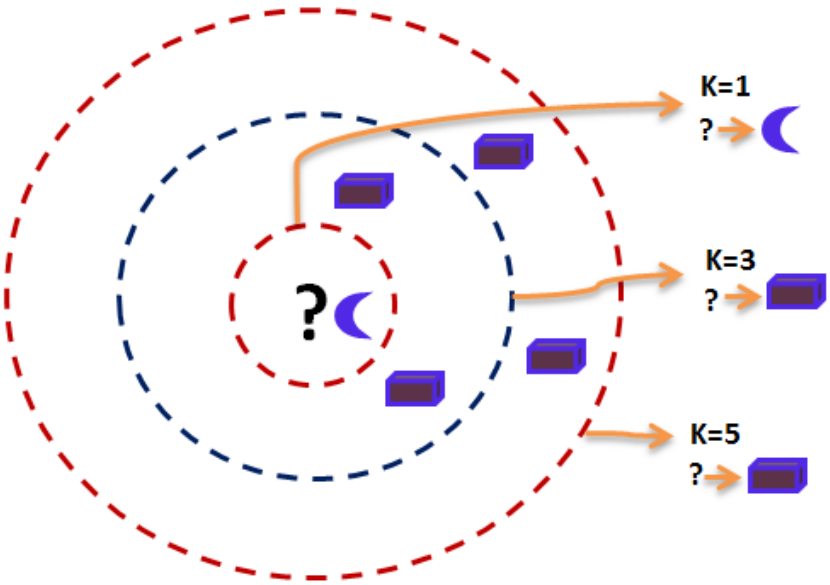


FIGURE 2.5 – Exemple d'utilisation de l'algorithme K-NN.

Apprentissage non supervisé

L'apprentissage non supervisé *(en anglais clustering)* [17] vise à construire des groupes *(clusters)* d'objets similaires à partir d'un ensemble hétérogène d'objets. Chaque cluster issu de ce processus doit vérifier les deux propriétés suivantes :

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

Le processus de «clustering» repose sur une mesure précise de similarité des objets qu'on veut regrouper. Cette mesure est appelée distance *(ou métrique)*. Il est utilisé dans plusieurs applications telles que le traitement d'images, les études démographiques, la recherche génétique, le forage des données et l'analyse des opinions. On distingue plusieurs algorithmes de clustering :

- *K-Moyennes (K-Means) :* un algorithme de partitionnement des données en K groupes ou clusters [17]. Chaque objet sera associé à un seul cluster. Le K est fixé par l'utilisateur.
- *Fuzzy K-Means :* il s'agit d'une variante du précédent algorithme [17] proposant qu'un objet ne soit pas associé qu'à un seul groupe.

— *Espérance Maximisation (EM) :* cet algorithme utilise des probabilités [17] pour décrire qu'un objet appartient à un groupe. Le centre du groupe et ensuite recalculé par rapport à la moyenne des probabilités de chaque objet du groupe.

— *Regroupement hiérarchique :* deux sous-algorithmes en découlent [17] : le «bottom up» qui a pour fonction d'agglomérer des groupes similaires, donc en réduire le nombre *(les rendre plus lisibles)* et d'en proposer un ordre hiérarchique et le «top down» qui fait le raisonnement inverse en divisant le premier groupe récursivement en sous-ensembles.

Travaux connexes effectués sur l'analyse de sentiments basée sur l'apprentissage automatique (non supervisé)

— En 2010, **Moghaddam** et **Ester** [22] proposent une solution nommée opinion digger qui est une méthode d'apprentissage automatique non supervisée avec une particularité d'extrait des aspects importants d'un produit et détermine la satisfaction globale du consommateur pour chacun d'entre eux, en estimant une évaluation de 1 à 5. Il détermine également en sortie un ensemble d'autres aspects et les notations de chaque aspect conformément à la ligne directrice.
Opinion digger fonctionne en deux étapes. Tout d'abord, il détermine l'ensemble des aspects. Après le pré-traitement, chaque phrase est étiquetée avec POS *(Part Of Speech)*. Il suppose que les aspects sont des noms, il isole d'abord les noms fréquents en tant qu'aspects potentiels, avec les phrases correspondantes aux aspects connus, ils déterminent les modèles d'opinion, séquence de balises POS qui expriment une opinion sur un aspect. Les modèles fréquents utilisés avec des aspects connus sont considérés comme des modèles d'opinion. La deuxième phase consiste à évaluer les aspects, il recherche deux synonymes dans le guide du graphique de synonymie *WordNet*[2]. Opinion Digger augmente la précision de la méthode d'apprentissage automatique non supervisée.

La TABLE 2.1 montre les avantages et inconvénients de quelques techniques de l'approche basée apprentissage automatique *(apprentissage supervisé et non supervisé)* [23][24][25] :

2. C'est une base de données, on peut la téléchargée sur : http ://wntw.sourceforge.net/download.htm

Technique	Avantages	Inconvénients
SVM *(supervisé)*	— Grande précision de prédiction. — Fonctionne bien sur de plus petits datasets.	— Ne convient pas à des jeux de données plus volumineux. — Moins efficace sur les jeux de données contenant du bruits.
NB *(supervisé)*	— Relativement simple à comprendre et à construire. — Facile à former.	— Implique que chaque fonctionnalité soit indépendante, ce qui n'est pas toujours le cas.
K-NN *(supervisé)*	— Rapide. — Facile à comprendre.	— Prédiction lente. — Méthode gourmande en place mémoire.
K-Means *(non supervisé)*	— Efficace. — Convient aux gros datasets.	— Manque de cohérence. — Limitation des calculs.

TABLE 2.1 – Avantages et inconvénients des techniques mentionnées.

Apprentissage semi supervisé

L'apprentissage semi supervisé [17] utilise un ensemble de données étiquetées et non étiquetées. Il se situe ainsi entre l'apprentissage supervisé qui n'utilise que des données étiquetées et l'apprentissage non supervisé qui n'utilise que des données non étiquetées. L'utilisation combinées de ces données, permet d'améliorer de façon significative la qualité de l'apprentissage. Un autre avantage vient du fait que l'étiquette de données nécessite

l'intervention d'un utilisateur humain. Lorsque les jeux de données deviennent très grands, cette opération peut s'avérer fastidieuse. Dans ce cas, l'apprentissage semi supervisé, qui ne nécessite que quelques étiquettes, revêt un intérêt pratique évident et indiscutable.

2.5.2 Approche basée lexique

Les mots d'opinion sont utilisés dans de nombreuses tâches de classification des sentiments [17]. Les mots d'opinion positives sont utilisés pour exprimer certains états souhaités, tandis que les mots d'opinion négatives sont utilisés pour exprimer certains états non désirés. Il existe également des expressions d'opinion et des idiomes qui s'appellent le lexique d'opinion.
Il existe trois approches principales pour compiler *(ou rassembler)* la liste de mots d'opinion. L'approche manuelle prend beaucoup de temps et n'est pas utilisée seule. Il est généralement associé aux deux autres approches automatisées en guise de vérification finale pour éviter les erreurs résultants des méthodes automatisées.

1. **Approche basée sur le dictionnaire :** un petit ensemble de mots d'opinion est collecté manuellement avec des orientations connues [17]. Cet ensemble est développé en recherchant dans les corpus bien connus WordNet ou dans le dictionnaire des synonymes et antonymes. Les mots récemment trouvés sont ajoutés à la liste de semences, puis la prochaine itération commence. Le processus itératif s'arrête lorsqu'aucun nouveau mot n'est trouvé. Une fois le processus terminé, une inspection manuelle peut être effectuée pour supprimer ou corriger les erreurs.

2. **Approche basée sur le corpus :** un corpus [17] est un grand corps de texte en langage naturel utilisé pour accumuler des statistiques sur le texte en langage naturel. Les corpus incluent souvent des informations supplémentaires comme une étiquette pour chaque mot indiquant sa partie de discours, et peut-être l'arbre d'analyse pour chaque phrase.

   Un lexique [17] est une collection d'informations sur les mots d'une langue à propos des catégories lexicales auxquelles ils appartiennent. Un lexique est généralement structuré comme une collection d'entrées lexicales. Une entrée lexicale inclura d'autres informations sur les rôles joués par le mot, tels que les informations sur les caractéristiques *(par exemple, si un verbe est transitif, intransitif etc, quelle forme prend le verbe : participe, présent, passé, etc...).*

3. **Méthode manuelle :** utilisé en combinaison avec des approches automatisées [26] telles que l'approche basée sur un dictionnaire et sur le corpus, l'annotation de sentiment manuelle prend beaucoup de temps .

Travaux connexes effectués sur l'analyse des sentiments basée lexique

— En 2003, une recherche illustre l'approche d'analyse de sentiments [27] permettant d'extraire à partir d'un document les sentiments associés à des polarités positives ou négatives pour des sujets spécifiques, au lieu de classer l'ensemble du document en positif ou en négatif. Les résultats indiquent qu'ils peuvent réellement extraient des informations utiles sur les sentiments de la plupart des textes. Les expériences initiales ont abouti à une précision d'environ 95%, aussi la précision peut descendre à environ 75% selon les domaines et les types de données développées et à un rappel d'environ 20%. Le système développe manuellement le lexique de sentiments, et il doit modifier et ajouter des termes de sentiments pour de nouveaux domaines. Bien que, le dictionnaire dépendant du domaine relativement, avec moins de 100 entrées chacun pour 5 domaines différents. En plus, il travaille à la génération automatisée des lexiques de sentiments afin de réduire l'intervention humaine dans la maintenance du dictionnaire et pour améliorer la précision de nouveaux domaines avec le rappel global.

— En 2010, une recherche sur l'analyse des sentiments [28] est faite où elle aborde la classification du texte à l'aide d'étiquettes. Le système @AM développé permettant de classer les phrases en utilisant des types d'attitude précis et de traiter de manière approfondie de la sémantique des verbes dans l'analyse des attitudes. L'évaluation de cette méthode sur 1000 phrases décrivant des expériences personnelles a donné des résultats comme suit : la précision moyenne sur le niveau de grain fin *(14 étiquettes)* était de 62%, sur le niveau moyen *(7 étiquettes)* 71% et le niveau supérieur *(3 étiquettes)* 88%.

— En 2013, une expérience a étudié l'utilité de l'analyse de polarité basée sur la RST*(Rhetorical Structure Theory)* dans la blogosphère [29] qui fournit des informations importantes des différentes tailles de texte d'un document. Toutefois, la RST n'a été étudiée que pour les problèmes de classification de polarité dans des scénarios restreints et à petite échelle. Les résultats montrent que RST fournit des

informations précieuses sur la structure des textes, qui peuvent être utilisées pour établir un classement plus précis des documents en termes d'estimation de leurs sentiments dans des blogs à thèmes multiples.

2.5.3 Approche basée hybride

Cette approche combine l'approche basée lexique et l'approche basée apprentissage automatique. L'utilisation de l'approche hybride permet d'annoter automatiquement le corpus d'apprentissage avec la méthode basée sur le lexique, et ensuite entrainer le classificateur sur ce corpus avec une méthode issue des méthodes de l'apprentissage automatique [17].

Travaux connexes effectués sur l'analyse des sentiments basée sur l'approche hybride

En 2009, une recherche [30] qui combine une classification basée sur des règles, un apprentissage supervisé et un apprentissage automatique dans une nouvelle méthode testée sur des critiques de films, des critiques de produits et des commentaires sur MySpace. Les résultats montrent qu'une classification hybride peut améliorer et entraîner une meilleure efficacité de classification en termes d'une mesure F1 *(la précision et le rappel)* en micro et macro-moyenne. De plus, il propose une approche semi-automatique et complémentaire dans chaque classificateur qui peut aider d'autres classificateurs à atteindre un bon niveau d'efficacité.
La TABLE 2.2 montre les avantages et limites des deux approches *(l'approche basée apprentissage automatique* [31]*, l'approche basée sur le lexique* [32]*)* :

Approche	Avantages	Inconvénients
L'apprentissage automatique	— Peut être transformé en ce que demande le domaine afin de mieux travailler. — Donne de meilleurs résultats en terme de haute précision de classification. — Dictionnaire n'est pas nécessaire.	— Peut être affectée par les variations de classes et par l'effet des changements linguistiques. — Les classificateurs qui se sont entrainés sur un domaine spécifique, ne fonctionnent pas avec un autre dans la plupart des cas.
Approche basée sur un lexique	— Produire moins d'opérations de calcul parce qu'elle ne demande aucune données d'entraînement.	— Moins de capacité de classification en fonction du contexte ou du domaine. — Exige l'existence de ressources linguistiques puissantes qui ne sont pas toujours disponibles.

TABLE 2.2 – Avantages et inconvénients des approches basées sur différents niveaux d'analyse du sentiments.

2.6 Sources des données

Les opinions des utilisateurs présentent le critère principal pour l'amélioration de la qualité des services fournis et la mise en valeur des produits livrés. Ces opinions se présentent sous différentes sources de données.

1. **Sites d'avis :** les opinions ont le rôle de décideur pour tout utilisateur durant la phase d'achat. Les avis générés par les utilisateurs sur les produits et les services sont largement disponibles sur Internet [33]. La classification de sentiments utilise les données de l'examinateur collectées à partir des sites web tels que :

— www.gsmarena.com (*revues de téléphone portable*).

— www.amazon.com (*revues des produits*).

— www.CNETdownload.com (*revues des produits*).

Ces sites accueillent des millions d'avis sur les produits par les consommateurs .

2. **Blogs :** un blog [34] où les personnes peuvent écrire les différents sujets dans un but de partager avec d'autres personnes sur le même site. La simplicité de la création des postes blogs ainsi que leurs forme libre à rendre le blogging un évènement accessible. La blogosphère contient un nombre important de messages relatifs à une panoplie des sujets d'intérêt. Les blogs sont utilisés comme sources d'opinions dans la plupart des études relatives à l'analyse des sentiments.
3. **Micro-blogs :** les micro-blogs [35] sont parmi les outils de communication très populaires des utilisateurs d'Internet. Chaque jour, des millions de messages apparaissent dans des sites web populaires pour les micro-blogging tels que : Twitter, Tumblr, Facebook. Parfois les messages Twitter expriment des opinions qui sont utilisées comme source de données pour classifier le sentiment.

2.7 Outils d'analyse des sentiments

Un outil d'analyse des sentiments est un logiciel qui analyse les conversations en texte et évalue le ton, l'intention et l'émotion de chaque message. Voici une liste non exhaustive des outils les plus connus :

— **Sentiment Analyzer** [3] **:** est un outil en ligne gratuit qui permet d'analyser les sentiments de tous les textes rédigés en anglais. Le système calcule un score de sentiment qui reflète le sentiment général, le ton ou le sentiment émotionnel du texte d'entrée. Les scores de sentiment vont de -100 *(indique un ton très négatif)* à +100 *(indique un ton très positif ou enthousiaste)*, cet outil utilise une approche lexicale.

— **Sentiment140 :** (*anciennement connu sous "Twitter sentiment"*) est un outil en ligne gratuit qui permet de connaître le sentiment d'une marque, d'un produit ou d'un sujet sur Twitter. Il utilise des classificateurs construits à partir d'algorithmes d'apprentissage automatique *(ME)*, il a été créé par **Alec Go**, **Richa Bhayani** et

3. https ://www.danielsoper.com/sentimentanalysis/default.aspx

Lei Huang, étudiants de l'université de Stanford *(projet académique)* [36].

— **SentiStrength** [4] **:** est un outil gratuit compatible avec Windows uniquement. Il propose une analyse de sentiment automatique qui passe au crible jusqu'à 16000 textes de réseaux sociaux par seconde. Il utilise deux pôles de tonalité :
  — De -1 (*non négatif*) à -5 (*extrêmement négatif*).
  — De 1 (*non positif*) à 5 (*extrêmement positif*).

— **AFINN** [5] **:** est un outil en ligne gratuit, qui permet l'évaluation *(entre -5 pour les négatives et 5 pour les positives)* du texte entré dans la boite du dialogue et on obtient une analyse sur le web. Le lexique AFINN est peut être l'un des lexiques les plus simples et les plus populaires [37] pouvant être largement utilisé pour l'analyse des sentiments, la version actuelle du lexique est **AFINN-en-165.txt**, il contient plus de 3300 mots avec un score de polarité associé à chaque mot. Il déduit la polarité d'une phrase donnée à travers la somme des scores du chaque mot.

