République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université A. Mira de Béjaïa

Faculté des Sciences Exactes

Département d'Informatique

Mémoire de Fin d'Etude

En vue de l'obtention d'un Master en Génie Logiciel

Thème

Analyse des sentiments sur les avis des clients dans le E-Commerce

Réalisé par M^{elles} :

ATMANIOU Siham

MILI Feriel

Soutenu le 06 Octobre devant le jury composé de :

Présidente :	M^{me} *BOUKERRAM* Samira	M.A.A Université de Béjaïa.
Examinatrice :	M^{me} *ADEL* Karima	M.C.A Université de Béjaïa.
Encadrante :	M^{me} *EL BOUHISSI* Houda	M.C.A Université de Béjaïa.

Promotion 2020 - 2021.

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 toute notre reconnaissance à Madame EL BOUHISSI Houda.
Nous la remercions de nous avoir encadré, orienté, aidé et conseillé.

Nous adressons nos sincères remerciements à tous les professeurs, intervenants et toutes les personnes qui par leurs paroles, leurs écrits, leurs conseils et leurs critiques ont guidé nos réflexions et ont accepté de nous rencontrer et de répondre à nos questions durant nos recherches.

Nous tenons à remercier également chacun des membres du jury pour l'intérêt porté à ce travail et d'avoir accepté de l'évaluer.

Dédicaces

Je remercie ma très chère mère, Khoukha, qui a toujours été là pour moi. Je remercie ma sœur Radia, et mon frère Dalim, pour leurs encouragements.

Enfin, je remercie mes amies Cycy, tata Nouria, Melissa, tata Samia, Sara, Hayat, Ryma, tata Houria, Mira, tata Fatiha, Widad, tata Ghania et tata Nadhira et tonton Tayeb qui ont toujours été là pour moi. Leur soutien inconditionnel et leurs encouragements ont été d'une grande aide.

Je souhaite personnellement remercier ma binôme et amie Siham, avec laquelle j'ai pris beaucoup de plaisir à travailler. Nous avons formé une belle équipe.

À la mémoire de mon oncle Karimou, mes grand-mères Chrifa et Khoukha et mon très cher grand-père Boualem, que Dieu les accueille dans son vaste paradis.
À tous ces intervenants, je présente mes remerciements, mon respect et ma gratitude.

Feriel

Dédicaces

Je dédie ce modeste travail :

À Ma très chère mère, pour ses sacrifices, son amour et son aide et à qui sa prière était le secret de mon succès.

À Mon très cher père, pour ses conseils et son soutien matériel et qui s'est toujours sacrifié pour me voir réussir.

À la mémoire de ma grand-mère Tassaadith et mon grand-père Abdu allah que Dieu les accueille dans son vaste paradis.

À mes grands-parents Tounes et Abdelkader.

À mes chers frères : Rachid, Wahid, Boubekeur et sa femme Rosa.

À mes chères sœurs : Kahina et Khadidja.

À mes neveux : Sérine, Mayas, Ayline et Ania.

À mes amies : Sylia, Fatma, Loubna, Yasmina, Ryma, Lynda et Hassiba.

À ma binôme Feriel et sa famille.

Siham

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

API	**A**pplication **P**rogramming **I**nterface
CNN	**C**onvolutional **N**eural **N**etwork
CNRC	**C**onseil **N**ational de **R**echerches du **C**anada
CNTE	**U**niform **C**entre **T**ele-**E**nseignement
CSV	**C**omma **S**eparated **V**alues
HTML	**H**yper **T**ext **M**arkup **L**anguage
IA	**I**ntelligence **A**rtificielle
IDE	**I**ntegrated **D**evelopment **E**nvironment
IDF	**I**nverse **D**ocument **F**requency
IMDB	**I**nternet **M**ovie **D**ata **B**ase
IOS	**I**nternal **O**perating **S**ystem
KNN	**K**-**N**earest N**N**eighbours
NB	**N**aïve **B**ayes
NLP	**N**atural **L**anguage **P**rocessing
NoSQL	**N**ot **o**nly **S**tructured **Q**uery **L**anguage
NUMPY	**NUM**erical **PY**thon
OM	**O**pinion **M**ining
PCA	**P**rincipal **C**omponent **A**nalysis
POS	**P**art **O**f **S**peech
SA	**S**entiment **A**nalysis
SGBDR	**S**ystème de **G**estion de **B**ases de **D**onnées **R**elationnelle
SVM	**S**upport **V**ector **M**achine
TALN	**T**raitement **A**utomatique du **L**angage **N**aturel
TF	**T**erm **F**requency
TLN	**T**raitement du **L**angage **N**aturel
TSV	**T**ab **S**eparated **V**alues
URL	**U**niform **R**esource **L**ocator
XML	**E**xtensible **M**arkup **L**anguage
XSL	**E**xtensible **S**tylesheet **L**anguage

Résumé

Le E-Commerce est devenu un domaine très large où acquérir des connaissances sur ses consommateurs ou anticiper leurs attentes est devenu une tâche cruciale pour les entreprises afin d'assurer la propérité de leur image et produits et tout cela est possible à partir de la collecte des avis de clients sur un produit, un service, une marque. D'où l'émergence de l'analyse des sentiments, qui est devenu une véritable industrie, toute aussi stratégique que celle des sondages et avec l'évolution de la technologie, la puissance de calcul de l'informatique permettrait de suivre toutes ces évolutions en temps réel quelque soit leur volume sur le web. Cependant des milliers d'avis sont postés chaque jour sur les sites d'achat et il est difficile d'analyser toutes ces données et les catégoriser selon leur polarité et différencier les vrais avis des faux et prendre en considération la langue utilisé, la syntaxe, le sarcasme, la neutralité de l'avis. De nombreuses techniques, la plus populaire étant le Machine Learning, et outils ont été développés pour assurer l'analyse des sentiments dans tous les domaines, dans notre cas le commerce éléctronique. Le présent mémoire passe en revue les principales approches liées à l'analyse des sentiments des avis des produits, propose une nouvelle approche utilisant le Traitement automatique des langues (NLP) et trois algorithmes : Support Vector Machine (SVM), K-Nearest Neighbours (KNN) et Naïve Bayes (NB) pour résoudre le problème de la polarité et le taux de précision des donnés. Nous décrivons une approche permettant d'analyser des commentaires de téléphones portables vérouillés obtenus d'Amazon sous forme d'un dataset, analyser ce dataset et obtenir des résultats précis.

Mots clés : Amazon, Analyse des sentiments, Dataset, KNN, Machine Learning, NB, NLP, Opinion Mining, Polarité, SVM.

Abstract

E-Commerce has now become a very wide field where acquiring knowledge about its consumers or anticipating their expectations has become a crucial task for companies in order to ensure the evolution of their turnover and the prosperity of their image and products and all this is possible from the collection of customers' opinions on a product, a service, a brand. Hence the emergence of sentiment analysis, which has become a real industry, just as strategic as that of surveys and with the evolution of technology, the computing power of computers would allow to follow all these developments in real time whatever their volume on the web. However, thousands of reviews are posted every day on shopping sites and it is difficult to analyze all these data and categorize them according to their polarity and differentiate the real reviews from the fake ones and take into consideration the language used, the syntax, the sarcasm, the neutrality of the review. Many techniques, the most popular being Machine Learning, and tools have been developed to ensure sentiment analysis in all domains, in our case e-commerce. Our presentation reviews the main approaches related to sentiment analysis product reviews, proposes a new approach using Natural Language Processing (NLP) and three algorithms : Support Vector Machine (SVM), K-Nearest Neighbours (KNN) and Naïve Bayes (NB) to solve the polarity problem and the accuracy rate of the data. We describe an approach to analyze locked cell phone reviews obtained from Amazon as a dataset, analyze this dataset and obtain accurate results.

Keywords : Amazon, Dataset, KNN, Machine Learning, Naïve Bayes, NLP, Opinion Mining, Polarity, Sentiment analysis, SVM.

Chapitre 1

Introduction

Sommaire

1.1 Introduction

Actuellement, l'Internet occupe une place importante dans la vie quotidienne d'une personne, il est devenu un outil incontournable d'échange d'informations tant au niveau professionnel que personnel. Le Web a offert et continue à offrir un monde de l'information considérable et a évolué de simples ensembles de pages statiques vers des services de plus en plus complexes. Parmi ces services, acheter des produits en ligne, lire des journaux en ligne, échanger des messages avec d'autres personnes, discuter sur de multiples forums ou la possibilité d'exprimer son opinion sur des blogs.

L'Internet contient un nombre colossal d'informations, et pour la plupart du temps c'est le premier lieu pour trouver ces informations, faire une réservation, acheter des produits, consulter les avis d'autres utilisateurs sur les produits qui nous intéressent, lire les critiques de films avant de choisir le film à voir au cinéma, etc. Le problème principal ne se trouve pas dans savoir si l'information se trouve dans le Web mais comment la

trouver car le flux informationnel est excessivement répandu. Un autre problème, non lié à Internet lui même mais plutôt à des considérations sociologiques qui est l'envahissement par la globalisation. Avoir l'accès à beaucoup plus de produits que l'on ne peut en connaître. Ces produits peuvent être de différents domaine : le commerce, la technologie, le travail, les films, la musique, l'école, etc. C'est là qu'Internet intervient et vient en aide aux utilisateurs et facilite énormément le référencement, la recherche et l'accès aux informations.

Pour fournir aux utilisateurs des alternatives de produits, les moteurs de prédictions ont été crée pour ça et aussi pour des raisons commerciales. Les gens aiment en général consulter les recommandations d'autres utilisateurs avant de se faire leurs propres opinions et acheter un article. Les prédictions en ligne sont devenues très utiles pour les clients et les fournisseurs. Les algorithmes de prédiction sont basés sur les expériences et les avis des autres utilisateurs. La plupart des fournisseurs du E-Commerce (par exemple Amazon) cherchent a augmenter l'efficacité et la fiabilité des ventes en encouragent les utilisateurs à commenter et évaluer leurs produits et services. Ces opinions exprimés aident les entreprises à améliorer leurs produits en question et les clients potentiels pour prendre des décisions d'achat. Les utilisateurs expriment maintenant leurs opinions et sentiments sur les sites Web sous forme d'un commentaire ou d'une évaluation. Pour cette raison, le besoin de collecter ces commentaires et les analyser pour déduire ce qui plaît ou déplaît aux clients. Le sujet de ce mémoire tente de répondre à ces besoins.

Le domaine de recherche présenté dans ce mémoire est *l'analyse des sentiments*. Le but est d'analyser les commentaires des clients et trouver la polarité de chaque avis. L'activité de recherche consiste à déduire la polarité d'une opinion.

L'analyse des sentiments est le processus permettant d'extraire les opinions ou les sentiments des avis des clients publiés sur des blogs, des forums, des sites Web d'achat, etc. Elle est utilisée pour identifier et extraire les informations subjectives contenues dans le texte. L'analyse et la classification des sentiments est une étude informatique qui tente de résoudre le problème de la polarité en extrayant des informations subjectives des textes donnés en langage naturel, comme les opinions et les sentiments. Différentes approches ont été utilisées pour aborder ce problème, depuis le traitement du langage naturel, l'analyse

de texte, la linguistique computationnelle et la biométrie. Ces dernières années, les méthodes d'apprentissage automatique sont devenues populaires dans l'analyse sémantique et l'Opinion Mining pour leur simplicité et leur précision.

1.2 Problématique

Comme le E-Commerce est devenu populaire au cours des dernières décennies, les vendeurs et les marchands en ligne demandent à leurs acheteurs de partager leurs opinions sur les produits qu'ils ont achetés. Chaque jour, des millions d'avis sont générés sur Internet à propos des différents produits.

Cependant, comme le nombre d'avis disponibles d'un produit augmente, il devient de plus en plus difficile pour un consommateur potentiel de prendre une bonne décision quant à l'achat du produit. Les opinions divergentes sur le même produit d'une part et les avis ambigus d'autre part rendent les clients plus confus pour prendre la bonne décision. La solution du problème de classification des sentiments semble cruciale pour toutes les entreprises du E-Commerce.

1.3 Objectifs

L'objectif de ce mémoire est de détecter automatiquement les sentiments des internautes et leurs avis positifs, négatifs ou neutres sur un produit. En développant un système d'analyse des sentiments pour classer les opinions en trois catégories : positif, négatif et neutre, et représenter les différentes techniques et approches proposées et leurs résultats.

L'analyse des avis des clients joue un rôle crucial afin de maintenir la qualité des produits, répondre aux attentes des clients et offrir une meilleure expérience d'achat. Cela aide les entreprises comme Amazon, SHEIN et AliExpress à augmenter leurs ventes. Elle permet aussi d'aider les entreprises à comprendre le point de vue des clients envers les stratégies de la marque.

1.4 Méthodologie de travail

Notre démarche de travail repose plus précisément sur les étapes suivantes :

— **Étape de recherche et d'analyse :** qui établit un état de l'art des différentes technologies proposées dans le cadre de l'analyse des sentiments et qui fait une comparaison des avantages et inconvénients de chaque approche proposée.

— **Étape d'identification du problème et de la proposition d'une solution :** qui permet de définir la problématique et la solution proposée en vigueur.

— **Étape d'implémentation et d'expérimentation du système proposé :** qui met en évidence le système proposé, son fonctionnement et son intérêt.

1.5 Organisation du mémoire

Le reste du mémoire est structuré comme suit :

Le deuxième chapitre est consacré pour le domaines d'analyse des sentiments et Opinion Mining. On présentera les différents niveaux d'analyse des sentiments, les disciplines en relation avec l'analyse des sentiments, les problèmes liés à l'analyse des sentiments et L'analyse des sentiments dans le E-Commerce.

Dans le troisième chapitre, nous élaborerons l'état de l'art qui représentera tous les travaux connexes que nous synthétiserons, nous présenterons ceci dans un tableau qui contiendra les grandes lignes de chaque document synthétisé, tout en suivant chaque travail par un bref paragraphe qui le résume, par la suite, nous procéderons à une analyse comparative entre les approches des documents connexes et notre approche.

Le quatrième chapitre porte sur l'expérimentation de notre approche. Il présente les différents aspects liés à l'implémentation du prototype que nous développerons et nous effectuerons dans le cadre du déploiement de ce prototype sur un cas réel.

Dans le cinquième chapitre,nous présentons les outils de programmation et l'implémentation de notre application, présentation des interfaces et les résultats d'exécution,

ainsi que les logiciels choisis pour l'implémentation de notre approche.

Enfin, nous conclurons ce mémoire dans le chapitre six qui donne les conclusions et perspectives de ce travail. Ce chapitre propose un bilan du travail effectué durant ce mémoire et un ensemble de perspectives liées notamment à la poursuite de ce travail.

Chapitre 2

Généralités sur l'analyse des sentiments

Sommaire

2.1 Introduction

L'analyse des sentiments et l'Opinion Mining est un domaine très populaire pour analyser et trouver des informations à partir de données textuelles provenant de diverses sources telles que Facebook, Twitter et Amazon, etc. Elle joue un rôle essentiel en permettant aux entreprises de travailler activement sur l'amélioration de la stratégie commerciale et d'obtenir un aperçu approfondi des commentaires des acheteurs sur leur produit. Elle implique l'étude computationnelle du comportement d'un individu en termes d'intérêt

d'achat, d'extraction de données et d'exploitation de ses opinions sur une entité commerciale d'une entreprise. Cette entité peut être visualisée comme un événement, un individu, un article de blog ou une expérience de produit.

L'analyse des sentiments fait appel à l'étude de l'analyse des textes, du traitement du langage naturel et de la linguistique informatique pour identifier, extraire et étudier scientifiquement les informations subjectives des données textuelles.

Le sentiment ou l'opinion est l'attitude des clients provenant des avis, des réponses aux enquêtes, des médias sociaux en ligne, etc. La signification générale de l'analyse des sentiments est de déterminer l'insolence d'un orateur, d'un écrivain ou d'un autre sujet par rapport à un thème particulier ou à une polarité contextuelle d'un événement spécifique, d'une discussion, d'un forum, d'une interaction ou de tout autre document, etc.

La polarité d'une opinion exprime la positivité, la négativité ou une information de cette dernière. On dit d'une opinion positive qu'elle possède une polarité positive, et inversement, on dit d'une opinion négative qu'elle possède une polarité négative ou neutre possède une information.

La tâche essentielle de l'analyse des sentiments est de déterminer cette polarité d'un texte donné au niveau de la caractéristique, de la phrase et du document. En raison de l'augmentation de l'utilisation d'Internet, chaque utilisateur est intéressé à mettre son opinion sur le Web par le biais de différents médias et les résultats des données d'opinions générés par cet opinion sur la toile. L'analyse des sentiments aide à analyser ces données d'opinions et d'en extraire des informations importantes qui aideront les autres utilisateurs à prendre une décision. Les données des médias sociaux peuvent être de différents types comme critiques de produits, critiques de films, critiques de compagnies aériennes, critiques d'hôtels, l'interaction avec les employés, les revues de santé, les nouvelles et les articles, etc.

Dans ce deuxième chapitre nous allons présenter quelques définitions du domaine d'études qu'est l'analyse des sentiments, ses caractéristiques, ses difficultés, les problèmes

liés à ce domaine et le Machine Learning et ses types.

2.2 Analyse des sentiments

L'analyse des sentiments est souvent désignée sous le nom d'extraction d'opinion, car l'opinion recueillie auprès du client sera extraite pour révéler la note du produit ensuite elle sera exploitée pour révéler l'évaluation du produit. Elle fait partie du Machine Learning. Étant donné que les données en ligne augmentent considérablement de jour en jour, elle est considérée comme très importante dans la situation actuelle, car de nombreux textes contenant l'opinion des utilisateurs sont disponibles sur le Web. L'analyse des sentiments est considérée comme l'étude des pensées et des sentiments des utilisateurs à l'égard d'un produit. Les deux termes **SA** (**S**entiment **A**nalysis) et **OM** (**O**pinion **M**ining) sont interchangeables.

L'importance de l'analyse des sentiments ou de l'extraction d'opinions augmente chaque jour, car les données s'accroissent de jour en jour. Les machines doivent être fiables et efficaces pour interpréter et comprendre les émotions et les sentiments humains [1].

L'analyse des sentiments est un domaine multidisciplinaire, qui englobe la psychologie, la sociologie, le traitement du langage naturel et le Machine Learning. Récemment, la croissance exponentielle des quantités de données et de la puissance de calcul a permis de mettre en place des formes d'analyse plus avancées. Le Machine Learning est donc devenu un outil dominant pour l'analyse des sentiments. Il existe une abondance de littérature scientifique sur l'analyse des sentiments et plusieurs études secondaires ont été menées sur le sujet [2].

2.2.1 Caracteristiques

L'analyse des sentiments est un domaine de recherche vaste et complexe. Dans ce qui suit, les principales caractéristiques qui constituent le processus d'analyse des sentiments sont décrites et discutées en détail.

A) Catégorisation des sentiments : *Phrases objectives versus phrases sub-*

jectives

Le premier objectif de l'analyse des sentiments consiste généralement à distinguer les phrases subjectives des phrases objectives. Si une phrase donnée est classée comme objective, aucune autre tâche fondamentale n'est requise, alors que si celle-ci est classée comme subjective, sa polarité (positive, négative ou neutre) doit être estimée. La classification de la subjectivité [3] est la tâche qui distingue les phrases qui expriment des informations objectives des phrases qui expriment des opinions et des avis subjectives.

Un exemple de phrase objective est L'iPhone est un smartphone, tandis qu'un exemple de phrase subjective est L'iPhone est génial. La classification de polarité est la tâche qui distingue les phrases qui expriment des polarités positives, négatives ou neutres. Notant qu'une phrase subjective peut ne pas exprimer un sentiment positif ou négatif (par exemple, Je suppose qu'il est arrivé), pour cette raison, il doit être classé comme neutre.

B) Niveaux d'analyse :

Comme mentionné précédemment, le but de l'analyse des sentiments est de définir des outils automatiques capables d'extraire des informations subjectives à partir des textes en langage naturel. Le premier choix lorsqu'on applique l'analyse des sentiments est de définir ce que signifie le texte (c'est-à-dire l'objet analysé) dans le cas d'étude considéré.

En général, l'analyse des sentiments dans le E-Commerce peut être étudiée principalement à trois niveaux :

- **Niveau de texte :** L'objectif est de détecter la polarité d'un texte d'opinion. Par exemple, dans le cadre d'une revue de produit, le système détermine si le texte exprime une opinion globale positive, négative ou neutre au sujet du produit. L'hypothèse est que l'ensemble du texte n'exprime qu'une seule opinion sur une seule entité (un seul produit).
- **Niveau phrase :** Le but est de déterminer la polarité de chaque phrase contenue dans un texte. L'hypothèse est que chaque phrase, dans un texte donné, désigne une seule opinion sur une seule entité.