— **SentiWordNet :** est une ressource lexicale pour l'extraction d'opinion. SentiWordNet attribue à chaque synset [6] de WordNet trois scores de sentiment : positivité, négativité et objectivité, sa version actuelle est 3.0, basée sur WordNet 3.0 [38].

2.8 Domaines d'application de l'analyse des sentiments

L'importance de la détection d'opinion est présente dans plusieurs domaines, ainsi plusieurs applications ont vu le jour dans ce contexte [39]. La FIGURE 2.6 les illustre.

4. http ://sentistrength.wlv.ac.uk/

5. https ://darenr.github.io/afinn/

6. C'est une fonction dans la bibliothèque NLTK sous la branche WordNet et SentiWordNet. Elle renvoie une liste de tous les synonymes d'un mot plus un positive_score, negative_score et objective_score de ce mot

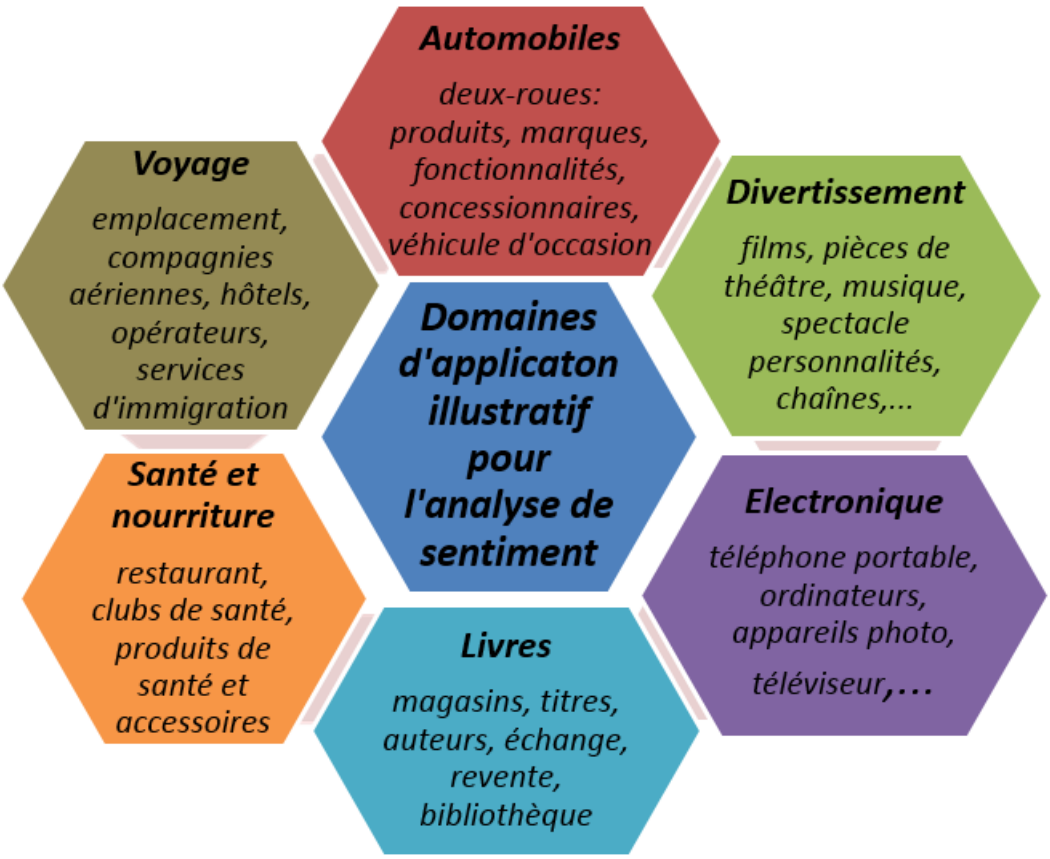


FIGURE 2.6 – Domaines d'application illustratifs de l'analyse des sentiments.

Ci-dessous nous citons brièvement quelques applications :

1. *La politique :* les acteurs politiques ont suivi la tendance d'opinion, tel qu'avant de promulguer une nouvelle loi, les politiciens essayent de récolter l'avis des internautes sur cette loi. Il est intéressant de connaître aussi l'avis des internautes sur un homme politique pour une élection présidentielle.
2. *Les entreprises :* à travers l'analyse des sentiments, les entreprises peuvent connaître l'opinion des clients sur leurs produits ou leurs services. Dans une perspective d'améliorer leurs produits et d'augmenter leurs chiffres d'affaires. Le marketing a rapidement compris l'intérêt de l'analyse de sentiments.
3. *Les clients :* l'analyse des sentiments fait partie aussi de vie des internautes. Les sondages dans ce domaine montrent que la majorité des clients avant qu'ils achètent un produit, ils font des recherches d'avis sur ce produit ou un service donné et même ils sont prêts à payer plus cher un produit dont l'avis est plus favorable qu'un autre.

2.9 Problèmes d'analyse du sentiment

L'analyse des sentiments s'applique sur une quantité énorme de données ce qui peut causer de nombreux défis empêchant l'exactitude de l'analyse .
Dans ce qui suit, nous citons brièvement quelques problèmes :

2.9.1 Problème de la langue

Le défi consiste à développer des ressources [40] comme des lexiques, des dictionnaires et des corpus pour les autres langues *(l'Arabe, le Chinois, ...etc)* puisque la langue la plus utilisée est la langue Anglaise.

2.9.2 Problème du sujet *(Aboutness en anglais)*

Le problème réside dans un mot, qui peut avoir un différent contexte lors de son utilisation mais le classificateur considère les deux situations comme une seule [40].

2.9.3 Détection des fausses opinions

Les utilisateurs se trompent dans la croyance d'un opinion négative ou positive fabriquée sur une entité spécifique dans le but de sous-estimer la réputation de cette entité [40].

2.9.4 Traitement de la négation

Les mots négatifs sont difficiles à manipuler correctement [40] parce que le fait de les mentionner dans une phrase avant un mot positif conduira à un sentiment différent par rapport à une autre phrase qui ne contient que ce mot positif.

2.9.5 Ambiguïté de sentiment

Il existe des phrases avec des mots positifs ou négatifs [26], n'expriment pas nécessairement un sentiment. Il y'a aussi des phrases qui ne possèdent ni des mots positifs ni négatifs alors qu'elles expriment un sentiment.

2.10 Conclusion

Dans ce chapitre, nous avons pu définir les principaux fondements, méthodes et techniques de classification des sentiments. Il est à noter que certains travaux dans l'approche basée sur le lexique travaillent sur une méthode dictionnaire et d'autres avec les mots du corpus. L'avantage de la méthode dictionnaire est d'englober un grand ensemble de mots mais elle a l'inconvénient d'être indépendante du domaine et du contexte.
L'approche basée sur l'apprentissage automatique donne des résultats excellents grâce aux algorithmes sophistiqués employés pour construire le modèle de classification. Cela fait que cette approche s'adapte aux mots employés dans le corpus. Cependant, son inconvénient majeur est la nécessité de l'annotation manuelle, ce qui est difficile à réaliser dans de grands corpus et ce qui est très problématique étant donné que théoriquement, tout dépend de ce que nous travaillons avec.
Dans le chapitre suivant, nous allons décrire, processus par processus, quelques améliorations que nous proposons afin d'améliorer les performances.

Chapitre 3

Conception, mise en œuvre et améliorations

3.1 Introduction

L'opinion est la clarification d'une personne à propos d'un objet ou d'un sujet particulier, tandis que le sentiment est un avis d'une personne sur un objet ou un sujet, cet avis caractérisé par une polarité soit positive, soit négative ou un mélange.
Ce serait bien si on connait le sentiment et l'opinion exprimés par des gens, cela peut être une tâche complexe, surtout s'il y a une quantité énorme de données à lire et à comprendre, c'est là qu'intervient l'analyse des sentiments qui permet de traiter et extraire rapidement des informations exploitables à partir d'énormes volumes de texte sans les avoir lues.
Ce chapitre est consacré à détailler l'objectif d'analyse de sentiments qui est la reconnaissance de polarité de l'opinion exprimé sur différents types des réseaux sociaux où nous avons comme données d'entrée des tweets*(en anglais)* extraits de deux datasets *sentiment140* et *US Airline* pour entraîner et tester le modèle naïf bayes et celui de la solution proposée sous formes d'améliorations de la méthode qui se base sur un dictionnaire.

3.2 Généralités sur twitter

Twitter a été créé en mars 2006 par les développeurs *(Jack Dorsey, Noah Glass, Biz Stone et Evan Williams)* [41]. Ce service américain est devenu rapidement populaire dans le monde. C'est un microblogging et réseau social en ligne, il permet au utilisateurs de publier et interagir avec des messages appelés *tweets*. Les tweets sont limités à 280

caractères, pour toutes les langues autres que le chinois, le japonais et le coréen.

3.2.1 Structure de Tweet

Un tweet aborde multiples informations :

— L'id du tweet,

— La date du tweet,

— Le nom de l'utilisateur qui a posté le tweet,

— Le texte du tweet,

— L'emoji *(les messages publiés peuvent désormais accueillir les émoticônes)*,

— La photo du profil,

— L'emplacement de l'utilisation.

La FIGURE 3.1 représente la structure de Tweet :

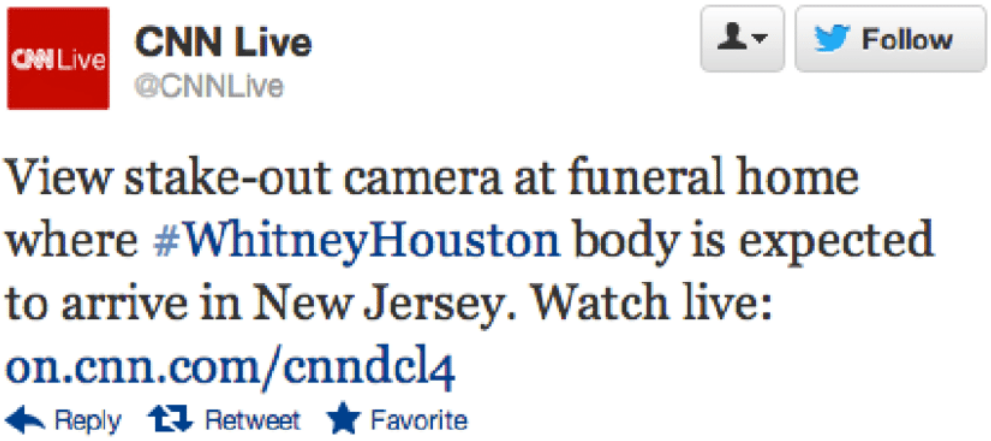


FIGURE 3.1 – Exemple d'un Tweet.

3.2.2 Fondements de Twitter

Dans un tweet, l'arobase « @ » indique le nom d'utilisateur et un lien direct vers un compte Tweeter. Chaque utilisateur, dégage un ensemble d'informations lors de son tweet, telle que (1) la langue du tweet, (2) le fuseau horaire de l'emplacement, (3) l'emplacement à partir duquel le tweet a été envoyé, (4) la photo du profil (5) l'emplacement de l'utilisation, (6) la page web, (7) une brève biographie, (8) les liens favoris. Il existe différentes terminologies [42] dans un tweet pour pouvoir interagir avec les utilisateurs et participer facilement à des conversations, elles sont citées comme suit :

1. Réponse : la possibilité de répondre à un tweet.
2. Retweet : la possibilité de partager le tweet ou le citer pour y ajouter un commentaire.
3. J'aime : la possibilité de monter qu'on apprécie un tweet.
4. Mention : la possibilité de motionner ou attirer l'attention d'un utilisateur **@nomdutilisateur** dans un message.
5. Hashtag : c'est un mot précédé du symbole # est utilisé pour organiser les conversations et faciliter la recherche sur un sujet donné.

3.3 Dataset

Dans ce qui suit, on présente deux datasets très répondus :

3.3.1 Dataset : "Sentiment140"

Le dataset **"Sentiment140"** est un projet de classe à l'université de Stanford [43] où les étudiants ont exploré divers aspects de la classification de l'analyse des sentiments dans les projets finaux pour les classes suivantes :

— CS224N Traitement du langage naturel au printemps 2009, enseigné par *Chris Manning*.

— CS224U Compréhension du langage naturel à l'hiver 2010, animé par Dan *Jurafsky* et Bill *MacCartney*.

— CS424P Signification sociale et sentiments à l'automne 2010, enseignés par *Chris Potts* et Dan *Jurafsky*.

Ce jeu de données contient 1600000 tweets qui sont collectés en fonction de la situation de tous les sujets en utilisant la technologie API Twitter où il inclut 6 champs : sentiment *(la polarité du tweet (0 = négatif, 2 = neutre, 4 = positif))*, id *(l'identifiant du tweet)*, date *(la date du tweet)*, query_string *(s'il n'y a pas de requête, alors cette valeur est NO_QUERY)*, user *(l'utilisateur qui a tweeté)*, text *(le texte du tweet)*.

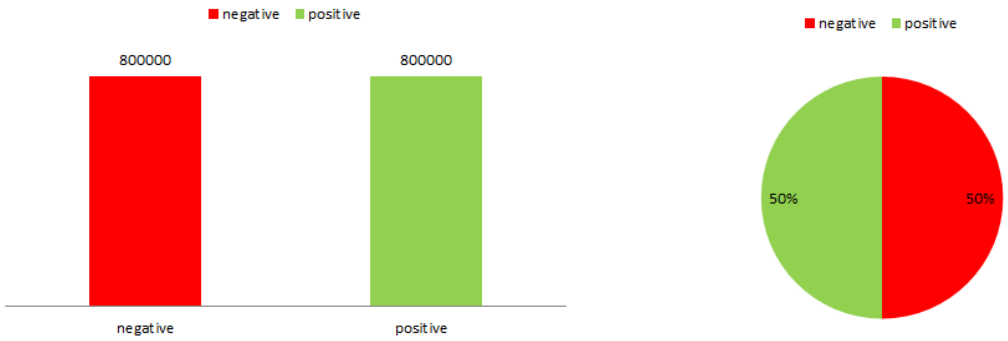
La FIGURE 3.2 illustre une portion de dataset "Sentiment140".

	sentiment	id	date	query_string	user	text
0	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...
2	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...
3	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
4	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....
...	...	...	...	...	...	...
1599995	4	2193601966	Tue Jun 16 08:40:49 PDT 2009	NO_QUERY	AmandaMarie1028	Just woke up. Having no school is the best fee...
1599996	4	2193601969	Tue Jun 16 08:40:49 PDT 2009	NO_QUERY	TheWDBoards	TheWDB.com - Very cool to hear old Walt interv...
1599997	4	2193601991	Tue Jun 16 08:40:49 PDT 2009	NO_QUERY	bpbabe	Are you ready for your MoJo Makeover? Ask me f...
1599998	4	2193602064	Tue Jun 16 08:40:49 PDT 2009	NO_QUERY	tinydiamondz	Happy 38th Birthday to my boo of alll time!!! ...
1599999	4	2193602129	Tue Jun 16 08:40:50 PDT 2009	NO_QUERY	RyanTrevMorris	happy #charitytuesday @theNSPCC @SparksCharity...

1600000 rows × 6 columns

FIGURE 3.2 – Extraction des données de dataset "Sentiment140".

Ces données *(voir la* FIGURE 3.3*)* sont répartis en deux polarités annotés (0 = négatif, 4 = positif), 800.000 tweets pour chacun.



(a) Nombre de tweets de l'ensemble de données. (b) Pourcentage des polarités des tweets.

FIGURE 3.3 – Visualisation de nombre de tweets du l'ensemble de données "Sentiment140".

3.3.2 Dataset : "Twitter US Airline Sentiment"

L'ensemble de données **"Twitter US Airline Sentiment"** Dataset de Kaggle[1] contient des tweets sur *US Airline* de février 2015 classés en 14640 tweets : positifs*(2363)*, négatifs*(9178)* et neutres*(3099)*.

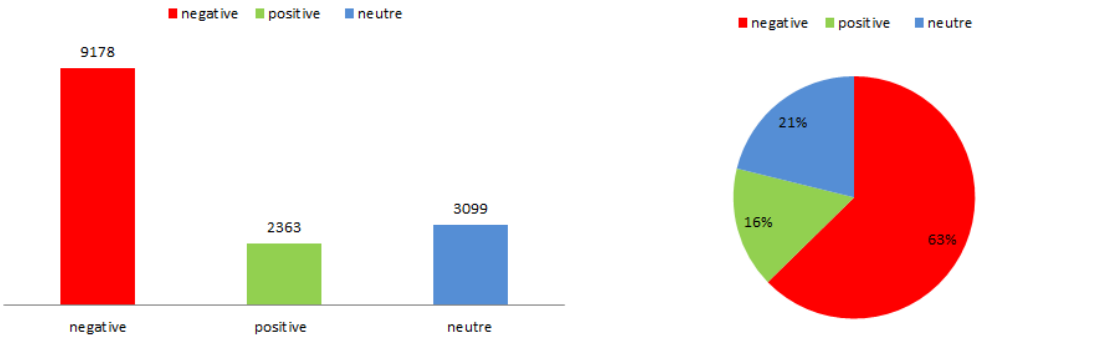
La FIGURE 3.4 représente une partie de cet ensemble de données :

1. https ://www.kaggle.com/

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gold	name	negativereason_gold	retweet_count	text	tweet_coord	tweet_created	tweet_location	user_timezone
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	NaN	cairdin	NaN	0	@VirginAmerica What @dhepburn said.	NaN	2015-02-24 11:35:52 -0800	NaN	Eastern Time (US & Canada)
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica plus you've added commercials t...	NaN	2015-02-24 11:15:59 -0800	NaN	Pacific Time (US & Canada)
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	NaN	yvonnalynn	NaN	0	@VirginAmerica I didn't today... Must mean I n...	NaN	2015-02-24 11:15:48 -0800	Lets Play	Central Time (US & Canada)
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica it's really aggressive to blast...	NaN	2015-02-24 11:15:36 -0800	NaN	Pacific Time (US & Canada)
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica and it's a really big bad thing...	NaN	2015-02-24 11:14:45 -0800	NaN	Pacific Time (US & Canada)

FIGURE 3.4 – Extraction des données de dataset "US Airline".

La répartition des classes de cet ensemble de données est représenté ci-dessous *(voir la* FIGURE 3.5*)*.



(a) Nombre de tweets de l'ensemble de données. (b) Pourcentage des polarités des tweets.

FIGURE 3.5 – Visualisation de nombre de tweets du l'ensemble de données "US Airline".

3.4 Prétraitement des données

Prétraitement *(preprocessing)* est une étape importante pour la préparation de dataset, qui vise à traiter les messages pour les structurer et faciliter leurs utilisations. Ci-dessous la fonction de nettoyage des données, l'ordre du nettoyage est :

— **Souping *(Décodage HTML)* :** l'information en ligne n'est pas obligatoirement informatives telles que balises HTML que ne se convertit pas en texte et il reste dans le champ de texte sous la forme " &amp", "&quot"...etc.

```
In [57]: dataframe140.text[400]

Out[57]: "#3 woke up and was having an accident - &quot;It's pushing, it's
         pushing!&quot; he was crying because he couldn't stop from wettin
         g his pants. "
```

(a) Exemple avant souping : Décodage HTML.

```
In [13]: dataframe140.text[400]

Out[13]: 'woke up and was having an accident it pushing it pushing he was
         crying because he could not stop from wetting his pants'
```

(b) Exemple aprés souping : Décodage HTML.

— **Suppression de BOM(Byte Order Mark) :** la nomenclature UTF-8 est une séquence d'octets (EF BB BF) qui permet au lecteur d'identifier un fichier comme étant codé en UTF-8.

```
In [53]: dataframe140.text[194]

Out[53]: '@JonathanRKnight I hate the limited letters,too.Hope you and the
         guys are fine?I pray for my dog,shei¿½s not well '
```

(a) Exemple avant la suppression de BOM.

```
In [22]: dataframe140.text[194]

Out[22]: 'hate the limited letters too hope you and the guys are fine pray
         for my dog she not well'
```

(b) Exemple aprés la suppression de BOM.

— **Suppression d'adresse URL (motif 'http :', 'www :') et de l'identifiant Twitter (motif '@') :** ces informations n'ajoutent aucune valeur pour exprimer un opinion.

```
In [56]: dataframe140.text[37]

Out[56]: '@MissXu sorry! bed time came here (GMT+1)   http://is.gd/fNge'
```

(a) Exemple avant la suppression des motifs ('@', 'http :').

```
In [16]: dataframe140.text[37]

Out[16]: 'sorry bed time came here gmt'
```

(b) Exemple aprés la suppression de des motifs ('@', 'http :').

— **Lower-case (minuscule) :** utilisé pour rendre le texte en minuscule.

```
In [58]: dataframe140.text[84]
Out[58]: "Damn... I don't have any chalk! MY CHALKBOARD IS USELESS "
```

(a) Exemple avant l'utilisation de lower-case.

```
In [17]: dataframe140.text[84]
Out[17]: 'damn do not have any chalk my chalkboard is useless'
```

(b) Exemple aprés l'utilisation de lower-case.

— **Traitement de la négation :** la négation peut changer complètement le sentiment d'un tweet, donc c'est important de la garder d'une façon séparable pour bien la montrer.

```
In [30]: dataframe140 .text[1999]
Out[30]: '2morw  I get my blasted wisdom teeth pulled! Need sleep...cnt st
         op worryng, I hate needles '
```

(a) Exemple d'abréviation de négation.

```
In [24]: dataframe140.text[1999]
Out[24]: 'morw get my blasted wisdom teeth pulled need sleep can not stop
         worryng hate needles'
```

(b) Exemple de négation d'une façon séparable.

— **Suppression du hashtag *(motif '#')*, les nombres et les caractères spéciaux :** nous avons supprimé le motif de hachtag en gardant le texte qui le suit car il peut fournir des informations utiles sur le tweet, en nettoyant aussi les caractères non-lettres y compris les nombres.

```
In [54]: dataframe140.text[193]
Out[54]: '@goodlaura What about Reese dying on #TTSC? And season finale ne
         xt week.  #24 boring, Madame President is a crazy woman.'
```

(a) Exemple d'un tweet contenant le motif hachtag(#), caractère(?) et un nombre.

```
In [21]: dataframe140.text[193]
Out[21]: 'what about reese dying on ttsc and season finale next week borin
         g madame president is crazy woman'
```

(b) Exemple d'un tweet après la suppresion du motif hachtag(#), caractère(?) et un nombre.

— **Tokenisation, enlever mots vides et rejoindre :** la tokenisation nous a permet de découper le tweet en tokens qui seront par la suite des entrées du processus des mots vides qui consiste à éliminer les mots qui n'influent pas sur l'opinion exprimée dans un tweet, après nous effectuons la jointure afin de reformuler le tweet du nouveau.

```
In [3]: dataframe140.text[10]
Out[3]: "spring break in plain city... it's snowing "
```

(a) Exemple d'un tweet contenant des mots vides.

```
In [70]: dataframe140.text[10]
Out[70]: 'spring break plain city snowing'
```

(b) Exemple d'un tweet après l'enlève des mots vides.

3.5 Description des améliorations proposées

La solution proposée s'est inspirée de la méthode basée sur un dictionnaire, aide à trouver une façon d'exploiter les tweets afin de les classifier selon des classes pertinentes. Pour réaliser notre solution nous avons suivi les étapes suivantes :

— Utilisation du dictionnaire AFINN *(précédemment expliqué dans le chapitre 2),*

— Saisie d'un tweet en entrée,

— Calcul du score de tweet,

— Classification du tweet selon le score calculé en classe *(positive, négative ou neutre).*

Nous avons téléchargé la dernière version du dictionnaire AFINN *(AFINN-en-165.txt)* qui contient 3300 mots en anglais, chacun étiqueté avec une force de sentiment et ciblant l'analyse de sentiments sur un texte court comme on le trouve en réseaux sociaux *(tweet).* Nous avons effectué une visualisation sur la polarité des mots du dictionnaire AFINN et nous avons obtenu le graphe illustré dans la FIGURE 3.6 :

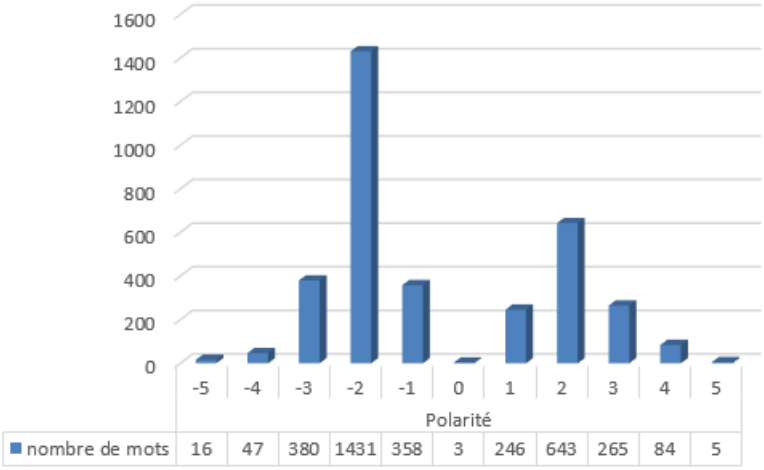


FIGURE 3.6 – Visualisation du nombre de mots pour chaque polarité *(AFINN)*.

Nous pouvons utiliser la fonction **afinn.score** de package *afinn* qui fait la somme des polarités de chacun des mots d'un tweet dans les deux dataset *(Sentiment140 et US Airline)*.

	sentiment	text	afinn_score
6	4	Almost an A is still a B... still, 69 % ain't ...	-3.0
0	0	@robots_ ...you forgot minger.	-1.0
13	4	Can you say no more school books...well till n...	-1.0
11	4	IDIOT: th*nks for the follow new friends!!! ho...	-1.0
18	0	I don't think this "Horny Kitty" is ...	-1.0
9	0	@trxiegirl71 -sorry about your guinea pig	-1.0
7	0	how dare you!!	0.0
8	0	Another weekend comes to an end	0.0
10	4	@caliblondie I'm SO glad. I know how bad I fee...	0.0
16	4	@fortja Thought you might. Tried to keep it go...	0.0
14	0	wishing i was at chain reaction seeing @myamer...	1.0
15	0	In a rather somber mood, I just want to be wit...	1.0
3	4	Sunny days make me smile more.	2.0
5	4	Watching Sweeny todd hopefully this movie wil...	2.0
17	0	Since the sun won't show itself I'm going tann...	3.0
1	4	the beach is less than two weeks away stressin...	4.0
2	4	@sploshy thanks will check it out from PC at w...	5.0
4	0	everyone wish me luck on my 15 hour work day t...	6.0
12	4	this business is growing on me.. covering mag...	6.0
19	4	@DangerErin Hurray!! So excited for you! And s...	9.0

(a) Échantillon d'utilisation de la fonction *afinn.score* sur le dataset "Sentiment140".

	airline_sentiment	text	afinn_score
2	negative	@JetBlue even @Citi responded quicker via Twit...	-3.0
6	positive	@JetBlue you don't remember our date Monday ni...	-3.0
9	negative	@JetBlue but by Cancelled Flighting my flight ...	-1.0
17	negative	@USAirways that seems unlikely without a crew ...	-1.0
5	negative	@SouthwestAir FIND A WAY TO Cancelled Flight F...	-1.0
16	negative	@AmericanAir How do I change my flight if the ...	0.0
13	neutral	@USAirways US 728. Refuel; we've sat for so lo...	0.0
19	negative	@USAirways I have been on hold for 4 hours my ...	0.0
1	neutral	@VirginAmerica pilot says we expect a choppy l...	0.0
8	positive	@united Honestly, I stopped trying to report t...	1.0
18	neutral	@united also checked email you have on file fo...	1.0
12	positive	@SouthwestAir filing it now. Thank you for you...	2.0
4	negative	@united Apparently they are asking 20 people t...	2.0
15	neutral	@JetBlue hi! Is it possible to upgrade to a Mi...	2.0
3	negative	@USAirways I've been on hold at the reservatio...	2.0
0	positive	@USAirways a big thanks to the gate agent flt5...	2.0
10	neutral	@united hey! think someone could meet me with ...	3.0
11	negative	@AmericanAir why would I pay $200 to reactivat...	4.0
7	positive	@VirginAmerica not worried, it's been a great ...	5.0
14	negative	@USAirways lots of fun to be removed from top ...	10.0

(b) Échantillon d'utilisation de la fonction *afinn.score* sur le dataset "US Airline".

FIGURE 3.7 – Échantillons d'utilisation de la fonction *afin.score*.

Nous avons remarqué que la fonction *afinn.score* fait la somme des polarités des mots, donc nous avons programmé une première fonction qui fait la somme des polarités des mots à partir du fichier déjà téléchargé.

$$score(tweet) = polarité_{mot1} + polarité_{mot2} + ... + polarité_{motn}$$

Cependant, en vue d'améliorer ses performances, nous avons décidé de la modifier dans quelques points.

— Premièrement, nous avons modifié le principe du calcul de la fonction programmée *(de la somme vers la somme de l'inverse des polarités)*. Mais, avant de l'appliquer nous avons modifié le dictionnaire par l'ignorance des mots neutres*(éviter le problème de division par zéro)*.

$$score(tweet) = \frac{1}{polarité_{mot1}} + \frac{1}{polarité_{mot2}} + ... + \frac{1}{polarité_{motn}}$$

— Deuxièmement, nous avons enrichi le fichier déjà téléchargé *(AFINN-en-165.txt)* avec les émoticônes *(exemple : ” :)” sa polarité : 2 et ” :(” sa polarité : -2)*.

- Nous avons appliqué la somme des polarités prenant en considération les émoticônes.

- Nous avons appliqué la somme de l'inverse des polarités prenant en considération les émoticônes.

Pour réaliser notre proposition, nous avons suivi les étapes que nous avons résumé dans l'architecture ci-dessus.

3.5.1 Architecture globale de la solution

La FIGURE 3.8 résume toutes les étapes de notre solution.

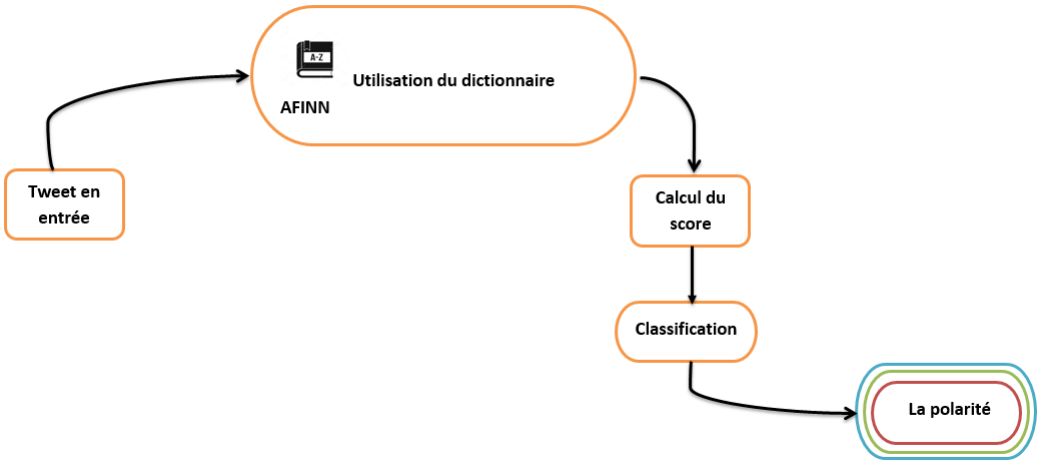


FIGURE 3.8 – Architecture de notre proposition.

3.5.2 Implémentation de la solution proposée

Dans cette partie, nous allons présenter l'application de l'analyse des sentiments à l'aide du dictionnaire *afinn*, en abordant les améliorations précédemment citées.