- **Niveau d'entité et d'aspect :** Effectue une analyse plus fine que le niveau du texte et de la phrase. Elle repose sur l'idée qu'une opinion est constituée d'un sentiment et d'une cible (d'opinion). Par exemple, la phrase L'iPhone est très bien, mais ils ont encore besoin de travailler sur la durée de vie de la batterie et la sécurité évalue trois aspects : iPhone (neutre), la durée de vie de la batterie (négatif) et la sécurité (négatif).

C) Opinion régulière versus opinion comparative :

Une opinion peut prendre différentes nuances et peut être assignée à l'un des groupes suivants :

- **Opinion régulière :** Une opinion régulière est souvent désignée dans la littérature comme une opinion standard et elle a deux sous-types principaux :
  - **Opinion directe :** Une opinion directe fait référence à une opinion exprimée directement sur une entité (par exemple, La luminosité de l'écran de l'iPhone est impressionnante).
  - **Opinion indirecte :** Une opinion indirecte est une opinion qui est exprimée indirectement sur une entité sur la base de ses effets sur d'autres entités. Par exemple, la phrase Après être passé à l'iPhone, j'ai perdu toutes mes données décrit un effet indésirable du passage à l'iPhone sur les données, ce qui donne indirectement un sentiment négatif à l'iPhone.
- **Opinion comparative :** Une opinion comparative exprime une relation de similitude ou de différence entre deux ou plusieurs entités et/ou une préférence du détenteur d'opinion basée sur certains aspects communs des entités [4]. Par exemple, les phrases iOS est plus performant qu'Android et iOS est le système d'exploitation le plus performant expriment deux opinions comparatives. Une opinion comparative est habituellement exprimée en utilisant la forme comparative ou superlative d'un adjectif ou d'un adverbe.

D) Opinions explicites versus opinions implicites

Parmi les différentes nuances qu'une opinion peut prendre, nous distinguons les opinions explicites et implicites :

- **Opinion explicite :** Une opinion explicite est une déclaration subjective qui donne une opinion régulière ou comparative (par exemple, La luminosité de l'écran de l'iPhone est impressionnante).
- **Opinion implicite :** Une opinion implicite est un énoncé objectif qui implique une opinion régulière ou comparative qui exprime habituellement un fait désirable ou indésirable (par exemple, « Samedi soir, j'irai au cinéma pour regarder 'I am legend'. J'ai hâte de le regarder ! » et « 'Saving Private Ryan' est plus violent que 'I am legend' »). Le premier exemple suggère qu'il y a de bonnes attentes à propos du film, bien qu'il ne soit pas expliqué en mots, alors que la compréhension de l'opinion cachée dans le second exemple est difficile même pour les humains. Pour certaines personnes, la violence dans les films de guerre pourrait être une bonne caractéristique qui rend le film plus réaliste, alors qu'elle pourrait être une caractéristique négative pour d'autres.

Il est clair que les opinions explicites sont plus faciles à détecter et à classer que les opinions implicites. Une grande partie de la recherche actuelle s'est concentrée sur des opinions explicites. Relativement moins de travail a été fait sur les opinions implicites.

2.2.2 Disciplines en relation avec l'analyse des sentiments

Plusieurs disciplines ont une relation directe ou moins directe avec l'analyse des sentiments, l'opinion mining, l'intelligence artificielle, le traitement automatique du langage naturel, le texte mining et même le data mining offrent des outils et algorithmes indispensables pour le traitement et la classification des sentiments.

2.2.2.1 Fouille de texte

La fouille de texte ou Text mining est l'analyse des données contenues dans un texte en langage naturel. L'application de techniques d'exploration de texte pour résoudre des problèmes métier est appelée analyse de texte.

Le Text Mining est un domaine passionnant qui englobe de nouvelles méthodes de recherche et des outils logiciels qui sont utilisés dans le milieu universitaire ainsi que par des entreprises et des organismes gouvernementaux. Aujourd'hui, les chercheurs utilisent

les outils d'exploration de texte dans des projets ambitieux pour tenter de prédire tout, de la direction des marchés boursiers [5] à l'occurrence de protestations politiques [6].

L'exploration de texte est également couramment utilisée dans la recherche marketing et de nombreuses autres applications commerciales, ainsi que dans le travail du gouvernement et de la défense.

Les processus de fouille de textes comprennent généralement la recherche d'informations (méthodes d'acquisition de textes) et des applications de méthodes statistiques avancées et de traitement du langage naturel (TLN) telles que le marquage des parties de la parole et l'analyse syntaxique. L'exploration de texte comprend aussi souvent la reconnaissance d'entités nommées (REN), qui est l'utilisation de techniques statistiques pour identifier les caractéristiques de textes nommés tels que les personnes, les organisations et les noms de lieux ; la désambiguïsation, qui est l'utilisation d'indices contextuels pour décider où les mots se réfèrent à l'une ou l'autre de leurs multiples significations et l'analyse des sentiments, qui implique de discerner le matériel subjectif et d'extraire des informations attitudinales telles que le sentiment, l'opinion, l'humeur et l'émotion [7].

2.2.2.2 Traitement automatique du langage naturel (TALN)

Le Deep Learning et le Machine Learning continuent de proliférer dans diverses industries, et a révolutionné le sujet abordé dans ce titre : Le traitement du langage naturel (TLN). Le TLN est un sous-domaine de l'informatique qui se concentre sur la possibilité pour les ordinateurs de comprendre un langage d'une manière naturelle, comme le font les humains. En général, il s'agit de tâches telles que la compréhension du sentiment d'un texte, la reconnaissance vocale et la génération de réponses à des questions.

Le TLN est devenu un domaine en évolution rapide, dont les applications ont représenté une grande partie en intelligence artificielle (IA). Quelques exemples d'applications utilisant le Deep Learning sont les Chatbots qui traitent les demandes du service clientèle, la vérification automatique de l'orthographe sur les téléphones portables et les assistants d'intelligence artificielle tels que Cortana et Siri, sur les smartphones. Pour ceux qui ont de l'expérience en Machine Learning et en Deep Learning, le traitement du langage natu-

rel est l'un des domaines les plus passionnants pour les individus qui souhaitent appliquer leurs compétences. Cependant, afin de fournir un contexte particulier, on se réfère au développement du traitement du langage naturel en tant que domaine [8].

2.3 Problèmes liés à l'analyse des sentiments

De nos jours, l'analyse de sentiments est un domaine de recherche très populaire. De nombreux travaux sont réalisés, mais il n'existe pas encore de méthode suffisamment bonne pour classer les sentiments. Pour de nombreux auteurs, la moyenne des résultats est légèrement supérieure à 85%, mais cela ne suffit pas si nous avons besoin de résultats plus précis.

L'objectif principal de l'analyse des sentiments est d'analyser les avis et de tester les scores des sentiments. Cette analyse est divisée en trois niveaux [9] : Niveau document [10], niveau phrase [11], niveau mot/terme [12] ou niveau aspect [13]. Les processus séquentiels sont l'évaluation de l'analyse des sentiments et la détection de la polarité des sentiments.

Plusieurs enjeux doivent être pris en compte lors de la conduite du l'AS [14]. Deux enjeux majeurs sont abordés. Premièrement, le point de vue (ou l'opinion) observé comme négatif dans une situation peut être considéré comme positif dans une autre situation. Deuxièmement, les gens n'expriment pas toujours leurs opinions de la même manière. La plupart des techniques de traitement de texte courantes utilisent le fait que des modifications mineures entre les deux fragments de texte ne sont pas susceptibles de changer le sens réel [14].

L'analyse des sentiments des données des médias sociaux a également été appliquée pour évaluer les produits, comme expliqué dans [15]. Chaque auteur propose ses propres méthodes pour évaluer les opinions. Malheureusement, la plupart des outils ou algorithmes d'analyse des sentiments sont encore au stade de la recherche. Jusqu'à présent, il n'existe aucun algorithme qui puisse fournir des résultats 100% précis pour l'analyse de sentiments. Il y a encore plusieurs débats entre différents chercheurs qui tentent de prouver que leur

solution est plus parfaite que les autres.

L'extraction du sentiment ou d'opinion consiste à déterminer la polarité de ce dernier. Dans ce qui suit nous citerons quelques difficultés de cette procédure :

— Ambiguïté de certains mots positifs ou négatifs selon les contextes et qui ne peut pas toujours être levée [16].

— Difficulté due aux structures syntaxiques et sémantiques d'une phrase et l'expression de l'opinion. Par exemple : l'histoire du film est intéressante mais les acteurs étaient mauvais. Dans ce cas la polarité de la deuxième partie est opposée à la première.

— Difficulté due au contexte : La nécessite d'une bonne analyse syntaxique du texte ; analyse qui peut se révéler particulièrement difficile dans des cas de coordination entre plusieurs parties d'une phrase. Par exemple : ma tante a bien préparé le gâteau, son décor est beau mais je n'ai pas aimé le goût, l'opinion de la dernière partie de la phrase est la plus importante.

— Difficulté due à l'analyse de la phrase par paquets de mots . Les deux phrases suivantes contiennent les mêmes paquets de mots sans pour autant exprimer les mêmes sentiments. La première phrase contient un sentiment positif alors que la deuxième est négative : Je l'ai apprécié pas seulement à cause de ..., Je ne l'ai pas apprécié seulement à cause de ... où se présente la gestion de négation.

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

— Difficulté de déterminer un lexique adapté à l'analyse de l'ensemble des textes d'opinions [17].

2.4 L'analyse des sentiments dans le E-Commerce

Aujourd'hui, le développement rapide de l'Internet et de ses utilisateurs a modifié la façon dont les individus communiquent dans le monde entier, en particulier lorsqu'ils font

des affaires. Les applications de l'Internet sur les opérations commerciales ont développé de nouvelles possibilités de vente de produits ou de services dans le monde entier. L'accessibilité des plateformes de médias sociaux a permis aux internautes d'exprimer et de partager leurs opinions sur différents types d'éléments basés sur leur expérience de vie, y compris les produits et services qu'ils apprécient.

L'analyse des sentiments est une technologie en plein essor qui exploite les demandes des clients sur la base du traitement du langage naturel. Cette motivation est généralement utilisée pour bien comprendre ce que les clients veulent, quand, pourquoi et comment ils le veulent. Les détaillants doivent se tourner vers l'analyse des sentiments, afin d'éviter de commettre les mêmes erreurs et de prendre les bonnes décisions en fonction des commentaires ou des critiques. Dans le cadre du commerce électronique, les achats en ligne sont un bon exemple de la manière dont les produits ou services sont vendus sur Internet.

Les grands distributeurs tels qu'Amazon et Alibaba ainsi que les petits distributeurs ont certainement connu des résultats décevants, l'un des principaux facteurs de la lenteur de leurs ventes étant le mauvais assortiment de produits. Ces distributeurs étaient essentiellement incapables de mettre les bons produits en rayon, et les clients les ont punis en dépensant leur argent ailleurs, comme cela s'est produit pour une entreprise technologique comme Fitbit en 2016 [1]. La compréhension des consommateurs a toujours figuré en bonne place sur la liste des tâches à accomplir par les distributeurs et l'utilisation de l'analyse des sentiments pour surveiller ces émotions a été la principale motivation des entreprises pour comprendre à quel point l'extraction d'opinions sur les avis des clients peut être diverse et approfondie.

L'Internet est un champ de mines de perspectives, être capable d'accéder à ces opinions sur une variété de plateformes différentes est un avantage significatif pour toute entreprise qui cherche à améliorer ses produits ou services [18].

1. T. Green, (11/01/2017). *Why Fitbit Stock Crashed 75% in 2016*. The Motley Fool.

2.5 Machine Learning

Même s'il est actuellement dopé par les nouvelles technologies et de nouveaux usages, le Machine Learning n'est pas un domaine d'étude récent. On en trouve une première définition dès 1959, due à Arthur Samuel, l'un des pionniers de l'intelligence artificielle, qui définit le Machine Learning comme le champ d'étude visant à donner la capacité à une machine d'apprendre sans être explicitement programmée [19]. En 1997, Tom Mitchell, de l'université de Carnegie Mellon, propose une définition plus précise : « A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E » [20].

« On dit d'un programme informatique qu'il apprend de l'expérience E concernant une certaine classe de tâches T et une mesure de performance P, si sa performance sur les tâches de T, telle que mesurée par P, s'améliore avec l'expérience E ».

Un robot qui apprend à conduire ? Cela peut paraître un peu loin de vos préoccupations... mais illustrons brièvement ce que peut faire le Machine Learning avec un cas simple, sans doute plus proche de votre quotidien : un filtre antispam. Dans un premier temps, on peut imaginer que la « machine » (votre service de messagerie) va « analyser » la façon dont vous allez classer vos mails entrants en spam ou pas. Grâce à cette période d'« apprentissage », la machine va déduire quelques grands critères de classification. Par exemple, la probabilité que la machine classe un mail en spam va augmenter si le mail contient des termes tels qu'« argent », « rencontre facile » et si l'expéditeur du mail n'est pas dans votre carnet d'adresses. A contrario, la probabilité de classement en spam va baisser si l'expéditeur est connu et que les mots du mail sont plus « classiques » [20].

Quand aux différents types du Machine Learning, ils en existent trois et sont détaillés en dessous :

2.5.1 Apprentissage supervisé

L'apprentissage supervisé est peut-être le type de Machine earning le plus facile à appréhender : son but est d'apprendre à faire des prédictions, à partir d'une liste d'exemples étiquetés, c'est-à-dire accompagnés de la valeur à prédire (voir Figure 2.1). Les étiquettes servent de « professeur » et supervisent l'apprentissage de l'algorithme.

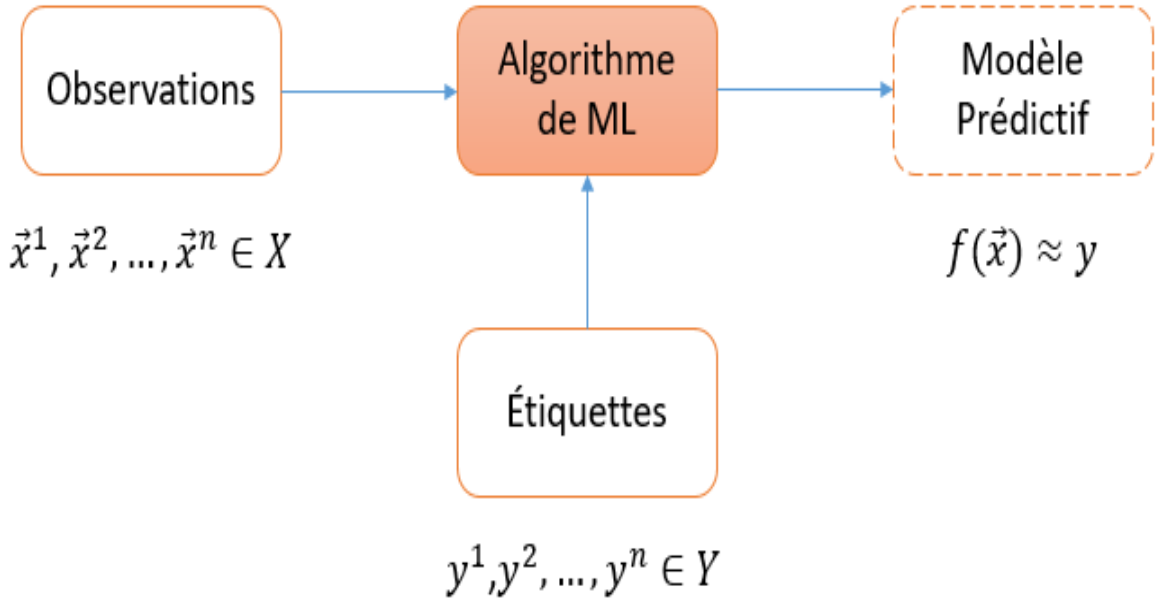

FIGURE 2.1 – Apprentissage supervisé [21].

Définition (Apprentissage supervisé) On appelle apprentissage supervisé la branche du machine learning qui s'intéresse aux problèmes pouvant être formalisés de la façon suivante : étant données n observations x_i , i=1,...,n décrites dans un espace X, et leurs étiquettes y_i , i=1,...,n décrites dans un espace Y, on suppose que les étiquettes peuvent être obtenues à partir des observations grâce à une fonction ϕ : X $\rightarrow$ Y fixe et inconnue : $y_i = \phi(x_i)$+i, où i est un bruit aléatoire. Il s'agit alors d'utiliser les données pour déterminer une fonction f : X $\rightarrow$ Y telle que, pour tout couple (x, ϕ(x)) $\in$ X $\times$ Y, f(x) $\approx$ ϕ(x) [21].

2.5.2 Apprentissage non supervisé

Dans le cadre de l'apprentissage non supervisé, les données ne sont pas étiquetées. Il s'agit alors de modéliser les observations pour mieux les comprendre (voir Figure 2.2).

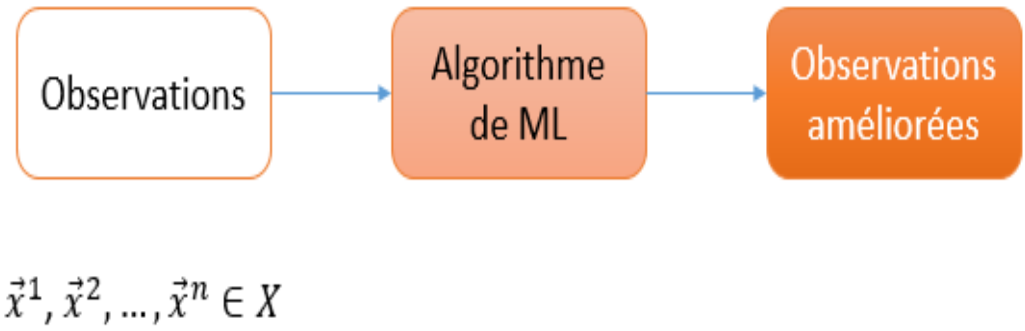

FIGURE 2.2 – Apprentissage non supervisé [21].

Définition (Apprentissage non supervisé) On appelle apprentissage non supervisé la branche du Machine Learning qui s'intéresse aux problèmes pouvant être formalisés de la façon suivante : étant données n observations x_i , i=1,...,n décrites dans un espace X, il s'agit d'apprendre une fonction sur X qui vérifie certaines propriétés [21].

2.5.3 Apprentissage avec renforcement

Dans le cadre de l'apprentissage par renforcement, le système d'apprentissage peut interagir avec son environnement et accomplir des actions. En retour de ces actions, il obtient une récompense, qui peut être positive si l'action était un bon choix, ou négative dans le cas contraire. La récompense peut parfois venir après une longue suite d'actions ; c'est le cas par exemple pour un système apprenant à jouer au go ou aux échecs. Ainsi, l'apprentissage consiste dans ce cas à définir une politique, c'est-à-dire une stratégie permettant d'obtenir systématiquement la meilleure récompense possible. Les applications principales de l'apprentissage par renforcement se trouvent dans les jeux (échecs, go, etc) et la robotique [21].

2.6 Conclusion

De nos jours, la recherche sur l'analyse des sentiments et l'extraction d'opinions est très importante. La plupart des industries créent différents types de données et ont besoin d'analyser ces données pour prendre des décisions qui sont bénéfiques pour l'industrie. Les médias sociaux génèrent également d'énormes quantités de données et il est nécessaire de les analyser et d'en tirer des enseignements de ces données en question.

Dans ce chapitre, nous avons présenté ce qu'est l'analyse des sentiments, ses caractéristiques comme la catégorisation des sentiments et le niveaux d'analyse ainsi que les différentes opinions, ses difficultés qui consistent en somme de déterminer la polarité d'un sentiment, les différentes disciplines en lien avec l'AS comme le Text Mining et le TALN et le Machine Learning, les problèmes liés à l'AS et ce dernier dans le E-Commerce et enfin parler du Machine Learning qui est le noyau de ce projet ainsi que ses différents types.

Dans le chapitre suivant, nous allons dresser un état de l'art des principales approches relatives au domaine du E-Commerce. Une étude comparative sera établie pour les principaux travaux déjà réalisés.

Chapitre 3

Etat de l'art

Sommaire

3.1 Introduction

Grâce au développement du E-Commerce, de plus en plus de sites de commerce éléctronique et de plateformes de réseaux sociaux encouragent les consommateurs ou les clients à noter et publier des avis en ligne sur les produits. On prend comme exemple le site **Amazon** qui a fourni un moyen permettant aux consommateurs de partager leurs évaluations concernant un produit en ligne.