— D'abord, nous présentons une partie du lexique *(AFINN-en-165.txt)* dans ce qui suit.

Entrée [13]:
```
import pandas as pd
afinn_wl_url = ('AFINN-en-165.txt')
afinn_wl_df = pd.read_csv(afinn_wl_url,
                          header=None, # no column names
                          sep='\t',  # tab sepeated
                          names=['term', 'value']) #new column names

afinn_wl_df.head(15)
```

Out[13]:

	term	value
0	abandon	-2
1	abandoned	-2
2	abandons	-2
3	abducted	-2
4	abduction	-2
5	abductions	-2
6	abhor	-3
7	abhorred	-3
8	abhorrent	-3
9	abhors	-3
10	abilities	2
11	ability	2
12	aboard	1
13	aborted	-1
14	aborts	-1

FIGURE 3.9 – Échantillon du dictionnaire *AFINN*.

— Ensuite, nous avons programmé une fonction qui fait la somme des polarités de chaque mot du tweet en entrée et comme une sortie elle renvoie son score de polarité ainsi son classement.

In [5]:
```
import pandas as pd
afinn_wl_url = ('AFINN-en-165.txt')
afinn_wl_df = pd.read_csv(afinn_wl_url,
                          header=None, # no column names
                          sep='\t',  # tab sepeated
                          names=['term', 'value'])
prod = 0
phrase = input("Entrer un tweet à classifier: ")
ph = phrase.split(" ")
for i in range (len(ph)):
    for index, row in afinn_wl_df.iterrows():
        if(ph[i] == row['term']):
            prod = prod + afinn_wl_df.value[index]
print("Le score (somme) selon le dictionnaire AFINN de: ","<<",phrase,">>",":",prod)
if(prod > 0):
    print("La classe du tweet","<<",phrase,">>","selon le dictionnaire AFINN est: positive")
elif (prod < 0):
    print("La classe du tweet","<<",phrase,">>","selon le dictionnaire AFINN est:  négative")
else:
    print("La classe du tweet","<<",phrase,">>","selon le dictionnaire AFINN est:  neutre")
```

```
Entrer un tweet à classifier: Best WTF moment is :kristen didn't kiss robert . it was my best WTF moment for the MTV movie awar
ds :)))) What the hell is she thinking
Le score (somme) selon le dictionnaire AFINN de:  << Best WTF moment is :kristen didn't kiss robert . it was my best WTF moment
for the MTV movie awards :)))) What the hell is she thinking  >> : 4.0
La classe du tweet << Best WTF moment is :kristen didn't kiss robert . it was my best WTF moment for the MTV movie awards :))))
What the hell is she thinking  >> selon le dictionnaire AFINN est: positive
```

FIGURE 3.10 – Utilisation de la fonction qui somme les polarités des mots et renvoie le classement du tweet.

— Puis, nous avons modifié la formule du calcul de la fonction précédente par la somme de l'inverse des valeurs associées à chaque mot du tweet.

In [6]:
```
import pandas as pd
afinn_wl_url = ('AFINN-en-165.txt')
afinn_wl_df = pd.read_csv(afinn_wl_url,
                          header=None, # no column names
                          sep='\t',  # tab sepeated
                          names=['term', 'value'])
prod = 0
phrase = input("Entrer un tweet à classifier: ")
ph =phrase.split(" ")
for i in range (len(ph)):
    for index, row in afinn_wl_df.iterrows():
        if(ph[i] == row['term']):
            prod = prod + (1 /afinn_wl_df.value[index])
print("Le score (somme de l'inverse des polarités) selon le dictionnaire AFINN de: ","<<",phrase,">>",":",prod)
if(prod > 0):
    print("La classe du tweet","<<",phrase,">>","selon le dictionnaire AFINN est : positive")
elif (prod < 0):
    print("La classe du tweet","<<",phrase,">>","selon le dictionnaire AFINN est:  négative")
else :
    print("La classe du tweet","<<",phrase,">>","selon le dictionnaire AFINN est:  neutre")
```

```
Entrer un tweet à classifier: Best WTF moment is :kristen didn't kiss robert . it was my best WTF moment for the MTV movie awar
ds :)))) What the hell is she thinking
Le score (somme de l'inverse des polarités) selon le dictionnaire AFINN de:  << Best WTF moment is :kristen didn't kiss robert
. it was my best WTF moment for the MTV movie awards :)))) What the hell is she thinking  >> : 0.9166666666666665
La classe du tweet << Best WTF moment is :kristen didn't kiss robert . it was my best WTF moment for the MTV movie awards :))))
What the hell is she thinking  >> selon le dictionnaire AFINN est : positive
```

FIGURE 3.11 – Utilisation de la fonction qui somme l'inverse des polarités des mots et renvoie le classement du tweet.

— Finalement, nous avons ajouté les émoticônes *(chacun étiqueté avec une polarité du sentiment)* au dictionnaire *Afinn*.

```
In [1]: import pandas as pd
        #afinn = Afinn(emoticons=True)
        afinn_wl_url = ('AFINN-en.txt')
        afinn_wl_df = pd.read_csv(afinn_wl_url,
                                  header=None, # no column names
                                  sep='\t',  # tab sepeated
                                  names=['term', 'value'])
        prod = 0
        phrase = input("Entrer un tweet à classifier:")
        ph = phrase.split(" ")
        for i in range (len(ph)):
            for index, row in afinn_wl_df.iterrows():
                if(ph[i] == row['term']):
                    prod = prod + afinn_wl_df.value[index]
        print("Le score (somme) avec émoticônes selon le dictionnaire AFINN de: ","<<",phrase,">>",":",prod)
        if(prod > 0):
            print("La classe du tweet","<<",phrase,">>","selon le dictionnaire AFINN est: positive")
        elif (prod < 0):
            print("La classe du tweet","<<",phrase,">>","selon le dictionnaire AFINN est: négative")
        else:
            print("La classe du tweet","<<",phrase,">>","selon le dictionnaire AFINN est: neutre")

        Entrer un tweet à classifier:Best WTF moment is :kristen didn't kiss robert . it was my best WTF moment for the MTV movie award
        s :)))) What the hell is she thinking
        Le score (somme) avec émoticônes selon le dictionnaire AFINN de:  << Best WTF moment is :kristen didn't kiss robert . it was my
        best WTF moment for the MTV movie awards :)))) What the hell is she thinking  >> : 7.0
        La classe du tweet << Best WTF moment is :kristen didn't kiss robert . it was my best WTF moment for the MTV movie awards :))))
        What the hell is she thinking  >> selon le dictionnaire AFINN est: positive
```

FIGURE 3.12 – Utilisation de la fonction qui fait la somme des polarités des mots *(émoticônes inclus)* et renvoie le classement du tweet.

```
In [2]: import pandas as pd
        #afinn = Afinn(emoticons=True)
        afinn_wl_url = ('AFINN-en.txt')
        afinn_wl_df = pd.read_csv(afinn_wl_url,
                                  header=None, # no column names
                                  sep='\t',  # tab sepeated
                                  names=['term', 'value'])
        prod = 0
        phrase = input("Entrer un tweet à classifier:")
        ph =phrase.split(" ")
        for i in range (len(ph)):
            for index, row in afinn_wl_df.iterrows():
                if(ph[i] == row['term']):
                    prod = prod + (1 /afinn_wl_df.value[index])
        print("Le score (somme de l'inverse des polarités) avec émoticônes selon le dictionnaire AFINN de: ","<<",phrase,">>",":",prod)
        if(prod > 0):
            print("La classe de la phrase","<<",phrase,">>","selon le dictionnaire AFINN est : positive")
        elif (prod < 0):
            print("La classe de la phrase","<<",phrase,">>","selon le dictionnaire AFINN est: négative")
        else :
            print("La classe de la phrase","<<",phrase,">>","selon le dictionnaire AFINN est: neutre")

        Entrer un tweet à classifier:Best WTF moment is :kristen didn't kiss robert . it was my best WTF moment for the MTV movie award
        s :)))) What the hell is she thinking
        Le score (somme de l'inverse des polarités) avec émoticônes selon le dictionnaire AFINN de:  << Best WTF moment is :kristen did
        n't kiss robert . it was my best WTF moment for the MTV movie awards :)))) What the hell is she thinking  >> : 1.24999999999999
        98
        La classe de la phrase << Best WTF moment is :kristen didn't kiss robert . it was my best WTF moment for the MTV movie awards
        :)))) What the hell is she thinking  >> selon le dictionnaire AFINN est : positive
```

FIGURE 3.13 – Utilisation de la fonction qui fait la somme de l'inverse des polarités des mots *(émoticônes inclus)* et renvoie le classement du tweet.

L'avantage des approches par dictionnaire est de disposer d'une référence, le corpus sur lesquels ils se constituent pouvant être précisément caractérisé.

3.6 Conclusion

Nous avons présenté dans ce chapitre une approche de classification qui contribue au problème de l'analyse du sentiment, où nous avons déterminé l'ensemble de données utilisées dans notre travail. Ensuite, nous avons appliqué un processus de prétraitement sur ces données qui sont des tweets de langue anglaise, pour minimiser le bruit et enlever l'ambiguïté afin d'améliorer la précision et la qualité de classification.
Dans le chapitre suivant, nous discuterons les résultats obtenus lors d'application de la méthode proposée ainsi qu'une comparaison avec d'autres modèles dont le but de choisir la meilleure méthode de classification.

Chapitre 4

Évaluation des performances

4.1 Introduction

Dans ce chapitre, nous présentons les outils et les langages utilisés pour implémenter le classificateur naïve bayes et la méthode basée sur dictionnaire améliorée. Par la suite, nous montrons les résultats obtenus après la comparaison entre les deux méthodes et l'utilisation d'un outil d'analyse des sentiment "Sentiment Analyzer".

4.2 Langages d'implémentation

Les méthodes implémentées dans notre travail sont réalisées avec le langage de programmation Python. Ce langage favorise la programmation impérative structurée, et orientée objet, il est utilisé pour (1) développement web (côté serveur), (2) développement de logiciels, (3) mathématiques, (4) script système. Il fonctionne sur différentes plates-formes *(Windows, Mac, Linux, ...etc.)*. Il a une syntaxe qui permet aux développeurs d'écrire des programmes avec moins de lignes que certains autres langages de programmation [44]. En effet, parmi ses qualités, il permet notamment aux développeurs de se concentrer sur ce qu'ils font plutôt que sur la manière dont ils le font.

4.3 Framework de programmation

1. **Anaconda :** est une distribution Python libre [45] qui intègre directement un grand nombre de package, le launcher donne accès aux applications disponibles parmi

eux nous avons choisi l'environnement de développement Jupyter qui permet de créer et de partager des documents contenant du code en direct, des équations, des visualisations et du texte. Les utilisations incluent : nettoyage et transformation de données, simulation numérique, modélisation statistique, visualisation de données, apprentissage automatique...etc.

4.4 Bibliothèques utilisées

La réalisation de notre travail consiste en installation des différents packages qui se trouvent dans Python, ces packages sont listés comme suit :

- **Pandas :** est une bibliothèque open source [46] fournissant des structures de données hautes performances et faciles à utiliser, ainsi que des outils d'analyse des données.
- **NumPy :** est le paquet du traitement de tableau [45] pour les nombres, les chaînes de caractères, les enregistrements et les objets.
- **Re :** est un module [47] qui forme un motif de recherche d'une séquence de caractères, il peut être utilisé pour vérifier si une chaîne contient le motif de recherche spécifié.
- **Bs4** : est une bibliothèque [48] qui permet facilement d'extraire des informations de pages Web. Il repose sur un analyseur HTML ou XML.
- **NLTK :** est une boite à outils permettant le traitement de texte [49] pour la classification, la tokenisation, la création de raccourcis, le balisage, l'analyse, le raisonnement sémantique, élimination des mots vides, ... etc.
- **Scikit-learn :** est un module permettant de convertir une collection de documents texte en une matrice de nombre de tokens [50].
- **Afinn :** est le paquet qui contient une approche basée sur une liste de mots pour l'analyse des sentiments.

4.5 Méthode Naïve Bayes

Après avoir compris les détails de cet algorithme *(expliqué précédemment dans le chapitre 2)*, nous avons choisis de l'implémenter nous mêmes spécifiquement sur les datasets "Sentiment140" et "US Airline".

La répartition des données est illustrée dans La TABLE 4.1 :

	text train	text test	total
Sentiment140	1568000	32000	1600000
US Airline	14347	293	14640

TABLE 4.1 – Description des dataset "Sentiment140" et "US Airline"

4.5.1 Architecture de l'implémentation

la FIGURE 4.1 représente les étapes nécessaires pour l'implémentation du classificateur naïve bayes :

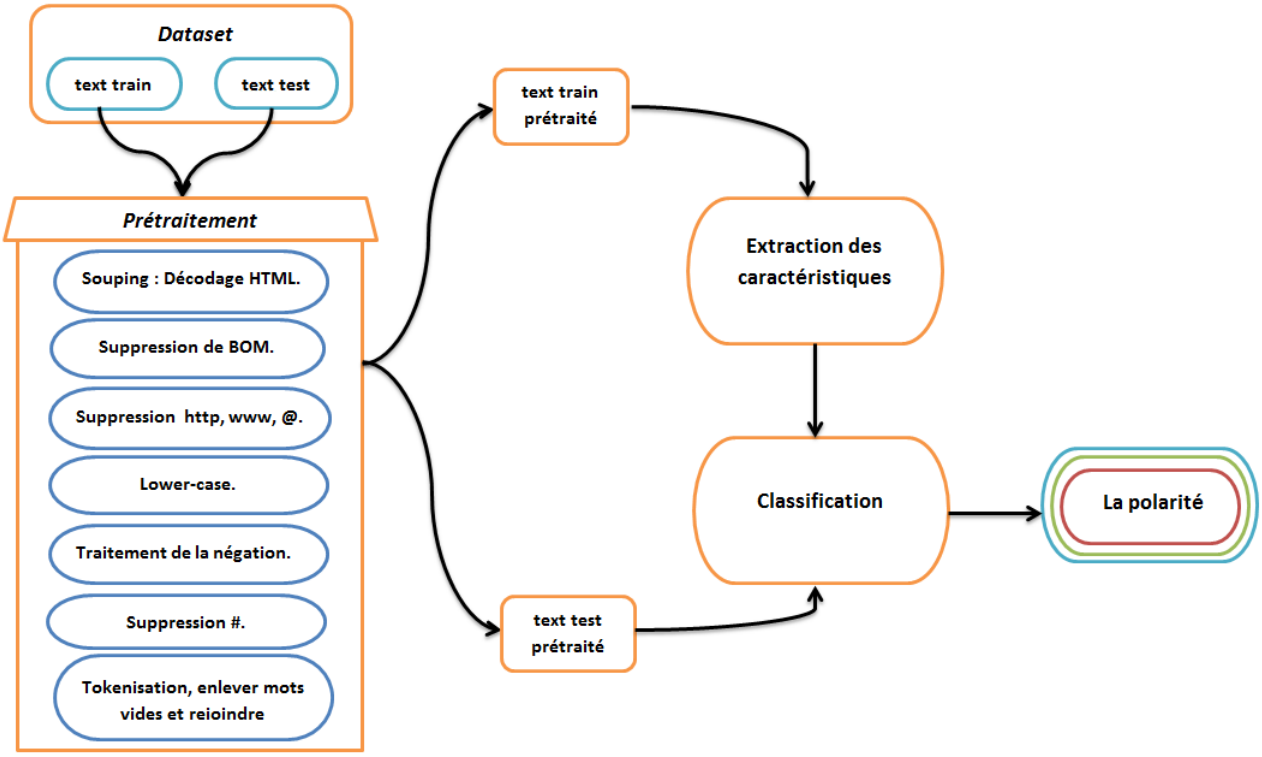


FIGURE 4.1 – Représentation graphique de l'architecture du modèle d'implémentation.