D'après certaines recherches, les résultats ont montré que les avis de produits en ligne ont un impact significatif sur les décisions d'achats des clients. En d'autres termes, avant de prendre leur décision d'achat, les consommateurs peuvent visiter les sites web concernés et lire les avis en ligne sur les produits alternatifs. Mais il est difficile pour le consommateur de prendre directement la décision d'achat en se basant sur les avis laissés par d'autres clients sur les produits car le nombre d'avis en ligne est énorme. Par conséquent,

pour soutenir la décision d'achat du client, il est nécessaire de développer une méthode de classement des produits par le biais des avis en ligne. Cette méthode qui est l'analyse des sentiments permet d'identifier automatiquement l'orientation du sentiment de chaque avis en ligne, les performances des produits alternatifs pour chaque caractéristique du produit et les résultats de l'évaluation, les performances des produits alternatifs concernant chaque caractéristique de ces derniers peuvent être analysées, et leur classement peut être déterminé.

Pour classer les produits par le biais des avis, il est nécessaire d'identifier les orientations des sentiments des critiques en ligne sur les produits concernant différentes caractéristiques en utilisant des techniques d'analyse des sentiments. Ensuite, sur la base des orientations de sentiment identifié, les études sur le classement des produits par le biais d'évaluations en ligne peuvent être menées. Ainsi, il existe plusieurs méthodes d'analyse des sentiments, nous verrons quelques-unes dans la synthèse des travaux connexes que nous allons voir dans ce chapitre.

Dans ce chapitre, nous élaborerons l'état de l'art des principeles contributions relatives au E-commerce qui représentera tous les travaux connexes que nous synthétiserons, nous présenterons ceci dans un tableau qui contiendra les grandes lignes de chaque approche synthétisée, tout en suivant chaque travail par un bref paragraphe qui le résume, par la suite, nous procéderons à une analyse comparative et une analyse.

3.2 Travaux connexes

Traditionnellement, l'analyse des sentiments porte sur des avis contradictoires, c'est-à-dire, un client peut avoir une opinion positive, neutre, ou négative à l'égard d'un produit et c'est pourquoi on a besoin de ces approches afin de classer ces avis et d'obtenir des résultats précis.

Le thème de l'opinion mining a beaucoup d'importance dans le domaine de la recherche.De nombreux travaux récents ont proposé différentes approches pour classer les avis en ligne et aider les clients à choisir leur produit et les entreprises à comprendre les besoins des consommateurs.

Dans cette section, nous présenterons les principaux travaux relatifs à l'analyse des sentiments dans le demaine de E-commerce.Nous avons établi une classification de ces travaux selon deux catégories : les approches basées sur le machine learning et les approches basées sur la sémantique.

3.2.1 Approche basée sur le Machine Learning

Le Machine Learning est un sous-domaine de l'intelligence artificielle qui s'intéresse à comprendre et reproduire la faculté de l'apprentissage humain par des systèmes artificiels. Dans cette partie, nous présenterons les travaux qui ont utilisé le *Machine Learning* comme technique. Il existe 3 types ou catégories de l'apprentissage machine. Le premier s'appelle *l'apprentissage supervisé* où, après avoir présenté les données et les résultats souhaités aux ordinateurs, ils auront la capacité de faire des prédictions pour de nouvelles données d'entrée, comme : La **classification Bayésienne**, la **machine à vecteurs de support (SVM)**, le **réseau neuronal**, l'**arbre de décision**, la **forêt d'arbres décisionnels (Random Forest)** et le **KNN Classificateur**. Le second est *l'apprentissage non supervisé* où l'on ne donne à l'ordinateur que les données et où il doit trouver une structure avec un sens par lui-même sans l'intervention d'une supervision extérieure. Elle dépend principalement du clustering comme : Le **K-Means**. Le troisième est *l'apprentissage par renforcement* où la machine se comporte comme un agent qui apprend de son environnement d'une manière interactive jusqu'à ce qu'il découvert les comportements qui produisent des récompenses. Il existe plusieurs algorithmes et techniques utilisés dans l'analyse des sentiments.

On distinguera quelques unes des techniques utilisées dans les travaux étudiés dans la partie suivante :

Devasia and Shreik [22], ont proposé une solution pour résoudre le problème d'identifier les commentaires parmi des milliers qui parlent de la caractéristique spécifique du produit que cherche le client. La solution en question est un système proposé qui suit une approche sémantique pour extraire les caractéristiques du produit. ils ont utilisé un algorithme, qui utilise des dépendances typées, est introduit à cet effet. Le modèle Deep récursif est utilisé pour déterminer l'orientation du sentiment des phrases de révision. Les

auteurs ont construit une matrice de révision pour déterminer l'importance et la polarité de chaque caractéristique du produit. Les résultats expérimentaux montrent que la méthode proposée est efficace et a atteint l'objectif souhaité.

Mubarok et al [23], ont utilisé l'analyse de sentiment pour analyser et extraire la polarité de sentiment sur les critiques de produits en fonction d'un aspect spécifique du produit. Ils ont mené leur travail en trois phases, telles que le prétraitement des données qui impliquent le Part-of-Speech (POS) tagging, la sélection des fonctionnalités à l'aide de Chi Square et la classification de la polarité du sentiment des aspects à l'aide de Naïve Bayes. D'après leurs résultats de l'évaluation, les auteurs ont conclus que le système est en mesure d'effectuer une analyse de sentiment basée sur les aspects avec sa mesure F1 la plus élevée de 78,12%.

Ray et Chakrabarti [24], ont proposé un cadre de travail pour l'analyse de sentiment à l'aide du logiciel R qui peut analyser le sentiment des utilisateurs sur les données Twitter en utilisant l'API Twitter. Leur méthode implique la collecte de données twitter, son prétraitement et suivi d'un lexique pour analyser le sentiment des utilisateurs.

Suganya et Vijayarani [25], ont utilisé la technique du Web Scraping pour collecter les commentaires en ligne sur les produits. Ensuite ces commentaires collectés ont été analysé à l'aide de l'opinion ou l'analyse des sentiments utilisant des modèles de classification tels que KNN, SVM, Random Forest, Convolutional Neural Network (CNN) et l'hybride proposé SVM-CNN. Des expériences ont déjà été faites sur les modèles de classification et ont réalisé des résultats prometteurs.

Jagdale et al [26], ont utilisé les techniques du Machine Learning, Naïve Bayes et le Support Vector Machine (SVM) en faisant une étude informatique du comportement d'un individu en matière d'intérêt d'achat suivi d'une étude de ses opinions sur l'entité commerciale d'une entreprise. Cette entité peut être visualisée comme un événement, un individu, un billet de blog ou une expérience produit. ils ont utilisé un Dataset qui a été pris d'Amazon qui contient des critiques de caméra, ordinateurs portables, téléphones mobiles, tablettes, TV, vidéosurveillance, ils ont d'abord appliqué le prétraitement puis des algorithmes du Machine Learning pour classer les avis qui sont positifs ou négatifs. Ils

ont conclu que les techniques du Machine Learning donnent les meilleurs résultats pour classifier les critiques de produits. Naïve Bayes a obtenu la précision 98,17% et le SVM a obtenu la précision 93,54% pour les commentaires de caméra.

Pankaj et al [27], ont utilisé l'Opinion Mining, Text Minig et les sentiments, qui ont affecté le monde environnant en changeant leur opinion sur un produit spécifique, pour présenter une évaluation des commentaires des clients sur un produit. Les données qu'ils ont utilisées dans leur étude sont des commentaires en ligne sur des produits collectés sur Amazon. Ils ont effectué une analyse comparative des sentiments des commentaires récupérés. Leur recherche fournit une analyse sentimentale de diverses opinions sur les téléphones intelligents en les divisant en comportements positifs, négatifs et neutres.

Bose et al [28], ont utilisé le lexique des émotions du CNRC qui peut être catégorisé en huit émotions de base et deux sentiments. World cloud a aidé également leur recherche à faire des comparaisons entre les huit catégories d'émotions. Ils ont suivi 568 454 critiques sur la nourriture de 74 258 produits et 256 059 utilisateurs sur Amazon sur une période de dix ans. Pour analyser le résultat, ils ont sélectionné six produits et utilisateurs les plus populaires basés sur des avis en texte claire. Leurs résultats montrent que l'analyse des sentiments a aidé à identifier les comportements des consommateurs et à surmonter ces risques pour satisfaire les consommateurs.

Dey et al [29], ont présenté une comparaison entre deux approches du Machine Learning pour analyser le sentiment des critiques des clients sur les produits Amazon. Dans leur travail de recherche, ils ont d'abord analysé le sentiment du consommateur à l'aide du Naïve Bayes Classifier. Parallèlement, ils ont classifié les sentiments des utilisateurs en catégories binaires grâce au SVM. Par conséquent, les données ont été passées par le Network modèle après la méthode de prétraitement nommée fréquence des termes (TF) et fréquence inverse des documents (IDF) pour évaluer la caractéristique. Ils ont eu pour objectif de trouver des approches du Machine Learning comparativement meilleures parmi les SVM et les Naïve Bayes classifiers basés sur la mesure statistique.

Nandal et al [30], ont introduit une nouvelle approche qui utilise la détection de sentiments au niveau d'aspects, qui se concentre sur les caractéristiques du produit. Ils

ont mis en oeuvre et testé leur travail sur les commentaires de clients d'Amazon (données explorées) où les termes d'aspect sont d'abord identifiés pour chaque commentaire. Leur système a effectué des opérations de prétraitement telles que l'abréviation, la tokénisation, le casing, la suppression des mots d'arrêt sur l'ensemble de données afin d'extraire des informations significatives et donner finalement le classifier en négatif ou positif.

3.2.2 Approche sémantique

La sémantique est une approche qui tire parti d'une ontologie pour fournir des scores de sentiment plus élaborés concernant les notions contenues dans un tweet. Une ontologie peut être définie comme une spécification explicite, lisible par une machine, d'une conceptualisation partagée [31]. Les ontologies sont utilisées pour modéliser les termes d'un domaine d'intérêt ainsi que les relations entre ces termes et sont maintenant appliquées dans divers domaines, comme les systèmes de gestion des agents et des connaissances et les plateformes de commerce électronique[32]. D'autres applications comprennent la génération de langage naturel, l'intégration intelligente de l'information, l'accès à Internet basé sur la sémantique et l'extraction d'informations à partir de textes. L'objectif de cette approche est de fournir des scores de sentiment pour chaque aspect/caractéristique d'un produit.

La technique utilisée dans le papier étudié, basée sur l'approche sémantique est expliquée ci-dessus :

Polsawat et al [33], ont proposé une solution qui vise à résoudre les problèmes que les entreprises rencontrent en s'efforçant d'analyser et d'interpréter la multitude d'opinions et de sentiments des clients, une évaluation précise devient problématique. De nombreuses recherches se heurtent à des conflits sémantiques entre des mots ou des synonymes, et des erreurs se produisent dans l'algorithme SentiWordNet, lors de l'évaluation des mots positifs et négatifs dans certaines phrases. Ces problèmes ont été résolu par le biais de DBpedia, en s'attaquant aux différences du sens des mots, et à créer une interface utilisateur permettant de retrouver des produits sous forme de mots-clés, afin d'aider les consommateurs à prendre des décisions dans le choix des produits. Les auteurs ont mesuré l'efficacité de l'analyse des sentiments dans le cadre de l'étude qui est de 94%.

3.3 Étude comparative et analyse

Les tableaux ci-dessous résument les principales caractéristiques des approches citées ci-dessus. Les tableaux contiennent sept (07) colonnes qui indiquent un critère de comparaison comme suit :

- La colonne **Approche** désigne l'approche de chaque papier synthétisé.
- La colonne **Catégorie de l'approche** désigne la catégorie de l'approche utilisée dans le papier.
- La colonne **Dataset** indique les sources de données utilisées pourl'analyse des sentiments.
- La colonne **Output** indique la production finale de l'approche.
- La colonne **Technique utilisée** indique les méthodes utilisées pour le processus d'analyse des sentiments.
- La colonne **Outil supporté** désigne si l'approche a été implémenté avec un outil logiciel.
- La colonne **Avantages** présente les principaux avantages de l'approche.

Approche	**Catégorie de l'approche**	**Data source**	**Output**	**Technique utilisée**	**Outil supporté**	**Avantages**
N. Devasia et R. Sheik (2016)	Machine Learning	Amazon	Candidate feature set	- Extraction des fonctionnalités. - Trouver la polarité des phrasesde la revue du produit. - Analyse des sentiments. - Recursive Deep Analyser	Oui	- Meilleurs résultats en termes de précision, de rappel et d'exactitude.
Mohamad Syahrul Mubarok et al. (2017)	Machine Learning	SemEval-2014 Task 4	Classification des opinions	- Prétraitement. - Etiquetage morpho-syntaxique (Part-of-speech tagging (POS)). - Test du X^2 (Chi Square). - Naïve Bayes.	Oui	- Meilleure performance et meilleur temps de réponse.
Paramita Ray et al. (2017)	Machine Learning	Twitter	Commentaires classifiés d'un IPhone	- Collecter les datas sur Twitter. - Prétraitement des données. - Classification.	Oui	- Utiliser une analyse au niveau du document, une analyse au niveau d'aspect. - Un dictionnaire d'acronymes a été utilisé par les auteurs pour identifier les acronymes.
Teerawat Polsawat et al. (2018)	Sémantique	Twitter	Score des laptops	- Conception d'une ontologie de produit. - Processus de l'analyse des sentiments. - Ontology-base.	Oui	- Meilleure efficacité. - Résout le problème synonymes.
E. Suganya et S. Vijayarani (2018)	Machine Learning	- Amazon - Flipcart - Snapdeal	Commentaires classifiés	- Web Scraping. - KNN algorithm. - SVM. - Random Forest. - Convolutional Neural Network (CNN). - Hybrid SVM-CNN.	Oui	- L'analyse de sentiments par les prédictions.

TABLE 3.1 – État de l'art des travaux connexes.

Approche	Catégorie de l'approche	Data source	Output	Technique utilisée	Outil supporté	Avantages
Rajkumar S. Jagdale et al. (2019)	Machine Learning	Amazon	Score des produits	- Classification. - Support Vector Machine (SVM). - Naïve Bayes. - Machine Learning.	Oui	- Meilleurs classements des commentaires.
Pankaj Thakur et al. (2019)	Machine Learning	Amazon	Commentaires classifiés	- POS Tagger. - Prétraitement. - Préfiltrage. - Précision des données. - Opinion Mining. - Text Mining.	Oui	- Intégration de divers algorithmes pour évaluer et donner du sens au corpus des données.
Rajesh Bose et al. (2020)	Machine Learning	Amazon	Commentaires et utilisateurs les plus populaires et leur polarité	-Enquête bibliographique. -Préparation des données. -Prétraitement. -Analyse des avis d'un produit. - NRC emotion lexicon.	Oui	- Approche robuste.
Sanjay Dey et al. (2020)	Machine Learning	Amazon	Taux de précision des avis positifs et négatifs	- SVM. - Term Frequency (TF). - Inverse Document Frequency (IDF). - Naïve Bayes. - Technique Lexicon-based. - Technique Corpus-based. - Technique Dictionary-based. - Machine Learning supervisé. - Arbre de décision.	Oui	- La valeur de précision moyenne du classificateur de quatre classes de sentiments montre que la méthode SVM est plus efficace et supérieure à d'autres méthodes telles que NBC. Cela montre que la méthode SVM est plus performante. - La méthode basée sur le lexique ont obtenu les meilleurs résultats individuellement.
Neha Nandal et al. (2020)	Machine Learning	Amazon	Commentaires classés avec une polarité précise	- Collection de données. - Identification d'aspect. - Prétraitement de données. - Evaluation et classification. - SVM Classifier.	Oui	- La méthode proposée est novatrice pour la détection des sentiments au niveau de l'aspect, en mettant l'accent sur les mots bipolaires. - L'utilisation d'un algorithme d'IA pour créer un meilleur système d'évaluation en automatisant le processus de filtrage, car un tel système d'IA est capable d'extraire les évaluations qui sont fausses, sarcastiques, négationnistes ou qui suivent le même schéma.

TABLE 3.2 – État de l'art des travaux connexes (suite).

Les travaux étudiés tentent d'analyser les données et obtenir de meilleurs résultats en terme de précision, de rappel et d'exactitude, plusieurs approches ont été utilisé comme le Machine Learning et la sémantique en tenant compte de plusieurs facteurs, dans ce cas l'ironie, la négation, les synonymes, les homonymes. Plusieurs techniques ont été utilisé comme le Text Mining, Web Scraping et d'autres techniques d'apprentissage supervisé comme SVM, Naïve Bayes, Hybrid SVM-CNN..., les avantages majeur de cette approche pour l'analyse des sentiments sont qu'elle est facile à interpréter et que les résultats sont calculés de manière efficace grâce aux différents algorithmes appliqués avec une meilleure performance et un meilleur temps de réponse ce qui augmente la robustesse du système d'évaluation. Par contre, ces approches présentent plusieurs inconvénients comme la dépendance de la classification des sentiments de la taille des données. Lorsque la taille des données augmente, l'approche utilisé devient plus erronée. L'hypothèse que les différents attributs des algorithmes sont indépendants est un inconvénient de ces techniques en question car elles ne peuvent pas être valides à tout moment.

Nous proposons une solution dont le but est d'améliorer le système d'analyse en utilisant la méthode de pondération TF-IDF dans le prétraitement des données pour appliquer par la suite les techniques de l'apprentissage automatique supervisé telles que SVM, Naïve Bayes et K-NN afin de rendre l'analyse optimale en un temps meilleur et obtenir une meilleure précision. La combinaison de ces trois algorithmes de classification offre une bonne stabilité par rapport aux autres algorithmes déjà utilisés dans les travaux précédents.

3.4 Conclusion

Dans ce chapitre, nous avons établi un état de l'art des principeles contributions dans le domaine du E-commerce qui représente tous les travaux connexes que nous avons synthétisés, nous avons présenté ceci dans un tableau qui contient les grandes lignes de chaque approche synthétisée, tout en suivant chaque travail par un bref paragraphe qui le résume.

Dans le chapitre suivant, nous allons détailler notre approche et ses différentes étapes.

Chapitre 4

Analyse des sentiments dans le E-Commerce

Sommaire

4.1 Introduction

La collecte et l'analyse des opinions des individus sont devenues des sources d'informations précieuses dans le E-Commerce ainsi que pour les entreprises.

Le marketing a rapidement compris l'intérêt de l'analyse des sentiments. Il existe même des agences qui vendent aux entreprises la traque des moindres mots sur leur image, sur leurs produits. À partir des sites d'opinions de consommateurs, ces derniers viennent y échanger des avis et trouver des conseils pour leurs décisions d'achat. L'analyse des sentiments permet de catégoriser les avis au sujet d'un produit, de les détailler à un niveau fin. Elle aide aussi à lutter contre le spam en détéctant les faux avis postés par des agences.

Dans ce chapitre, nous présenterons en détail notre approche qu'on a utilisé au cours de notre projet ainsi que ses différentes étapes pour effectuer une analyse des sentiments à partir des commentaires.

4.2 Approche proposée

Notre projet consiste en l'analyse des sentiments dans le E-Commerce, c'est-à-dire appliquer une analyse sur les commentaires des produits. Pour cela, il y a plusieurs étapes qui doivent être effectuées pour obtenir de meilleurs résultats. Ces étapes sont les suivantes : Collecte des données et étiquetage, prétraitement, classification des sentiments et évaluation des sentiments.

La figure **4.1** donne un aperçu de l'approche proposée et les différentes étapes qui la composent : importer les données depuis Amazon sous forme d'un Dataset, convertir les données en DataFrame, ensuite le DataFrame est divisé en deux catégories : Données textuelles (Reviews) et Données numériques (Rating et Review Votes), nettoyer les données textuelles qui consistent à filtrer les données importées en supprimant toute donnée dupliquée, manquante ou aberrante et appliquer le prétraitement des données, traiter les données en utilisant la fouille de texte (Text Mining), calculer la précision, le rappel, le F1 score et dessiner la courbe ROC en utilisant les 3 classifieurs de notre système : SVM, KNN et Naïve Bayes.

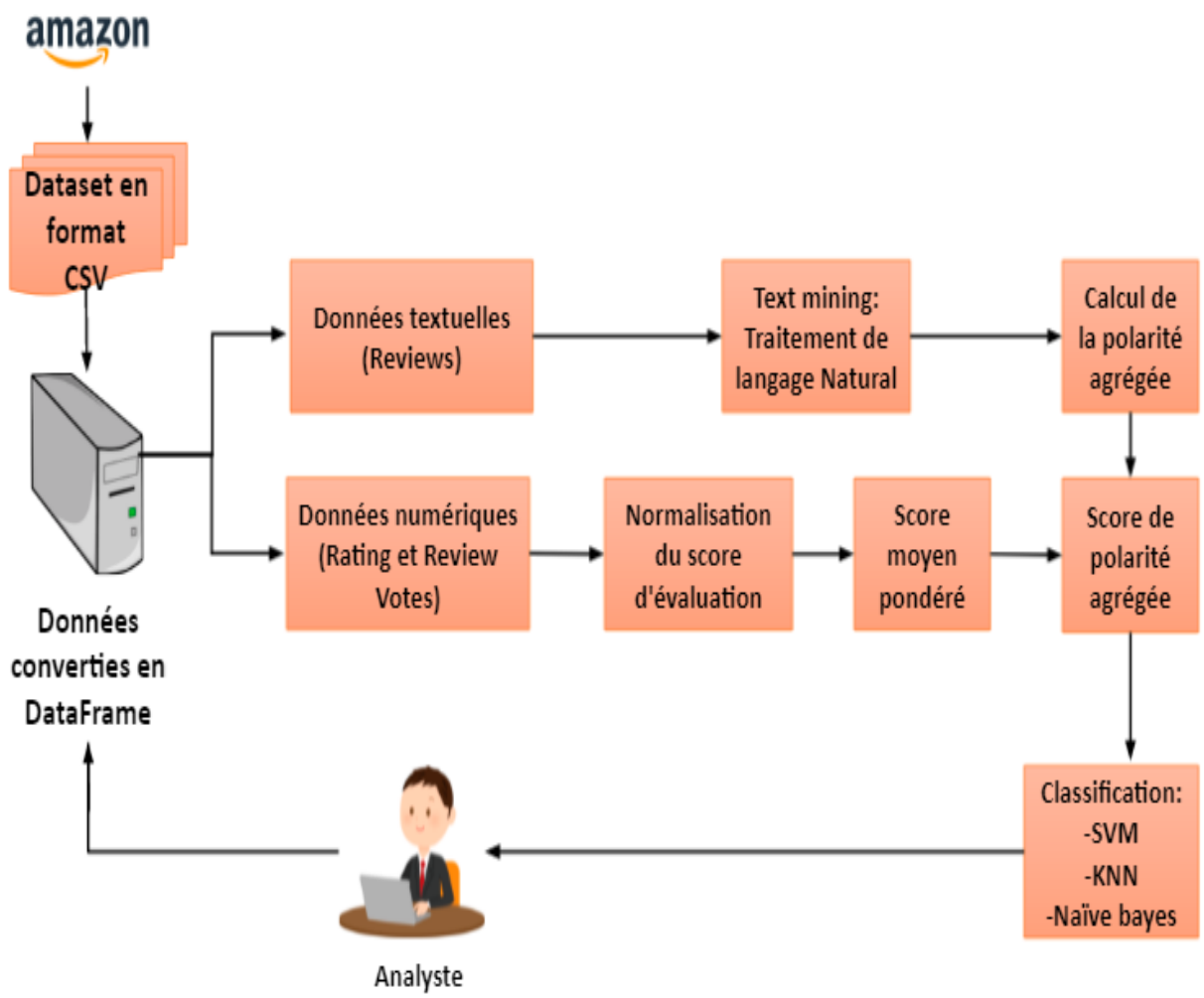

FIGURE 4.1 – Schéma global de l'approche.

Nous détaillons ci-dessous chaque étape comme suit :

4.2.1 Collecte des données

La première étape de l'analyse des sentiments est la collecte des données. Les données ont été extraites du site Amazon.

L'objectif le plus important de la collecte de données est de s'assurer que des données riches en informations et fiables sont collectées pour l'analyse statistique afin que des décisions fondées sur les données puissent être prises pour la recherche.

4.2.2 Prétraitement

Après la collecte des données, l'étape suivante est le prétraitement. Le prétraitement est une étape importante dans l'analyse des sentiments, c'est un outil puissant pour traiter les données textuelles en général. Cette étape est celle où les données sont préparées pour devenir des données prêtes à être analysées, en supprimant ou en modifiant les données qui sont incorrectes, incomplètes, non pertinentes, dupliquées ou mal formatées, dans le but de donner une importance à la sémantique des avis. Il y a plusieurs étapes dans ce prétraitement, dont le nettoyage, la conversion de la négation, la conversion des émoticônes, le pliage de la casse, la tokenisation, le filtrage des mot d'arrêt et le stemming. Voici une description détaillée des étapes de prétraitement ci-dessus :

- **Nettoyage**

  Le nettoyage est un processus de préparation des données en vue de leur analyse où les caractères et la ponctuation qui ne sont pas nécessaires sont supprimés du texte. Cette étape permet de réduire le risque dans l'ensemble de données. Exemples des caractères qui sont omis tels que les URL, les hashtags (#), la ponctuation comme les points (.), les virgules (,) et autres signes de ponctuation [34].

  Exemple : #Phone good# ... thank you for the great deal!!
  Après le nettoyage : Phone good thank you for the great deal

- **Conversion de la négation**

  Une autre caractéristique importante pour déterminer la polarité de l'opinion est la négation. Le seul lemme décrivant la négation peut changer complètement le sens et donc la polarité de la phrase.

Une difficulté avec la négation est qu'elle peut être décrite de manière très subtile, ainsi le sarcasme et l'ironie sont très difficiles à détecter [35].

Exemple : "Dont" et "Don't" sera "Do Not"
"Couldn't" sera "Could Not"

- **Remplacer les majuscules**

  Dans l'écriture d'un commentaire, il doit y avoir une forme différente de lettres, cette étape est un processus d'uniformisation des lettres, qu'elles soient minuscules ou majuscules.

  Exemple : "Nice PHONE" sera "nice phone"

- **Tokenisation**

  La Tokenisation est un processus effectué pour couper ou décomposer les phrases en parties ou en mots. Le résultat de cette déduction est appelé un jeton. Dans certains cas, le processus de tokenisation est également effectué en supprimant la ponctuation qui n'est pas nécessaire. Il existe plusieurs modèles de tokénisation qui peuvent être utilisés, à savoir unigramme, bigramme, trigramme et n-gramme[36].

  La tokenisation cherche à transformer un texte en une série de tokens individuels. Dans l'idée, chaque token représente un mot, et identifier des mots semble être une tâche relativement simple.

  Exemple : "It's battery life is great. It's very responsive to touch. "
  Cela serait divisé en :
  "It's" "battery" "life" "is" "great" "It's" "very" "responsive" "to" "touch"

  On peut également appliquer une tokenisation par phrase afin d'identifier les différentes phrases d'un texte. Cette étape peut à nouveau sembler facile, puisque a priori, il suffit de couper chaque phrase lorsqu'un point est rencontré (ou un point d'exclamation ou d'interrogation).

- **Filtrage**

  Le filtrage est l'étape qui consiste à éliminer les mots qui apparaissent en grand nombre mais qui sont considérés comme n'ayant aucune signification (mots vides). Fondamentalement, la liste des mots-clés est un ensemble de mots qui sont largement

utilisés dans diverses langues. La raison de la suppression des mots d'arrêt dans de nombreux programmes d'application liés à l'exploration de texte est parce que leur utilisation est trop générale, les utilisateurs peuvent donc se concentrer sur d'autres mots qui sont beaucoup plus importants[34].

Exemple : " This is a great product it came after two days of ordering it "

sera

" great product came two days ordering "

- **Stemming**

  Stemming (Racinisation) consiste à ne conserver que la racine des mots étudiés. Le but du stemming est de regrouper de nombreuses variantes d'un mot comme un seul et même mot. L'idée étant de supprimer les suffixes, préfixes et autres des mots afin de ne conserver que leur origine selon les règles d'anglais correctes[37].
  le mot résultant est le même. Cela permet notamment de réduire la taille du vocabulaire dans les approches de type sac de mots ou Tf-IdF.

Exemple : Had this phone before and loved it but was not working so I got this phone.

après le Stemming :

Have this phone before and love it but be not work so I get this phone.

- **Pondération des mots TF-IDF**

  La pondération des mots est un mécanisme permettant de donner un score sur la fréquence d'occurrence d'un mot dans un texte document. L'une des méthodes populaires de pondération des mots est TF-IDF (Term Frequency-Inverse Document Frequency), qui est une technique de recherche d'information qui pondère la fréquence d'un terme (TF) et la fréquence inverse des documents (IDF).

  L'idée d'avoir TF-IDF est de réfléchir à l'importance d'un mot pour un document d'une collection.

  Term frequency (TF) : qui est simplement le rapport entre le nombre de mots présents dans une phrase par rapport à la longueur de celle-ci.

  TF capture essentiellement l'importance du mot indépendamment de la longueur du document. Par exemple, un mot dont la fréquence est de 3 et la longueur de la

phrase de 10 n'est pas le même que lorsque la longueur de la phrase est de 100 mots. Il devrait avoir plus d'importance dans le premier scénario ; c'est ce que fait la TF.

Inverse Document Frequency (IDF) : L'IDF de chaque mot est *le logarithme* du rapport entre le nombre total de lignes et le nombre de lignes d'un document particulier dans lequel un mot spécifique est présent.

IDF = log(N/n), où **N** est le nombre total de lignes et **n** le nombre de lignes dans lesquelles le mot est présent.

L'IDF mesure la rareté d'un terme. Des mots comme «**a**» et «**the**» apparaissent dans tous les documents du corpus, mais les mots rares ne seront pas présents dans tous les documents. Donc, si un mot apparaît dans presque tous les documents, alors ce mot ne nous est d'aucune utilité puisqu'il ne nous aide pas à classer ou à la recherche d'informations. L'IDF annule ce problème.

TF-IDF est le produit simple de TF et IDF, de sorte que les deux inconvénients sont traités, ce qui rend les prédictions et la recherche d'informations relatives.

Chaque mot ou phrase a son score pour TF et IDF. De plus, le résultat du produit TF et IDF d'un terme font référence au poids TF-IDF de ce terme. Ainsi, on peut dire que plus le terme est rare et vice versa, plus le score TF-IDF (poids) est élevé. Par conséquent, le TF d'un mot est la fréquence d'un mot, tandis que le TF-IDF d'un mot est une mesure de l'importance de ce terme dans le corpus. Si les mots ont un poids TF-IDF élevé, leur contenu apparaîtra toujours parmi les premiers résultats de recherche. On peut donc arrêter de s'embêter avec les mots vides. En plus de trouver des mots avec une quantité de recherche plus élevée et une concurrence réduite avec succès[29].

4.2.3 Classification des sentiments

Après le prétraitement des données, l'étape suivante est la classification de l'analyse des sentiments. Cette étape est celle qui permet de fournir une formation et mettre en œuvre divers algorithmes d'exploration de données.

4.2.3.1 Classifieur SVM

En machine learning, les machines à vecteurs de support (SVM, ou réseaux de vecteurs de support) sont des modèles d'apprentissage supervisé avec des algorithmes d'apprentissage associés qui analysent les données utilisées pour la classification et l'analyse de régression. Dans cet algorithme, on représente chaque élément de données comme un point dans un espace à n dimensions (où n est le nombre de caractéristiques des utilisateurs) avec la valeur de chaque caractéristique étant la valeur d'une coordonnée particulière [38].

SVM est un classifieur discriminant formellement défini par un hyperplan de séparation. En d'autres termes, étant donné des données d'apprentissage étiquetées (apprentissage supervisé), l'algorithme produit un hyperplan optimal qui catégorise les nouveaux exemples.

Un modèle SVM est une représentation des exemples sous forme de points dans l'espace, cartographiés de manière à ce que les exemples des catégories distinctes soient divisés par un écart clair aussi large que possible. En plus d'effectuer une classification linéaire, les SVM peuvent efficacement effectuer une classification non linéaire, en cartographiant implicitement leurs entrées dans des espaces de caractéristiques à haute dimension.

Étant donné un ensemble d'exemples d'apprentissage, chacun étant marqué comme appartenant à l'une ou l'autre de deux catégories, un algorithme d'apprentissage SVM construit un modèle qui affecte les nouveaux exemples à l'une ou l'autre catégorie, ce qui en fait un classificateur linéaire binaire non probabiliste.

4.2.3.2 Classifieur K-NN

L'algorithme K-Nearest Neighbor (K-NN) est l'un des algorithmes de classification les plus basiques et pourtant essentiels dans le machine learning. Il appartient au domaine de l'apprentissage supervisé et trouve une application intense dans la reconnaissance des formes, l'exploration des données et la détection des intrusions.

L'algorithme des K-voisins les plus proches (KNN) est un algorithme de machine learning supervisé simple et facile à mettre en œuvre, qui peut être utilisé pour résoudre les problèmes de classification et de régression. L'algorithme KNN part du principe que les éléments similaires existent à proximité les uns des autres. En d'autres termes, les choses similaires sont proches les unes des autres. KNN capture l'idée de similarité (parfois appelée distance ou proximité) avec des mathématiques que nous avons peut-être apprises

dans notre enfance, le calcul de la distance entre des points sur un graphique. Il existe d'autres façons de calculer la distance, et l'une d'entre elles peut être préférable en fonction du problème à résoudre. Cependant, la distance en ligne droite (également appelée distance euclidienne) est un choix populaire et familier [39].

Elle est largement utilisable dans des scénarios de la vie réelle car elle est non paramétrique, ce qui signifie qu'elle ne fait aucune hypothèse sous-jacente sur la distribution des données.

4.2.3.3 Classifieur Naïve Bayes

L'algorithme Naïve Bayes est un machine learning qui utilise des calculs de probabilité selon le concept de l'approche bayésienne. L'utilisation du théorème de Bayes dans l'algorithme de Naïve Bayes consiste à combiner la probabilité antérieure et la probabilité conditionnelle dans une formule qui peut être utilisée pour calculer la probabilité de chaque classification possible [40].

L'algorithme utilise le théorème de Bayes et suppose que tous les attributs sont indépendants étant donné la valeur de la variable de classe. Cette hypothèse d'indépendance conditionnelle se vérifie rarement dans les applications du monde réel, d'où la caractérisation de l'algorithme comme naïf, mais il tend à être performant et à apprendre rapidement dans divers problèmes de classification supervisée. Cette "naïveté" permet à l'algorithme de construire facilement des classifications à partir de Dataset sans recourir à des schémas compliqués d'estimation itérative des paramètres.

4.3 Conclusion

Dans ce chapitre, nous avons présenté en détail notre approche d'analyse des sentiments dans le E-commerce en utilisant les principaux fondements, méthodes et techniques de classification des sentiments et d'opinion mining.

Notre approche est inspirée des travaux relatif à l'analyse des sentiments est permet d'utiliser plusieurs métriques pour la classification.

Dans le chapitre suivant, nous procéderons à l'explication de tous les aspects liés à l'implémentation de notre approche.

Chapitre 5

Expérimentation

Sommaire

5.1 Introduction

Dans le cadre de notre recherche, nous avons traité les textes anglais pour l'analyse du sentiment. Ce traitement permet d'extraire la polarité des opinions qui s'exprime en négatif et positif dans le cas d'une classification binaire et en négatif, neutre et positif

dans le cas d'une classification multiple. Les données d'entrée que nous avons utilisées sont des tweets, extraits de dataset pour entraîner et tester notre approche en temps réel sur des tweets extraits du site Amazon.

Dans ce chapitre, nous introduirons les différents aspects liés à l'implémentation de notre approche que nous avons développée, à savoir, les technologies, les logiciels et les langages choisis en utilisant différentes sources de données pour l'implémentation de notre approche.

5.2 Description du Dataset

Dataset : Pour simplifier, un Dataset est une source de données (comme un fichier texte) qui contient des lignes et des colonnes de données. Chaque ligne est généralement appelée un **point de données**, et chaque colonne est appelée une **caractéristique**. Un ensemble de données peut être un fichier CSV, TSV, une feuille de calcul Excel, une table dans un SGBDR, un document dans une base de données NoSQL, un output d'un service Web, etc. Il doit être analyser pour déterminer les caractéristiques les plus importantes et celles qui peuvent être ignorées en toute sécurité afin d'entraîner un modèle avec le Dataset donné.

Un ensemble de données peut être très petit (quelques caractéristiques et 100 lignes) ou très grand (plus de 1 000 caractéristiques et plus d'un million de lignes). Il peut y avoir des difficultés quant à la détermination des caractéristiques les plus importantes dans un grand ensemble de données. Dans cette situation un ***expert du domaine*** est nécessaire, qui comprend l'importance des caractéristiques, leurs interdépendances (le cas échéant) et la validité des valeurs des données pour les caractéristiques. En outre, il existe des algorithmes (appelés algorithmes de réduction de la dimensionnalité) qui peuvent aider à déterminer les caractéristiques les plus importantes. Par exemple, Le PCA (Principal Component Analysis) est l'un de ces algorithmes [41].

Le Dataset utilisé est extrait auprès du site Amazon et il est au format CSV car il est plus pratique pour Python de traiter ce type de fichiers dans le domaine de l'analyse des sentiments. La taille du Dataset est 128789 Ko et comporte 413825 avis sur des téléphones

mobiles non verrouillés vendus sur Amazon afin d'obtenir des informations sur les avis, les évaluations, le prix. Il peut être téléchargé [42] et utilisé par n'importe quel analyste ayant besoin d'un Dataset seulement on doit prendre en considération la structure et le type de données se trouvant dans le Dataset.

Le Dataset est composé de six (6) colonnes :

- **Product Name** qui est le nom du produit (Ex : Sprint EPIC 4G Galaxy SPH-D7).
- **Brand Name** qui est le nom de l'entreprise mère (Ex : Samsung).
- **Price** le prix du produit (Min : 1.73$, Max : 2598$, Moy : 226.86$).
- **Rating** qui est la note du produit (Entre 1 et 5).
- **Reviews** qui est les avis des clients (Description de l'expérience client).
- **Review Votes** qui est les votes d'évaluation (Nombre de personnes ayant voté la critique ; Min : 0, Max : 645, Moyenne : 1.50) [42].

La figure 5.1 représente l'étape de la collecte des données, nous avons importé ce dataset du site **Kaggle**, après l'importation, les données sont sous forme d'un dataframe (structure de données de deux dimentions ; plusieurs lignes et colonnes).

	Product Name	Brand Name	Price	Rating	Reviews	Review Votes
0	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	5	I feel so LUCKY to have found this used (phone...	1.0
1	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	4	nice phone, nice up grade from my pantach revu...	0.0
2	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	5	Very pleased	0.0
3	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	4	It works good but it goes slow sometimes but i...	0.0
4	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	4	Great phone to replace my lost phone. The only...	0.0
5	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	1	I already had a phone with problems... I know ...	1.0
6	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	2	The charging port was loose. I got that solder...	0.0
7	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	2	Phone looks good but wouldn't stay charged, ha...	0.0
8	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	5	I originally was using the Samsung S2 Galaxy f...	0.0
9	"CLEAR CLEAN ESN" Sprint EPIC 4G Galaxy SPH-D7...	Samsung	199.99	3	It's battery life is great. It's very responsi...	0.0

FIGURE 5.1 – Collecte des données sous forme de Dataset.

5.3 Environnement de développement

5.3.1 Anaconda

C'est une distribution libre et open source des langages de programmation Python et R, appliqué au développement d'applications dédiées à la science des données et à l'apprentissage automatique .

Implémentation et tests Une distribution est un langage de programmation, certaines librairies et autres fonctionnalités. Anaconda est donc une distribution Python, faite pour la Data Science.

5.3.2 Jupyter Notebook

Jupyter Notebook est un outil open-source, basé sur un navigateur, qui fonctionne comme un Lab Notebook virtuel pour prendre en charge les flux de travail, le code écrit en Python, les données et les visualisations détaillant le processus de recherche. Il est lisible par la machine et par l'homme, ce qui facilite l'interopérabilité et la communication scientifique. Ces notebooks peuvent vivre dans des dépôts en ligne et fournir des connexions à des objets de recherche tels que des ensembles de données, du code, des

documents de méthodes, des flux de travail et des publications qui se trouvent ailleurs [43].

5.4 Langage de programmation

Python : est un langage de programmation open source créé par le programmeur **Guido van Rossum** en **1991**. Il tire son nom de l'émission *Monty Python's Flying Circus*.

Il s'agit d'un **langage de programmation interprété**, qui ne nécessite donc pas d'être compilé pour fonctionner. Un programme " interpréteur " permet d'exécuter le code Python sur n'importe quel ordinateur. Ceci permet de voir rapidement les résultats d'un changement dans le code. En revanche, ceci rend ce langage plus lent qu'un langage compilé comme le C [44].

5.5 Bibliothèques de Python

5.5.1 Pandas

Acronyme de Python Data Analysis Library, Pandas est un package de Python qui est compatible avec d'autres packages Python, tels que NumPy, Matplotlib, etc. Son installation se fait en ouvrant un shell de commande et en invoquant cette commande pour Python 3.x : *pip3 install pandas*.