4.5.2 Description de classificateur naïve bayes

Nous pouvons décrire ce classificateur par les étapes suivantes :

1. **Définition des classes du dataset :** après la phase du pré-traitement, l'étape qui suit est l'identification des classes du dataset (text train). Dans notre cas, on a utilisé le dataset "Sentiment140" qui contient deux classes *(0 : négative, 4 : positive)* et le dataset "US Airline" avec trois classes *(négative, positive, neutre)* .

2. **Indication de la probabilité antérieure des classes :** déduire la probabilité de chacune des classes positive(P), négative(N) et neutre(NT) donnée par le calcul suivant :

$$Probabilité(P) = \frac{nombreD'objetsDeLaClassePositive}{nombreD'objetsTotale}$$

$$Probabilité(N) = \frac{nombreD'objetsDeLaClasseNégative}{nombreD'objetsTotale}$$

$$Probabilité(NT) = \frac{nombreD'objetsDeLaClasseNeutre}{nombreD'objetsTotale}$$

3. **Extraction des caractéristiques :** est le concept le plus important dans la mise en œuvre d'un classificateur, le vecteur de caractéristiques est utilisé pour créer un modèle que le classificateur apprend à partir des données d'apprentissage. Dans les tweets, nous pouvons utiliser la présence/absence de mots apparaissant dans le tweet comme caractéristique. Dans les données d'apprentissage, composées de tweets positifs, négatifs et neutres, nous pouvons scinder chaque tweet en mots et ajouter chaque mot au vecteur de caractéristiques. Le succès de classificateur dépend de ce dernier.

4. **Classification :** le processus de classification est effectué en prenant en entrée un nouveau tweet (text test), réalise en premier lieu le prétraitement de ce dernier, ensuite il prend la probabilité de chaque mot par rapport aux classes et il fait le produit de ces probabilité multiplié par la probabilité de la classe en question, comme un résultat nous allons avoir des probabilités pour chacune des classes positive, négative et neutre.

   Après le calcul de la probabilité pour les classes positive, négative et neutre, il fait la comparaison pour attribuer la classe ayant la probabilité la plus élevée au nouveau tweet.

4.5.3 Implémentation de classificateur naïve bayes

Nous allons abordé ci-dessous le processus d'implémentation de classificateur naïve bayes.

D'abord, nous présentons le code qui nous a permet d'identifier les classes des deux datasets, illustré comme suit :

```
separitionnegative = clean_sent140[clean_sent140.sentiment == 0]
separitionpositive = clean_sent140[clean_sent140.sentiment == 4]
print(separitionnegative, separitionpositive)
```

FIGURE 4.2 – Définition des classes du dataset "Sentiment140".

```
separitionnegative = clean_usairline [clean_usairline['sentiment'].str.contains("negative")]
separitionpositive = clean_usairline [clean_usairline['sentiment'].str.contains("positive")]
separitionneutral = clean_usairline [clean_usairline['sentiment'].str.contains("neutral")]
print(separitionnegative, separitionpositive, separitionneutral)
```

FIGURE 4.3 – Définition des classes du dataset "US Airline".

Et pour continuer avec l'indication de la probabilité antérieure des classes, nous avons déclaré les définitions suivante :

```
def totalClass(negative_tweets, positive_tweets):
    total_pro = len(negative_tweets) + len(positive_tweets)
    print("total",total_pro)
    return total_pro
def ProbabilityClassP(positive_tweets, total_pro):
    probability_positive = len(positive_tweets) / total_pro
    print("probability_positive",probability_positive)
    return probability_positive
def ProbabilityClassN(negative_tweets, total_pro):
    probability_negative = len(negative_tweets) / total_pro
    print("probability_negative",probability_negative)
    return probability_negative
```

FIGURE 4.4 – Indication de la probabilité antérieure des classes du dataset "Sentiment140".

```
def totalClass(negative_tweets, positive_tweets, neutral_tweets):
    total_pro = len(negative_tweets) + len(positive_tweets) + len(neutral_tweets)
    print("total",total_pro)
    return total_pro
def ProbabilityClassP(positive_tweets, total_pro):
    probability_positive = len(positive_tweets) / total_pro
    print("probability_positive",probability_positive)
    return probability_positive
def ProbabilityClassN(negative_tweets, total_pro):
    probability_negative = len(negative_tweets) / total_pro
    print("probability_negative",probability_negative)
    return probability_negative
def ProbabilityClassNT(neutral_tweets, total_pro):
    probability_neutral = len(neutral_tweets) / total_pro
    print("probability_neutral",probability_neutral)
    return probability_neutral
```

FIGURE 4.5 – Indication de la probabilité antérieure des classes du dataset "US Airline".

Ensuite, nous devons convertir les mots en nombres, c'est l'extraction des caractéristiques qui peuvent être facilement créées à l'aide de la fonction *CountVectorizer* de sklearn. Les lignes de code ci-dessous définissent la base d'extraction :

```
from sklearn.feature_extraction.text import CountVectorizer
def freqMotParClass(dataset,p, n):
    cvector = CountVectorizer(min_df = 0.0, max_df = 1.0)
    cvector.fit(dataset.text)
    n = cvector.transform(dataset[dataset.sentiment == 0].text)
    p = cvector.transform(dataset[dataset.sentiment == 4].text)
    neg_words = n.sum(axis=0)
    neg_words_freq = [(word, neg_words[0, idx]) for word, idx in cvector.vocabulary_.items()]
    neg_tf = pd.DataFrame(list(sorted(neg_words_freq, key = lambda x: x[1], reverse=True)),columns=['Terms','negative'])
    pos_words = p.sum(axis=0)
    pos_words_freq = [(word, pos_words[0, idx]) for word, idx in cvector.vocabulary_.items()]
    pos_words_tf = pd.DataFrame(list(sorted(pos_words_freq, key = lambda x: x[1], reverse=True)),columns=['Terms','positive'])
    neg_tf_df = neg_tf.set_index('Terms')
    pos_words_tf_df = pos_words_tf.set_index('Terms')
    term_freq_df = pd.concat([neg_tf_df,pos_words_tf_df],axis=1)
    term_freq_df['total'] = term_freq_df['negative'] + term_freq_df['positive']
    term_freq_df.sort_values(by='total', ascending=False)
    term_freq_df['propabilite_mot_pos'] = term_freq_df['positive'] * 1./term_freq_df['total']
    term_freq_df.sort_values(by='propabilite_mot_pos', ascending=False)
    term_freq_df['propabilite_mot_neg'] = term_freq_df['negative'] * 1./term_freq_df['total']
    term_freq_df.sort_values(by='propabilite_mot_neg', ascending=False)
    term_freq_df['total'] = term_freq_df['negative'] + term_freq_df['positive']
    term_freq_df.sort_values(by='total', ascending=False).to_csv("freq1568000.csv")
```

FIGURE 4.6 – Extraction des caractéristiques du dataset "Sentiment140".

```
from sklearn.feature_extraction.text import CountVectorizer
def freqMotParClass(dataset,p, n, nt):
    cvector = CountVectorizer(min_df = 0.0, max_df = 1.0)
    cvector.fit(dataset.text)
    n = cvector.transform(dataset [dataset['sentiment'].str.contains("negative")].text)
    p = cvector.transform(dataset [dataset['sentiment'].str.contains("positive")].text)
    nt = cvector.transform(dataset [dataset['sentiment'].str.contains("neutral")].text)
    neg_words = n.sum(axis=0)
    neg_words_freq = [(word, neg_words[0, idx]) for word, idx in cvector.vocabulary_.items()]
    neg_tf = pd.DataFrame(list(sorted(neg_words_freq, key = lambda x: x[1], reverse=True)),columns=['Terms','negative'])
    pos_words = p.sum(axis=0)
    pos_words_freq = [(word, pos_words[0, idx]) for word, idx in cvector.vocabulary_.items()]
    pos_words_tf = pd.DataFrame(list(sorted(pos_words_freq, key = lambda x: x[1], reverse=True)),columns=['Terms','positive'])
    neut_words = nt.sum(axis=0)
    neut_words_freq = [(word, neut_words[0, idx]) for word, idx in cvector.vocabulary_.items()]
    neut_words_tf = pd.DataFrame(list(sorted(neut_words_freq, key = lambda x: x[1], reverse=True)),columns=['Terms','neutral'])
    neg_tf_df = neg_tf.set_index('Terms')
    pos_words_tf_df = pos_words_tf.set_index('Terms')
    neut_words_tf_df = neut_words_tf.set_index('Terms')
    term_freq_df = pd.concat([neg_tf_df,pos_words_tf_df,neut_words_tf_df],axis=1)
    term_freq_df['total'] = term_freq_df['negative'] + term_freq_df['positive'] + term_freq_df['neutral']
    term_freq_df.sort_values(by='total', ascending=False)
    term_freq_df['probabilite_mot_pos'] = term_freq_df['positive'] * 1./term_freq_df['total']
    term_freq_df.sort_values(by='probabilite_mot_pos', ascending=False)
    term_freq_df['probabilite_mot_neg'] = term_freq_df['negative'] * 1./term_freq_df['total']
    term_freq_df.sort_values(by='probabilite_mot_neg', ascending=False)
    term_freq_df['probabilite_mot_neut'] = term_freq_df['neutral'] * 1./term_freq_df['total']
    term_freq_df.sort_values(by='probabilite_mot_neut', ascending=False)
    term_freq_df['total'] = term_freq_df['negative'] + term_freq_df['positive'] + term_freq_df['neutral']
    term_freq_df.sort_values(by='total', ascending=False).to_csv("freq14347.csv")
```

FIGURE 4.7 – Extraction des caractéristiques du dataset "US Airline".

Nous terminons avec les tests du classification, les figures *(*FIGURE 4.8 *et* FIGURE4.9*)* illustrent le code appliqué pour les classifications afin d'avoir les résultats des polarités des tweets entrés.

```
def nouvtt(probapos,propaneg):
    towrow = pd.read_csv("freq1568000.csv",encoding="ISO-8859-1")
    towrow.rename(columns={'Unnamed: 0': 'terms'}, inplace=True)
    prod = 1
    prod1 = 1
    phrase = input("Entrer un tweet à classifier: ")
    phraseclean = tweet_cleaner(phrase)
    print("le tweet après le prétraitement: ","<<",phraseclean,">>.")
    ph = phraseclean.split(" ")
    for i in range (len(ph)):
        for index, row in towrow.iterrows():
            if(ph[i] == row['terms']):
                prod = prod * towrow.propabilite_mot_neg[index]
                prod1 = prod1 * towrow.propabilite_mot_pos[index]
    prodneg = prod* propaneg
    prodpos = prod1* probapos
    print("la probabilité négative du tweet","<<",phrase,">>",":",prodneg)
    print("la probabilité positive du tweet","<<",phrase,">>",":",prodpos)
    if prodneg < prodpos :
        print("La classe du tweet","<<",phrase,">>","est: positive")
    else:
        print("La classe du tweet","<<",phrase,">>","est: négative")
```

FIGURE 4.8 – Classification du dataset "Sentiment140".

```
def nouvtt(probapos,probaneg,probaneut):
    towrow = pd.read_csv("freq14347.csv",encoding="ISO-8859-1")
    towrow.rename(columns={'Unnamed: 0': 'terms'}, inplace=True)
    prod =  1
    prod1 = 1
    prod2 = 1
    phrase = input("Entrer un tweet à classifier: ")
    phraseclean = tweet_cleaner(phrase)
    print("le tweet après le prétraitement: ","<<",phraseclean,">>.")
    ph = phraseclean.split(" ")
    for i in range (len(ph)):
        for index, row in towrow.iterrows():
            if(ph[i] == row['terms']):
                prod1 = prod1 * towrow.probabilite_mot_pos[index]
                prod = prod * towrow.probabilite_mot_neg[index]
                prod2 = prod2 * towrow.probabilite_mot_neut[index]
    prodpos = prod1 * probapos
    prodneg = prod * probaneg
    prodneut = prod2 * probaneut
    print("la probabilité positive du tweet","<<",phrase,">>",":",prodpos)
    print("la probabilité négative du tweet","<<",phrase,">>",":",prodneg)
    print("la probabilité neutre du tweet","<<",phrase,">>",":",prodneut)
    if (prodneut > prodneg) & (prodneut > prodpos):
        print("La classe du tweet","<<",phrase,">>","est: neutre")
    elif (prodneg > prodneut) & (prodneg > prodpos):
        print("La classe du tweet","<<",phrase,">>","est: négative")
    else :
        print("La classe du tweet","<<",phrase,">>","est: positive")
```

FIGURE 4.9 – Classification du dataset "US Airline".

4.6 Évaluation

4.6.1 Matrice de confusion

La matrice de confusion *(voir la* TABLE 4.2*)* permet la prédiction sur un problème de classification. Les prédictions justes et injustes sont réparties par classe. Les résultats sont ainsi comparés avec les valeurs réelles.

Cette matrice permet de comprendre de quelle façon le modèle de classification est trouble lorsqu'il effectue des prédictions.

		Valeurs de prédictions	
		Positive	Negative
Valeurs réelles	Positive	VP	FP
	Negative	FN	VN

TABLE 4.2 – Matrice de confusion.

En référence à la matrice de confusion de la TABLE 4.2, les paramètres d'évaluation

peuvent être définis comme suit :

— Vrai positive VP : classe positive considérée positive.

— Vrai négative VN : classe négative considérée négative.

— Faux positive FP : classe négative considérée positive.

— Faux négative FN : classe positive considérée négative.

Accuracy, précision, et le rappel *(Recall)* ont été utilisés comme paramètres dans l'évaluation de l'algorithme de classification naïve bayes et notre proposition.

1. Rappel *(Recall)* : de ceux qui existent, combien l'algorithme a pu trouver.

$$Rappel = \frac{VP}{VP+FN}$$

2. Précision : de ceux que l'algorithme a pu classer, combien sont corrects.

$$Precision = \frac{VP}{VP+FP}$$

3. Accuracy : pourcentage de prédictions correctes.

$$Accuracy = \frac{VP+VN}{VP+VN+FP+FN}$$

4.6.2 Résultat d'évaluation

Les résultats obtenus par la méthode NB

— ***Dataset : "Sentiment140" :***

La FIGURE 4.10 montre l'évolution de temps en fonction de la taille du dataset, nous constatons que la valeur du temps surcroît avec l'augmentation de la taille du dataset, parce que le modèle effectue en premier lieu le prétraitement ensuite l'extraction des caractéristiques, calcul des probabilités et enfin la classification c'est pour cela qu'il prend de temps pour répondre.

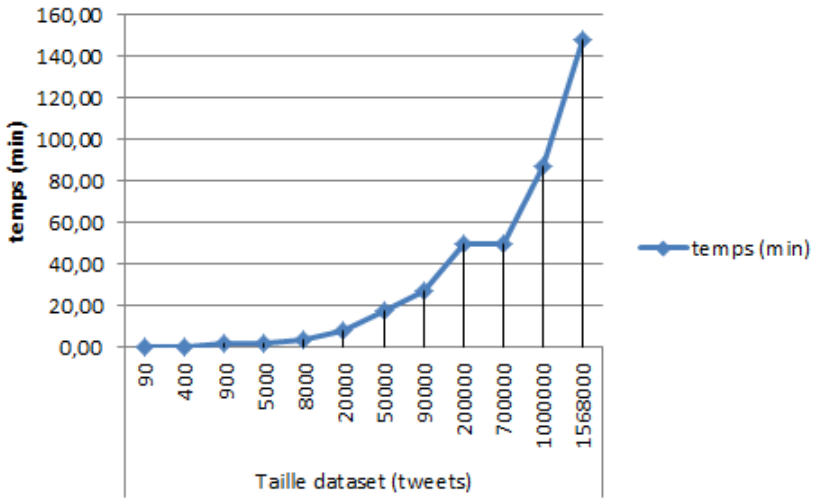


FIGURE 4.10 – Variation du temps en fonction de la taille du dataset "Sentiment140".

La FIGURE 4.11 présente l'écart entre les deux probabilités positive et négative en fonction de la taille du dataset, nous remarquons tout d'abord une grande distance entre les deux probabilités au niveau de la taille 900, puisque les tweets sont distribués aléatoirement pour chaque taille, ainsi pour la taille 900 la présence des tweets négative est plus élevée par rapport aux tweets positifs c'est pour cela la probabilité positive est inférieure à la probabilité négative.

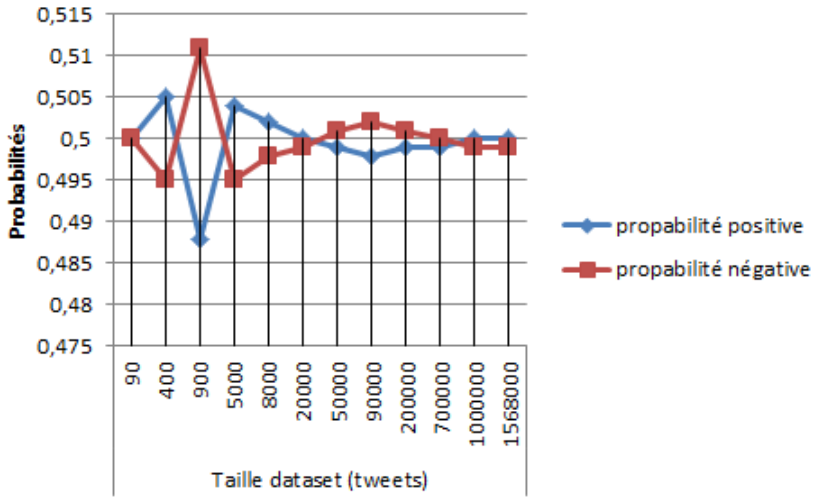


FIGURE 4.11 – Variation des probabilités en fonction de la taille du dataset "Sentiment140".

La FIGURE 4.12 présente la variation du rappel, précision et accuracy en fonction du nombre de tests, nous remarquons tout d'abord la valeur cernée de la précision dans les 10 premiers tests vaut 1 *(les prédictions sont justes à 100%)*, parce que le faux positive est nulle, pour le rappel les prédictions justes sont à 80%, aussi pour le pourcentage de prédictions justes *(accuracy)* est à 90%, donc nous déduisons que le classificateur est plus performant dans les 10 premiers tests, ensuite pour le reste des tests nous remarquons une légère diminution des valeurs du rappel, précision et accuracy cela dû à cause de l'augmentation du taux de faux positive et faux négative *(le grand pourcentage des prédictions erronées)*.

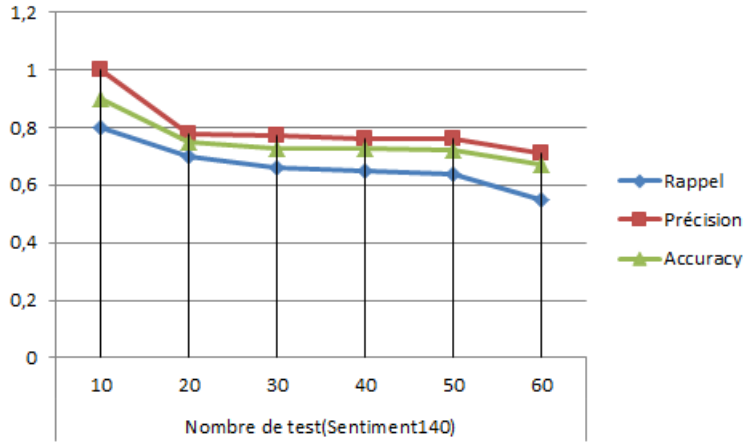


FIGURE 4.12 – Variation du rappel, précision et accuracy en fonction du nombre de tests "Sentiment140".

— ***Dataset : "US Airline" :***

Après l'application de l'algorithme sur ce dataset, nous remarquons tout d'abord que sa taille arrive à 14347 tweets, d'où le classificateur prend quelques minutes à rendre les résultats, quant aux probabilités nous remarquons aussi que la probabilité négative est supérieure par rapport aux autres *(positive et neutre)*, parce que la plupart des tweets dans ce dataset ont une polarité négative.

La FIGURE 4.13 montre la variation du rappel, précision et le pourcentage de prédictions correctes en fonction du nombre de tests effectués, ainsi nous remarquons que les valeurs cernées des trois paramètres sur le point des 40 tests sont maximales, car les prédictions fausses sont supérieures aux prédictions justes *(faux positive inférieur au faux négative)*, quant au pourcentage des prédictions justes nous remarquons une augmentation de ses valeurs jusqu'à 70%.

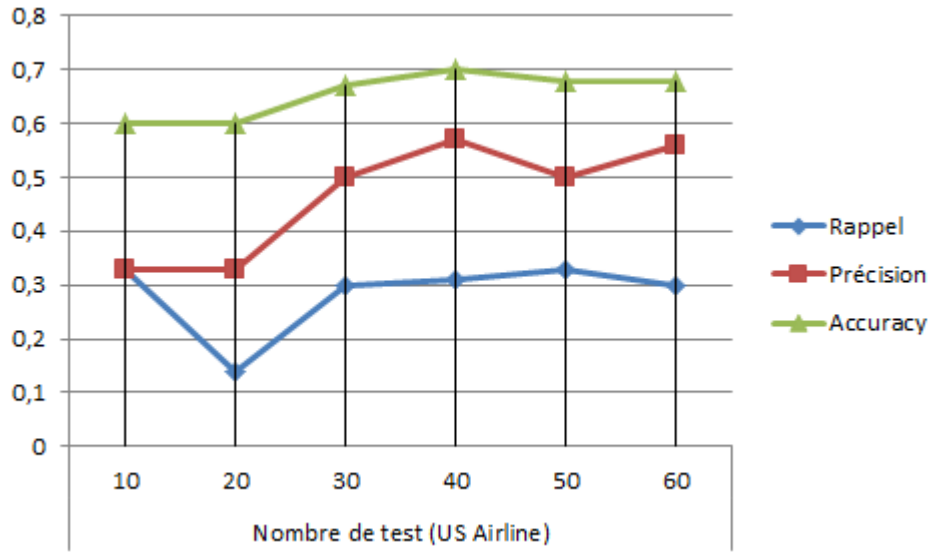


FIGURE 4.13 – Variation du rappel, précision et accuracy en fonction du nombre de tests "US Airline".

Les résultats obtenus par la méthode basée dictionnaire AFINN

Nous avons appliqué la méthode basée sur le dictionnaire AFINN sur les deux dataset "Sentiment140" et "US Airline". Nous présentons les résultats obtenus ci-après.

— **Dataset : "Sentiment140" :**

La FIGURE 4.14 présente une variation du rappel, précision et accuracy en fonction du nombre de tests, nous remarquons tout d'abord dans les 30 premiers tests les valeurs des paramètres cernées sont supérieures par rapport aux autres tests. Dans les 10 premiers tests effectués, la valeur du rappel arrive jusqu'à 60% *(à cause du nombre de faux négative)*, quant à la valeur de précision qui arrive jusqu'à 75% *(à cause du nombre de faux positive)* et le pourcentage des prédictions justes *(accuracy)* estime une valeur de 70%. En augmentant le nombre des tests, nous remarquons une diminution de ces valeurs et leurs stabilité, car le faux positive et négative augmentent. Cela signifie que l'algorithme se trompe dans la plupart de ses prédictions et il arrive à un pourcentage de 50% pour chacun des paramètres. Donc dans les 10 premiers tests l'algorithme rend des résultats meilleurs.

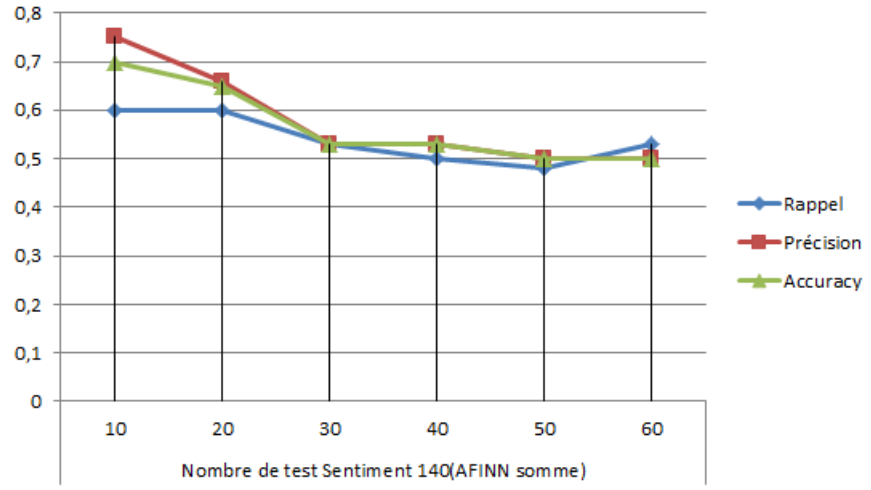


FIGURE 4.14 – Variation du rappel, précision et accuracy en fonction du nombre de test (Somme-AFINN) de dataset "Sentiment140".

La FIGURE 4.15 montre le changement de valeurs des métriques d'évaluation *(rappel, précision et accuracy)* en fonction du nombre de tests effectués. Tout d'abord, nous constatons que dans les 30 premiers tests les valeurs des métriques sont égales car les valeurs du faux positive et faux négative sont égales, quant aux reste des tests en faisant comparaison avec la FIGURE 4.14, nous remarquons que ces valeurs sont supérieures à celles cernées par l'algorithme de la somme.

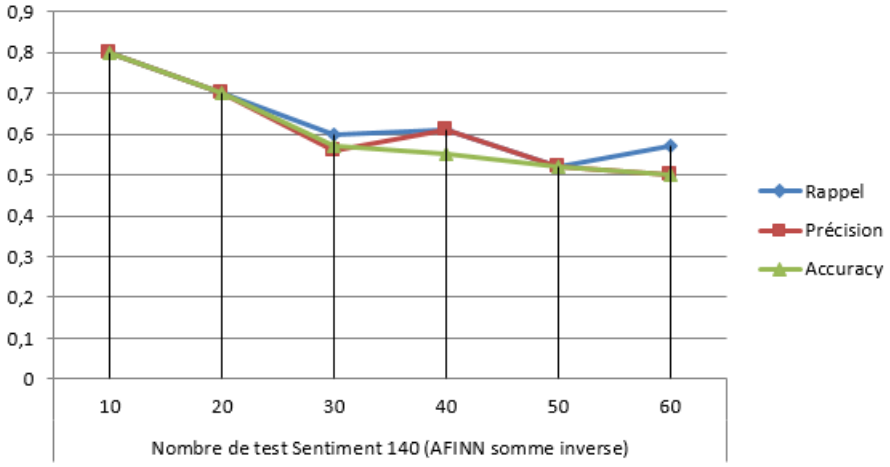


FIGURE 4.15 – Variation du rappel, précision et accuracy en fonction du nombre de tests (Somme inverse-AFINN) de dataset "Sentiment140".

La FIGURE 4.16 montre la variation des métriques d'évaluation en fonction de nombre de tests, nous remarquons tout d'abord que le rappel arrive à une valeur maximale dans les 30 premiers tests effectués à cause du nombre minimale du faux négative *(prédictions positives erronées)*, en comparaison avec la FIGURE 4.14, nous constatons que les valeurs rendues par l'algorithme de la somme prenant en considération les émoticônes sont supérieures aux valeurs rendues par l'algorithme de la somme des polarités.

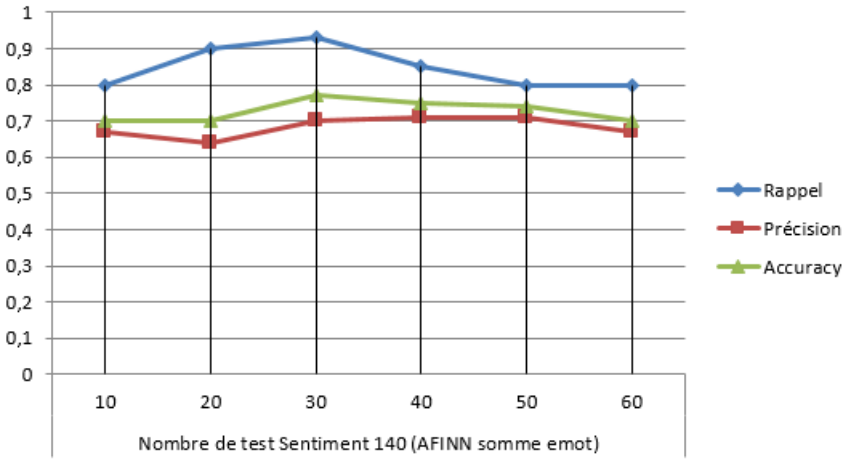


FIGURE 4.16 – Variation du rappel, précision et accuracy en fonction du nombre de tests (Somme avec emoticônes-AFINN) de dataset "Sentiment140".

La FIGURE 4.17 montre la variation des métriques d'évaluation en fonction de nombre de tests, nous remarquons tout d'abord dans les 30 premiers tests que le rappel arrive à une valeur maximale 1, cela signifie que les prédictions justes sont à 100% *(les valeurs du faux négative sont nulles)* et elle diminue dans le reste des tests effectués, quant à la précision et accuracy arrivent à des valeurs *(0.83, 0.90 respectivement)* dans les 10 premiers tests et ils diminuent dans le reste des tests effectués, en faisant comparaison avec la FIGURE 4.16 nous remarquons que les valeurs des métriques de la somme inverse avec émoticônes sont supérieures à celles retournées par la somme avec émoticônes.

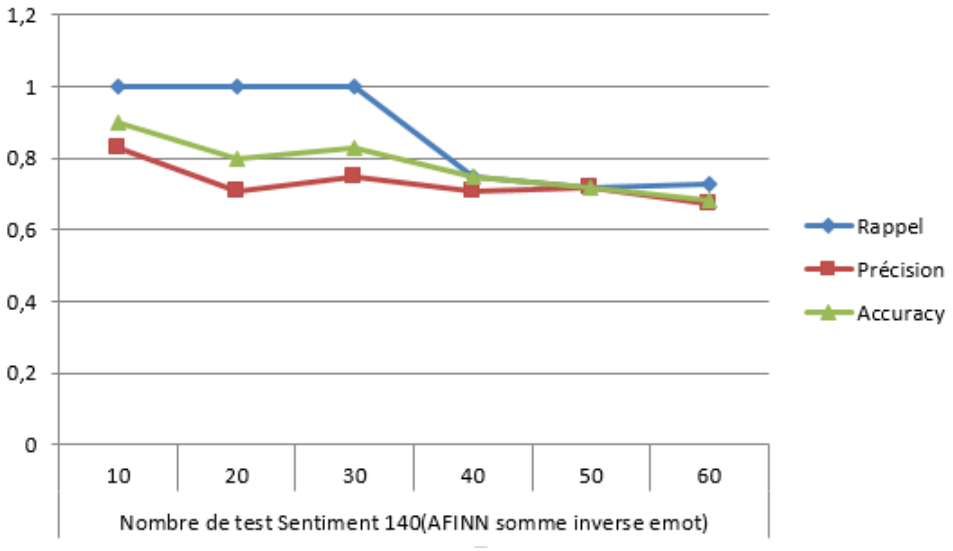


FIGURE 4.17 – Variation du rappel, précision et accuracy en fonction du nombre de tests (Somme inverse avec emoticônes-AFINN) de dataset "Sentiment140".

— ***US Airline :***

La FIGURE 4.18 présente la variation des métriques d'évaluation en fonction de nombre de tests, ainsi nous remarquons dans les 30 premiers tests que le rappel et la précision sont égaux à cause de l'égalité du faux positive et faux négative et dans le reste des tests effectués, le rappel est bien supérieur à la précision *(faux négative est inférieur au faux positive)*, quant au pourcentage des prédictions justes il arrive à une valeur maximale *(0.67)* dans les 30 tests effectués, donc l'algorithme trouve les prédictions justes à 67% dans les 30 tests, elle prend presque des valeurs stables dans le reste des tests effectués.