À bien des égards, le package Pandas a la sémantique d'une feuille de calcul, et il fonctionne également avec divers types de fichiers, tels que les fichiers xsl, xml, html, CSV et TSV. Pandas prend des données stockées dans des fichiers CSV ou TSV et fournit un type de données appelé Dataframe (similaire à un dictionnaire Python) avec des fonctionnalités extrêmement puissantes (similaires à celles d'un feuille de calcul) [45].

5.5.2 Numpy

NumPy est une bibliothèque Python qui fournit de nombreuses méthodes pratiques et aussi de meilleures performances. NumPy fournit une bibliothèque de base pour le

calcul scientifique en Python, avec des tableaux multidimensionnels performants et de bonnes fonctions mathématiques vectorielles, ainsi que le support de l'algèbre linéaire et des nombres aléatoires vectorisées, ainsi qu'un support pour l'algèbre linéaire et les nombres aléatoires.

NumPy est calqué sur MATLAB, avec un support pour les listes, les tableaux, etc. NumPy est plus facile à utiliser que MATLAB, et il est très courant dans le code TensorFlow ainsi que dans le code Python [45].

5.5.3 Tkinter

Tkinter, ou "interface Tk", est une librairie de Python qui fournit une interface à la boîte à outils GUI de tk, développée en TCL (Tool Command Language) et multiplateforme, avec un support pour Linux, MAC OS et MS Windows. Tk est présent en natif dans Linux et MAC OS, et peut être facilement installé sur MS Windows. Elle fait partie de Python et s'appelle "Tkinter" dans les versions antérieures à 3, et "tkinter" dans les suivantes [46].

5.5.4 Matplotlib

Matplotlib est l'une des bibliothèques de visualisation de données les plus utilisées de Python. Cette bibliothèque a été créée par un certain **John Hunter** qui, avec plusieurs contributeurs, ont mis plus de temps à inciter cette librairie à être utilisée par les scientifiques et philosophes du monde entier .

Matplotlib est une bibliothèque graphique pour la visualisation de données en Python qui fait partie intégrante de la pile de Data Science, elle est facilement prise en charge par NumPy, Pandas et d'autres bibliothèques pertinentes[47].

5.5.5 Sklearn

Il s'agit d'une bibliothèque Python à code source ouvert qui met en œuvre une série d'algorithmes de machine learning, de prétraitement, de validation croisée et de visualisation à l'aide d'une interface unifiée. Scikit-learn expose une grande variété d'algorithmes de machine learning, supervisé et non supervisé. Ce qui est important, les algorithmes sont implémentés dans un langage de haut niveau[48].

5.6 Mise en service

La figure 5.2 représente l'interface d'« Accueil » de notre application avec une barre de menu, le premier menu « **Accueil** » contient 7 boutons :

- Le bouton **Importer le Dataset :** en cliquant dessus une fenêtre s'ouvre affichant les fichiers avec l'extension CSV existant dans nos dossiers. On séléctionne le fichier (Dataset) qu'on veut utiliser dans notre système d'analyse des sentiments comme le montre la **la figure 5.3**.
- Le bouton **Nombre d'avis :** en cliquant dessus le système nous compte le nombre d'avis total se trouvant dans notre Dataset en supprimant les données dupliquée, manquante ou aberrante, il compte aussi le nombre d'avis positifs, négatifs et neutre comme dans **la figure 5.4**.
- Le bouton **Naïve Bayes :** en cliquant dessus le système applique l'algorithme NB sur les données et affiche les résulats comme la précison, le rappel et le f1 score de ce classifieur ainsi que son graphe ROC comme dans **la figure 5.5**.
- Le bouton **K-NN :** en cliquant dessus le système applique l'algorithme K-NN sur les données et affiche les résulats comme la précison, le rappel et le f1 score de ce classifieur ainsi que son graphe ROC comme dans **la figure 5.6**.
- Le bouton **SVM :** en cliquant dessus le système applique l'algorithme SVM sur les données et affiche les résulats comme la précison, le rappel et le f1 score de ce classifieur ainsi que son graphe ROC comme dans **la figure 5.7**.
- Le bouton **Comparer les classifieurs :** en cliquant sur ce bouton le système compare les trois classifieurs qu'on a utilisé et démontre cette comparaison en dessinant Un diagramme à barres avec leur taux de précision comme dans **la figure 5.8**.
- Le bouton **Quitter :** en cliquant dessus une autre fenêtre s'affiche pour confirmer si on veut quitter l'application comme dans **la figure 5.14**.

La première interface lors du lancement de notre application et le premier menu qui s'affiche est le menu « Accueil ».

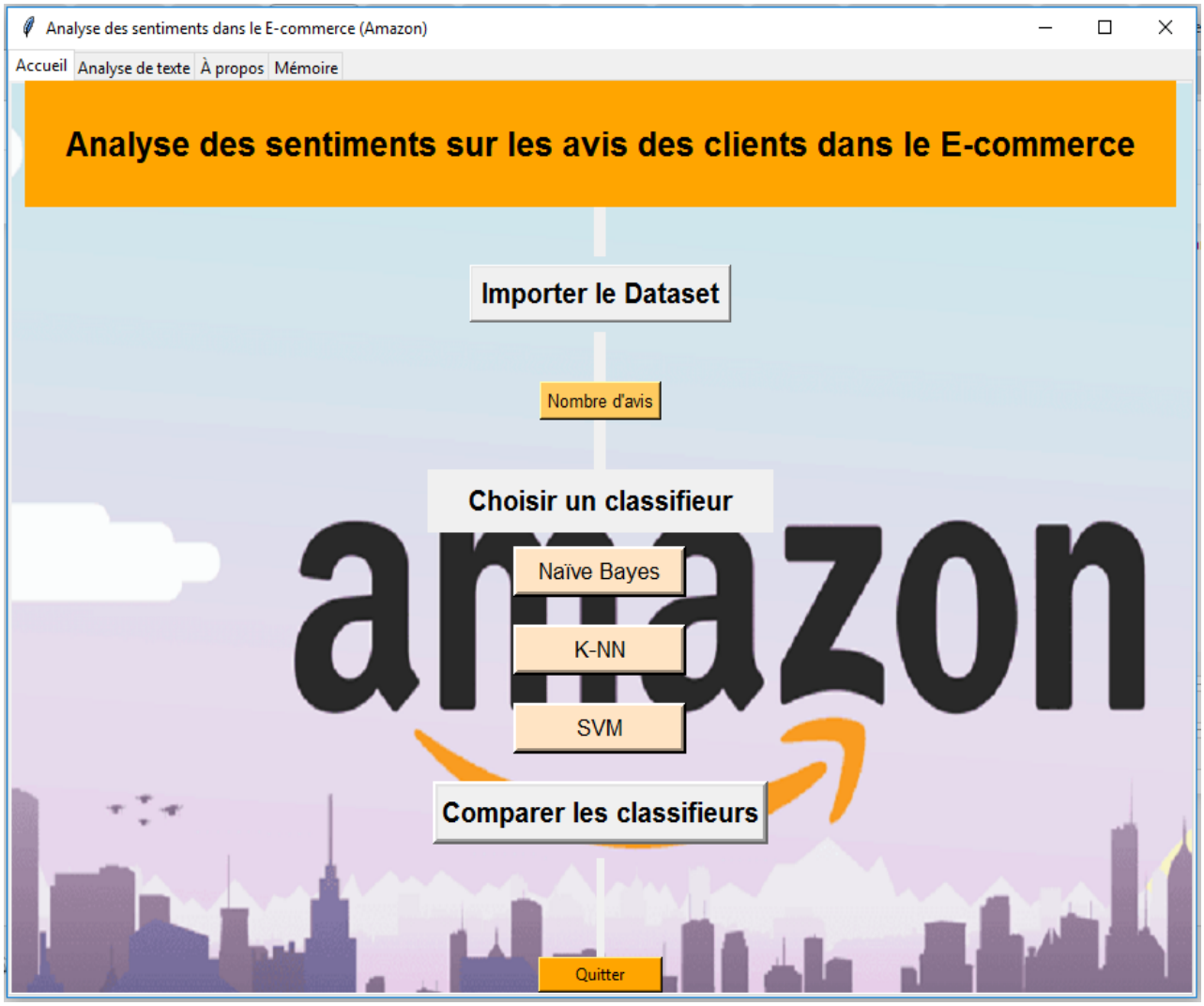

FIGURE 5.2 – Interface « Accueil ».

La figure 5.3 : Après avoir cliquer sur le bouton « **Importer le Dataset** », une nouvelle fenêtre s'ouvre affichant les fichiers existants dans notre disque dur, on sélectionne le Dataset qui contient les données, dans notre cas c'est fichier CSV nommé Amazon_Mobiles.csv et on l'importe pour l'utiliser dans notre système d'évaluation.

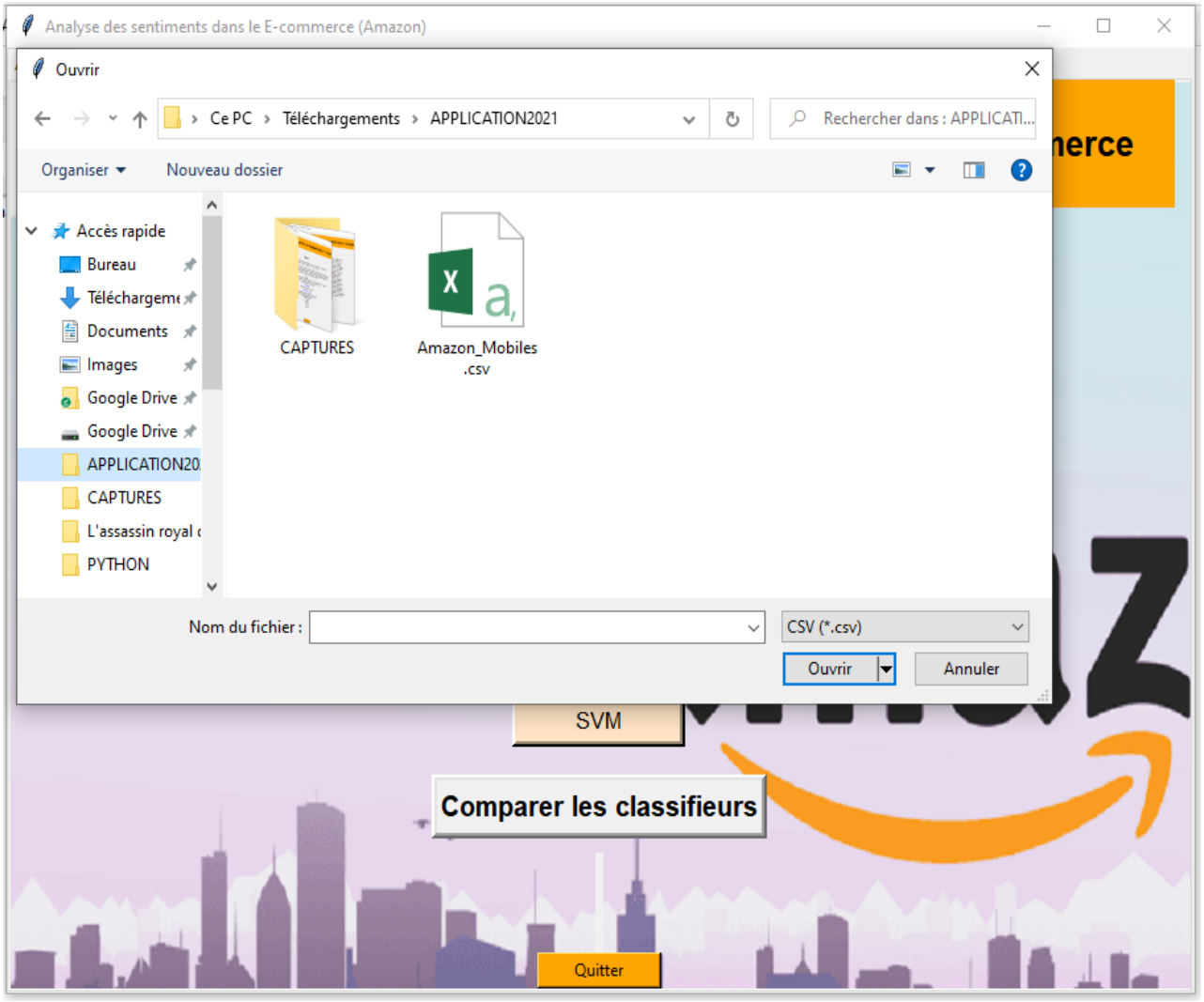

FIGURE 5.3 – Interface pour importer le Dataset.

La figure 5.4 : Après avoir cliquer sur le bouton « **Nombre d'avis** », une nouvelle fenêtre s'ouvre affichant le nombre d'avis dans le Dateset, le nombre d'avis positifs, le nombre d'avis négatifs et le nombre d'avis neutres après le nottoyage des données.

FIGURE 5.4 – Interface « Nombre d'avis ».

La figure 5.5 : Après avoir cliquer sur le bouton « **Naïve Bayes** », une nouvelle fenêtre s'ouvre affichant les résulats de cet algorithme appliqué sur les données comme la précision, le rappel, le f1-score ainsi que son graphe ROC.

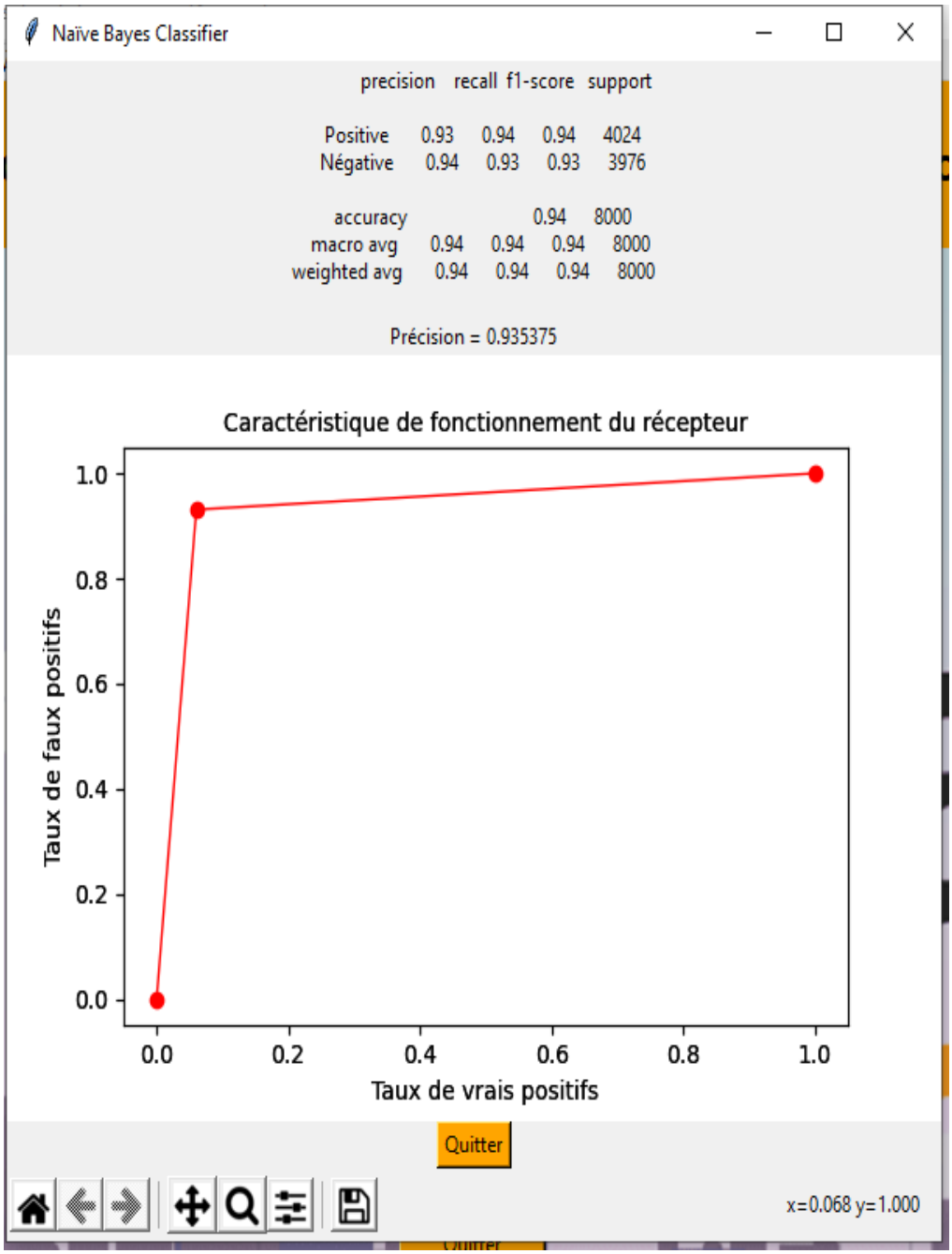

FIGURE 5.5 – Interface « Naïve Bayes Classifier ».

La figure 5.6 : Après avoir cliquer sur le bouton « **KNN** », une nouvelle fenêtre s'ouvre affichant les résulats de de cet algorithme appliqué sur les données comme la précision, le rappel, le f1-score ainsi que son graphe ROC.

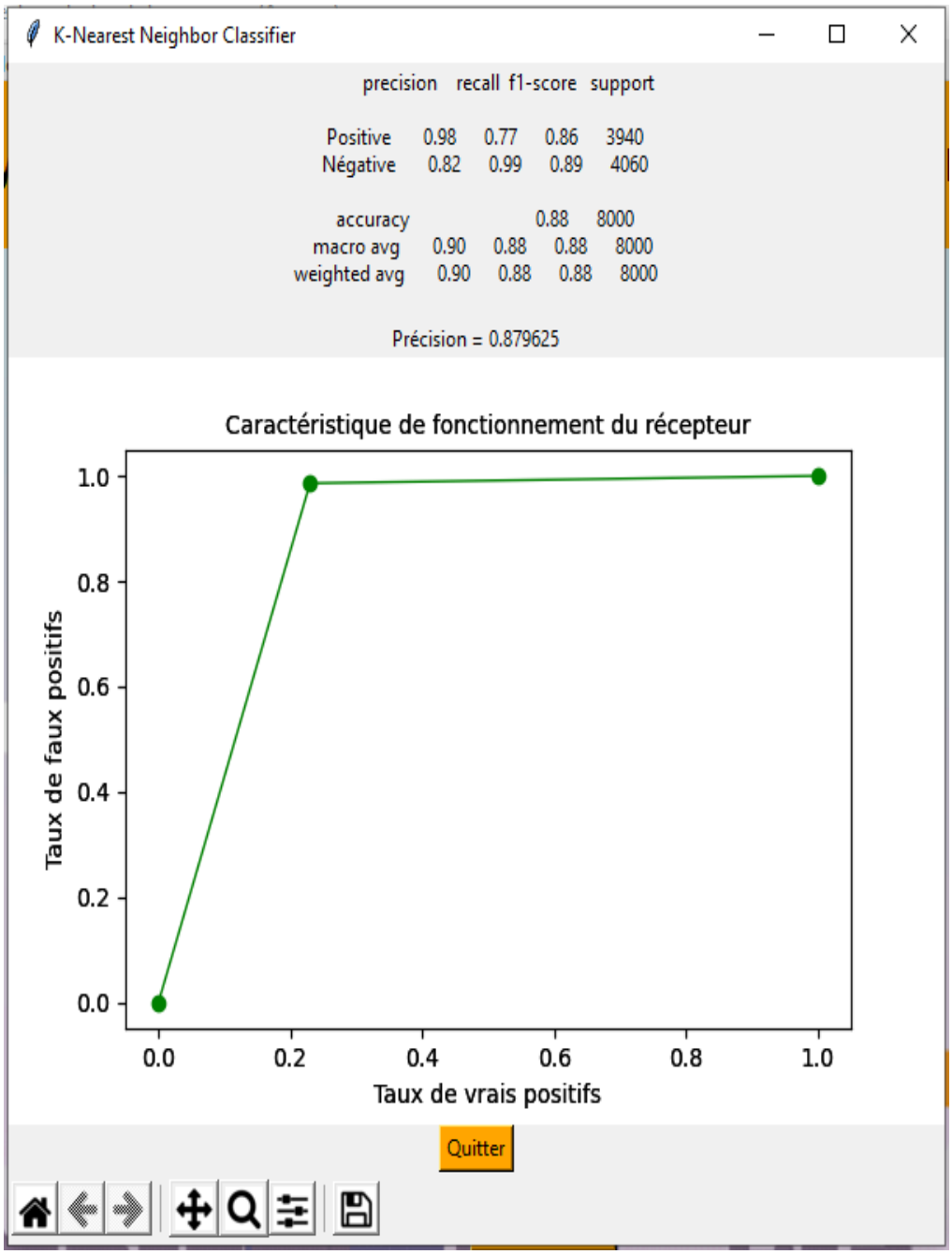

FIGURE 5.6 – Interface « K-Nearest Neighbor Classifier »

La figure 5.7 : Après avoir cliquer sur le bouton « **SVM** », une nouvelle fenêtre s'ouvre affichant les résulats de de cet algorithme appliqué sur les données comme la précision, le rappel, le f1-score ainsi que son graphe ROC.