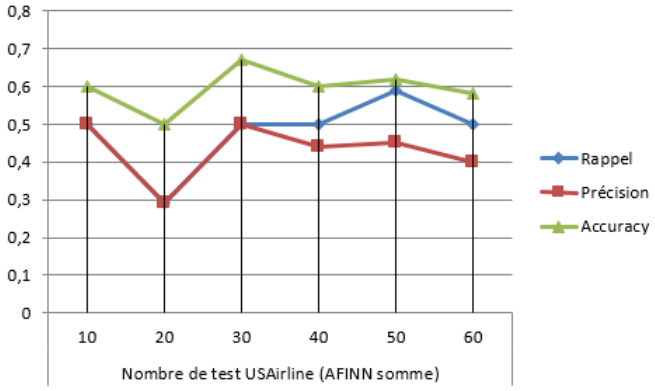


FIGURE 4.18 – Variation du rappel, précision et accuracy en fonction du nombre de tests (Somme-AFINN) de dataset "US Airline".

La FIGURE 4.19 présente le changement des valeurs des paramètres d'évaluation en fonction du nombre de tests, tout d'abord nous remarquons que les taux dans les 50 tests effectués arrivent à des valeurs maximales, ainsi elles diminuent dans le reste des tests. Aussi, nous remarquons que dans les 20 premiers tests le rappel et la précision sont égaux à cause de l'égalité du faux positive et faux négative, ainsi dans le reste des tests le rappel est supérieur à la précision *(faux négative est inférieur au faux positive)*, quant au pourcentage des prédictions justes arrive à une valeur maximale 63% dans les 30 premiers tests.

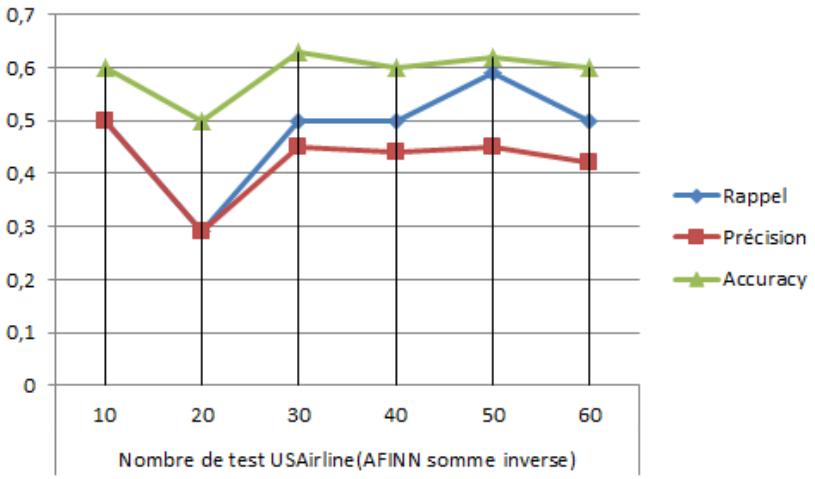


FIGURE 4.19 – Variation du rappel, précision et accuracy en fonction du nombre de tests (Somme inverse-AFINN) de dataset "US Airline".

La FIGURE 4.20 présente la variation des métriques d'évaluation en fonction de nombre de tests, nous remarquons tout d'abord que la valeur du rappel est constante à 1 durant les 60 tests *(signifie la valeur du faux négative est nulle)* et que l'algorithme a pu trouver toutes les prédictions (100%) dans ceux qui existe. Pour la précision, elle atteint des valeurs moyennes (0.5) *(cela à cause de l'augmentation du faux positive)* quant au pourcentage des prédictions justes atteint sa valeur maximale dans les 20 tests effectués (0.7) donc l'algorithme a pu trouver ses prédictions à 70%.

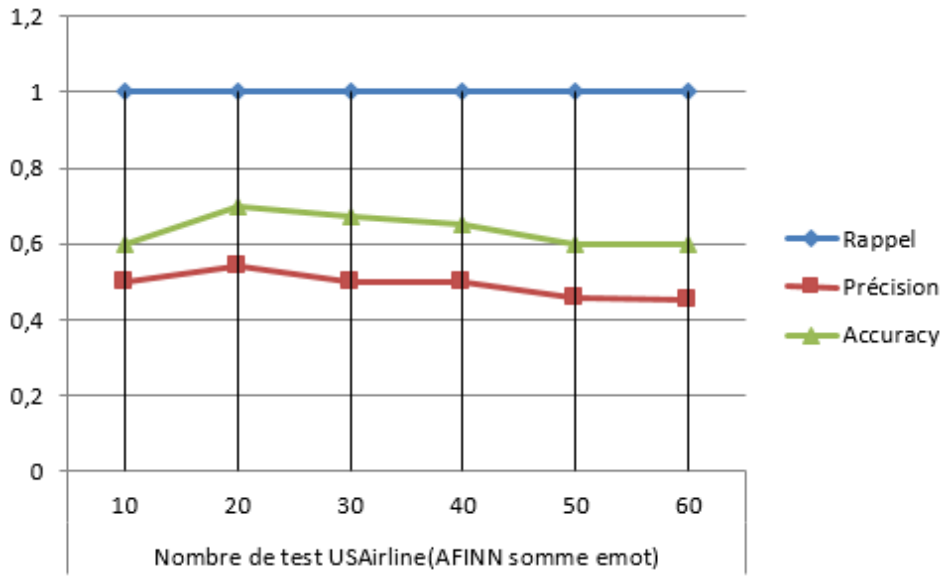


FIGURE 4.20 – Variation du rappel, précision et accuracy en fonction du nombre de tests (Somme avec émoticônes-AFINN) de dataset "US Airline".

La FIGURE 4.21 présente la variation des métriques d'évaluation en fonction de nombre de tests, nous remarquons tout d'abord que la valeur du rappel est constante à 1 durant les 60 tests *(signifie la valeur du faux négative est nulle)* et que l'algorithme a pu trouver toutes les prédictions (100%) dans ceux qui existe. Pour la précision, elle atteint des valeurs moyennes (0.5) *(cela à cause de l'augmentation du faux positive)* quant au pourcentage des prédictions justes il atteint sa valeur maximale dans les 20 tests effectués (0.7) donc l'algorithme a pu trouver ses prédictions à 70%.

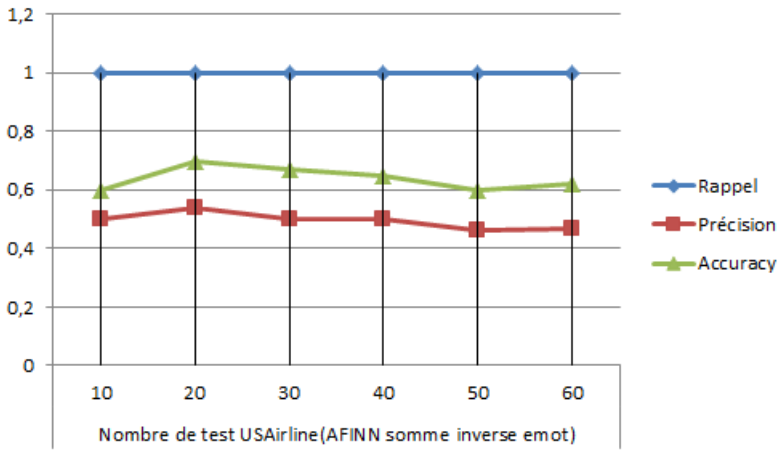


FIGURE 4.21 – Variation du rappel, précision et accuracy en fonction du nombre de tests (Somme inverse avec émoticônes-AFINN) de dataset "US Airline".

Les résultats obtenus par l'utilisation de "Sentiment Analyzer"

Nous avons utilisé l'outil d'analyse de sentiments "Sentiment Analyzer" pour comparer les résultats obtenus.

— **Dataset : "Sentiment140" :**

La FIGURE 4.22 montre la variation des paramètres d'évaluation en fonction de nombre de tests, nous remarquons que dans les 30 premiers tests les valeurs de ces métriques augmentent jusqu'au maximum *(diminution du faux positives et négatives)*, mais dans les autres tests nous touchons une diminution et une stabilité de ces valeurs *(augmentation du faux positives et négatives)*.

FIGURE 4.22 – Variation du rappel, précision et accuracy en fonction du nombre de tests de dataset "Sentiment140".

— **Dataset : "US Airline" :**

La FIGURE 4.23 montre la variation des paramètres d'évaluation en fonction de nombre de tests, nous remarquons que le rappel estime sa valeur maximale dans les 10 premiers tests à cause du nombre de faux négatifs qui vaut 3, quant à la précision qui estime sa valeur maximale dans les 10 et 40 tests à cause du nombre de faux positives, pour le pourcentage des prédictions justes s'augmente avec l'augmentation de nombre de tests.

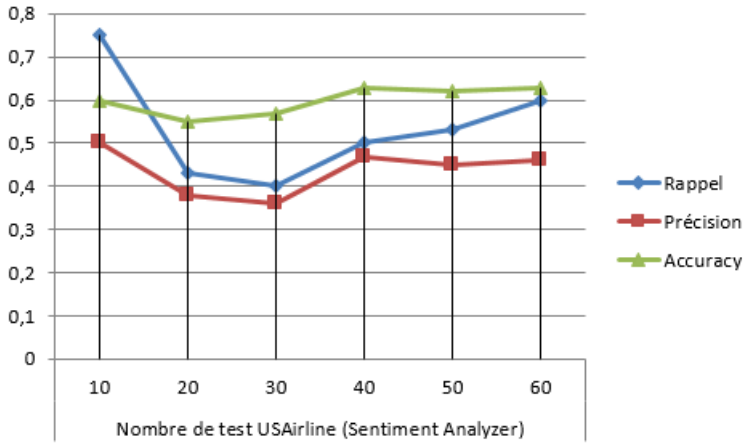


FIGURE 4.23 – Variation du rappel, précision et accuracy en fonction du nombre de tests de dataset "US Airline".

4.7 Discussion

Dans cette partie, nous présentons la superposition des graphes *(faux positive et faux négative)* des modèles *(NB, la solution améliorée et l'outil d'analyse existé)*, et nous déduirons à la fin quel modèle jugé le plus performant, et comparer les résultats avec ceux obtenus avec l'utilisation de l'outil d'analyse des sentiments existé.
Comme nous l'avons indiqué précédemment, que nous avons appliqué les modèles sur deux datasets "Sentiment140" et "US Airline", dans ce qui suit nous montrons les résultats obtenus par chaque dataset.

1. **Modèle NB :**

   Selon la FIGURE 4.24, nous remarquons que les valeurs de faux négative du "US Airline" sont supérieures par rapport aux celles du "Sentiment140", ce qui signifie que NB trouve ses prédictions positives justes dans "Sentiment140", quant aux valeurs de faux positive du "US Airline" sont inférieures aux celles du "Sentiment140", ce qui signifie que NB trouve ses prédictions négatives justes dans "US Airline".

   À partir de là, nous déduisons que le modèle NB est performant sur le "Sentiment140" qui est un dataset orienté social *(contient des différents sujets)*.

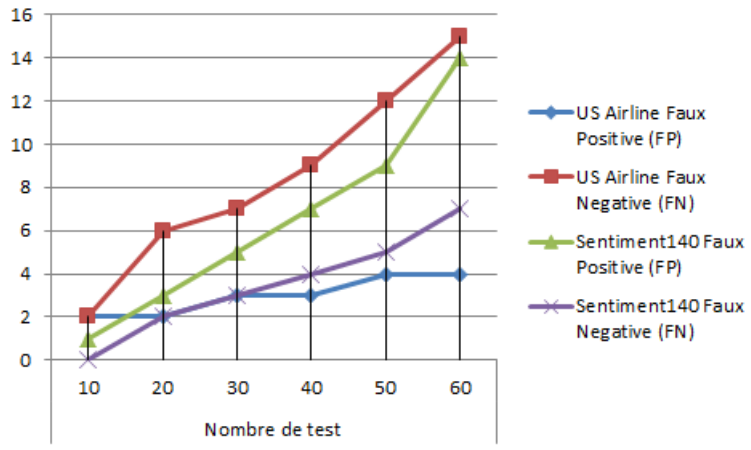


FIGURE 4.24 – Superposition du taux de faux positive et faux négative de modèle NB.

2. **Modèle basée dictionnaire AFINN :**

— ***Dataset : "Sentiment140" :***

Selon la FIGURE 4.25, nous remarquons que les valeurs du faux positive de la somme inverse avec émoticônes sont inférieures aux autres courbes, cela signifie que la précision est supérieure d'où l'algorithme de la somme inverse avec émoticônes a pu classer correctement ses prédictions.

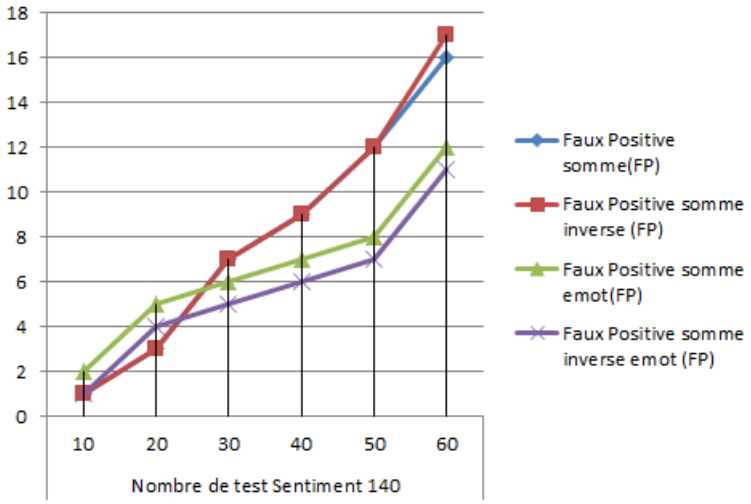


FIGURE 4.25 – Superposition du taux de faux positive de dataset "Sentiment140".

Selon la FIGURE 4.26, nous remarquons que tout d'abord dans les 30 premiers tests les valeurs de faux négative sont nulles dans la somme inverse avec émoticônes ce qui signifie que cette formule rend des résultats meilleurs par rapport aux autres, quant aux autres tests nous notons que la somme avec émoticônes est légèrement inférieure à la somme inverse avec émoticônes, d'où nous constatons que l'algorithme de la somme inverse avec émoticônes rend des résultats meilleurs.

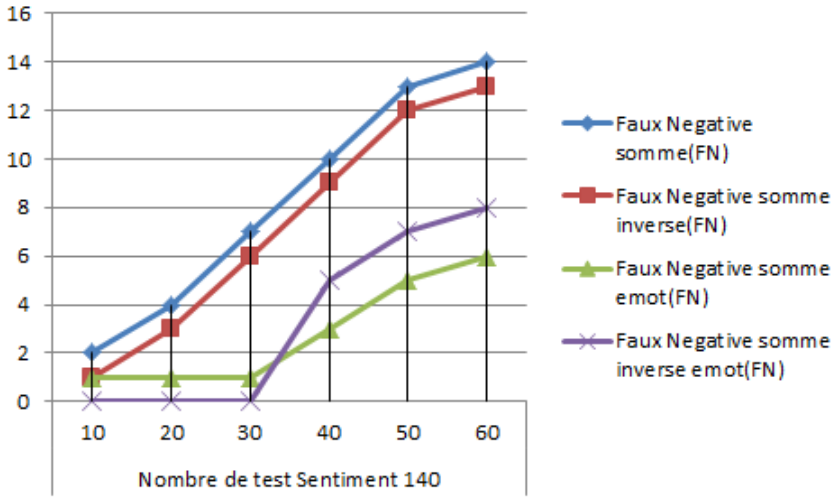


FIGURE 4.26 – Superposition du taux de faux négative de dataset "Sentiment140".

— ***Dataset : "US Airline" :***

Selon la FIGURE 4.27, nous remarquons que les courbes qui présentent les valeurs du faux positive des algorithmes qui utilisent la somme et la somme inverse sont inférieures aux autres, donc nous constatons que ces algorithmes rendent des résultats meilleurs en terme de précision par rapport aux autres.