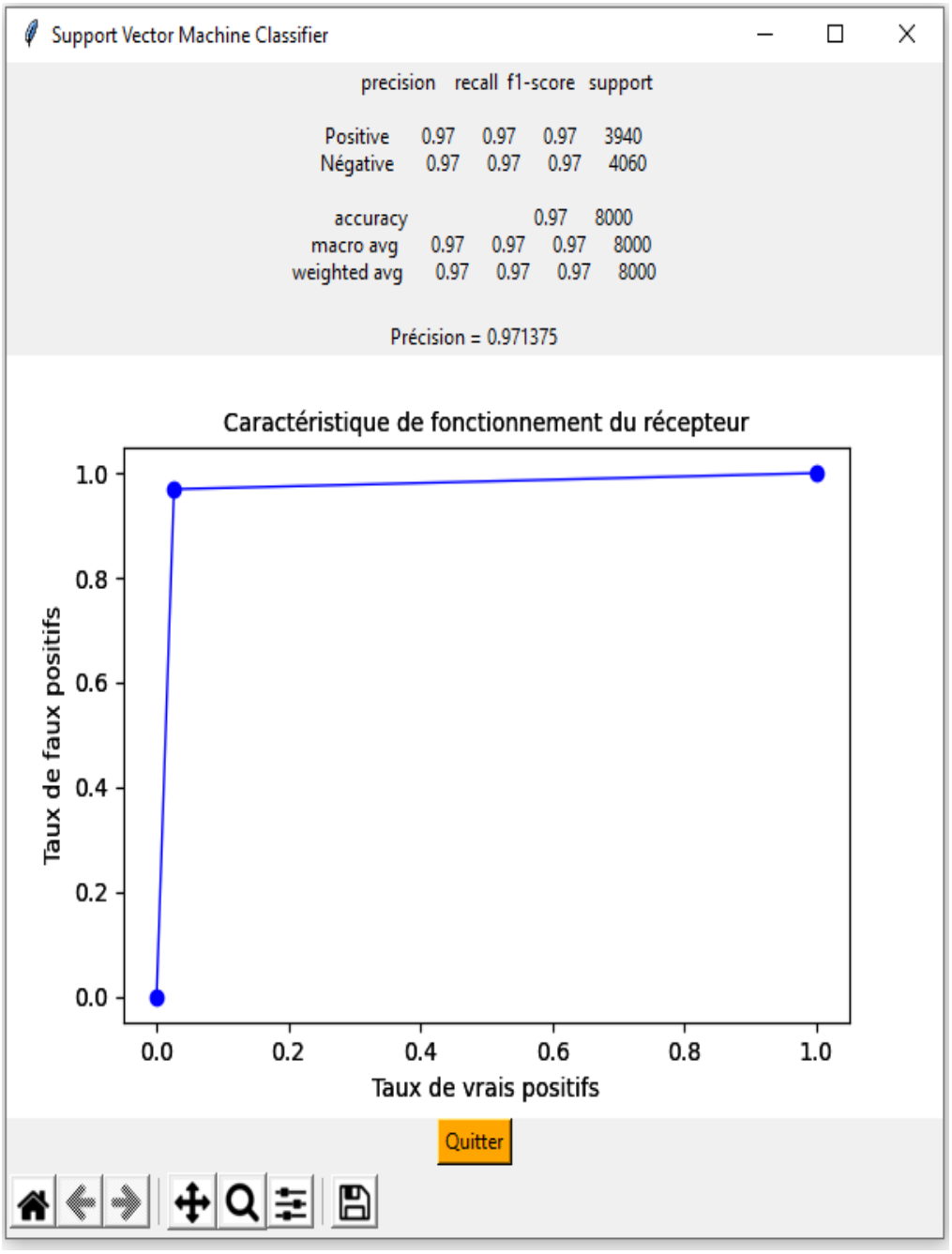

FIGURE 5.7 – Interface « Support Vector Machine Classifier ».

La figure 5.8 : Après avoir cliquer sur le bouton « **Comparer les classifieurs** », une nouvelle fenêtre s'ouvre affichant Une courbe à barres qui compare les trois classifieurs d'après leur taux de précision .

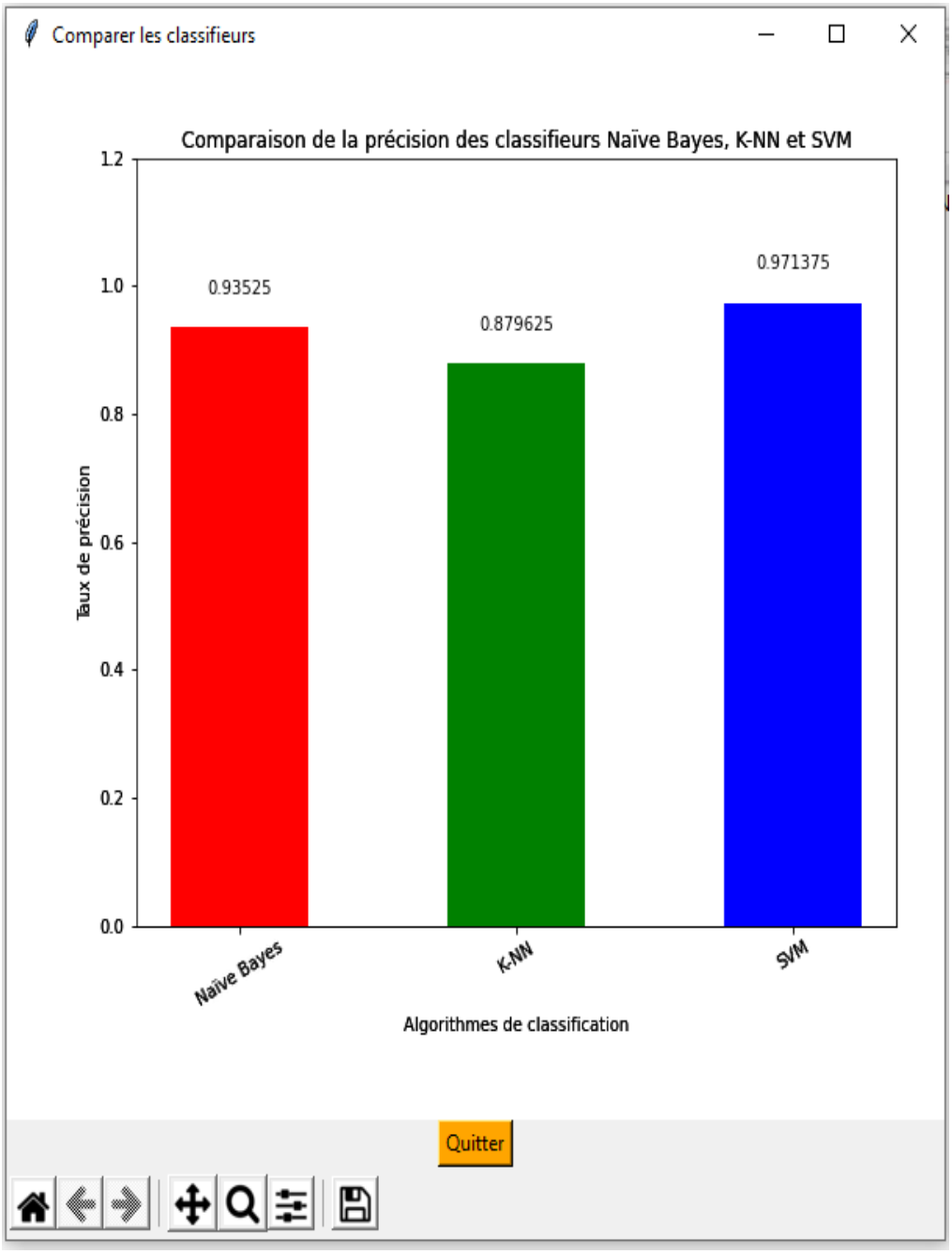

FIGURE 5.8 – Interface « Comparer les classifieurs ».

La figure 5.9 représente l'interface du deuxième menu « Analyse de texte » de l'application qui se compose de :

— **Un champ de texte** où saisir le texte à analyser.

— **Un bouton « Réinitialiser »** pour vider le champ de texte de tout saisi.

— **Un bouton « Sentiment »** pour calculer le pourcentage du sentiment positif, négatif et neutre et afficher le sentiment du texte saisi.

— **Un bouton « Tokénisation »** pour découper le texte en mots.

— **Un bouton « Analayse »** pour analyser et calculer la subjectivité et la polarité du texte saisi.

— **Un bouton « Réinitialiser le résultat »**pour vider le champ de texte de tout résultat affiché.

— **Un bouton « Quitter »**pour quitter l'application.

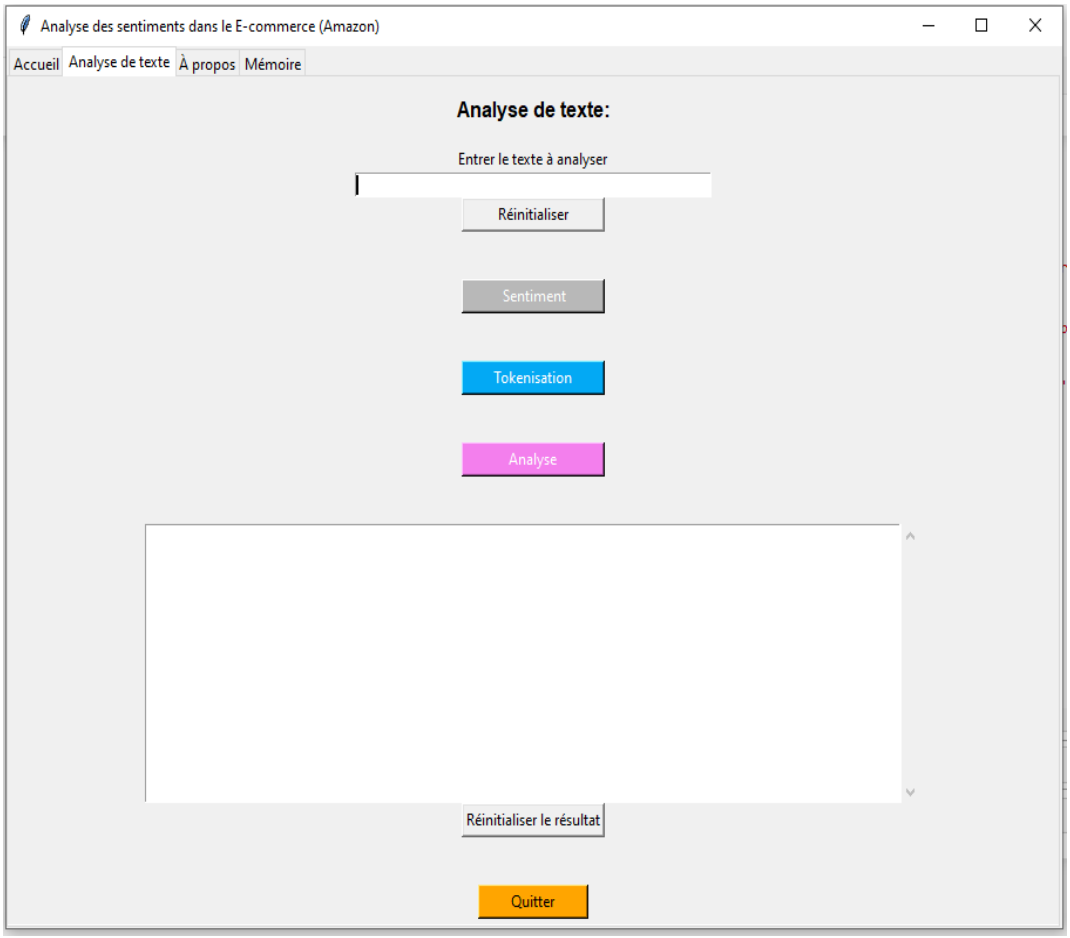

FIGURE 5.9 – Interface du menu « Analyse de texte ».

La figure 5.10 représente un exemple d'« Analyse de texte » :

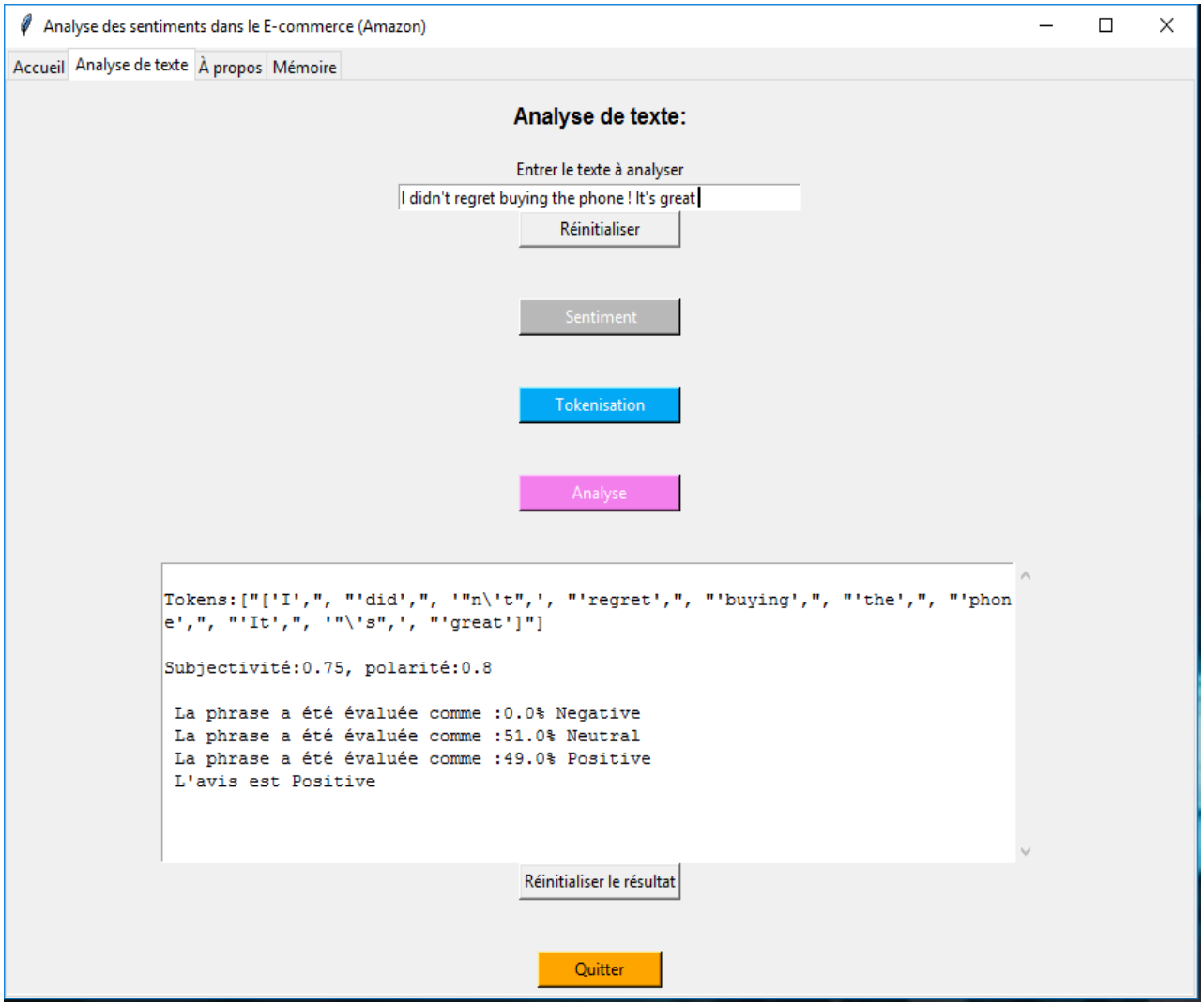

FIGURE 5.10 – Interface « Exemple d'Analyse de texte ».

La figure 5.11 représente l'interface du menu « À propos » :
Une petite description de notre application, et la source du Dataset utilisé ainsi que les outils et techniques utilisé pour développer le système d'analyse des sentiments.

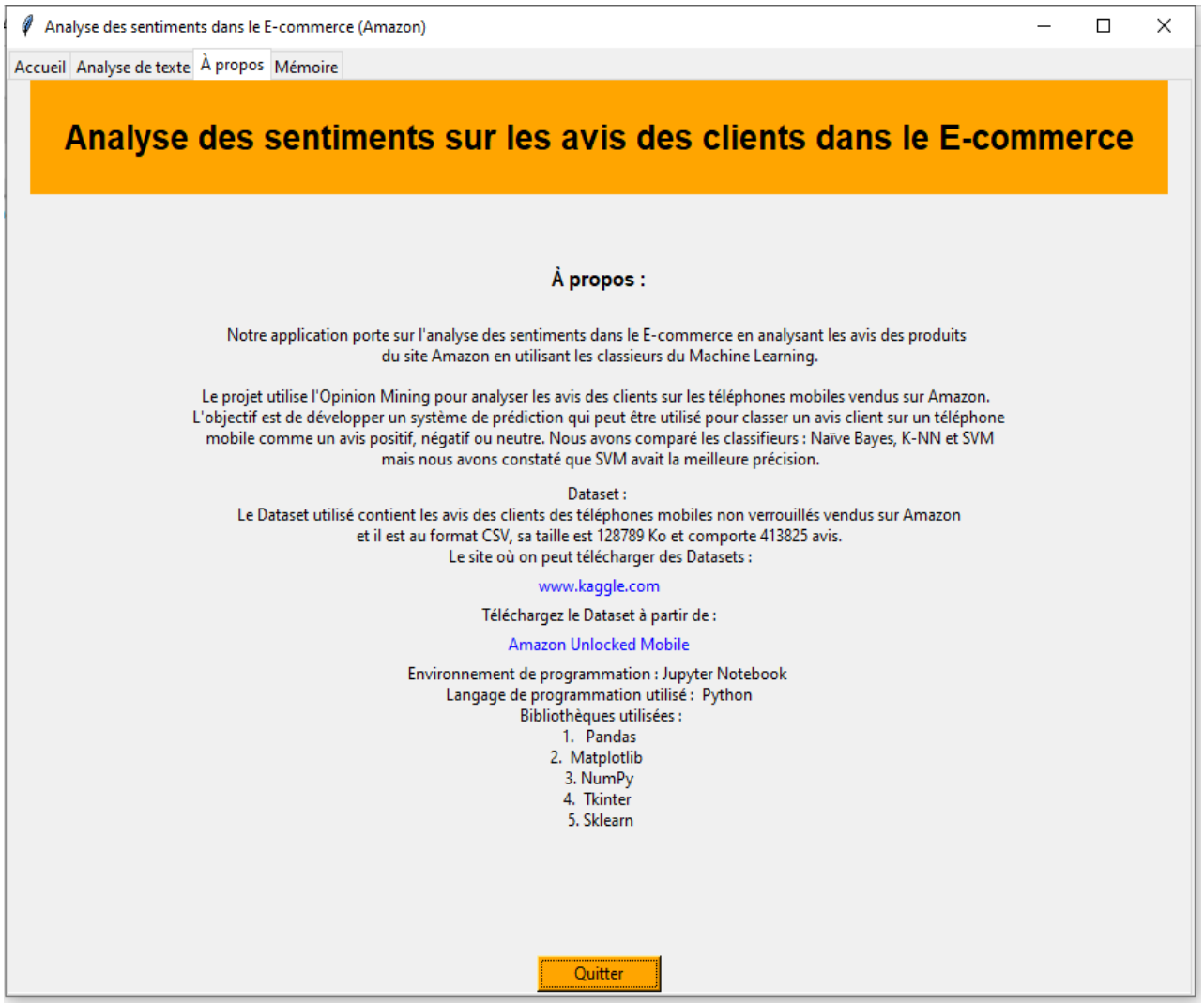

FIGURE 5.11 – Interface du menu « À propos ».

La figure 5.12 représente l'interface du menu « Mémoire » :
On pourra lire le résumé de notre mémoire ainsi que le lien pour accéder et lire le mémoire en cliquant sur le bouton en dessous.