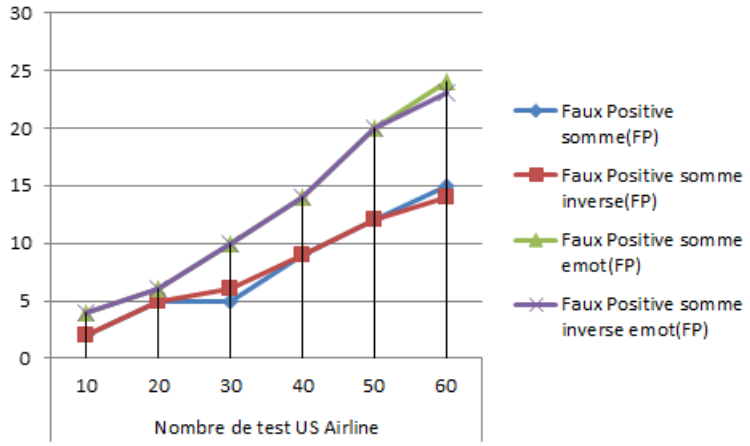


FIGURE 4.27 – Superposition du taux de faux positive de dataset "US Airline".

Selon la FIGURE 4.28, nous remarquons que les courbes qui présentent les valeurs du faux négative des algorithmes qui utilisent la somme avec émoticônes et la somme inverse avec émoticônes sont nulles *(dans ceux que les algorithmes ont pu tous trouver)*, donc nous constatons que ces algorithmes rendent des résultats meilleurs par rapport aux autres.

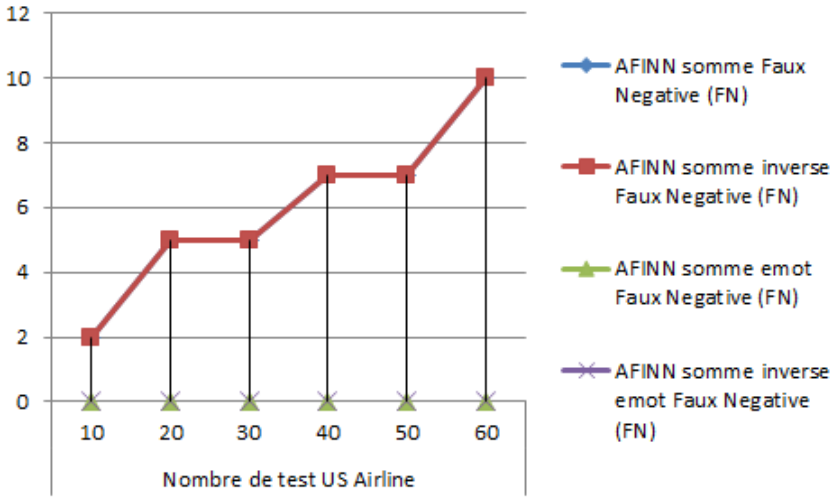


FIGURE 4.28 – Superposition du taux de faux négative de dataset "US Airline".

3. **Outil "Sentiment Analyzer" :**

   Selon la FIGURE 4.29, nous remarquons que les courbes qui présentent les valeurs du faux négatives du dataset "Sentiment140" sont inférieures aux autres dans les 40 tests effectués, donc nous constatons que l'outil "Sentiment Analyzer" rend des résultats meilleurs sur ce dataset en comparaison avec le dataset "US Airline".

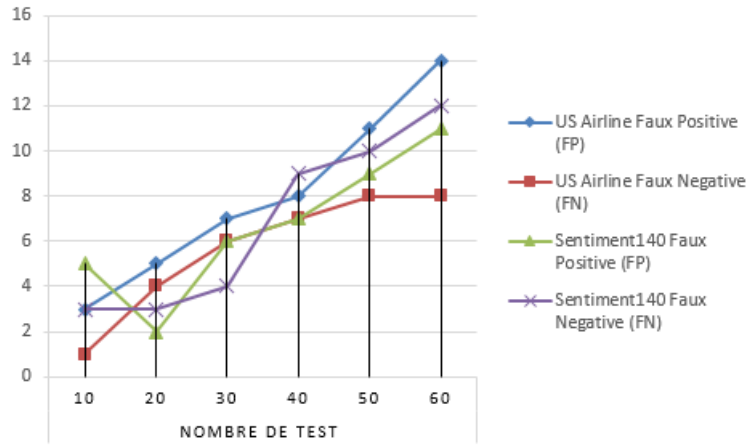


FIGURE 4.29 – Superposition du taux de faux négative et faux positive sur différents datasets.

Modèles	Dataset	Rappel	Précision	Accuracy
NB	Sentiment140	0.8	1	0.9
	US Airline	0.33	0.57	0.7
Dictionnaire somme	Sentiment140	0.6	0.75	0.7
	US Airline	0.59	0.5	0.67
Dictionnaire somme inverse	Sentiment140	0.8	0.8	0.8
	US Airline	0.59	0.5	0.63
Dictionnaire somme émoticônes	Sentiment140	0.93	0.71	0.77
	US Airline	1	0.54	0.7
Dictionnaire somme inverse émoticônes	Sentiment140	1	0.83	0.9
	US Airline	1	0.54	0.7
Outil Sentiment Analyzer	Sentiment140	0.73	0.65	0.67
	US Airline	0.75	0.5	0.63

TABLE 4.3 – Meilleurs résultats de modèle NB, dictionnaire et l'outil "Sentiment Analyzer" sur les deux datasets.

La TABLE 4.3 illustre les meilleurs résultats obtenus en utilisant les deux datasets : Sentiment140 et US Airline pour les différents modèles.

4.8 Conclusion

Dans ce chapitre, nous avons illustré les différents langages et outils de développement utilisés pour implémenter les deux méthodes, le classificateur naïve bayes et la solution proposée. Nous avons aussi présenté les paramètres d'évaluation appliqués sur ces algorithmes, avec un objectif de les tester dans le coté de fonctionnement et la faisabilité des réponses sur l'analyse des sentiments.

Les résultats d'évaluation obtenus indiquent que le modèle NB est plus performant en l'appliquant sur "Sentiment140", et le modèle basé dictionnaire est plus performant en l'appliquant sur "Sentiment140" avec la somme inverse des polarités prenant en considération les émoticônes, même si nous les avons comparer avec les résultats obtenus lors d'utilisation de l'outil d'analyse de sentiments "Sentiment Analyzer".

Conclusion générale et perspectives

Le domaine de l'analyse des sentiments est un nouvel axe de recherche passionnant en raison du grand nombre d'applications du monde réel dans lesquelles la découverte de l'opinion de la population est importante pour une meilleure prise de décision.

Afin d'atteindre notre objectif, nous avons essayé de comprendre les concepts et comment appliquer les méthodes naïve bayes et l'approche basée dictionnaire AFINN à notre problématique.

Ainsi, nous avons présenté la conception de notre système d'analyse des sentiments en présentant toutes les étapes des processus. Nous avons indiqué les étapes importantes qui sont le prétraitement des données, ensuite nous avons utilisés le classificateur naïve bayes pour classifier les données prétraitées. Aussi, nous avons exploité le dictionnaire AFINN *(mots avec leurs polarités)* afin de classifier les tweets en trois classes : positive, négative et neutre.

D'après la comparaison entre les deux datasets sur lesquels nous avons appliqué les méthodes de classification naïve bayes, l'approche basée sur dictionnaire et l'outil "Sentiment Analyzer" avec l'évaluation des résultats obtenus, nous avons remarqué que le dataset "Sentiment140" donne des meilleurs résultats en terme d'accuracy *(pourcentage des prédictions justes)* dans les méthodes.

Comme perspectives nous pouvons citer :

— Appliquer ces méthodes sur d'autres ensembles de données.

— Inclure dans le dictionnaire un certain nombre de mots fréquemment utilisés sur Internet avec leur score de polarité, tels que LOL *(rire fort)*.

— Utiliser les approches présentées pour classer les sentiments dans d'autres langues.

— Tester d'autres modèles de classification .

Bibliographie

[1] Rémi Bachele. Réseaux sociaux. Centrale LILLE, 4 mai 2016.

[2] https ://www.reseau-canope.fr/savoirscdi/cdi-outil-pedagogique/reflexion/les-reseaux-sociaux-au-cdi/typologie-des-reseaux-sociaux.html, consulté le 27/11/2019.

[3] https ://lewebpedagogique.com/jddreseauxsociaux/2011/04/15/les-reseaux-sociaux%c2%a0-caracteristiques-donnees-chiffrees-principaux-reseaux-sociaux-connus/, consulté le 27/11/2019.

[4] PMTIC. Communication medias sociaux. LabSET ULg, 2017.

[5] Elsevier. Social media guide for google+.

[6] SOUALMI Samiha BENSALEM Khadidja. *Détection des rumeurs dans les réseaux sociaux*. Université Abderrahmane Mira, Bejaia, Mémoire de fin de cycle, 2016/2017.

[7] Mary Grammont. Wizbii, la 1ère start-up à obtenir le pass french tech à grenoble. Communiqué de presse.

[8] http ://www.youthvillage.co.za/2013/10/advantages-disadvantages-social-networks/, consulté le 27/11/2019.

[9] https ://www.murielle-cahen.com/publications/facebook-reseaux.asp, consulté le 25/11/2019.

[10] https ://fr.m.wikipedia.org/wiki/sentiment, consulté le 27/11/2019.

[11] Tiejian Luo, Su Chen, Guandong Xu, and Jia Zhou. *Sentiment Analysis*, pages 53–68. Juin 2013.

[12] Grzegorz Dziczkowski. *Analyse des sentiments : système autonome d'exploration des opinions exprimées dans les critiques cinématographiques*. PhD thesis, École Nationale Supérieure des Mines de Paris, 2008.

[13] Benoît Matthieu, Gaëtan and Stépaliane. Fouille de réseaux sociaux en ligne. Synthèse bibliographique sur le Data Mining, 2012.

[14] François Yvon. Une petite introduction au traitement automatique des langue naturelles.

[15] Benmammar. L'apprentissage automatique. Université de Tlemcen.

[16] Ludovic Arnold, Sébastien Rebecchi, Sylvain Chevallier, and Hélène Paugam-Moisy. An introduction to deep learning. volume 1, pages 477–488, January 2011.

[17] Hoda Karashy Walaa Medhat, Ahmed Hassan. Sentiment analysis algorithms and applications : A survey. Ain Shams Engineering Journal, 2014.

[18] Narasinga Rao Kavya Suppala. Sentiment analysis using naïve bayes classifier. *International Journal of Innovative Technology and Exploring Engineering (IJITEE).*

[19] https ://fr.wikipedia.org/wiki/classifieur_lin%c3%a9aire, consulté le 27/11/2019.

[20] Joseph Awange, Bela Palancz, and Lajos Volgyesi. *Neural Networks*, pages 293–411. 01 2020.

[21] Padraig Cunningham and Sarah Delany. k-nearest neighbour classifiers. *Mult Classif Syst*, 04 2007.

[22] Samaneh Moghaddam and Martin Ester. Opinion digger : an unsupervised opinion miner from unstructured product reviews. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1825–1828. ACM, 2010.

[23] Ricco Rakotomalala. Le classifieur bayesien naif (modèle d'indépendance conditionnelle). Université Lumière Lyon 2.

[24] Ricco Rakotomalala. Svm support vector machine. Université Lumière Lyon 2.

[25] Wassim Lahbibi. Algorithme knn. Ecole supérieure de commerce, 2012-2013.

[26] Goudjil Ayyoub Hasni Bachir. *Détection d'opinion à partir de Twitter*. Université de Djilali Bounaama, Khemis Miliana, Mémoire Master, Juin 2019.

[27] Yi Nasukawa, Tetsuya and Jeonghee. Sentiment analysis : Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77. ACM, 2003.

[28] Prendinger Helmut Ishizuka Neviarouskaya, Alena and Mitsuru. Recognition of affect, judgment, and appreciation in text. In *Proceedings of the 23rd international conference on computational linguistics*, pages 806–814. Association for Computational Linguistics, 2010.

[29] Jose M Chenlo, Alexander Hogenboom, and David E Losada. Sentiment-based ranking of blog posts using rhetorical structure theory. In *International Conference on Application of Natural Language to Information Systems*, pages 13–24. Springer, 2013.

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

[31] Ram Mohana Monisha Kanakaraj and Reddy Guddeti. Nlp based sentiment analysis on twitter data using ensemble classifiers. 3rd International Conference on Signal Processing Communication and Networking (ICSCN), 2015.

[32] Erik Cambria, Björn Schuller, Bing Liu, Haixun Wang, and Catherine Havasi. Knowledge-based approaches to concept-level sentiment analysis. *IEEE intelligent systems*, 28(2) :12–14, 2013.

[33] Dr. M. B. Chandak Arti Buche and Akshay Zadgaonkar. Opinion mining and analysis : a survey. International Journal on Natural Language Computing (IJNLC), 2013.

[34] Debanjan Mahata Vivek Kumar Sigh. A clustering and opinion mining approach to socio-political analysis of the blogosphere. Computational Intelligence and Computing Research (ICCIC), 2010.

[35] Patrick Paroubek Alexander Pak. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, 2010.

[36] Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. Evaluation datasets for twitter sentiment analysis : a survey and a new dataset, the sts-gold. 2013.

[37] Finn Arup Nielsen. afinn project.

[38] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0 : an enhanced lexical resource for sentiment analysis and opinion mining.

[39] D Boullier and A Lohard. Opinion mining et sentiment analysis : Methodes et outils. OpenEdition, 2012.

[40] Medjdoubi Abdelkader. *L'analyse du sentiment utilisant le deep learning*. Université Dr.Tahar Moulay Saida, Mémoire Master, 2017/2018.

[41] Boey yew tung. *Twitter the next big lap or lapse*.

[42] Ouserir Amina Beghdad Abdelkrim. *Une approche Deep Learning pour l'analyse des Sentiments Sur Twitter*. Université de Djilali Bounaama, Khemis Miliana, Mémoire Master, 2018.

[43] http ://help.sentiment140.com/for-students, consulté le 27/11/2019.

[44] https ://www.w3schools.com/python/python_intro.asp, consulté le 26/11/2019.

[45] http ://eric.univ-lyon2.fr/ ricco/cours/slides/pf - distribution anaconda.pdf, consulté le 27/11/2019.

[46] https ://pandas.pydata.org/, consulté le 27/11/2019.

[47] https ://www.w3schools.com/python/python_regex.asp, consulté le 27/11/2019.

[48] https ://pypi.org/project/beautifulsoup4/, consulté le 27/11/2019.

[49] https ://www.nltk.org, consulté le 27/11/2019.

[50] https ://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.countvectorizer consulté le 27/11/2019.

Résumé

L'analyse des sentiments ou l'opinion mining est l'étude informatique des opinions, sentiments, attitudes et émotions exprimés dans un langage écrit. C'est l'un des domaines de recherche les plus actifs dans le traitement du langage naturel et l'extraction de texte au cours des dernières années. Ce problème a beaucoup de solutions proposées avec différentes techniques de classification de l'apprentissage machine. Nous avons utilisé l'une des technique de l'apprentissage supervisé et une autre méthode basée dictionnaire et nous avons appliqué certaines modifications afin d'avoir des meilleurs résultats.

Mots clés : analyse des sentiments, opinion mining, apprentissage machine, apprentissage supervisé, dictionnaire.

Abstract

Sentiment analysis or opinion mining is the computational study of people's opinions, sentiments, attitudes, and emotions expressed in written language. It is one of the most active research areas in natural language processing and text mining in recent years. This problem has many proposed solutions with different machine learning classification techniques. We used one of the supervised learning technique and another dictionary-based method and we applied some modifications in order to have better results.

Key words : Sentiment analysis, opinion mining, machine learning, supervised learning, dictionary.

تلخيص

تحليل المشاعر أو التنقيب عن الرأي هو الدراسة الحسابية لآراء الناس, والمشاعر, والمواقف, والعواطف المعبر عنها بلغة مكتوبة. إنها واحدة من أكثر مجالات البحث نشاطا في معالجة اللغات الطبيعية واستخراج النصوص في السنوات الأخيرة. هذه المشكلة لها العديد من الحلول المقترحة مع تقنيات تعلم الآلة المختلفة. استخدمنا أحد أساليب التلقين الذاتي وطريقة أخرى تعتمد على القاموس وطبقنا بعض التعديلات من أجل الحصول على نتائج أفضل.

الكلمات المفتاحية : تحليل المشاعر, التنقيب عن الرأي, تعلم الآلة, التلقين الذاتي, القاموس.