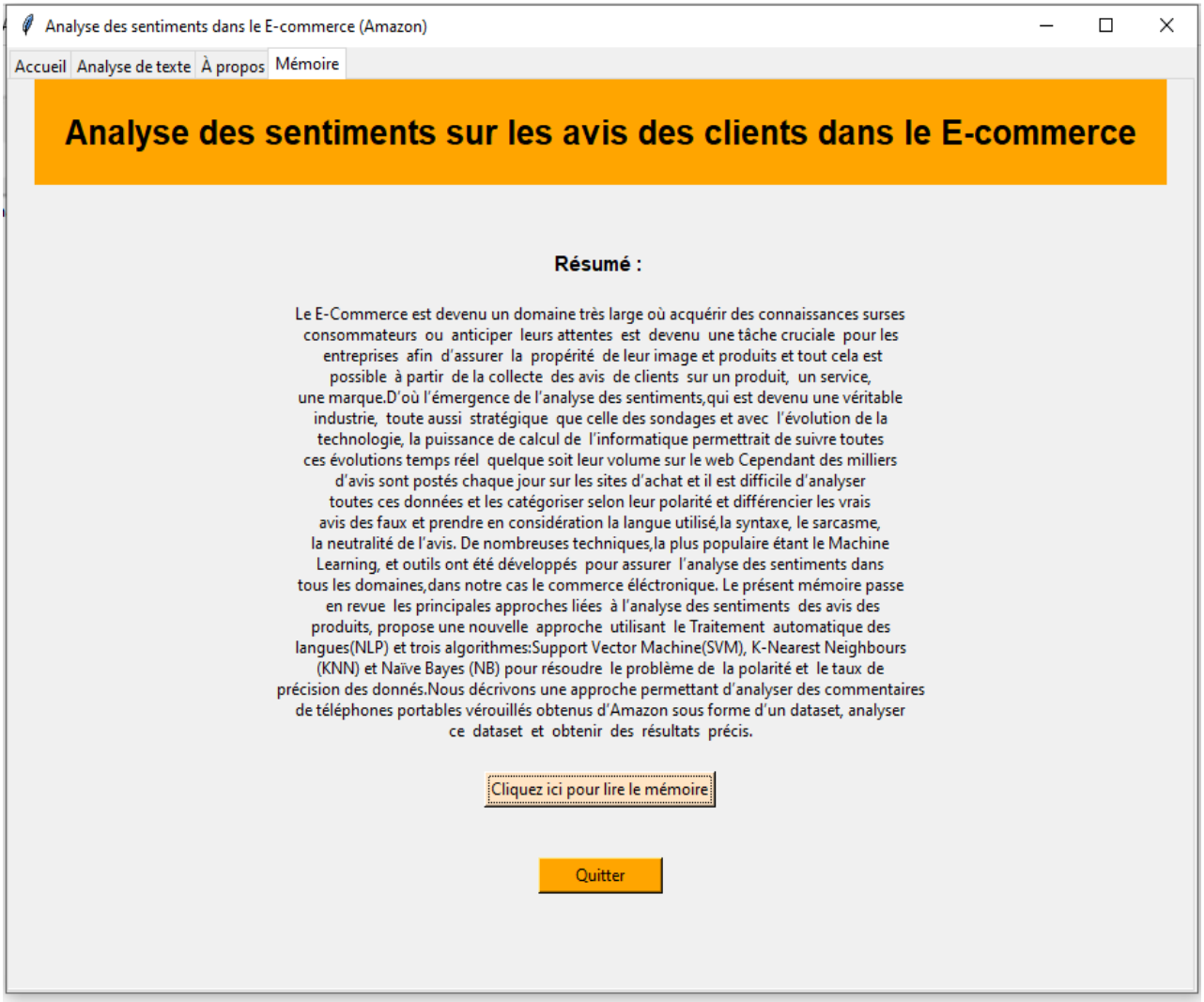

FIGURE 5.12 – Interface du menu « Mémoire ».

Le mémoire qui porte sur l'analyse des sentiments et l'application développée.

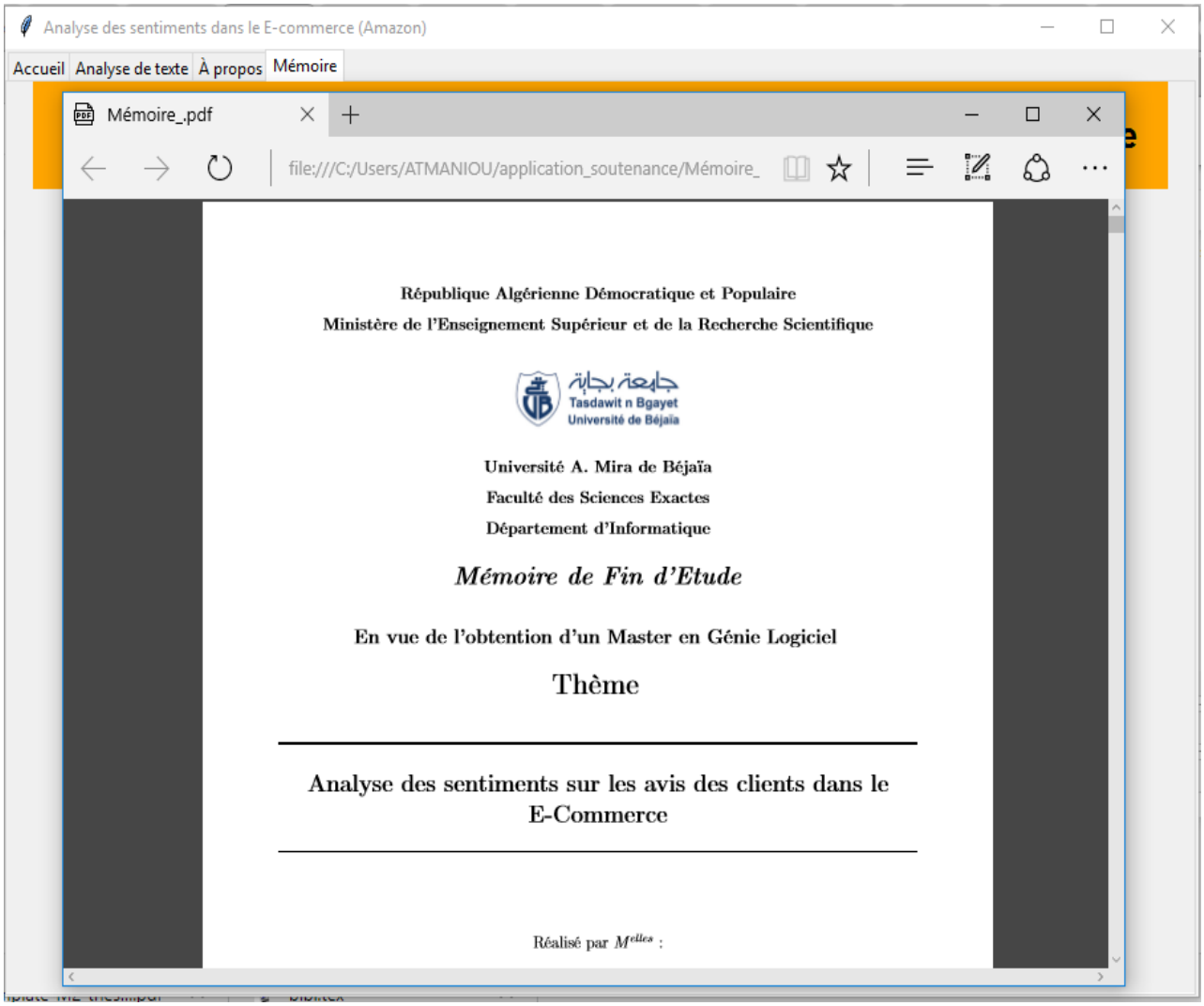

FIGURE 5.13 – Interface pour « Lire le Mémoire ».

La figure 5.14 : Après avoir cliquer sur le bouton « **Quitter** », une nouvelle fenêtre s'ouvre affichant une boîte de dialogue pour confirmer si on veut quitter l'application.

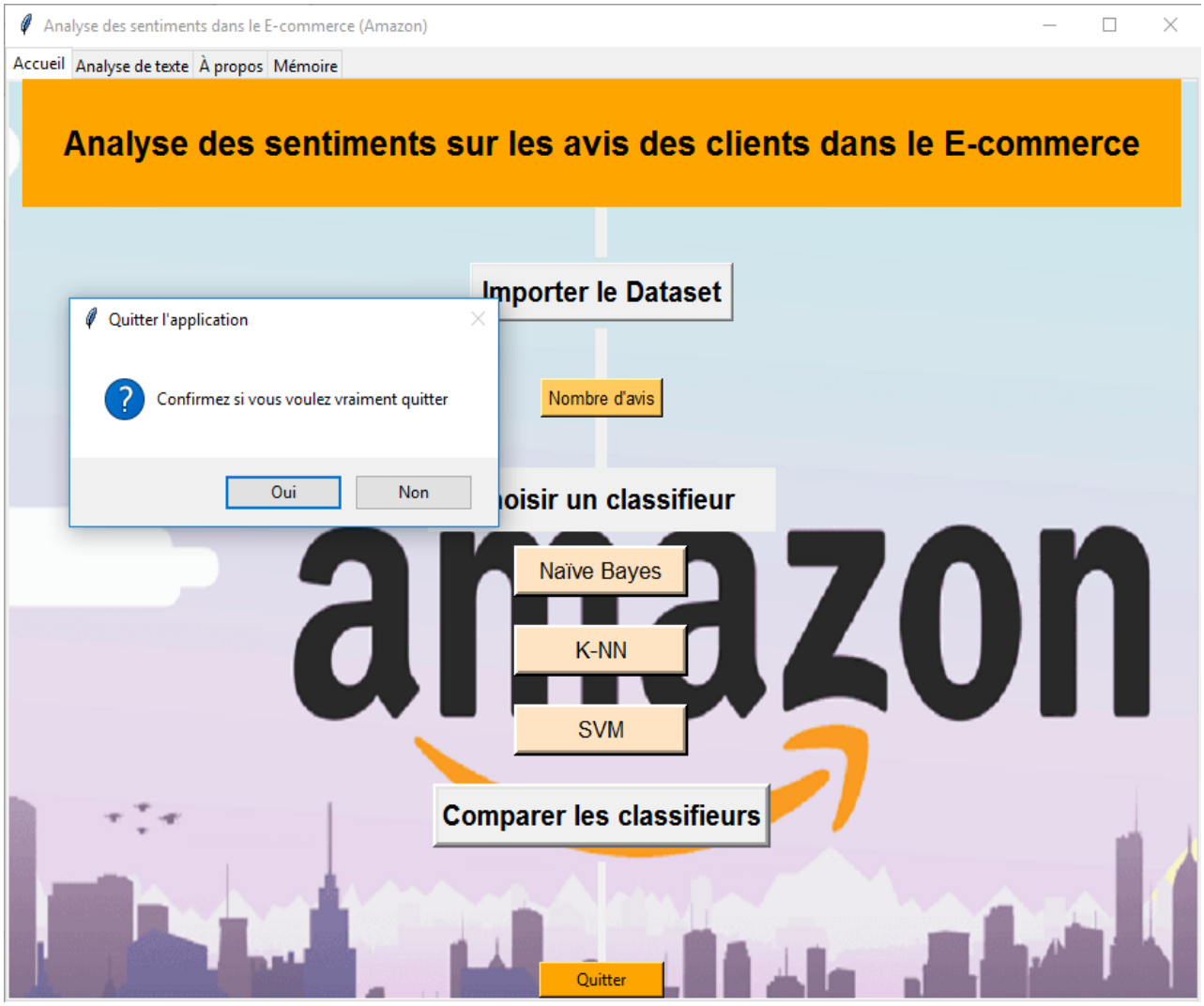

FIGURE 5.14 – Interface pour « quitter » l'application.

5.6.1 Évaluation des sentiments

Une fois le processus de classification de l'analyse des sentiments une étape supplémentaire est nécessaire pour déterminer la qualité du processus qui a été effectué, à savoir l'évaluation des résultats. A ce stade, les performances des calculs qui ont été effectués seront testées avec les paramètres de exactitude, précision et rappel.

L'évaluation d'un modèle est une partie essentielle de la construction d'un modèle de machine learning efficace. Pour démontrer la performance de la méthode proposée, certaines mesures ont été évaluées comme suit :

5.6.1.1 Précision du classificateur

La précision mesure l'exactitude d'un classificateur. Une précision plus élevée signifie moins de faux positifs , tandis qu'une précision plus faible signifie plus de faux positifs. Ceci est souvent en contradiction avec le rappel, car un moyen facile d' améliorer la précision est de diminuer le rappel[49].

Précision = Vrai Positif / (Vrai Positif + Faux Positif)

5.6.1.2 Rappel du classificateur

Le rappel mesure l'exhaustivité, ou la sensibilité , d'un classificateur. Un rappel plus élevé signifie moins de faux négatifs , tandis qu'un rappel plus faible signifie plus de faux négatifs. L'amélioration du rappel peut souvent diminuer la précision car il devient de plus en plus difficile d'être précis à mesure que l'espace d'échantillonnage augmente[49].

Rappel = Vrai Positif / (Vrai Positif + Faux Négatif)

5.6.1.3 Score F1

Le score F1 est une mesure de la précision d'un test. Il prend en compte à la fois la précision et le rappel du test pour calculer le score, F-Score est la moyenne harmonique de la précision et du rappel. Cela vous indiquera comment votre système fonctionne.

F1-Score= [2 * (Précision * Rappel) / (Précision + Rappel)]

5.6.1.4 La courbe ROC

La courbe ROC est connu comme le taux de faux positifs et la sensibilité est également connue comme le taux de vrais positifs. La courbe est elle-même le rapport entre la surface sous la courbe et la surface totale. Elle prend en compte les probabilités prédites pour déterminer la performance de notre modèle. Elle est largement utilisée lorsque l'ensemble de données est déséquilibré car la précision n'est pas une mesure de performance fiable pour les données déséquilibrées[50].

La courbe ROC est généralement une bonne métrique sous forme de graphique pour digérer la qualité du classifieur. Elle montre de meilleures prédictions lorsque la ligne dépasse la ligne de base diagonale. **Les figures 5.15**, **5.16** et **5.17** montrent les courbes ROC des classifieurs nb, knn et svm. La courbe ROC du classifieurs NB couvre 93 % de la zone qui dépasse la ligne de base diagonale. La courbe ROC du classifieurs KNN couvre 87 % de la zone qui dépasse la ligne de base diagonale. La courbe ROC du classifieurs svm couvre 97 % de la zone qui dépasse la ligne de base diagonale.

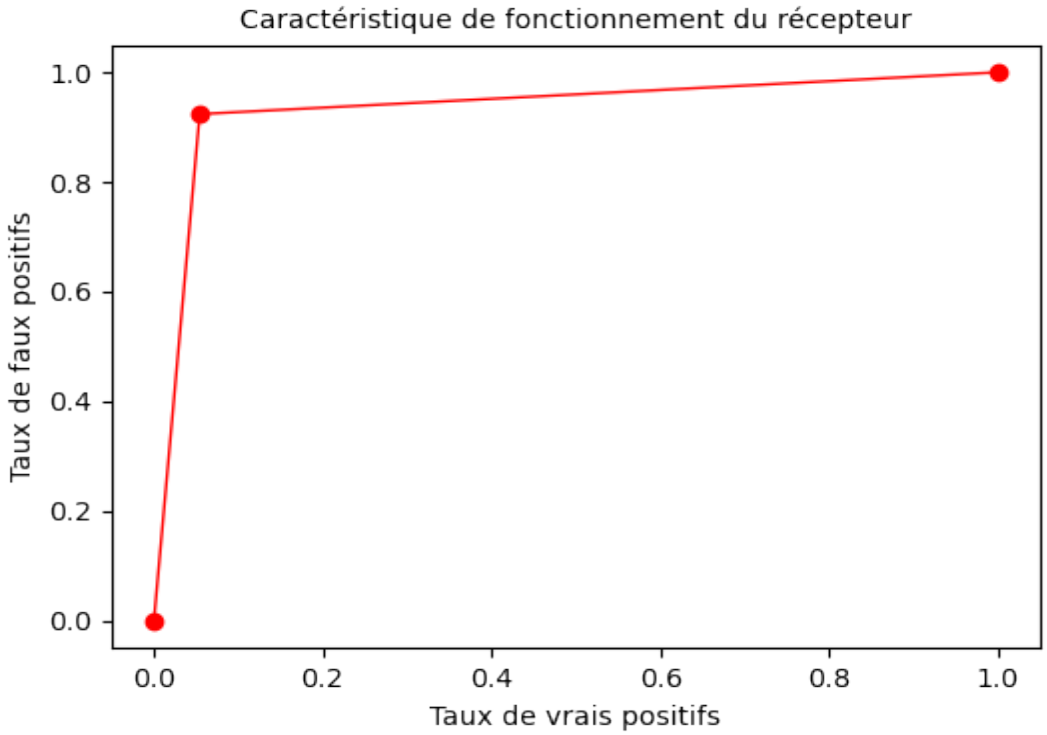

FIGURE 5.15 – La courbe ROC du classifieur Naïve bayes

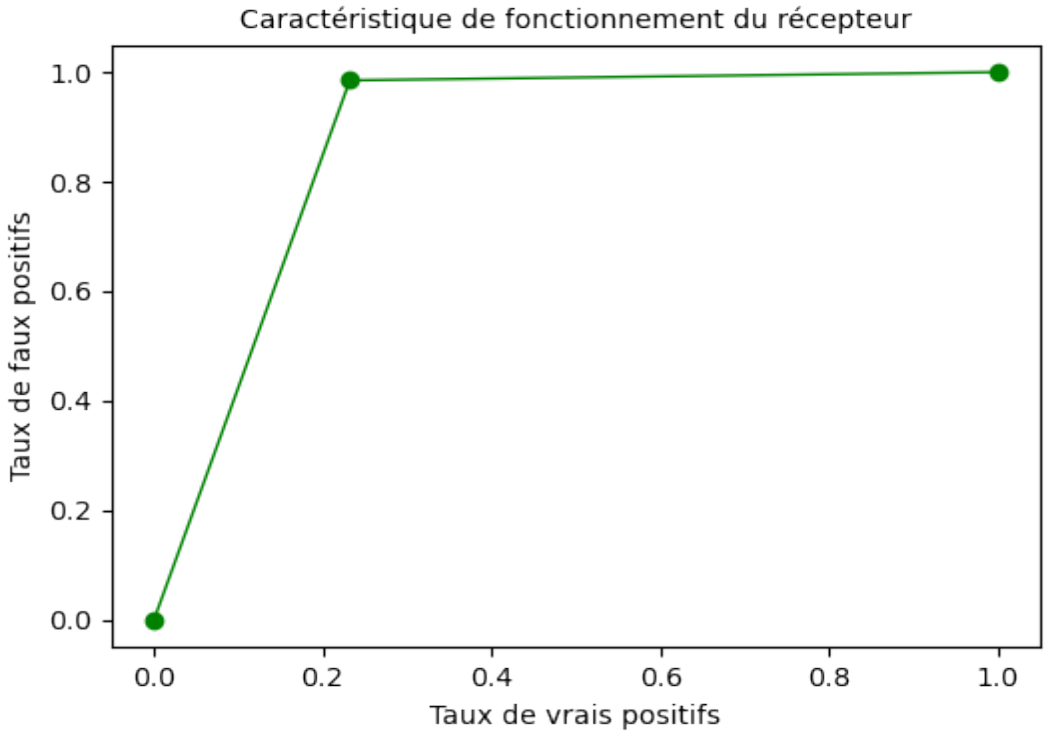

FIGURE 5.16 – La courbe ROC du classifieur KNN

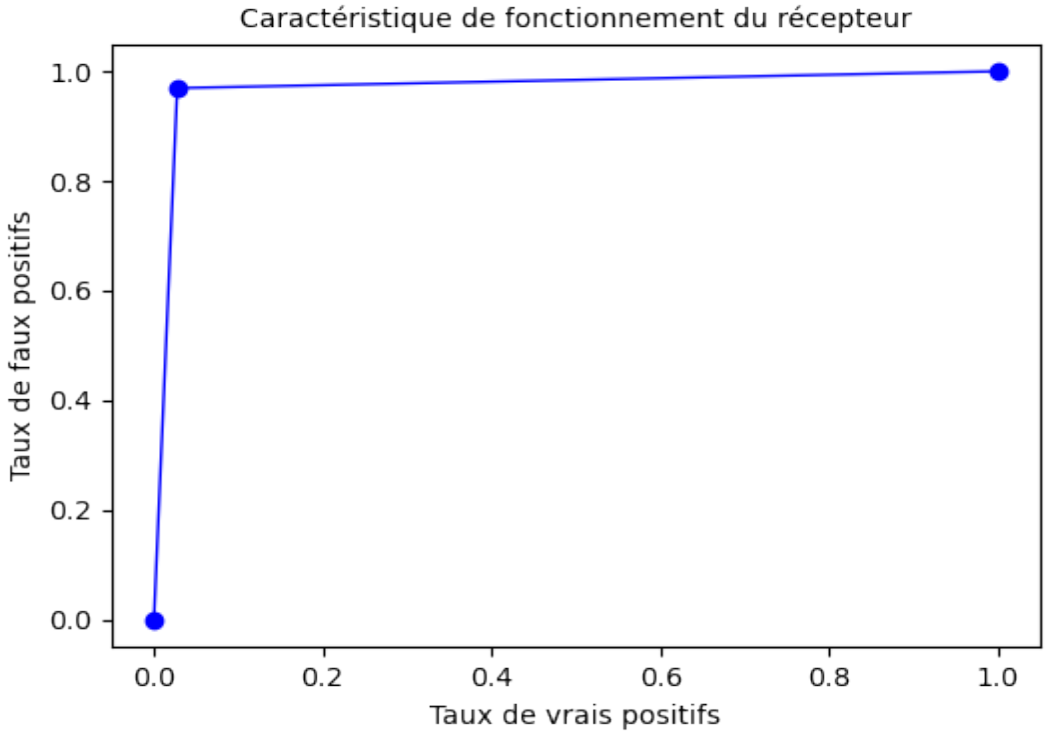

FIGURE 5.17 – La courbe ROC du classifieur SVM

Le tableau ci-dessous représente la précision, le rappel et le F1-score pour les classifieurs naïve bayes, KNN et SVM.

Classifieur	Sentiment	Précision	Rappel	F1-score
Naïve Bayes	Positif	0.93	0.94	0.94
	Négatif	0.94	0.93	0.93
KNN	Positif	0.98	0.77	0.86
	Négatif	0.82	0.99	0.89
SVM	Positif	0.97	0.97	0.97
	Négatif	0.97	0.97	0.97

TABLE 5.1 – Tableau de comparaison de la précision, du rappel et du F1 score des 3 classifieurs NB, KNN et SVM.

Il est souhaitable d'avoir à la fois une haute précision et un rappel élevé pour obtenir une précision finale élevée. Le score F1 tient compte à la fois de la précision et du rappel et donne un nombre unique à comparer entre les produits. Sur la base de la comparaison des scores F1, ce qui suit est arrivé pour l'ensemble de données donné.
Le classifieur SVM est meilleur que KNN et NB pour trouver un sentiment positif.
Le classifieur SVM est légèrement meilleur que KNN et NB pour trouver un sentiment négatif.

La figure 5.18 : montre la comparaison des performances des algorithmes de classification en utilisant la courbe barres. Plus la barre est élevée, plus l'algorithme est capable de distinguer les avis positifs et les avis négatifs. Le classifieur SVM atteint une précision moyenne de 97 % meilleur que NB (93 %) et KNN (87 %).

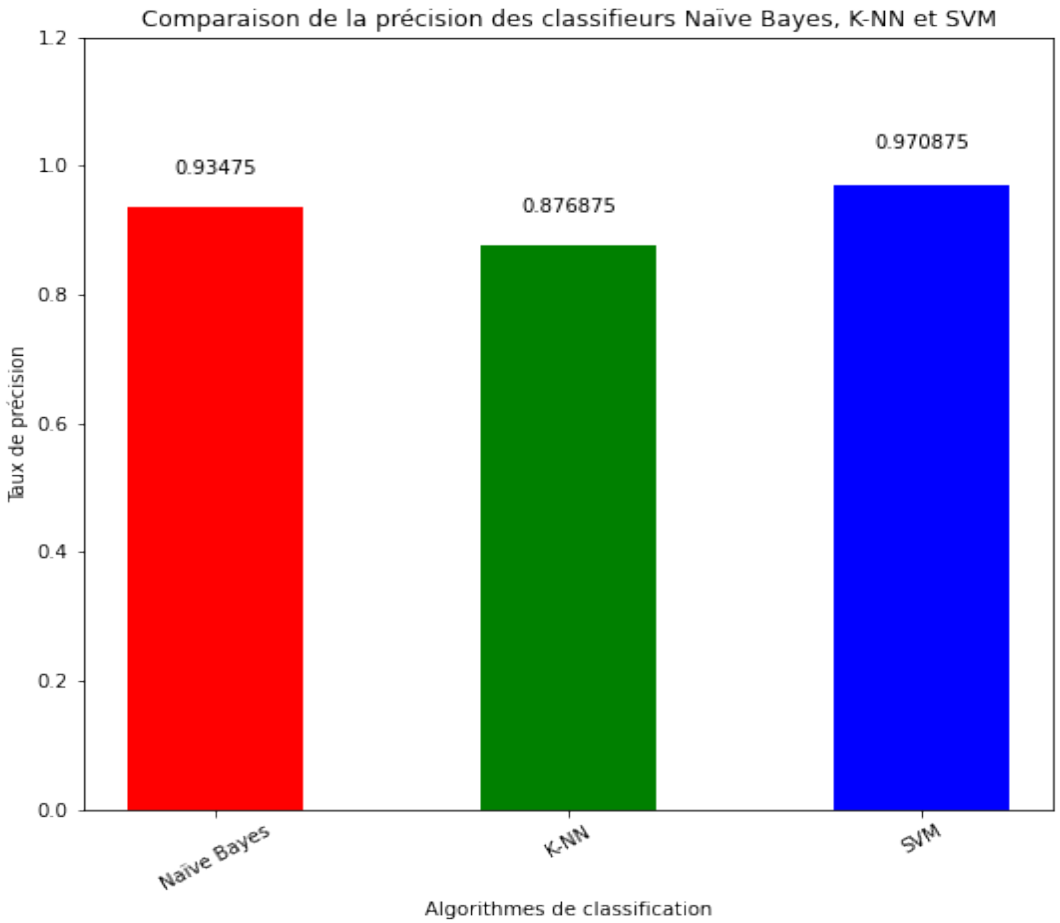

FIGURE 5.18 – Comparaison de la précision des classifieurs

Pour conclure, nous comprenons que pour l'ensemble de données donné, le classifieur SVM fonctionne mieux que les classifieurs KNN et naïve bayes. Il est important que nous devions considérer plusieurs ensembles de données et prendre la précision moyenne pour connaître le classement final du produit.

5.7 Conclusion

Dans ce chapitre, nous avons présenté l'essentiel de notre travail qui consiste à créer un système d'analyse d'opinion pour détecter les sentiments des clients dans le E-Commerce. Pour l'implémentation, nous avons utilisé des méthodes de classification les plus connue : SVM, KNN et NB. Pour l'implémentation, nous avons choisi le site d'Amazon en examinant les tweets pour les classifier en : positive, négative ou neutre. Notre système s'intègre dans le domaine d'intelligence artificielle précisément (Machine Learning). Car la précision de la classification augmente à chaque fois quand exécute l'algorithme de classification.

Chapitre 6

Conclusion générale

Ce travail a été réalisé dans le cadre de notre projet de fin de cycle Master en informatique option Génie logiciel. Il a consisté en une approche du Machine Learning pour l'analyse des sentiments des avis des clients dans le E-Commerce (avis extraits d'Amazon).

L'analyse des sentiments et l'opinion Mining est un domaine émergent, dans ces dernières années plusieurs recherches s'intéressent à la tâche de l'analyse des sentiments, en particulier dans le domaine du micro-blogging. Notre travail se concentre sur l'analyse des avis des clients du site Amazon.

Le Dataset qu'on a utilisé sur les produits a été pris sur le site Web d'Amazon contenant les avis des téléphones mobiles vérouillés. L'analyse des sentiments a été effectuée sur chaque revue de produit et ensuite classifier en utilisant des algorithmes du Machine Learning.

Le Machine Learning est la technique la plus pratique. Il serait donc beaucoup plus facile de passer en revue des milliers de commentaires si un modèle était adopté pour polariser ces derniers et apprendre à partir d'eux. Dans ce travail de recherche, le sentiment du consommateur a été analysé par le Naïve Bayes, KNN classifiers et le SVM.

Notre mémoire est constituée de six (6) chapitres organisés comme suit :

Dans le premier chapitre, nous avons défini notre contexte et problématique ainsi que nos objectifs , nous avons également détaillé notre méthodologie de travail.

Dans le deuxième chapitre, nous avons présenté quelques définitions du domaine d'études qu'est l'analyse des sentiments, ses caractéristiques, ses difficultés, les problèmes liés à ce domaine et le Machine Learning et ses types.

Dans le troisième chapitre, nous avons élaboré l'état de l'art qui représente tous les travaux connexes que nous avons synthétisés, nous avons présenté ceci dans un tableau qui contient les grandes lignes de chaque approche synthétisé, tout en suivant chaque travail par un bref paragraphe qui le résume, par la suite nous avons procédé à une analyse comparative entre les approches des documents connexes et notre approche.

Dans le quatrième chapitre, nous avons présenté en détail notre approche qu'on a utilisé au cours de notre projet ainsi que ses différentes étapes pour effectuer une analyse des sentiments à partir des commentaires.

Dans le cinquième chapitre, nous avons abordé les divers aspects liés à l'implémentation de notre approche que nous avons développés, à savoir, les technologies, les logiciels et les langages choisis en utilisant differentes sources de données pour l'implémentation de notre approche.

Nous avons poussé le projet aussi loin que possible, il reste cependant, de nombreuses étapes à ajouter. Notamment utiliser le Deep Learning à la place du Machine Learning et utiliser d'autres algorithmes de classification comme Random Forest, etc. Nous pensons à enrechir notre approche et l'implémenter dans de meilleures conditions matérielles et logicielles.

La réalisation de ce projet a été riche d'enseignements sous plusieurs aspects. Elle nous a permis d'acquérir de nouvelles compétences, et de mettre en pratique les connaissances théoriques que nous avons acquis le long de notre cursus universitaire. Nous avons progressé dans de nombreux domaines notamment dans la programmation en Python, l'analyse des sentiments, l'élaboration de l'état de l'art...

Bibliographie

[1] Raheesa Safrin, K.R. Sharmila, T.S.Shri Subangi, E.A.Vimal ,(2017), SENTIMENT ANALYSIS ON ONLINE PRODUCT REVIEW, *International Research Journal of Engineering and Technology* (IRJET), Vol.04, N° 04, p. 2381.

[2] Alexander Ligthart, Cagatay Catal and Bedir Tekinerdogan, (2021), Systematic reviews in sentiment analysis : a tertiary study, *Artificial Intelligence Review*, Information Technology Group, Wageningen University Research, Wageningen, The Netherlands, p.2.

[3] F. Barbieri,H. Saggion, (2014), *Modelling irony in Twitter*, in : Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Sweden, Gothenburg, pp. 56-64.

[4] A. Reyes, P. Rosso, and T. Veale, (2013), *A multidimensional approach for detecting irony in Twitter*, Language Resources and Evaluation, 47(1), p. 239-268.

[5] Johan Bollen, Huina Mao, Xiaojun Zeng, (2011), *Twitter mood predicts the stock market*,Journal of Computational Science.

[6] Nathan Kallus, (2014), *On the Predictive Power of Web Intelligence and Social Media The Best Way to Predict the Future Is to tweet It*, Massachusetts Institute of Technology, USA.

[7] Gabe Ignatow, Rada Mihalcea, *An Introduction to Text Mining*, 2018 by SAGE Publications, Inc, p. 12.

[8] Aditya Jain, Gandhar Kulkarni, Vraj Shah, (2018), *Natural Language Processing*,International Journal of Computer Sciences and Engineering, Vol.06, N°1, pp.161-167.

[9] Penubaka Balaji, O.Nagaraju, D.Haritha, (2017),*levels of sentiment analysis and its challenges : A literateure review*,Conference on big data analytics and computational intelligence (ICBDAC).

[10] Hitesh Parmar, Sanjay Bhanderi, Glory Shah, (2014), *Sentiment Mining of Movie Reviews using Random Forest with Tuned Hyperparameters*,Conference : International Conference on Information ScienceAt : Kerala.

[11] John Wieting, Mohit Bansal, Kevin Gimpel, Karen Livescu, (2016), *Towards universal paraphrastic sentence embeddings*, In Proceedings of International Conference on Learning Representations.

[12] BOJANOWSKI P, GRAVE E, JOULIN A, MIKOLOV T. (2016), *Enriching word vectors with subword information*, arXiv preprint arXiv :1607.04606.

[13] H. Zhou, F. Song, (2015), *Aspect-level sentiment analysis based on a generalized probabilistic topic and syntax model*, The Twenty-Eighth International Flairs Conference.

[14] G. Vinodhini, R. Chandrasekaran, (2012), *Sentiment analysis and opinion mining : a survey*, International Journal, 2(6), p.282–292.

[15] D. Oelke, M. Hao, C. Rohrdantz, D.A. Keim, U. Dayal, L.-E. Haug, (2009), *Visual opinion analysis of customer feedback data*, IEEE Symposium on Visual Analytics Science and Technology, p.187–194.

[16] Kamalakshi V.Deshmukh , Sankirti S.Shiravale, (2018), *Ambiguity Resolution in English Language for Sentiment Analysis*,IEEE punecon.

[17] : R. K. Jena, (2019), *Sentiment mining in a collaborative learning environment : capitalising on big data*, Behaviour Information Technology.

[18] Muhammad Marong, Nowshath K Batcha et Raheem Mafas, (2020), *Sentiment Analysis in E-Commerce : A Review on The Techniques and Algorithms*, Journal of Applied Technology and Innovations, Vol.04, N°1, pp. 6-7.

[19] E. Biernat, M. Lutz, (2015), *Well-posed learning problems*, Machine Learning, p.2-3.

[20] Tom M. Mitchell, (2017), *Data science : fondamentaux et études de cas Machine learning avec Python et R*, Eyrolles, p.23-24.

[21] Chloé-Agathe Azencott, (2018), *Introduction au Machine Learning*, Dunod.

[22] Neena Devasia and Reshma Sheik (2016). Feature extracted sentiment analysis of customer product reviews. *2016 International Conference on Emerging Technological Trends (ICETT)*, pp. 1-6.

[23] Mohamad Syahrul Mubarok, Adiwijaya and Muhammad Dwi Aldhi (2017), *Aspect-based Sentiment Analysis to Review Products Using Naïve Bayes*,International Conference on Mathematics : Pure, Applied and Computation, 1867-020060.

[24] Paramita Ray and Amlan Chakrabarti, *Twitter Sentiment Analysis for Product Review Using Lexicon Method*, International Conference on Data Management, Analytics and Innovation (ICDMAI) Zeal Education Society, Pune, India, pages 211-216, 2017.

[25] E. Suganya and S. Vijayarani, *Sentiment Analysis for Scraping of Product Reviews from Multiple Web Pages Using Machine Learning Algorithms*,Department of Computer Science, Bharathiar University, Coimbatore 641046, India, pages 677–685, 2018.

[26] R. S. Jagdale, , V. S. Shirsat et S. N. Deshmukh. (2018). *Sentiment Analysis on Product Reviews Using Machine Learning Techniques*. Cognitive Informatics and Soft Computing, 639–647. `https://doi.org/10.1007/978-981-13-0617-4_61`

[27] Pankaj, Prashant Pandey, Muskan and Nitasha Soni, *Sentiment Analysis on Customer Feedback Data : Amazon Product Reviews*, International Conference on Machine

Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), pages 320-322, 2019.

[28] Rajesh Bose, Raktim Kumar Dey, Sandip Roy and Debabrata Sarddar, *Sentiment Analysis on Online Product Reviews*, pages 559-569, 2020.

[29] Sanjay Dey, Sarhan Wasif, Dhiman Sikder Tonmoy, Subrina Sultana, Jayjeet Sarkar, and Monisha Dey, *A Comparative Study of Support Vector Machine and Naive Bayes Classifier for Sentiment Analysis on Amazon Product Reviews*,International Conference on Contemporary Computing and Applications (IC3A), pages 217-220, 2020.

[30] Neha Nandal, Rohit Tanwar and Jyoti Pruthi, *Machine learning based aspect level sentiment analysis for Amazon products*, Department of Computer Science and Technology, Manav Rachna University, Faridabad, India, 2020.

[31] Studer, R. Benjamins and D. Fensel, (1998), *Knowledge Engineering : Principles and Methods*, IEEE Trans. on Data and Knowledge Eng., 25(1–2), 161–197.

[32] A. Gómez-Pérez, and O. Corcho, (2002), *Ontology Languages for the Semantic Web*, IEEE Intelligent Systems, 17(1), 54-60.

[33] Teerawat Polsawat, Ngamnij Arch-int, Somjit Arch-int, and Apisak Pattanachak, *Sentiment Analysis Process for Product's Customer Reviews Using Ontology-Based Approach*, 2018.

[34] Achmad Bayhaqy, Sfenrianto Sfenrianto, Kaman Nainggolan, Emil R. Kaburuan, (2018), *Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes.*

[35] Grzegorz Dziczkowski. *Analyse des sentiments : système autonome d'exploration des opinions exprimées dans les critiques cinématographiques*. Automatique / Robotique. École Nationale Supérieure des Mines de Paris, 2008. Français. ffNNT : 2008ENMP1637ff. fftel-00408754f

[36] Atanu Dey, Mamata Jenamani and Jitesh J.Thakkar, (2018), Senti-N-Gram : An n-gram lexicon for sentiment analysis, *Department of Industrial and Systems Engineering, Indian Institute of Technology*, Kharagpur 721302, India.

[37] Yannis Chaouche, MAJ le (30/04/2021), Nettoyez et normalisez les données, `https://openclassrooms.com/fr/courses/4470541-analysez-vos-donnees-textuelles/4854971-nettoyez-et-normalisez-les-donnees`, consulé le 07/07/2021.

[38] Moloud Abdar and Vladimir Makarenkov, (2019),CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer, *Département d'Informatique, Université du Québec à Montreal, 201, av. Président-Kennedy*, Montréal, QC H2X 3Y7, Canada.

[39] S. Wahyuningsih, D. R. Utari, U. B. Luhur, D. Tree, and K. Validation, (2018), *Perbandingan Metode K-Nearest Neighbor , Naïve Bayes dan Decision Tree untuk Prediksi Kelayakan Pemberian Kredit*, Konf. Nas. Sist. Inf. 2018, pp. 8–9.

[40] P. Tripathi, S. K. Vishwakarma, and A. Lala, (2015), *Sentiment Analysis of English Tweets Using Rapid Miner*, International Conference on Computational Intelligence and Communication Networks (CICN), 2015, pp. 668–672.

[41] O. Campesato, (2021), Python 3 and Data Analytics Pocket Primer. *Mercury Learning Information*, pp. 31-32.

[42] Amazon Reviews : Unlocked Mobile Phones, `https://www.kaggle.com/PromptCloudHQ/amazon-reviews-unlocked-mobile-phones`. Consulté le 16/06/2021.

[43] Bernadette M. Randles, Irene V. Pasquetto, Milena S. Golshan and Christine L. Borgman,(2017) Using the Jupyter Notebook as a Tool for Open Science : An Empirical Study.

[44] L. Bastien, (17/06/2021), Python : tout savoir sur le principal langage Big Data et Machine Learning, `https://www.lebigdata.fr/python-langage-definition`, consulté le 20/06/2021.

[45] O. Campesato, (2021), Python 3 and Data Analytics Pocket Primer. *Mercury Learning Information*, pp. 31-32, 49-50.

[46] Gery J.Sumual, Benny Pinontoan, Luther A.Latumakulita, (2021), *GUI Application to Setup Simple Graph on the Plane using Tkinter of Python*, Jurnal Matematika dan Aplikasi, p-ISSN :2302-4224.

[47] Ali Hassan Sial, Syed Yahya Shah Rashdi and Abdul Hafeez Khan, (2021), *Comparative Analysis of Data Visualization Libraries Matplotlib and Seaborn in Python*, International Journal of Advanced Trends in Computer Science and Engineering, vol.10, ISSN 2278-3091.

[48] Jayesh Tembhekar, Jayesh Katkar, Darshan Dahekar and Prasad Daf, (2021), *Building and Training a Supervised Machine Learning Model using Scikit- Learn for Elevating Business Sales*, International Research Journal of Engineering and Technology (IRJET),vol.08, ISSN : 2395-0056.

[49] JACOB, (17/05/2010),CLASSIFICATION DE TEXTE POUR L'ANALYSE DES SENTIMENTS – PRÉCISION ET RAPPEL, `https://streamhacker.com/2010/05/17/text-classification-sentiment-analysis-precision-recall`, consulté le 21/08/2021.

[50] E Sujatha and R Radha, (2021), *A Hybrid of Proposed Filtration and Feature Selections to Enhance the Model Performance*, INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY.