RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE Dr. TAHAR MOULAY SAIDA

FACULTE : TECHNOLOGIE

DEPARTEMENT : INFORMATIQUE

MÉMOIRE DE MASTER

Option : Master 2 Modélisation Informatique des Connaissances et du Raisonnement

---

Une Approche hybride pour l'analyse des sentiments basée sur SVM et KNN.

Présenté par :

- Medjahdi Farouk
- Berkanne Abdel illah

Encadré par :

Enseignant : Dr. Abdelkrim Latreche

Année Universitaire 2020-2021

Remerciement

Merci à Dieu d'avoir terminé ce travail que je dédie à :

Mon cher pays l'Algérie

Qui m'a inculqué de nobles valeurs et de nobles idéaux avec leurs nobles significations

Mon cher père

Qui a fait le paradis sous ses pieds, celle au grand cœur battant de tendresse

Chère mère...

Et à tous ceux qui me sont chers. Et aux faiseurs de mon sourire et de mon soutien de tous les instants...

A la famille : Medjahdi et mes frères

Avec gratitude et reconnaissance...

A mes professeurs et collègues...

Et à tous ceux que j'ai oublié de mentionner, mes remerciements et mon respect...

Remerciement

Merci à Dieu d'avoir terminé ce travail que je dédie à :

Qui m'a fait passer d'un petit enfant à un prince

Mon cher père

Un symbole de générosité et de fidélité... Vous avez les plus belles femmes...

Ma mère bien-aimée

A ceux avec qui je partage l'atmosphère d'amour familial

A la famille : Berkanne mes frères

Avec gratitude et reconnaissance...

A mes professeurs, collègues et tous mes amis...

Et à tous ceux que j'ai oublié de mentionner, mes remerciements et mon respect...

Dédicaces

Louange à Dieu et merci à Dieu Tout-Puissant, qui nous a donné les connaissances, les connaissances et la capacité de mener à bien cet humble effort, et nous sommes heureux d'exprimer nos sincères remerciements et notre gratitude à la famille : ***Medjahdi***, ***Berkanne***, *à notre estimé Dr. :* ***LATTRACH***, *qui a accompagné cet effort depuis était une idée jusqu'à ce qu'elle devienne une réalité qui est apparue grâce à son patronage béni et à ses conseils avisés.*

Nous adressons également nos grands remerciements et notre gratitude aux honorables professeurs de l'Université du ***Dr. Taher Moulay Saida*** *et ses administrateurs et travailleurs pour leur bon traitement, et nous remercions également tous ceux qui ont contribué à la réalisation de cet effort, que ce soit par des encouragements ou du soutien.*

Table des matières

Introduction générale 03

Chapitre 1 : Analyse des sentiments et détection d'opinions

1.1. Introduction …… 06
1.2. Les réseaux sociaux …… 06
1.2.1. Définitions des réseaux sociaux …… 07
1.2.2. Différents types de réseaux sociaux …… 07
1.2.3. Sources des données …… 11
1.2.4. Twitter …… 12
1.3. Analyse des sentiments et détection d'opinions …… 12
1.3.1. Définitions …… 12
1.3.2. Historique de l'analyse des sentiments …… 13
1.3.3. Niveaux d'analyse des sentiments …… 13
1.3.4. Disciplines en relation avec l'analyse des sentiments …… 15
1.3.5. Domaines d'application de l'analyse des sentiments …… 15
1.3.6. Les approches de l'analyse des sentiments …… 16
1.3.6.1. Approches d'apprentissage automatique …… 17
1.3.6.2. Approche lexique …… 18
1.3.6.3. Approches hybride …… 20
1.3.6.4. Travaux de l'analyse des sentiments avec Twitter …… 21
1.3.7 Les problèmes de l'analyse des sentiments …… 22

Chapitre 2 : Conception

2-1. Introduction …… 25
2-2. Twitter et Tweet …… 25
2-2.1 Twitter …… 25
2-2.2. Les tweets …… 26
2-2.3. Caractéristique d'un tweet …… 26
2-3. Source de données (Dataset) …… 27
2-3.1. Sentiment140 dataset …… 27
2-3.2. Twitter US Airline Sentiment Dataset …… 28
2-3.3. SuperFetch dataset …… 29
2-3.4. SemEval (évaluation sémantique) …… 29
2-4. Architecture du système …… 29
2-4.1. Etape d'acquisition …… 30
2-4.2. Prétraitement …… 31
2-4.2.1. Filtrage …… 33

2-4.2.2. Tokenisation ("Découpage en mots ") 33
2-4.2.3. Pos tagger 34
2-4.2.4. Stop Words 35
2-4.2.6. Lemmatization 36
2-4.2.7. Emoticônes 37
2-4.3. Extraction et présentation des descripteurs (TF-IDF) 37
2-4.4. Classification des sentiments 38
2-4.4.1. KNN 39
2-4.4.2. SVM 40
2-4.4.2. KNN + SVM 41
Chapitre 3 : Réalisation
3-1. Environnement du travail..... 45
3-1.1 Environnement matériel..... 45
3-1.2 Environnement logiciel..... 45
3-1.2 Autres outils..... 46
3-2 Expérimentations et interprétations..... 46
3-3 Les résultants..... 47
3-4 Conclusion..... 51
Conclusion générale 52
Bibliographie

Table des figures

Figure 1-1 : Enchainement des réseaux sociaux 1978-2015	07
Figure 1-2 : Workflow de l'analyse des sentiments.	14
Figure 1-3 : Les niveaux d'analyse	15
Figure 1-4 : Les approches de l'analyse des sentiments	21
Figure 1-5 : les étapes d'analyse des sentiments sur les données Twitter...	22
Figure 2-1 : Extrait du data set Sentiment140	28
Figure 2-2 : Extrait du data set de Twitter US. Arline Sentiment	28
Figure 2-3 : Architecture du système	30
Figure 2-4 : Les processus de prétraitement	32
Figure 2-5 : Exemple d'un tweet contenant l'hashtag, l'URL du lien, les symboles et le nom d'utilisateur	33
Figure 2-6 : Tweet composé du texte et des signes de ponctuation	34
Figure 2-7 : Exemple d'affichage des Pos tagger	34
Figure 2-8 : Ensemble d'étiquettes.....	35
Figure 2-9 : Exemple de Tweet contentant des mots vides (stop Word).....	36
Figure 2-10 : Exemple de Tweet des mots sans lemmatisation	36
Figure 2-11 : Exemple de tweet contenant des émoticônes	37
Figure 2-12 : deux catégories différentes sont classées à l'aide d'un hyperplan.....	40
Figure 2-13 : Notre méthode entraîne un SVM sur les 50 voisins les plus proches.....	43
Figure 3-1 : Logo de Python.....	45
Figure 3-2 **:** Interface graphique.	47
Figure 3-3 : Précision KNN avec K=100 (Dataset US. Airlines Sentiment) ……	48
Figure 3-4 : Précision SVM avec K=100 (Dataset US. Airlines Sentiment) ……	50
Figure 3-5 : Précision KNN+SVM avec K=100 (Dataset US. Airlines Sentiment) .	52
Figure 3-6 comparaisons entre les modèles en termes de temps d'exécution	53

Liste des tableaux

Table 2.1 : Ensemble d'étiquettes …… 34
Table 2.2 : Tableau qui contenant des mots vides (stop Word) …….. 35
Table 2.3 : Tableau qui affiche des mots sans lemmatisation et avec lemmatisation …… 36
Table 3.1 : Environnement matériel…….. 45
Table 3.2 : Table de confusion…… 46
Table 3.3 : le nombre de tweets pour chaque phase…….. 47
Table 3.4 : les résultats de modèles KNN (Sentiment140)…… 48
Table 3.5 : les résultats de modèles KNN (US. Airlines Sentiment)…… 48
Table 3.6 : les résultats de modèles SVM (Sentiment140)…….. 49
Table 3.7 : les résultats de modèles SVM (US. Airlines Sentiment)…… 49
Table 3.8 : les résultats de modèles KNN+SVM (Sentiment140)…… 49
Table 3.9 : les résultats de modèles KNN+SVM (US. Airlines Sentiment)….. 50

List des Acronyme

Acronyme	Indication
TAL	Traitement Automatique De Langage
CNN	Convolutional Neural Network
ML	Machine Learning
RNN	Recurrent Neural Network
SVM	Support Vector Machine
KNN	K Nearest Neighbor
VP	Vrai Positif
FP	Faux Positif
VN	Vrai Négatif
FN	Faux Négatif
GA	Algorithme Génétique
ANN	Artificiel Neural Network
SA	L'analyse Des Sentiments
PNL	Traitement Du Langage Naturel
UL	Unsupervised Learning
VADER	Valence Aware Dictionary And Sentiment Reasoner

ملخص :

وقد أدى ظهور تكنولوجيا ويب 2.0 إلى توليد كمية هائلة من البيانات الأولية من خلال السماح لمستخدمي الإنترنت بنشر آرائهم ومشاعرهم وتعليقاتهم على الشبكة. ومعالجة هذه البيانات الخام لاستخلاص معلومات مفيدة يمكن أن تكون مهمة صعبة للغاية. وفي الوقت الراهن، فإن الشبكات الاجتماعية المليئة بالنصوص التي يعبر فيها مستخدمو الإنترنت عن أنفسهم في مواضيع مختلفة، فإن اهتمام آرائهم كبير، حيث يشكل فهم المحتوى الذي تحمله هذه النصوص عنصرا أساسيا. تحليل المشاعر أو تحليل الآراء هو دراسة حسابية لآراء الناس ومشاعرهم ومواقفهم وعواطفهم المعرب عنها بلغة مكتوبة. تحليل المشاعر هو أحد أكثر مجالات البحث نشاطا في معالجة اللغة الطبيعية وتحليل النصوص في السنوات الأخيرة. في هذا العمل، ركزنا بشكل رئيسي على تحليل مشاعر التغريدات. لتحليل وتصنيف التغريدات، استخدمنا ومقارنة SVM، KNN وطريقة هجينة على أساس مزيج من خوارزميات SVM وKNN. ونهدف إلى تحسين خوارزمية SVM وتبسيط عملية التعلم، عن طريق دمج خوارزمية KNN في خوارزمية SVM. إن الطبقات التي حددناها هي: الطبقة الإيجابية أو السلبية أو المحايدة. وتثبت الدراسات التجريبية الأولى التي أجريت فعالية الطريقة المقترحة من حيث الدقة وتمثل تعقيدا حسابيا معقولا لكل من التعلم والتنفيذ.

الكلمات المفتاحية: تحليل المشاعر، معالجة اللغة الطبيعية، KNN، SVM، KNN + SVM، التعلم الآلي.

Abstract

L’émergence de la technologie Web 2.0 a généré une énorme quantité de données brutes en permettant aux utilisateurs d’Internet de publier leurs opinions, avis et commentaires sur le Web. Le traitement de ces données brutes pour extraire des informations utiles peut être une tâche très difficile. Actuellement, les réseaux sociaux pleins des textes dans lesquelles, les internautes s’expriment en différents sujets, l'intérêt de leurs opinions est considérable, où la compréhension du contenu véhiculé par ces textes est un élément essentiel. L'analyse des sentiments ou l'analyse des opinions est l'étude computationnelle des opinions, des sentiments, des attitudes et des émotions des personnes exprimées dans le langage écrit. L'analyse des sentiments est l'un des domaines de recherche les plus actifs dans le traitement du langage naturel et l'analyse des textes ces dernières années. Dans ce travail, nous nous sommes concentrés principalement sur l'analyse des sentiments des tweets. Pour analyser et classifier les tweets, avons utilisés et comparés les algorithmes SVM, KNN et une méthode hybride basée sur la combinaison des algorithmes SVM et KNN. Nous visons à optimiser l'algorithme SVM et à simplifier le processus d’apprentissage, en intégrant l'algorithme KNN dans l’algorithme SVM. Les classes que nous avons définies sont : la classe positive, négative ou neutre. Les premières études expérimentales

réalisées prouvent l'efficacité de la méthode proposée en ce qui concerne la précision et présentent une complexité de calcul raisonnable tant pour l'apprentissage que pour l'exécution.

Mots clés : analyse des sentiments, traitement du langage naturel, KNN, SVM, KNN + SVM, apprentissage automatique.

Abstract

The emergence of Web 2.0 technology has generated a huge amount of raw data by allowing Internet users to post their opinions, opinions and comments on the Web. Processing this raw data to extract useful information can be a very difficult task. Currently, the social networks full of texts in which the Internet users express themselves in different topics, the interest of their opinions is considerable, where the understanding of the content conveyed by these texts is an essential element. The analysis of feelings or the analysis of opinions is the computational study of the opinions, feelings, attitudes and emotions of people expressed in written language. Feeling's analysis is one of the most active areas of research in natural language processing and text analysis in recent years. In this work, we focused mainly on analyzing the feelings of tweets. To analyze and classify tweets, we used and compared SVM, KNN and a hybrid method based on the combination of SVM and KNN algorithms. We aim to optimize the SVM algorithm and simplify the learning process, by integrating the KNN algorithm into the SVM algorithm. The classes we have defined are: the positive, negative or neutral class. The first experimental studies carried out prove the effectiveness of the proposed method in terms of precision and present a calculation complexity that is reasonable for both learning and execution.

Key Words: Sentiment Analysis, Natural Language Processing, KNN, SVM, KNN + SVM, Machine Learning.

Introduction générale

Avec l'avènement du web, il est beaucoup plus facile de recueillir des opinions diverses de différentes personnes à travers le monde. Les gens cherchent à passer en revue des sites (par exemple, CNET, Epinions.com), des sites de commerce électronique (Amazon, eBay), des sites d'opinion en ligne (TripAdvisor, Rotten Tomatoes, Yelp) et des médias sociaux (Facebook, Twitter). Obtenir des commentaires sur la façon dont un produit ou service particulier peut être perçu sur le marché. L'explosion des sources des données telles que les sites d'avis, les blogs et les microblogs est apparu la nécessité d'analyser des millions des postes, de tweets ou d'avis afin de savoir ce que pensent les internautes. Actuellement, l'un des meilleurs exemples de réseaux sociaux permettant d'observer l'évolution de ces opinions est Twitter, L'analyse des sentiments est une technologie d'analyse automatique des discours, écrits ou parlés et d'en faire ressortir les différentes opinions exprimées sur un sujet précis comme une marque, une actualité ou un produit. L'importance de l'analyse des sentiments est présente dans plusieurs domaines, à savoir politique, marketing, gestion de la réputation, médecine, etc. L'analyse des sentiments relève de plusieurs disciplines en l'occurrence d'une part du traitement automatique du langage naturel (NLP) et d'autre part de l'apprentissage automatique (Machine Learning).

Dans notre travail, on va analyser des différents tweets sur le réseau social Twitter. Notre objectif consiste à dévoiler les secrets de l'analyse des sentiments en adoptant une approche d'apprentissage automatique. Pour ce faire, avons utilisés et comparés les algorithmes SVM, KNN et une méthode hybride basée sur la combinaison des algorithmes SVM et KNN. Nous visons à optimiser l'algorithme SVM et à simplifier le processus d'apprentissage, en intégrant l'algorithme KNN dans l'algorithme SVM. Les classes que nous avons définies sont : la classe positive, négative ou neutre.

Les premières études expérimentales réalisées prouvent l'efficacité de la méthode proposée en ce qui concerne la précision et présentent une complexité de calcul raisonnable tant pour l'apprentissage que pour l'exécution.

Le reste du mémoire est organisé en deux chapitres : nous consacrons un premier chapitre à présenter des généralités sur le domaine d'analyse des sentiments en particulier Twitter comme source d'opinions. Notre

deuxième chapitre présente l'implémentation et l'expérimentions. Nous concluons avec une synthèse de travail et des perspectives.

Chapitre 1

1-1. Introduction

Les sentiments, les évaluations, les attitudes et les émotions sont les sujets d'étude de l'analyse des sentiments et de l'exploration d'opinions. La création et la croissance rapide du domaine coïncident avec celles des médias sociaux sur le Web, par exemple les critiques, les forums de discussion, les blogs, les microblogs, Twitter et les réseaux sociaux, car pour la première fois dans l'histoire de l'humanité, nous avons un volume énorme de données d'opinion enregistrées sous forme numérique. Depuis le début des années 2000, l'analyse des sentiments est devenue l'un des domaines de recherche les plus actifs dans le traitement du langage naturel. Il est également largement étudié dans l'exploration de données, l'exploration Web et l'exploration de texte. En fait, il s'est propagé de l'informatique aux sciences de gestion et aux sciences sociales en raison de son importance pour les entreprises et la société dans son ensemble. Ces dernières années, les activités industrielles liées à l'analyse des sentiments ont également prospéré. De nombreuses startups ont vu le jour. De nombreuses grandes entreprises ont développé leurs propres capacités internes. Les systèmes d'analyse des sentiments ont trouvé leurs applications dans presque tous les domaines commerciaux et sociaux.

1-2. Les réseaux sociaux

Les réseaux sociaux est l'utilisation de sites de médias sociaux sur Internet pour rester en contact avec des amis, la famille, des collègues, ou des clients. Les réseaux sociaux peuvent avoir un objectif social, un objectif commercial ou les deux, via des sites tels que Facebook, Twitter, LinkedIn et Instagram, entre autres. Les réseaux sociaux sont devenus une base importante pour les spécialistes du marketing qui cherchent à engager les clients.

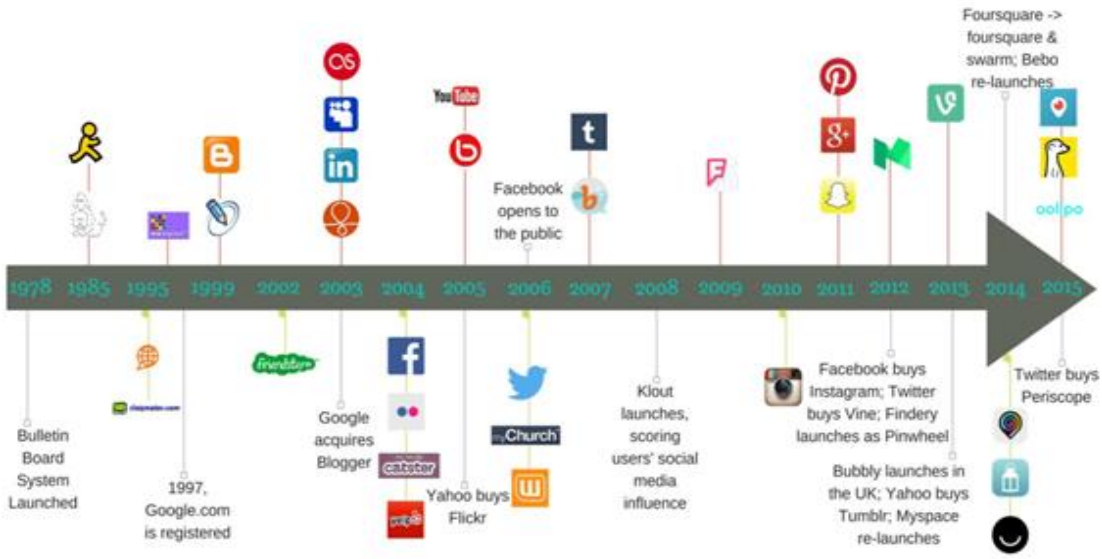

Figure 1-1 : Enchainement des réseaux sociaux 1978-2015

1-2.1. Définitions des réseaux sociaux

Le réseautage social consiste à augmenter le nombre de ses contacts professionnels et / ou sociaux en établissant des liens via des individus, souvent via des sites de médias sociaux tels que Facebook, Twitter, LinkedIn et Google+.

Le réseautage social établit des communautés en ligne interconnectées qui aident les gens à établir des contacts qu'il serait bon qu'ils connaissent, mais qu'il est peu probable qu'ils se soient rencontrés autrement.

1-2.2. Différents types de réseaux sociaux

Bien que Facebook, Twitter et LinkedIn soient les premiers sites qui viennent à l'esprit lorsque l'on pense aux réseaux sociaux, ces sites Web populaires ne représentent pas toute la gamme des réseaux sociaux existants. En savoir plus sur les différentes options disponibles pour les utilisateurs d'interagir et de collaborer en ligne [1].

1-2.2.1. Liens sociaux

Rester en contact avec les amis et les membres de la famille est l'un des plus grands avantages des réseaux sociaux. Voici une liste des sites Web les plus utilisés pour établir des liens sociaux en ligne.

- **Facebook** : sans doute l'utilitaire de médias sociaux le plus populaire, Facebook permet aux utilisateurs de créer des liens et de partager des informations avec les personnes et les organisations avec lesquelles ils choisissent d'interagir en ligne.
- **Twitter** : partagez vos réflexions et restez informé des autres via ce réseau d'information en temps réel.
- **Google +** : cet entrant relativement nouveau sur le marché des connexions sociales est conçu pour permettre aux utilisateurs de créer des cercles de contacts avec lesquels ils peuvent interagir et qui sont intégrés à d'autres produits Google.
- **Myspace** : Bien qu'il ait d'abord commencé comme un site de médias sociaux général, Myspace a évolué pour se concentrer sur le divertissement social, offrant un lieu pour les connexions sociales liées aux films, aux jeux musicaux et plus encore.

1-2.2.2. Partage multimédia

Les réseaux sociaux facilitent le partage de contenu vidéo et photographique en ligne. Voici quelques-uns des sites de partage multimédia les plus populaires.

- **YouTube** : plateforme de médias sociaux qui permet aux utilisateurs de partager et de voir du contenu vidéo
- **Flickr** : ce site offre une option puissante pour gérer les photos numériques en ligne, ainsi que pour les partager avec d'autres.

1-2.2.3. Professionnelle

Les réseaux sociaux professionnels sont conçus pour offrir des opportunités de croissance professionnelle. Certains de ces types de réseaux fournissent un forum général pour que les professionnels se connectent, tandis que d'autres se concentrent sur des professions ou des intérêts spécifiques. Quelques exemples de réseaux sociaux professionnels sont listés ci-dessous.

- **LinkedIn** : En novembre 2011, LinkedIn comptait plus de 135 millions de membres, ce qui en fait le plus grand réseau professionnel en ligne. Les participants ont l'occasion de nouer

des relations en établissant des liens et en rejoignant des groupes pertinents.

- **Classroom** : Réseau social spécialement conçu pour aider les enseignants à se connecter, à partager et à s'entraider sur des questions spécifiques à la profession.

1-2.2.4. Informatif

Les communautés informationnelles sont constituées de personnes cherchant des réponses aux problèmes quotidiens. Par exemple, lorsque vous songez à démarrer un projet de rénovation domiciliaire ou que vous souhaitez apprendre à devenir vert à la maison, vous pouvez effectuer une recherche sur le Web et découvrir d'innombrables blogs, sites Web et forums remplis de personnes qui recherchent le même type d'information. Quelques exemples incluent:

- **The Nature Conservancy** : communauté en ligne où les personnes intéressées à adopter des pratiques de vie vertes et à protéger la terre peuvent interagir
- **Diy chatroom** : ressource de médias sociaux pour permettre aux passionnés de bricolage d'interagir les uns avec les autres.

1-2.2.5. Éducative

Les réseaux éducatifs sont l'endroit où de nombreux étudiants se rendent afin de collaborer avec d'autres étudiants sur des projets académiques, de mener des recherches pour l'école ou d'interagir avec des professeurs et des enseignants via des blogs et des forums en classe. Les réseaux sociaux éducatifs deviennent aujourd'hui extrêmement populaires dans le système éducatif. Quelques exemples de tels réseaux sociaux éducatifs sont énumérés ci-dessous.

- **The Student Room** : communauté étudiante basée au Royaume-Uni avec un babillard modéré et des ressources utiles liées à l'école
- **Blog de l'école ePALS** : Ce réseau social international pour les élèves de la maternelle à la 12e année est conçu pour établir des liens internationaux pour promouvoir la paix dans le monde.

1-2.2.6. Loisirs

L'une des raisons les plus populaires pour lesquelles de nombreuses personnes utilisent Internet est de mener des recherches sur leurs projets préférés ou sur des sujets d'intérêt liés à leurs loisirs personnels. Lorsque les gens trouvent un site Web basé sur leur passe-temps préféré, ils découvrent toute une communauté de personnes du monde entier qui partagent la même passion pour ces intérêts. C'est ce qui est au cœur de ce qui fait fonctionner les réseaux sociaux, et c'est pourquoi les réseaux sociaux axés sur les loisirs sont parmi les plus populaires. Voici quelques exemples de sites de réseautage social axés sur les loisirs :

- **Grow It !** Application de réseau de médias sociaux spécialement pour les amateurs de jardinage.
- **Ma place sur Scrapbook.com** : conçu spécialement pour les amateurs de scrapbooking, les utilisateurs peuvent créer des profils, partager des informations, publier des mises à jour et plus encore.

1-2.2.7. Académique

Les chercheurs universitaires qui souhaitent partager leurs recherches et examiner les résultats obtenus par leurs collègues peuvent trouver le réseautage social spécifique à l'université comme étant très utile. Voici quelques-unes des communautés en ligne les plus populaires pour les universitaires :

- **Academia.edu** : Les utilisateurs de ce réseau social académique peuvent partager leurs propres recherches, ainsi que suivre les recherches soumises par d'autres.
- **ResearchGate** : ressources en ligne permettant aux scientifiques et aux chercheurs de trouver, d'organiser et de partager des informations utiles ainsi que de créer un réseau professionnel.

1-2.3. Sources des données :

Les sources de données sociales étaient les informations que vous collectiez à partir des nombreux canaux de médias sociaux que vous utilisez. Maintenant, c'est devenu un terme plus fourre-tout, englobant non seulement les médias sociaux, mais d'autres sources - blogs, analyses Web, critiques, etc. - que vous devez surveiller pour recueillir des informations sur les décisions commerciales fondées sur les données.

Quelques-unes des plus grandes sources de données sur les réseaux sociaux :

1-2.3.1. Facebook

Facebook est actuellement la plus grande plateforme de médias sociaux au monde, avec 2,41 milliards d'utilisateurs actifs par mois. Récemment, il s'est adapté pour être plus axé sur la communauté, avec une augmentation du soutien aux groupes.

Facebook offre l'audience la plus large, en raison de son large attrait. Le public typique est d'environ 18 à 34 ans (57% du total), mais il y a encore une utilisation importante des moins de 18 ans et des 34 plus.

1-2.3.2. Instagram

Instagram offre des informations visuelles précieuses sur la manière dont votre marque est partagée en ligne. Une image peint mille mots, et dans ce cas, elle vous donne une chance de voir votre public interagir avec les produits et fournit des détails sur comment, pourquoi et quand ils le font.

Instagram est également plus populaire auprès des jeunes, avec 36% de leur public publicitaire âgé de 13 à 24 ans, contre 30% pour Facebook.

1-2.3.3. Blogs

Les blogs donnent aux gens le nombre de mots nécessaire pour fournir un examen approfondi. Ce qui peut être bénéfique (car ils reflètent toutes vos caractéristiques uniques) ou préjudiciable (car ils

mettent en lumière tous les défauts ou problèmes que vous ne voudrez peut-être pas révéler).

Alors que les actualités ajoutent du prestige à une marque - mais ne tombez pas dans le piège de penser que toute nouvelle est une bonne nouvelle.

1-2.3.4. Twitter

En mars 2006, Twitter a été créé par le développeur Jack Dorsey pour rester en contact avec sa famille et ses amis. Twitter est un service de micro-blogging de réseautage social gratuit qui permet aux membres enregistrés de diffuser de courts messages appelés tweets. Les membres de Twitter peuvent diffuser des tweets et suivre les tweets d'autres utilisateurs en utilisant plusieurs plates-formes et appareils. Les tweets et les réponses aux tweets peuvent être envoyés par SMS sur téléphone portable, client de bureau ou en publiant sur le site Web Twitter.com.

1-3. Analyse des sentiments et détection d'opinions

L'analyse des sentiments (SA) peut être définie comme la tâche de détection, d'extraction et de classification des opinions sur quelque chose. Il s'agit d'un type de traitement du langage naturel (PNL) pour suivre l'humeur du public vis-à-vis d'une certaine loi, politique ou marketing, etc.

1-3.1. Définitions

1-3.1.1. Définition de sentiment

Sentir est la nominalisation du verbe sentir. Utilisé à l'origine pour décrire la sensation physique du toucher à travers l'expérience ou la perception, le mot est également utilisé pour décrire d'autres expériences, telles que « une sensation de chaleur » et de sensibilité en général.

1-3.1.2. Définition d'opinion

Une croyance, un jugement ou une façon de penser à quelque chose, ce que quelqu'un pense d'une chose particulière.

1-3.1.3. Définition Analyse des sentiments

L'analyse des sentiments est le processus d'identification et de catégorisation par ordinateur d'un morceau de texte, en particulier pour déterminer si le sentiment général du texte est positif, négatif ou neutre.

1-3.2. Historique de l'analyse des sentiments :

L'origine de l'analyse des sentiments remonte aux années 1950, lorsque l'analyse des sentiments était principalement utilisée sur des documents écrits sur papier.

Hatzivassiloglou et McKeown en 1997, travaillaient au niveau de document et utilisaient « World Street Journal » comme source de données. Dans le même niveau document Pang et al. Nigam et Hurst en 2004, travaillaient au niveau des expressions en se basant sur le lexique des phrases polaires et leurs parties du discours avec un modèle basé sur des règles syntaxiques en utilisant Usenet1 message board et autres sources en ligne comme source de donnes.

Concernant l'analyse des sentiments sur Twitter, Pak et Paroubek en 2010, Barbosa et Feng en 2010, ont travaillé au niveau des phrases des messages Twitter.

1-3.3. Niveaux d'analyse des sentiments

L'analyse des sentiments peut se produire à différents niveaux : **niveau document**, **niveau phrase** ou **niveau aspect/fonctionnalité** [2].

1-3.3.1. Niveau document

Dans ce processus, le sentiment est extrait de l'ensemble de l'examen et une opinion entière est classée en fonction du sentiment général du détenteur de l'opinion. L'objectif est de classer un avis comme positif, négatif ou neutre [3].

Exemple :

« J'ai acheté un Samsung S21 ultra 5G il y a quelques jours. C'est un très beau téléphone, bien qu'un peu gros. L'écran tactile est fascinant. La qualité de la voix est également claire. J'adore ça ! La classification des avis est-elle positive ou négative ?

Le niveau du document fonctionne mieux lorsque le document est rédigé par une seule personne et exprime une opinion / un sentiment sur une seule entité.

1-3.3.2. Niveau de la phrase

Ce processus comprend généralement deux étapes :

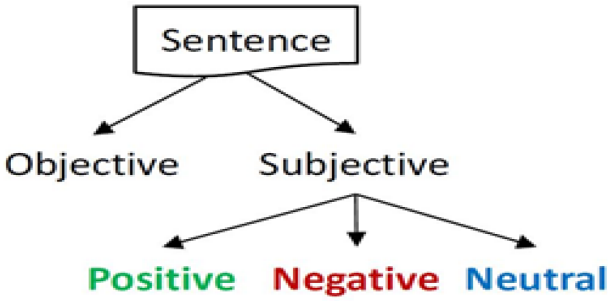

Figure 1-2 : Workflow de l'analyse des sentiments.

- Subjectivité d'une phrase dans l'une des deux classes : objective et subjective.
- Sentiment de phrases subjectives en deux classes : positive et négative et neutral.

Une phrase objective présente des informations factuelles, tandis qu'une phrase subjective exprime des sentiments, des opinions, des émotions ou des croyances personnelles.

L'identification subjective des phrases peut être réalisée par différentes méthodes. Cependant, il ne suffit pas de savoir que les phrases ont une opinion positive ou négative. Il s'agit d'une étape intermédiaire qui permet de filtrer les phrases sans opinion et de déterminer dans une certaine mesure si les sentiments sur les entités et leurs aspects sont positifs ou négatifs. Une phrase subjective peut contenir plusieurs opinions et clauses subjectives et factuelles [4].

Exemple :

« Les ventes d'iPhone se portent bien dans cette mauvaise économie. »

Niveaux d'analyse du document et de la phrase est utile, mais elle ne trouve pas ce que les gens aiment ou n'aiment pas, ni n'identifie des cibles d'opinion.

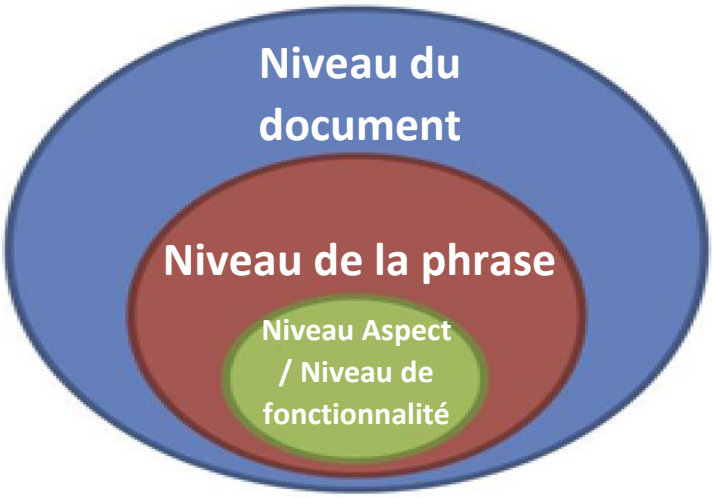

Figure 1-3 : Les niveaux d'analyse.

1-3.3.3. Aspect / Niveau de fonctionnalité

Dans ce processus, le but est d'identifier et d'extraire les caractéristiques de l'objet qui ont été commentées par le détenteur de l'opinion et de déterminer si l'opinion est positive, négative ou neutre. Les synonymes des fonctionnalités sont regroupés et un résumé basé sur les fonctionnalités de plusieurs avis est produit.

Le premier choix lorsque l'on applique l'analyse des sentiments est de définir le texte qui va être analysé dans le cas d'une étude considérée. En général, il existe trois niveaux d'analyse : le niveau du document, le niveau de la phrase et le niveau Aspect / Niveau de fonctionnalité.

1-3.4. Disciplines en relation avec l'analyse des sentiments

L'analyse des sentiments est une approche utile à un certain nombre de problèmes différents posés dans un certain nombre de disciplines différentes, comme la psychologie, l'éducation, la sociologie, les affaires, les sciences politiques et l'économie, ainsi que des domaines de recherche comme l'exploration de données et la recherche documentaire.

1-3.5. Domaines d'application de l'analyse des sentiments

L'analyse des sentiments est le processus automatisé d'analyse de texte pour déterminer le sentiment exprimé (positif, négatif ou

neutre). Certaines applications d'analyse des sentiments populaires, Nous citons brièvement ci-dessous :

- **Politique :** Aujourd'hui, les acteurs politiques ont suivi la tendance de l'analyse des sentiments, car avant de déclarer une nouvelle loi, les politiciens tentent de recueillir l'opinion des utilisateurs de médias sociaux sur cette loi. Il est hautement stratégique de connaître également l'opinion des internautes sur un politicien lors d'une élection présidentielle.

- **Économie :** Avant d'acheter un produit, la majorité des clients demandent conseil sur un produit ou un service donné et sont même disposés à payer plus pour un produit dont l'opinion est plus favorable qu'un autre, ce qui peut augmenter les ventes. Grâce à l'analyse des sentiments, les entreprises peuvent connaître l'opinion des clients sur leurs produits ou leurs services. Dans une perspective d'amélioration de leurs produits et d'augmentation de leurs ventes et revenus.

- **Éducation :** L'analyse des sentiments peut être utilisée pour extraire des informations utiles sur la méthodologie d'enseignement d'un enseignant et également sur le programme du cours. Il identifie le degré d'apprentissage des étudiants, comprend leurs besoins, prévoit leurs performances et apporte des changements effectifs dans le style. Les résultats de l'analyse des sentiments aident les enseignants et les établissements à prendre des mesures correctives.

1-3.6. Les approches de l'analyse des sentiments

Il existe de nombreuses méthodes différentes d'analyse des sentiments, qu'ils soient positifs, négatifs ou même neutres. Par conséquent, nous mentionnons dans le rapport suivant trois méthodes :

- Approches d'apprentissage automatique.
- Approche lexique.
- Approches hybride.

1-3.6.1. Approches d'apprentissage automatique

L'apprentissage automatique, sous une large direction scientifique appelée intelligence artificielle (IA), consiste à exploiter le potentiel des idées d'intelligence artificielle. La principale prédiction liée à l'apprentissage automatique prend en compte la mise en œuvre de la demande d'algorithmes ou de méthodes de calcul flexibles et adaptatifs [5]. Cela se traduit par de nouveaux systèmes et fonctionnalités logicielles. La disponibilité d'opportunités d'apprentissage automatique, c'est-à-dire la capacité d'apprendre et de formuler des recommandations au niveau des experts dans un domaine d'application restreint. L'objectif et le principal objectif de l'apprentissage automatique est de classer les groupes et les connaissances et de déterminer les relations entre eux, comme il est divisé en deux grands groupes :

1-3.6.1.1. Apprentissage supervisé (SL)

La majorité de l'apprentissage automatique pratique utilise l'apprentissage supervisé.

L'apprentissage supervisé est l'endroit où vous avez des variables d'entrée (x) et une variable de sortie (Y) et vous utilisez un algorithme pour apprendre la fonction de mappage de l'entrée vers la sortie.

$$Y = f(x)$$

L'objectif est de si bien approcher la fonction de mappage que lorsque vous avez de nouvelles données d'entrée (x), vous pouvez prédire les variables de sortie (Y) pour ces données.

C'est ce qu'on appelle l'apprentissage supervisé parce que le processus d'apprentissage d'un algorithme à partir de l'ensemble de données de formation peut être considéré comme un enseignant supervisant le processus d'apprentissage. Nous connaissons les bonnes réponses, l'algorithme fait des prédictions itératives sur les données d'entraînement et est corrigé par l'enseignant. L'apprentissage s'arrête lorsque l'algorithme atteint un niveau de performance acceptable.

1-3.6.1.2. Apprentissage non supervisé (UL)

L'apprentissage non supervisé est celui où vous n'avez que des données d'entrée (X) et aucune variable de sortie correspondante.

L'objectif de l'apprentissage non supervisé est de modéliser la structure ou la distribution sous-jacente dans les données afin d'en savoir plus sur les données.

Celles-ci sont appelées apprentissage non supervisé car, contrairement à l'apprentissage supervisé ci-dessus, il n'y a pas de bonnes réponses et il n'y a pas d'enseignant. Les algorithmes sont laissés à eux-mêmes pour découvrir et présenter une structure intéressante dans les données.

1-3.6.1.3. Utilisation des techniques d'apprentissage

Vous pouvez utiliser des techniques d'apprentissage non supervisées pour découvrir et apprendre la structure dans les variables d'entrée.

Vous pouvez également utiliser des techniques d'apprentissage supervisé pour faire de meilleures prédictions pour les données non étiquetées, réintégrer ces données dans l'algorithme d'apprentissage supervisé en tant que données d'entraînement et utiliser le modèle pour faire des prédictions sur de nouvelles données invisibles.

1-3.6.2. Approche lexique

L'approche basée sur le lexique utilise un dictionnaire des sentiments avec des mots d'opinion et les met en correspondance avec les données pour déterminer la polarité. Il existe trois techniques pour construire un lexique des sentiments : la **construction manuelle**, les méthodes basées sur le **corpus** et les méthodes basées sur le **dictionnaire**. La construction manuelle est une tâche difficile et longue. Les méthodes basées sur le corpus peuvent produire des mots d'opinion avec une précision relativement élevée. Enfin, dans les techniques à base de dictionnaires, l'idée est d'abord de collecter manuellement un petit ensemble de mots d'opinion avec des orientations connues, puis agrandir cet ensemble en recherchant dans le dictionnaire WordNet leurs synonymes et antonymes.

1-3.6.2.1. La construction manuelle

La construction d'un lexique peut être effectuée dans la construction manuelle nécessite principalement un ou plusieurs lexicographes qui maîtrisent une ou plusieurs langues particulières pour élaborer le lexique à la main [6]. A généralement estimé que le temps moyen nécessaire pour construire manuellement une entrée lexicale dans un lexique est d'environ 30 min. Ainsi, la construction manuelle est lente, coûteuse et encombrante. Dans un sens, cependant, la construction manuelle permet de rendre le contrôle sur son contenu utile pour certaines applications et sur son format pour minimiser la manipulation du lexique. Cette méthode a été utilisée dans la construction de plusieurs lexiques tels que LEXiTRON, WordNet, LDOCE et COBUILD.

1-3.6.2.2. Le corpus

La linguistique de corpus est une méthode pour étudier la langue dans laquelle des bases de données appelées 'corpus' - c'est-à-dire de grandes collections de mots (énoncés transcrits ou textes écrits) - sont utilisées, qui sont généralement stockées sur un ordinateur dans un format lisible par machine et sont accessibles. En ligne ou en utilisant des programmes informatiques [7] les corpus sont construits dans le but d'atteindre un degré éventuellement élevé de représentativité d'une langue donnée et de permettre à un chercheur de rechercher, lire et analyser de grandes quantités de données orales et écrites naturelles dans leur contexte linguistique [8].

1-3.6.2.3. Le dictionnaire

L'utilisation d'un dictionnaire composé de lexiques pré-étiquetés. Le texte d'entrée est converti en jetons par le Tokenizer. Chaque nouveau jeton rencontré est alors mis en correspondance avec le lexique du dictionnaire. En cas de correspondance positive, le score est ajouté au pool total de scores pour le texte d'entrée. Par exemple, si « dramatique » est une correspondance positive dans le dictionnaire, le score total du texte est incrémenté. Sinon, le score est décrémenté ou le mot est marqué comme négatif. Bien que cette technique semble être de nature amateur, ses variantes se sont avérées valables.

Il existe des outils permettant d'identifier le sentiment dégagé par un texte. Voici les deux plus connus outils :

- **WordNet** : permet de savoir à l'aide de groupe de synonymes si un mot est positif ou non.
- **AFINN :** est une liste de mots évalués pour la valence avec un entier compris entre moins cinq (négatif) et plus cinq (positif).

1-3.6.3. Approches hybride

L'approche hybride, la combinaison à la fois de l'apprentissage automatique et des approches basées sur le lexique a le potentiel d'améliorer les performances de classification des sentiments. Les principaux avantages des approches hybrides sont la symbiose lexique / apprentissage, la détection et la mesure du sentiment au niveau du concept et la moindre sensibilité aux changements dans le domaine thématique.

1-3.6.3.1. Le choix de l'hybride

En apprentissage automatique, le principal problème est la loi applicable aux nouvelles données car il y a un besoin de disponibilité de données étiquetées qui pourraient être coûteuses, voire prohibitives. Pour cette raison, nous essayons de trouver une nouvelle façon d'éliminer ce genre de problème, la même chose pour le lexique, le problème est un nombre fini de mots dans les lexiques et l'attribution d'une orientation sentimentale et d'un score fixe aux mots. Pour cela, les principaux avantages des approches hybrides sont la symbiose lexicale / apprentissage, la détection et la mesure du sentiment au niveau du concept et une moindre sensibilité aux changements dans le domaine thématique.

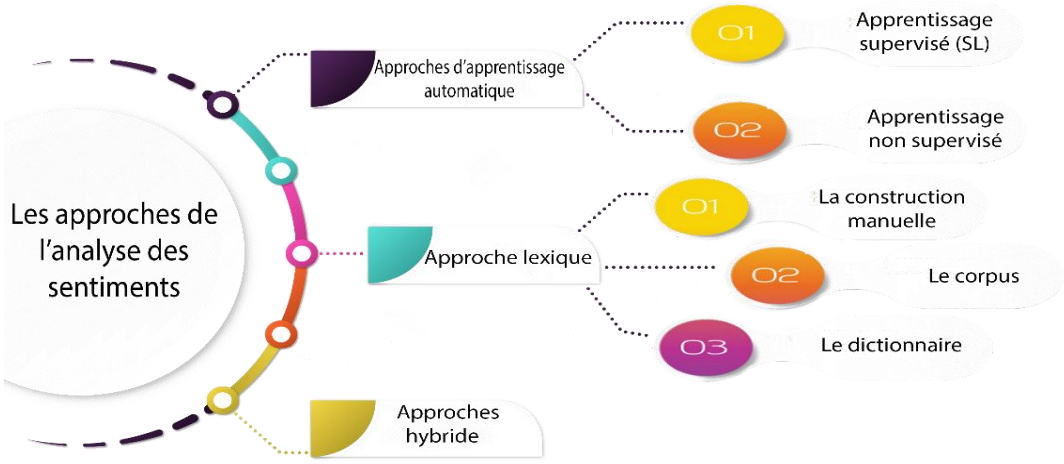

Figure 1-4 : Les approches de l'analyse des sentiments.

1-3.6.4. Travaux de l'analyse des sentiments avec Twitter :

L'analyse des sentiments sur les données Twitter comprend cinq étapes :

- **Première étape** : collecte de données - à ce stade, les données à analyser sont explorées à partir de diverses sources.
- **Deuxième étape** : pré-traitement - Dans cette étape, les données acquises sont nettoyées et préparées pour être introduites dans le classificateur. Le nettoyage comprend l'extraction de mots-clés et de symboles. Par exemple, les émoticônes sont les smileys utilisés sous forme textuelle pour représenter les émotions, par ex. ":-)", ":)", "=)", " : D", ":-(", ":(", "= (", " ; (", etc ... Correction de toutes les majuscules et tout en minuscules pour une cause commune, en supprimant les textes non anglais (ou en langue proposée), en supprimant les espaces blancs et les tabulations inutiles, etc.
- **Troisième étape** : données d'apprentissage - Une collection de données étiquetées manuellement est préparée par la méthode de crowdsourcing la plus couramment utilisée. Ces données sont le carburant du classificateur ; il sera alimenté à l'algorithme à des fins d'apprentissage.
- **Quatrième étape** : classification - C'est le cœur de toute la technique. En fonction des besoins de l'application, SVM ou KNN

est déployé pour l'analyse. Le classificateur (après avoir terminé la formation) est prêt à être déployé sur les tweets / texte en temps réel à des fins d'extraction de sentiments.

- **Cinquième étape** : Résultats - Les résultats sont tracés en fonction du type de représentation sélectionné, c'est-à-dire des tableaux, des graphiques, etc.

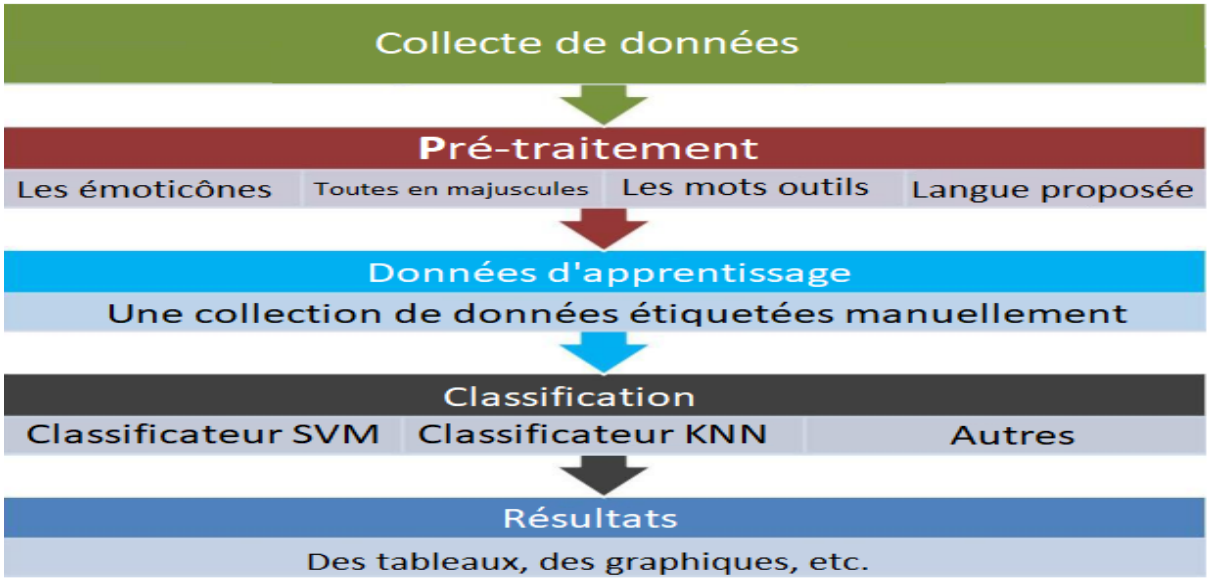

Figure 1-5 : les étapes d'analyse des sentiments sur les données Twitter.

1-3.7. Les problèmes de l'analyse des sentiments :

La recherche dans le domaine a commencé par la classification des sentiments et de la subjectivité, qui a traité le problème comme un problème de classification de texte. La classification des sentiments classe une phrase qui exprime une opinion positive ou négative [9]. La classification de subjectivité détermine si une phrase est subjective ou objective [10]. Cependant, de nombreuses applications réelles nécessitent une analyse plus détaillée car l'utilisateur veut souvent savoir quelles opinions ont été exprimées [11].

Les principaux défis de l'analyse des sentiments sont :

- **Reconnaissance d'entité nommée** : De quoi parle la personne, par exemple 300 Spartiates est-il un groupe de Grecs ou un film ?

- **Résolution de l'anaphore** : le problème de la résolution de ce à quoi un pronom ou une phrase nominale fait référence. "Nous

avons regardé le film et sommes allés dîner ; c'était horrible." À quoi fait-il référence ?

- **Analyse** : Quel est le sujet et l'objet de la phrase, auquel se réfèrent réellement le verbe et / ou l'adjectif ?
- **Sarcasme** : Si vous ne connaissez pas l'auteur, vous ne savez pas si « mauvais » signifie mauvais ou bon.
- **Twitter** : abréviations, manque de majuscules, mauvaise orthographe, mauvaise ponctuation, mauvaise grammaire, ...
- **Faux avis** : Il est également appelé faux avis et fait référence à des avis faux ou faux qui induisent en erreur les lecteurs ou les clients en leur fournissant des opinions négatives ou positives mensongères liées à tout objet et afin de réduire la réputation de tout objet. Ces spams rendent les opinions sentimentales inutiles dans divers domaines d'application.

Chapitre 2

2-1. Introduction

De nos jours, les réseaux sociaux, les blogs et autres médias produisent une énorme quantité de données sur le World Wide Web. Cette énorme quantité de données contient des informations cruciales liées à l'opinion qui peuvent être utilisées au profit des entreprises et d'autres aspects des industries commerciales, scientifiques et politiques. Le suivi manuel et l'extraction de ces informations utiles à partir de cette énorme quantité de données sont presque impossibles. L'analyse des sentiments des messages des utilisateurs est nécessaire pour aider à prendre des décisions commerciales. Il s'agit d'un processus qui extrait les sentiments ou les opinions des avis émis par les utilisateurs sur un sujet, un domaine ou un produit particulier en ligne.

Dans le cadre de notre travail, l'objectif principal est de détecter le plus correctement possible le sentiment des tweets et nous traitons les textes anglais pour l'analyse des sentiments. 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, positif et neutre dans le cas d'une classification multiple. Les données d'entrée que nous avons utilisées sont des tweets, extraits de deux dataset pour entraîner et tester notre modèle et aussi nous testons le modèle proposé en temps réel sur des tweets extraits du Twitter lui-même. Ces tweets représentent les statuts postés sur ce réseau social. Notre démarche d'analyse de sentiments s'inscrit dans l'approche d'apprentissage automatique supervisée. Nous utilisons et comparons trois différentes méthodes d'apprentissage supervisé pour l'analyse des sentiments, qui sont : KNN, SVM et une méthode hybride basée sur la combinaison des algorithmes KNN et SVM (KNN-SVM). Dans les sections suivantes nous présentons en détaille les algorithmes de classification KNN et SVM et ensuite la méthode hybride KNN-SVM.

2-2. Twitter et Tweet

2-2.1. Twitter : est un service gratuit de microblogging de réseautage social qui permet aux membres inscrits de diffuser de courts messages appelés tweets. Les membres de Twitter peuvent diffuser des tweets et suivre les tweets des autres utilisateurs en utilisant plusieurs plateformes et appareils. Les gazouillis et les réponses aux gazouillis peuvent être envoyés par message texte de téléphone portable, par client de bureau ou par affichage sur le site Web Twitter.com.

2-2.2. Les tweets : qui peuvent inclure des hyperliens, sont limités à 140 caractères, en raison des contraintes du système de livraison SMS de Twitter. Étant donné que les tweets peuvent être envoyés aux abonnés en temps réel, ils peuvent ressembler à des messages instantanés pour l'utilisateur novice. Mais contrairement aux IM qui disparaissent lorsque l'utilisateur ferme l'application, des tweets sont également publiés sur le site Twitter. Ils sont permanents, consultables et publics. Tout le monde peut rechercher des tweets sur Twitter, qu'il soit membre ou non.

Termes que vous devez savoir pour utiliser Twitter bien, vocabulaire spécifique est plus couramment utilisé sur Twitter :

Followers : Les abonnés sont des personnes qui reçoivent vos Tweets. Si quelqu'un vous suit : il apparaîtra dans la liste de vos abonnés. Ils verront vos Tweets dans leur chronologie d'accueil chaque fois qu'ils se connecteront à Twitter.

Following : Suivre quelqu'un sur Twitter signifie: vous vous abonnez à ses Tweets en tant qu'abonné. Leurs mises à jour apparaîtront dans votre chronologie d'accueil. Cette personne peut vous envoyer des messages privés.

Friends : sont tous les utilisateurs suivis par l'utilisateur spécifié.

Twittos : est un utilisateur actif de twitter.

Tweet : Un message publié sur Twitter contenant du texte, des photos, un GIF et / ou une vidéo.

2-2.3. Caractéristique d'un tweet

Pour comprendre ce qu'est Twitter et comment l'utiliser, nous mentionnons les termes les plus importants :

- **Twitter** : user : est chaque utilisateur enregistré. Il est représenté par @UserName.
- **Time Line** : c'est la partie de votre compte où vous pouvez voir, par ordre chronologique, les messages des utilisateurs que vous suivez.
- **Retweet (RT)** : c'est la republication d'un tweet lancé par un autre utilisateur.

- **Liste** : c'est une liste que vous pouvez configurer avec vos comptes favoris. Vous pouvez créer autant de listes que vous le souhaitez et leur donner un nom.
- **Like** : il est représenté par une icône en forme de cœur. On clique dessus si on a aimé un tweet.
- **Hashtag** : le roi de Twitter. Il est représenté par une icône dièse (#) et nous permet d'ajouter les termes que nous voulons après. Il est utilisé pour faciliter les recherches.
- **Sujet tendance** : Ce sont les sujets les plus discutés du moment, c'est-à-dire les mots avec le plus de mentions du réseau social sur une certaine période de temps.
- **Mention** : est un Tweet qui contient le nom d'utilisateur d'une autre personne n'importe où dans le corps du Tweet @username.

2-3. Source de données (Dataset)

2-3.1. Sentiment140 dataset :

Il s'agit de l'ensemble de données sentiment140.

Il contient 1 600 000 tweets extraits à l'aide de l'API Twitter. Les tweets ont été annotés (0 = **négatif**, 2 = **neutre**, 4 = **positif**) et ils peuvent être utilisés pour détecter les sentiments.

Il contient les 6 champs suivants :

- **Cible** : la polarité du tweet (0 = négatif, 2 = neutre, 4 = positif).
- **Ids**: l'identifiant du tweet (2087).
- **Date** : la date du tweet (sam 16 mai 23:58:44 UTC 2009).
- **Flag** : La requête (lyx). S'il n'y a pas de requête, cette valeur est NO_QUERY.
- **User** : l'utilisateur qui a tweeté (robotickilldozr).
- **Texte** : le texte du tweet (Lyx est cool).

	A	B	C	D	E	F
1	Cible	Ids	Date	Flag	User	Texte
2	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You should
3	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by texting it... and might cry as a re
4	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Managed to save 50% The rest go
5	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
6	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all. i'm mad. why am i here? becaus
7	0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	@Kwesidei not the whole crew
8	0	1467811592	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	mybirch	Need a hug
9	0	1467811594	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	coZZ	@LOLTrish hey long time no see! Yes.. Rains a bit ,only a bit LOL , I'm fine tha
10	0	1467811795	Mon Apr 06 22:20:05 PDT 2009	NO_QUERY	2Hood4Hollywood	@Tatiana_K nope they didn't have it
11	0	1467812025	Mon Apr 06 22:20:09 PDT 2009	NO_QUERY	mimismo	@twittera que me muera ?
12	0	1467812416	Mon Apr 06 22:20:16 PDT 2009	NO_QUERY	erinx3leannexo	spring break in plain city... it's snowing
13	0	1467812579	Mon Apr 06 22:20:17 PDT 2009	NO_QUERY	pardonlauren	I just re-pierced my ears
14	0	1467812723	Mon Apr 06 22:20:19 PDT 2009	NO_QUERY	TLeC	@caregiving I couldn't bear to watch it. And I thought the UA loss was embarr
15	0	1467812771	Mon Apr 06 22:20:19 PDT 2009	NO_QUERY	robrobbierobert	@octolinz16 It it counts, idk why I did either. you never talk to me anymore
16	0	1467812784	Mon Apr 06 22:20:20 PDT 2009	NO_QUERY	bayofwolves	@smarrison i would've been the first, but i didn't have a gun. not really thou
17	0	1467812799	Mon Apr 06 22:20:20 PDT 2009	NO_QUERY	HairByJess	@iamjazzyfizzle I wish I got to watch it with you!! I miss you and @iamlilnicki
18	0	1467812964	Mon Apr 06 22:20:22 PDT 2009	NO_QUERY	lovesongwriter	Hollis' death scene will hurt me severely to watch on film wry is directors cut
19	0	1467813137	Mon Apr 06 22:20:25 PDT 2009	NO_QUERY	armotley	about to file taxes
20	0	1467813579	Mon Apr 06 22:20:31 PDT 2009	NO_QUERY	starkissed	@LettyA ahh ive always wanted to see rent love the soundtrack!!

Figure 2-1 : Extrait du data set Sentiment140.

2-3.2. Twitter US Airline Sentiment Dataset:

Un travail d'analyse des sentiments sur les problèmes de chaque grande compagnie aérienne américaine. Les données Twitter ont été récupérées à partir de février 2015 et les contributeurs ont été invités à classer d'abord les tweets positifs, négatifs et neutres, puis à catégoriser les raisons négatives (telles que « vol tardif » ou « service impoli »).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	tweet_id	airline_sentiment	airline_sentiment_confi	negativereason	negativer	airline	airline_s	name	negativer	retweet_c	text	tweet_co	tweet_created	tweet_loc	user_timezone		
2	5.70306E+17	neutral	1			Virgin America		cairdin		0	@VirginAmerica Wh		2/24/2015 11:35		Eastern Time (US & Canada)		
3	5.70301E+17	positive	0.3486		0	Virgin America		jnardino		0	@VirginAmerica plu		2/24/2015 11:15		Pacific Time (US & Canada)		
4	5.70301E+17	neutral	0.6837			Virgin America		yvonnalynn		0	@VirginAmerica I d		2/24/2015 11:15	Lets Play	Central Time (US & Canada)		
5	5.70301E+17	negative	1	Bad Flight	0.7033	Virgin America		jnardino		0	@VirginAmerica it's		2/24/2015 11:15		Pacific Time (US & Canada)		
6	5.70301E+17	negative	1	Can't Tell	1	Virgin America		jnardino		0	@VirginAmerica an		2/24/2015 11:14		Pacific Time (US & Canada)		
7	5.70301E+17	negative	1	Can't Tell	0.6842	Virgin America		jnardino		0	@Virgin		2/24/2015 11:14		Pacific Time (US & Canada)		
8	5.70301E+17	positive	0.6745		0	Virgin America		cjmcginnis		0	@VirginAmerica yes		2/24/2015 11:13	San Franc	Pacific Time (US & Canada)		
9	5.703E+17	neutral	0.634			Virgin America		pilot		0	@VirginAmerica Re		2/24/2015 11:12	Los Angel	Pacific Time (US & Canada)		
10	5.703E+17	positive	0.6559			Virgin America		dhepburn		0	@virginamerica We		2/24/2015 11:11	San Diegc	Pacific Time (US & Canada)		
11	5.70295E+17	positive	1			Virgin America		YupitsTate		0	@VirginAmerica it v		2/24/2015 10:53	Los Angel	Eastern Time (US & Canada)		
12	5.70294E+17	neutral	0.6769		0	Virgin America		idk_but_youtube		0	@VirginAmerica dic		2/24/2015 10:48	1/1 loner	Eastern Time (US & Canada)		
13	5.7029E+17	positive	1			Virgin America		HyperCamiLax		0	@VirginAmerica I &		2/24/2015 10:30	NYC	America/New_York		
14	5.7029E+17	positive	1			Virgin America		HyperCamiLax		0	@VirginAmerica Thi		2/24/2015 10:30	NYC	America/New_York		
15	5.70287E+17	positive	0.6451			Virgin America		mollanderson		0	@VirginAmerica @v		2/24/2015 10:21		Eastern Time (US & Canada)		
16	5.70286E+17	positive	1			Virgin America		sjespers		0	@VirginAmerica Thi		2/24/2015 10:15	San Franc	Pacific Time (US & Canada)		
17	5.70282E+17	negative	0.6842	Late Flight	0.3684	Virgin America		smartwatermelon		0	@VirginAmerica SF(		2/24/2015 10:01	palo alto,	Pacific Time (US & Canada)		
18	5.70278E+17	positive	1			Virgin America		ItzBrianHunty		0	@VirginAmerica So		2/24/2015 9:42	west covi	Pacific Time (US & Canada)		
19	5.70277E+17	negative	1	Bad Flight	1	Virgin America		heatherovieda		0	@VirginAmerica I f		2/24/2015 9:39	this place	Eastern Time (US & Canada)		
20	5.70271E+17	positive	1			Virgin America		thebrandiray		0	I âĺ, flying @Virgin		2/24/2015 9:15	Somewhe	Atlantic Time (Canada)		
21	5.70268E+17	positive	1			Virgin America		JNLpierce		0	@VirginAmerica you		2/24/2015 9:04	Boston	Quito		
22	5.70266E+17	negative	0.6705	Can't Tell	0.3614	Virgin America		MISSGJ		0	@VirginAmerica wh		2/24/2015 8:55				
23	5.70264E+17	positive	1			Virgin America		DT_Les		0	@VirginA [40.74804		2/24/2015 8:49				
24	5.70259E+17	positive	1			Virgin America		ElvinaBeck		0	@VirginAmerica I lo		2/24/2015 8:30	Los Angel	Pacific Time (US & Canada)		
25	5.70259E+17	neutral	1			Virgin America		rjlynch21086		0	@VirginAmerica wil		2/24/2015 8:27	Boston, M	Eastern Time (US & Canada)		
26	5.70257E+17	negative	1	Customer Service Issue	0.3557	Virgin America		ayeevickiee		0	@VirginAmerica you		2/24/2015 8:18	714	Mountain Time (US & Canada)		

Figure 2-2 : Extrait du data set de Twitter US. Airlines Sentiment.

2-3.3. SuperFetch dataset

L'ensemble de données SuperFetch est utilisé pour rechercher les données d'entrée brutes. ww.osnews.com est un site Web qui fournit des informations sur SuperFetch et d'autres articles liés au système d'exploitation. Le site Web contient des informations détaillées sur les critiques positives et négatives de SuperFetch données par divers experts du système d'exploitation. Les critiques de SuperFetch sont également collectées auprès des différents professionnels des universités / collèges dans le format approprié, y compris la date, le nom de l'université / du collège, sous la forme d'une feuille Excel. Après la collecte des avis, les données sont enregistrées dans un ordre séquentiel et des informations volumineuses ont été tirées de cette revue, note et efficacité en forme de feuille avec la date.

2-3.4. SemEval (évaluation sémantique)

SemEval (évaluation sémantique) est une série d'évaluations en cours de systèmes d'analyse sémantique computationnelle ; il a évolué à partir de la série d'évaluation du sens des mots de Senseval. Les évaluations visent à explorer la nature du sens dans la langue.

Le SemEval fournit un mécanisme pour examiner les problèmes dans l'analyse sémantique des textes. Les sujets d'intérêt ne répondent pas à la rigueur logique que l'on trouve dans la sémantique informatique formelle, qui tente d'identifier et de caractériser les types de problèmes pertinents pour la compréhension humaine du langage. L'objectif principal est de reproduire le traitement humain au moyen de systèmes informatiques.

2-4. Architecture du système

La Figure 2-3, montre l'architecture du système proposé qui est basé sur les classificateurs SVM et KNN :

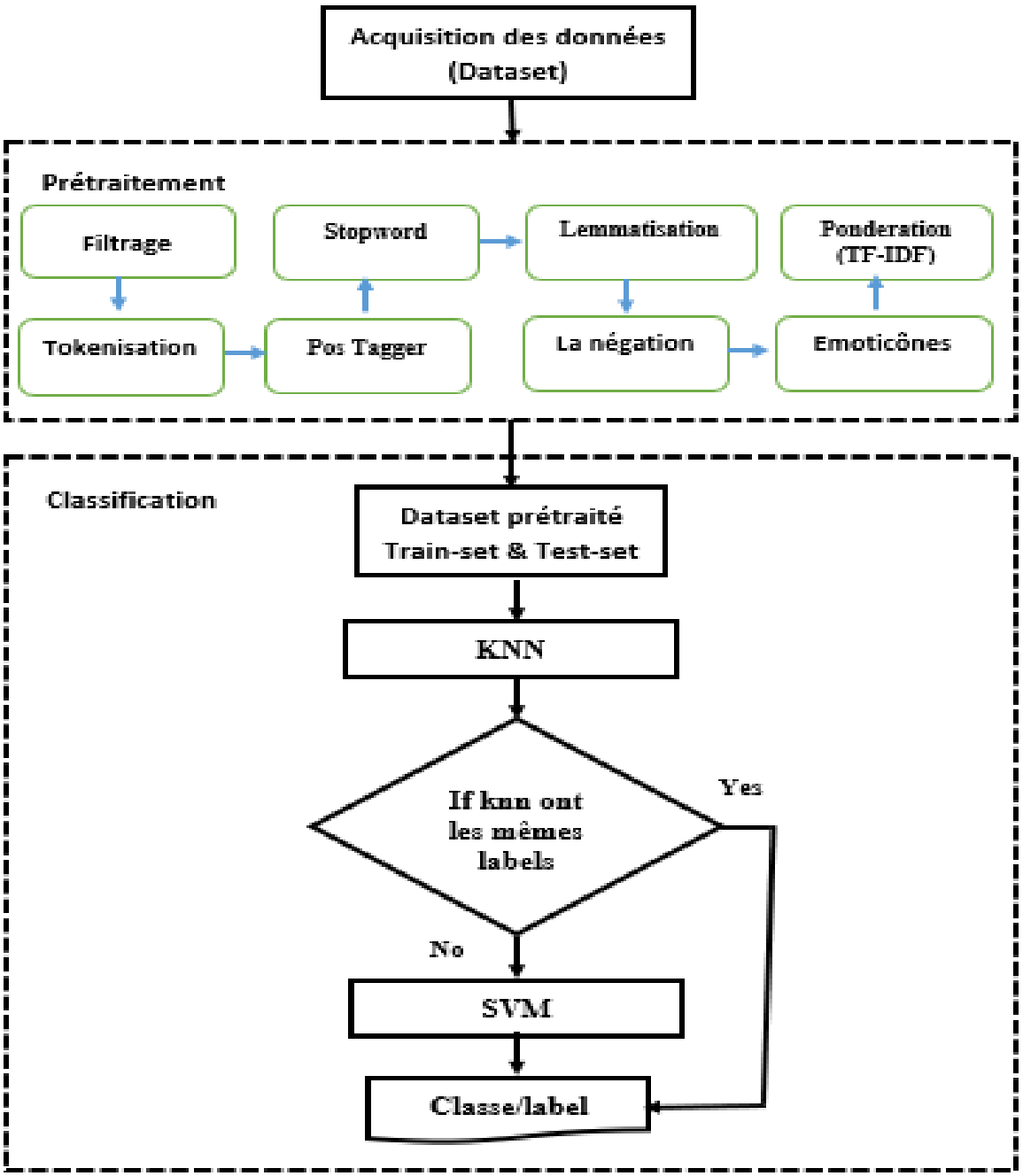

Figure 2-3 : Architecture du système.

2-4.1. Etape d'acquisition

Notre application fonctionne uniquement en anglais, ce processus permet de saisir des données et ce sont des tweets sur Twitter, et le processus de saisie se fait via la base de données ou manuellement.

2-4.2. Prétraitement

Le prétraitement des données est une technique d'exploration de données qui consiste à transformer les données brutes en un format compréhensible. Les données du monde réel sont souvent incomplètes, incohérentes, absentes de certains comportements ou tendances et sont susceptibles de contenir de nombreuses erreurs. Le prétraitement des données est une méthode éprouvée pour résoudre ces problèmes.

Lors de la collecte de données, on peut rencontrer trois principaux facteurs qui contribueraient à la qualité des données :

- **Exactitude** : valeurs erronées qui s'écartent des attentes. Les causes des données inexactes peuvent être diverses, notamment :
    - Erreurs humaines / informatiques lors de la saisie et de la transmission des données
    - Les utilisateurs soumettant délibérément des valeurs incorrectes (appelées données manquantes déguisées)
    - Formats incorrects pour les champs de saisie
    - Duplication d'exemples de formation
- **Complétude** : absence de valeurs d'attributs / caractéristiques ou de valeurs d'intérêt. L'ensemble de données peut être incomplet en raison de:
    - Indisponibilité des données
    - Suppression des données incohérentes
    - Suppression de données jugées non pertinentes initialement.
- **Cohérence** : l'agrégation des données est incohérente.
  Certaines autres caractéristiques qui affectent également la qualité des données incluent la rapidité (les données sont incomplètes jusqu'à ce que toutes les informations pertinentes soient soumises après certaines périodes), la crédibilité (dans quelle mesure les données sont-elles fiables par l'utilisateur) et l'interprétabilité (la facilité avec laquelle les données sont comprises par toutes les parties prenantes).

Pour garantir des données de haute qualité, il est essentiel de les prétraiter. Pour faciliter le processus, le prétraitement des données est divisé en sept étapes :

- Filtrage.
- Tokenisation.
- Pos tagger.
- Stop Word.
- Lemmatisation.
- La négation.
- Emoticônes.

Figure 2-4 : Les processus de prétraitement.

2-4.2.1. Filtrage

Le processus de nettoyage et de préparation du texte pour la classification. Les textes en ligne contiennent généralement beaucoup de bruit et de parties non informatives telles que des **balises HTML**, des **scripts** et des **publicités**. De plus, au niveau des mots, de nombreux mots du texte n'ont pas d'impact sur l'orientation générale de celui-ci.

L'ensemble du processus comprend plusieurs étapes : nettoyage du texte en ligne, suppression des espaces blancs, extension de l'abréviation, extraction, suppression des mots vides, gestion de la négation.

Nous suggérons quelques exemples de Twitter statuts contenant de grandes quantités de données indépendantes pour la suppression : l'hashtag, l'URL du lien, les symboles et le nom d'utilisateur

Figure 2-5 : Exemple d'un tweet contenant l'hashtag, l'URL du lien, les symboles et le nom d'utilisateur.

2-4.2.2. Tokenisation ("Découpage en mots ")

La tokenisation est l'une des tâches les plus courantes lorsqu'il s'agit de travailler avec des données texte. ... La tokenisation consiste essentiellement à diviser une phrase, une phrase, un paragraphe ou un document texte entier en unités plus petites, telles que des mots ou des termes individuels. Chacune de ces petites unités est appelée jetons.

Figure 2-6 : Tweet composé du texte et des signes de ponctuation.

Le processus de tokenisation transformer le texte en liste de tokens, donc le tweet de la figure 2-6 devient :

('Mercedes' , '-' , 'Maybach' , 'S' , '-' , 'Class' , ':' , 'A' , 'success' , 'story' , 'itself' '.')

2-4.2.3. Pos tagger

Pos tagger (ou la balise d'une partie du discours) est une désignation spéciale qui est attribuée à chaque jeton (mot) dans un groupe de texte pour désigner la partie du discours et souvent aussi d'autres catégories grammaticales telles que verbe, nombre (pluriel / singulier), adjectif, pronom, adverbe, etc. Les balises POS sont utilisées dans les recherches de groupe, les outils d'analyse de texte et les algorithmes.

Figure 2-7 : Exemple d'affichage des Pos tagger.

text = He has worked hard.

Tokens= (He, has, worked, hard).

Après POS TAG la phrase devient : [(He, PP), (has, VHZ), (worked, VVN),

(hard, RB)].

POS TAG	Description
PP	Pronom personnel
VHZ	Le verbe avoir, présent à la 3e personne
VVN	Verbe, participe passer
RB	Adverbe

Table 2.1: Ensemble d'étiquettes.

POS Tag	Description	Example
CC	coordinating conjunction	*and, but, or, &*
CD	cardinal number	*1, three*
DT	determiner	*the*
EX	existential there	*there is*
FW	foreign word	*d'œuvre*
IN	preposition/subord. conj.	*in,of,like,after,whether*
IN/that	complementizer	*that*
JJ	adjective	*green*
JJR	adjective, comparative	*greener*
JJS	adjective, superlative	*greenest*
LS	list marker	*(1).*
MD	modal	*could, will*
NN	noun, singular or mass	*table*
NNS	noun plural	*tables*
NP	proper noun, singular	*John*
NPS	proper noun, plural	*Vikings*
PDT	predeterminer	*both the boys*
POS	possessive ending	*friend's*
PP	personal pronoun	*I, he, it*
PPS	possessive pronoun	*my, his*
RB	adverb	*however, usually, here, not*
RBR	adverb, comparative	*better*
RBS	adverb, superlative	*best*
RP	particle	*give up*
SENT	end punctuation	*?, !, .*
SYM	symbol	*@, +, *, ^, |, =*
TO	*to*	*to go, to him*
UH	interjection	*uhhuhhuhh*
VB	verb *be*, base form	*be*
VBD	verb *be*, past	*was|were*
VBG	verb *be*, gerund/participle	*being*
VBN	verb *be*, past participle	*been*
VBZ	verb *be*, pres, 3rd p. sing	*is*
VBP	verb *be*, pres non-3rd p.	*am|are*
VD	verb *do*, base form	*do*
VDD	verb *do*, past	*did*
VDG	verb *do* gerund/participle	*doing*
VDN	verb *do*, past participle	*done*

Figure 2-8 : Ensemble d'étiquettes.

2-4.2.4. Stop Words

Mots vides : un mot vide est un mot couramment utilisé (comme “the”, “a”, “an”, “in”) qu'un moteur de recherche a été programmé pour ignorer, à la fois lors de l'indexation des entrées pour la recherche et lors de leur récupération. À la suite d'une requête de recherche, cela ne prend donc pas de place dans notre base de données, ni ne prend un temps de traitement précieux. Des exemples :

Les conjonctions de coordination	Les déterminants	Les prépositions
For	A/an	At
And	The	In
Nor	This	To
But	That	
Or	These	
Yet	Those	
So		

Table 2.2 : Tableau qui contenant des mots vides (stopwords).

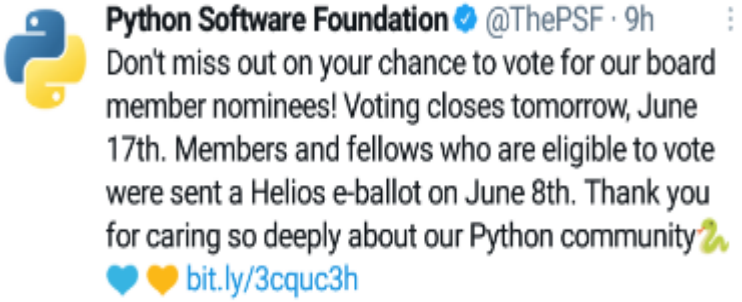

Figure 2-9 : Exemple de Tweet contentant des mots vides (stopwords).

2-4.2.5. Lemmatisation

La lemmatisation fait généralement référence à faire les choses correctement avec l'utilisation d'un vocabulaire et une analyse morphologique des mots, visant normalement à supprimer uniquement les fins flexionnelles et à renvoyer la forme de base ou de dictionnaire d'un mot, qui est connue sous le nom de lemme.

Figure 2-10 : Exemple de Tweet des mots sans lemmatisation.

Mots sans lemmatisation	Mots avec lemmatisation
Humbled	Humble
Builds	Build

Table 2.3 : Tableau qui affiche des mots sans lemmatisation et avec lemmatisation.

2-4.2.6. La négation

C'est un processus basé sur la négation qui intervient par mots successifs, lorsqu'un mot est suivi d'un mot négatif dans la formulation du sens, ce sens devient positif, lorsqu'un mot de négation apparaît dans une phrase, il est nécessaire de détecter sa portée, c'est-à-dire le nombre de termes suivants qui sont affectés par la négation. En général, il n'y a pas de fenêtre de négation fixe, car elle dépend de la structure particulière de la phrase. Les phrases complexes peuvent avoir plusieurs clauses dépendantes, reliées par de nombreuses conjonctions. Par exemple: “I do not like vegetables but they are healthy”.

2-4.2.7. Emoticones

Pour l'analyse de texte, nous pouvons avoir besoin de gérer des émoticônes. Parfois, les émoticônes donnent des informations fortes sur un texte, comme l'expression de sentiments. La meilleure approche pour traiter les emojis consiste à convertir les emojis en mots afin qu'il soit utile de conserver les informations à l'aide d'un dictionnaire d'émoticônes.

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

2-4.3. Extraction et présentation des descripteurs (TF-IDF)

TF-IDF (Term Frequency-Inverse Document Frequency) est une mesure statistique qui évalue la pertinence d'un mot pour un document dans une collection de documents. A été inventé pour la recherche de documents et la recherche d'informations. Cela fonctionne en augmentant proportionnellement au nombre de fois qu'un mot apparaît dans un document, mais est compensé par le nombre de documents qui contiennent le mot. Ainsi, les mots qui sont communs dans tous les documents, tels que ceci, quoi et si, se classent bas même s'ils peuvent apparaître plusieurs fois, car ils ne signifient pas grand-chose pour ce document en particulier.

TF-IDF pour un mot dans un document est calculé en multipliant deux métriques différentes :

- **Le terme fréquence d'un mot dans un document** : Il existe plusieurs façons de calculer cette fréquence, la plus simple étant le décompte brut des occurrences d'un mot dans un document. Ensuite, il existe des moyens d'ajuster la fréquence, par la longueur d'un document, ou par la fréquence brute du mot le plus fréquent dans un document.
- **La fréquence inverse du document du mot sur un ensemble de documents** : Cela signifie à quel point un mot est commun ou rare dans l'ensemble du jeu de documents. Plus il est proche de 0, plus un mot est courant. Cette métrique peut être calculée en prenant le nombre total de documents, en le divisant par le nombre de documents contenant un mot et en calculant le logarithme.

Ainsi, si le mot est très courant et apparaît dans de nombreux documents, ce nombre se rapprochera de 0. Sinon, il se rapprochera de 1.

La multiplication de ces deux nombres donne le score TF-IDF d'un mot dans un document. Plus le score est élevé, plus ce mot est pertinent dans ce document particulier.

Le score TF-IDF pour le mot t dans le document d de l'ensemble de documents D est calculé comme suit :

$$tf - idf\ (t, d, D) = tf(t, d).idf\ (t, D)$$

Où :

$$tf\ (t, d)\ =\ log\ (1\ +\ freq\ (t, d))$$

$$idf\ (t, D) = \log\ (\frac{N}{count\ (\ d\ \in D: t \in d)})$$

2-4.4. Classification des sentiments

Après la phase de prétraitement, l'étape suivante est la classification des sentiments. La classification des sentiments de Twitter, qui identifie la polarité des sentiments des tweets courts et informels, a suscité un intérêt croissant dans la recherche ces dernières années. Les méthodes basées sur l'apprentissage pour la classification des sentiments Twitter qui traitent la classification des sentiments des textes comme un cas particulier de problème de catégorisation de texte. De nombreuses méthodes d'apprentissage existantes sur la classification des sentiments sur Twitter que les performances du classificateur de sentiment dépendent fortement du choix de la représentation des fonctionnalités des tweets.

Pour la classification des sentiments en négatif et positif dans le cas d'une classification binaire et en négatif, positif et neutre dans le cas d'une classification multiple, de nombreux algorithmes d'apprentissage ont été utilisés, comme Naïve Bayes, Support Vector Machines (SVM), k-Nearest Neighbors (kNN), etc. Les techniques mentionnées ci-dessus ne sont pas suffisantes pour obtenir des résultats satisfaisants si elles sont utilisées séparément. Pour cette raison, de nouvelles techniques basées sur le principe de la combinaison de classificateurs sont apparues. Dans les sections suivantes nous présentons les algorithmes de classification KNN et SVM et ensuite la méthode hybride basée sur la combinaison des algorithmes KNN et SVM (KNN-SVM).

2-4.4.1. KNN

L'algorithme K-plus proche voisins (KNN) est un type d'algorithme de ML supervisé qui peut être utilisé à la fois pour les problèmes de classification et de prédiction de régression. Cependant, il est principalement utilisé pour les problèmes prédictifs de classification dans l'industrie. Les deux propriétés suivantes définiraient bien KNN :

- Algorithme d'apprentissage paresseux - KNN est un algorithme d'apprentissage paresseux car il n'a pas de phase d'entraînement spécialisée et utilise toutes les données pour l'entraînement lors de la classification.
- Algorithme d'apprentissage non paramétrique - KNN est également un algorithme d'apprentissage non paramétrique car il ne suppose rien sur les données sous-jacentes.

L'algorithme des K-plus proche voisins (KNN) utilise la « similarité des caractéristiques » pour prédire les valeurs des nouveaux points de données, ce qui signifie en outre que le nouveau point de données se verra attribuer une valeur en fonction de sa correspondance avec les points de l'ensemble d'apprentissage. Nous pouvons comprendre son fonctionnement à l'aide des étapes suivantes :

Étape 1 - Nous devons charger la formation ainsi que les données de test.

Étape 2 - Ensuite, nous devons choisir la valeur de K, c'est-à-dire les points de données les plus proches. K peut être n'importe quel entier.

Étape 3 - Pour chaque point des données de test, **procédez comme suit -**

3.1 - Calculer la distance entre les données de test et chaque ligne de données d'apprentissage à l'aide de l'une des méthodes à savoir : distance euclidienne, Manhattan ou Hamming. La méthode la plus couramment utilisée pour calculer la distance est euclidienne.

3.2 - Maintenant, en fonction de la valeur de la distance, triez-les par ordre croissant.

3.3 - Ensuite, il choisira les K premières lignes du tableau trié.

3.4 - Maintenant, il attribuera une classe au point de test en fonction de la classe la plus fréquente de ces lignes.

Étape 4 – Fin.

2-4.4.2. SVM

Support Vector Machine ou SVM est l'un des algorithmes d'apprentissage supervisé les plus populaires, utilisé pour la classification.

Le but de l'algorithme SVM est de créer la meilleure ligne ou frontière de décision qui puisse séparer l'espace à n dimensions en classes afin que nous puissions facilement placer le nouveau point de données dans la catégorie correcte à l'avenir. Cette meilleure frontière de décision est appelée un hyperplan.

SVM choisit les points / vecteurs extrêmes qui aident à créer l'hyperplan. Ces cas extrêmes sont appelés vecteurs de support, et par conséquent, l'algorithme est appelé Support Vector Machine. Considérez le Figure 2-12 ci-dessous dans lequel il existe deux catégories différentes qui sont classées à l'aide d'une frontière de décision ou d'un hyperplan :

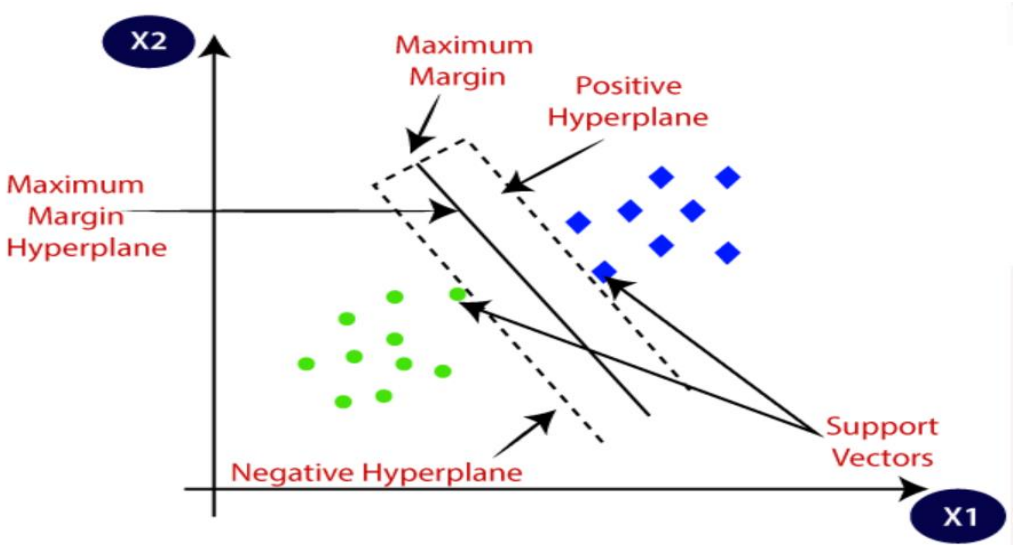

Figure 2-12 : deux catégories différentes sont classées à l'aide d'un hyperplan.

SVM peut être de deux types

- **SVM linéaire :** le SVM linéaire est utilisé pour les données linéairement séparables, ce qui signifie que si un ensemble de données peut être classé en deux classes en utilisant une seule ligne droite, ces données sont appelées données linéairement séparables, et le classificateur est appelé classificateur SVM linéaire.
- **SVM non linéaire :** SVM non linéaire est utilisé pour les données séparées non linéairement, ce qui signifie que si un ensemble de données ne peut pas être classé en utilisant une ligne droite, alors ces données sont

appelées données non linéaires et le classificateur utilisé est appelé non-classificateur SVM linéaire.

Hyperplan : Il peut y avoir plusieurs lignes / frontières de décision pour séparer les classes dans un espace à n dimensions, mais nous devons trouver la meilleure frontière de décision qui aide à classer les points de données. Cette meilleure frontière est connue sous le nom d'hyperplan de SVM.

Les dimensions de l'hyperplan dépendent des entités présentes dans le jeu de données, ce qui signifie que s'il y a 2 entités (comme indiqué sur l'image), alors l'hyperplan sera une ligne droite. Et s'il y a 3 entités, alors l'hyperplan sera un plan à 2 dimensions.

Nous créons toujours un hyperplan qui a une marge maximale, ce qui signifie la distance maximale entre les points de données.

Vecteurs de soutien : Les points de données ou vecteurs les plus proches de l'hyperplan et qui affectent la position de l'hyperplan sont appelés vecteur de support. Étant donné que ces vecteurs prennent en charge l'hyperplan, donc appelé vecteur de support.

2-4.4.3. L'algorithme hybride : SVM + SVM

Dans notre travail, nous utilisons une méthode hybride basée sur la combinaison des algorithmes SVM et KNN. Notre objectif est d'optimiser l'algorithme SVM et de simplifier le processus d'apprentissage, en intégrant l'algorithme KNN dans le SVM. Dans la littérature, nous avons constaté que l'approche de classification du SVM a de bonnes performances mais qu'elle souffre d'une forte consommation de temps d'exécution et d'une utilisation importante de la mémoire physique, en raison de ses processus d'apprentissage et de classification alambiqués, surtout lorsque la dimensionnalité des données est élevée, en ajoutant que lorsque le nombre de données pour l'apprentissage est moindre par rapport aux données de test. En outre, l'approche de classification KNN est remarquable par sa simplicité et son processus de formation à faible coût. Ensuite, dans notre travail sur la classification de sentiments, nous visons à simplifier le processus d'apprentissage et à optimiser l'algorithme SVM pour obtenir de meilleurs résultats, par l'intégration de l'algorithme KNN dans l'algorithme SVM. L'idée principale du classificateur hybride basé sur la combinaison de KNN et de SVM (KNN-SVM) est de trouver les voisins les plus proches d'un échantillon de requête par l'algorithme KNN, et de former un SVM local qui préserve la fonction de distance sur l'ensemble

des voisins. Notre algorithme de classification nommé « KNN-SVM » est résumé par le pseudo-code suivant :

Algorithme : KNN-SVM

1. Calculez la distance euclidienne de la requête pour tous les exemples d'apprentissage et sélectionnez les k voisins les plus proches.

2. Si les k voisins ont tous les mêmes étiquettes (sont de la même classe), la requête est étiquetée et fin ;

3. Sinon, calculez les distances par paires entre les k voisins et stockez-les dans une matrice ;

4. Convertissez la matrice de distance en une matrice de noyau en utilisant le "kernel trick" et appliquez le SVM multi classe.

5. Utilisez le classificateur résultant pour étiqueter la requête.

Par exemple, prenez $\mathbf{K(x,y) = exp(-d(x,y)/\sigma 2)}$, dans une fonction de noyau à base radiale. Ou la fonction de noyau polynomial. $\boldsymbol{K(x_i\ x_j) = (x_i^T\ x_i)^P}$

La figure 2.20 montre un exemple de classification KNN-SVM.

En termes de classification, les résultats obtenus montrent que la combinaison des algorithmes SVM et KNN améliore encore les performances de la classification par rapport aux cas où le SVM ou le KNN sont utilisés individuellement. La méthode proposée consomme moins de mémoire et prend moins de temps de calcul que le SVM.

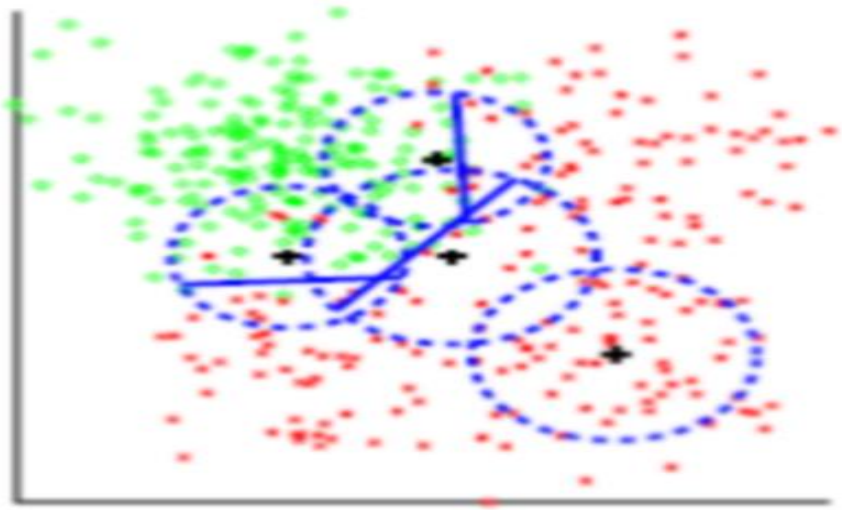

Figure 2-13 : Notre méthode entraîne un SVM sur les 50 voisins les plus proches.

Chapitre 3

3-1 Environnement du travail

D'abord, nous donnons une description de l'environnement de notre expérimentation :

3-1.1 Environnement matériel :

Afin de mener notre expérimentation et évaluation, nous avons utilisé un PC de caractéristiques suivantes :

PyCharm est un IDE multiplateforme qui offre une expérience cohérente sur les systèmes d'exploitation Windows, MacOs et Linux.

PyCharm est disponible en trois éditions : Professional, Community et Edu. Les éditions Community et Edu sont des projets open source et elles sont gratuites, mais elles ont moins de fonctionnalités. PyCharm Edu propose des cours et vous aide à apprendre la programmation avec Python. L'édition professionnelle est commerciale et fournit un ensemble exceptionnel d'outils et de fonctionnalités.

Exigence	Minimum	Recommandé
RAM	4 Go de RAM libre	8 Go de RAM système totale
CPU	Tout processeur moderne	Processeur multicœur. PyCharm prend en charge le multithreading pour différentes opérations et processus, ce qui le rend plus rapide plus il peut utiliser de cœurs de processeur.
Espace disque	2,5 Go et 1 Go supplémentaire pour les caches	Disque SSD avec au moins 5 Go d'espace libre
Résolution du moniteur	1024x768	1920×1080
Système opérateur	Versions 64 bits officiellement publiées	Dernière version 64 bits de Windows, MacOs ou Linux

Table 3.1 : Tableau qui affiche Exigence de PyCharm.

PyQt5 est un ensemble complet de liaisons Python pour Qt v5. Il est implémenté sous forme de plus de 35 modules d'extension. Il permet ainsi de créer des interfaces graphiques en Python.

3-1.2 Environnement logiciel :

Python : Python est un langage de programmation interprété, orienté objet et de haut niveau avec une sémantique dynamique. La syntaxe simple et facile à apprendre de Python met l'accent sur la lisibilité et réduit donc le coût de la maintenance du programme. Python prend en charge les modules et les packages, ce qui encourage la modularité du programme et la réutilisation du code.

Figure 3-1 : Logo de Python.

Nltk : NLTK est une bibliothèque permettant de créer des programmes Python pour travailler avec des données en langage humain. Il fournit des interfaces faciles à utiliser pour plus de 50 corpus et ressources lexicales telles que WordNet, ainsi qu'une suite de bibliothèques de traitement de texte pour la classification, la tokenisation, la recherche de racines, le marquage, l'analyse et le raisonnement sémantique.

Textblob : Textblob est une bibliothèque Python pour le traitement de données textuelles. Il fournit une API simple pour plonger dans les tâches courantes de traitement du langage naturel (NLP) telles que le balisage d'une partie du discours, l'extraction de phrases nominales, l'analyse des sentiments, la classification, la traduction, etc.

Scikit-learn : une bibliothèque pour le langage de programmation Python généralement utilisé dans les projets d'apprentissage automatique. Scikit-learn se concentre sur les outils d'apprentissage automatique, y compris les algorithmes mathématiques, statistiques et à usage général qui constituent la base de nombreuses technologies d'apprentissage automatique.

Jupiter : Jupiter Notebook est une application Web open source qui vous permet de créer et de partager des documents contenant du code en direct, des équations, des visualisations et du texte narratif. Les utilisations comprennent :

le nettoyage et la transformation des données, la simulation numérique, la modélisation statistique, la visualisation des données, l'apprentissage automatique et bien plus encore.

3-2 Expérimentations et interprétations

Après l'étape d'apprentissage, nous passons à l'étape de test afin de valider notre modèle et de voir l'intérêt de cette approche et ses avantages par rapport aux techniques classiques d'apprentissage automatique. Nous réalisons une série d'expérimentations pour vérifier la faisabilité et l'efficacité de notre modèle de classification. Dans la première série d'expérimentations, nous avons opté pour le SVM comme classifieur. Dans la deuxième série d'expérimentations, nous avons appliqué le KNN comme classifieur. Dans la troisième série d'expériences, nous avons utilisé notre classifieur SVM+KNN. Enfin, nous comparons les performances de trois modèles.

Pour la validation des performances de notre modèle, nous utilisons la méthode 70% 30%, telle que 70% utilisée pour la phase d'apprentissage et 30% pour la phase de test.

Pour nos expérimentations, nous n'avons considéré que 10% du premier ensemble de données (Sentiment140) et 100% du second ensemble de données que nous pouvons résumer dans le tableau suivant :

	Dataset size	Train size (70%)	Test size (30%)
DS 1	160000	112000	48000
DS 2	14000	9800	4200

Table 3.2 : le nombre de tweets pour chaque phase

Les mesures de performance utilisées sont la précision, le rappel et l'échelle F1 dont leurs bases de calcul se fait par rapport au tableau 3.2.

		Réel	
		Positive	Négative
Prédite	Positive	VP	FP
	Négative	FN	VN

Table 3.3 : Table de confusion

Sachant que : VP : Vrai Positif, FP : Faux Positif, VN : Vrai négatif, FN : Faux négatif.

Tel que les métriques que nous avons étudiées sont présentées sous les formes suivantes :

- $Accuracy = \frac{VP+VN}{VP+VN+FP+FN}$
- $Precision = \frac{VP}{VP+FP}$
- $Rappel = \frac{VP}{VP+FN}$
- $F1\_mesure = 2 \times \frac{Precision \times Rappel}{Precision+Rappel}$

Résultats et discussion

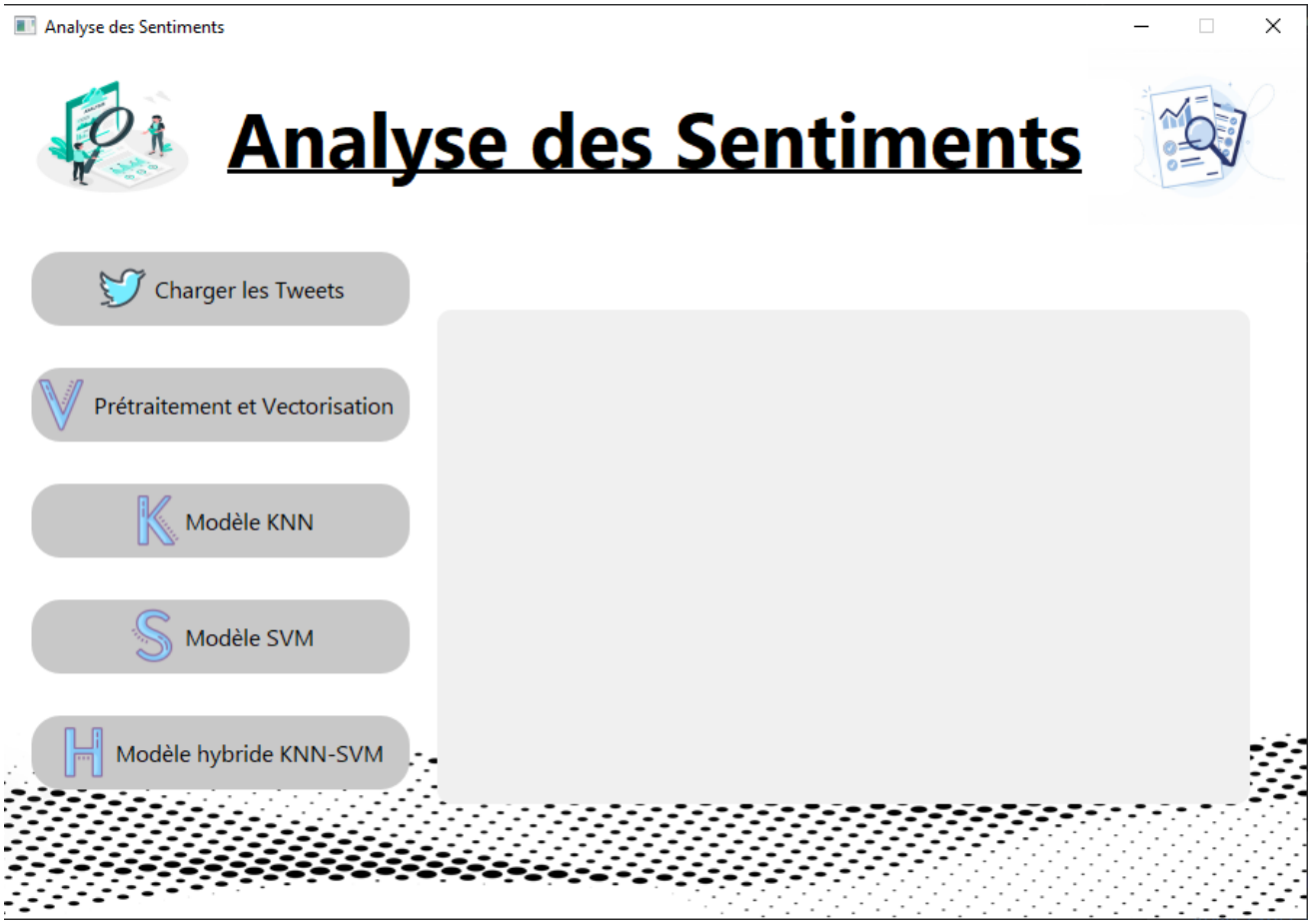

Figure 3-2 : Interface graphique.

➢ Résultats du modèle KNN

La Table 3.4 résume les résultats du dodèle KNN pour différentes valeur de K sur le premier ensemble de données (US. Airlines Sentiment).

	Mesure (%)	Positive	Négative	Accuracy (%)
K = 5	Précision	**58**	**62**	**60**
	Rappel	**60**	**60**	
	F-mesure	**59**	**61**	
K = 20	Précision	**47**	**72**	**59**
	Rappel	**62**	**57**	
	F-mesure	**53**	**64**	
K = 50	Précision	**41**	**75**	**58**
	Rappel	**62**	**56**	
	F-mesure	**49**	**64**	
K = 100	Précision	**33**	**81**	**57**
	Rappel	**64**	**55**	
	F-mesure	**44**	**65**	

Table 3.4 : les résultats de modèles KNN (Sentiment140)

Précision KNN avec K=100 (Dataset US. Airlines Sentiment) sur interface :

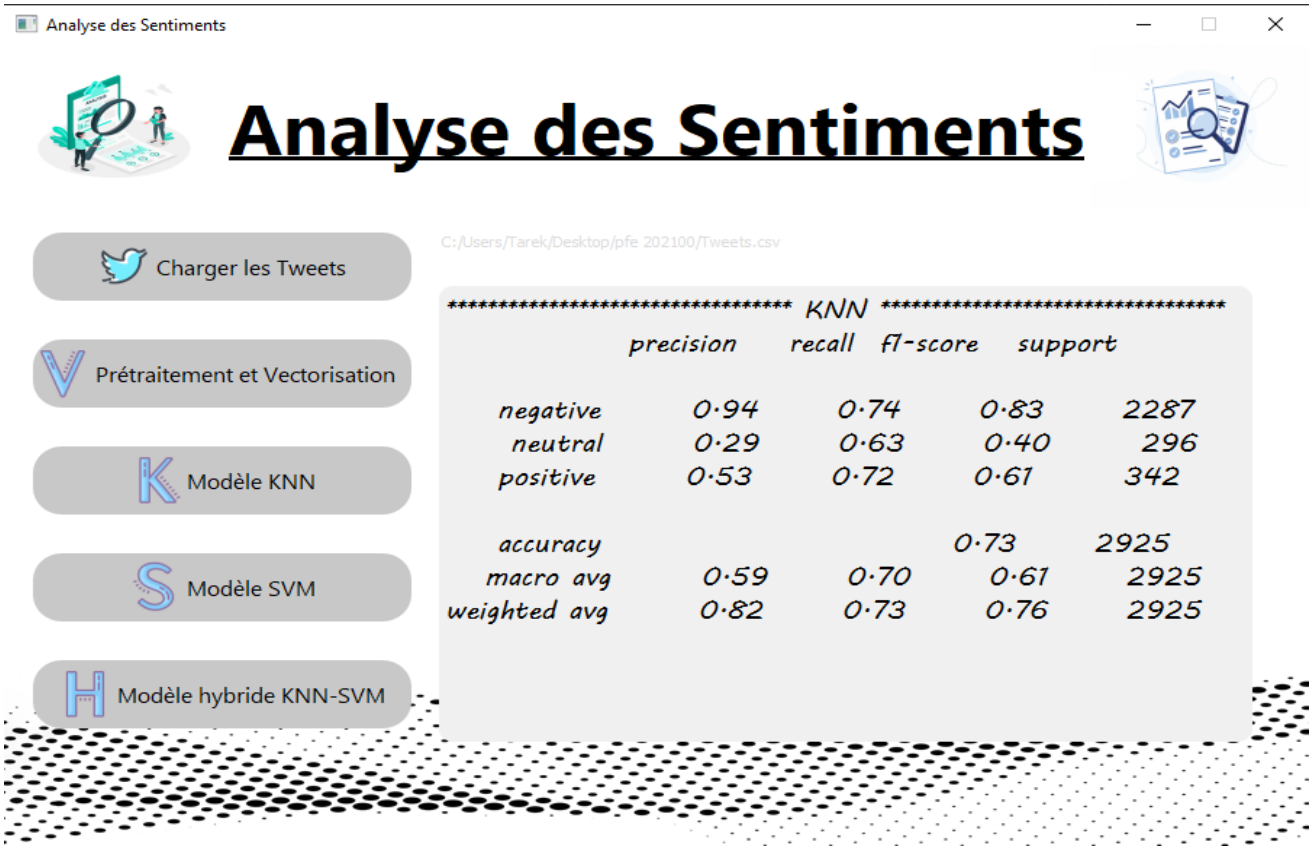

Figure 3-3 : Précision KNN avec K=100 (Dataset US. Airlines Sentiment).

La Table 3.5 résume les résultats du modèle KNN pour différentes valeurs de K sur le second ensemble de données.

	Mesure (%)	Positive	Neutral	Négative	Accuracy (%)
K = 5	Précision	**21**	**85**	**20**	**34**
	Rappel	**79**	**23**	**74**	
	F-mesure	**33**	**36**	**32**	
K = 20	Précision	**13**	**97**	**03**	**24**
	Rappel	**84**	**21**	**88**	
	F-mesure	**23**	**35**	**05**	
K = 50	Précision	**27**	**69**	**67**	**61**
	Rappel	**84**	**33**	**84**	
	F-mesure	**41**	**45**	**75**	
K = 100	Précision	**53**	**29**	**94**	**73**
	Rappel	**72**	**63**	**74**	
	F-mesure	**61**	**40**	**83**	

Table 3.5 : les résultats de modèles KNN (US. Airlines Sentiment).

- **Résultat du modèle SVM**

La Table 3.6 résume les résultats du modèle SVM sur le premier ensemble de données.

Mesure (%)	Positive	Négative	Accuracy (%)
Précision	**78**	**74**	**76**
Rappel	**75**	**77**	
F-mesure	**77**	**76**	

Table 3.6 : les résultats de modèles SVM (Sentiment140)

Précision SVM avec K=100 (Dataset US. Airlines Sentiment) sur interface :

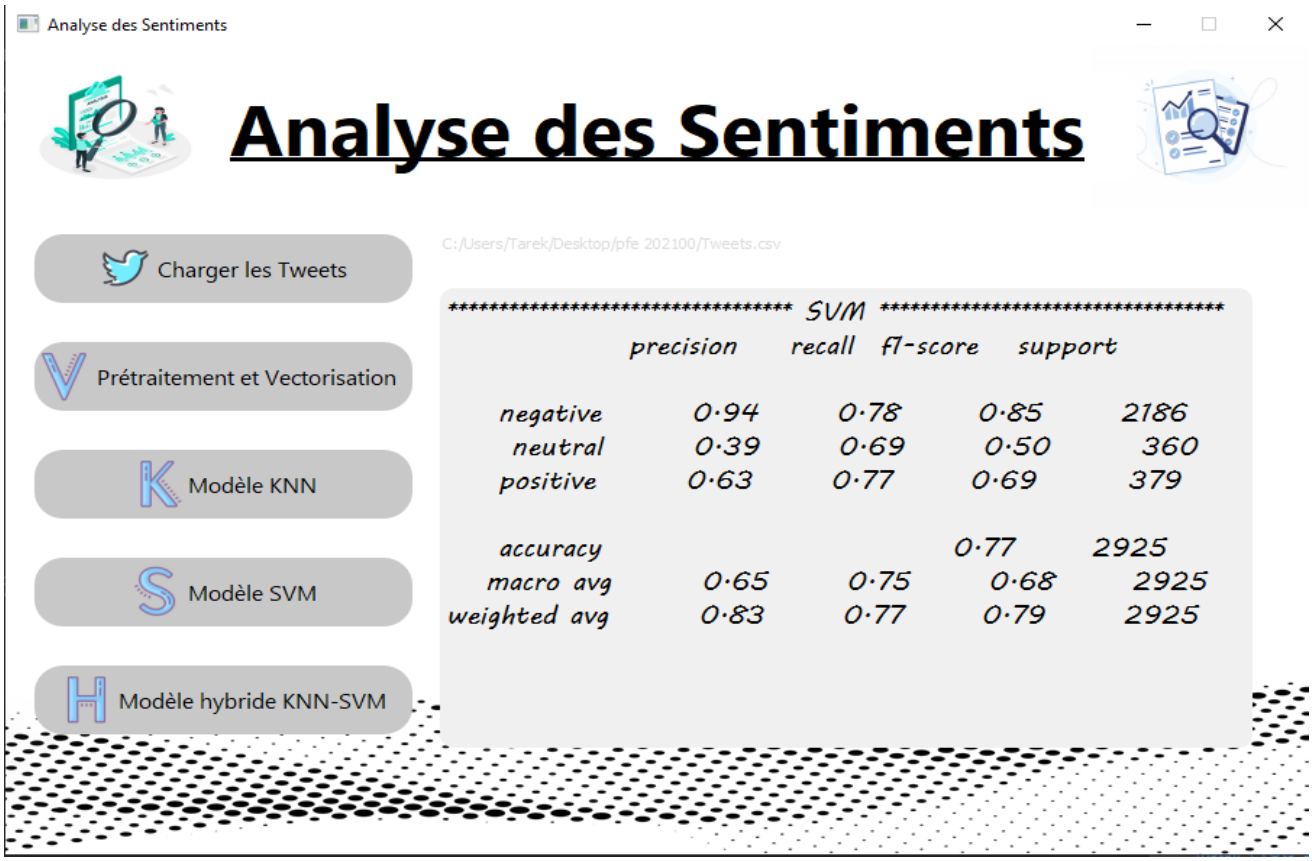

Figure 3-4 : Précision SVM K=100 (Dataset US. Airlines Sentiment).

La Table 3.7 résume les résultats du modèle SVM sur le second ensemble de données.

Mesure (%)	Positive	Neutral	Négative	Accuracy (%)
Précision	**63**	**39**	**94**	**77**
Rappel	**77**	**69**	**78**	
F-mesure	**69**	**50**	**85**	

Table 3.7 : les résultats de modèles SVM (US. Airlines Sentiment)

- **Résultats du modèle KNN+SVM**

La Table 3.8 résume les résultats du notre modèle KNN+SVM sur le premier ensemble de données avec différentes valeur de K.

	Mesure (%)	Positive	Négative	Accuracy (%)
K = 5	Précision	**59**	**62**	**61**
	Rappel	**61**	**60**	
	F-mesure	**60**	**61**	
K = 20	Précision	**59**	**64**	**62**
	Rappel	**62**	**61**	
	F-mesure	**61**	**63**	
K = 50	Précision	**56**	**67**	**62**
	Rappel	**63**	**61**	
	F-mesure	**60**	**64**	
K = 100	Précision	**52**	**72**	**62**
	Rappel	**65**	**60**	
	F-mesure	**57**	**65**	

Table 3.8 : les résultats de modèles KNN+SVM (Sentiment140)

Précision KNN+SVM avec K=100 (Dataset US. Airlines Sentiment) sur interface :

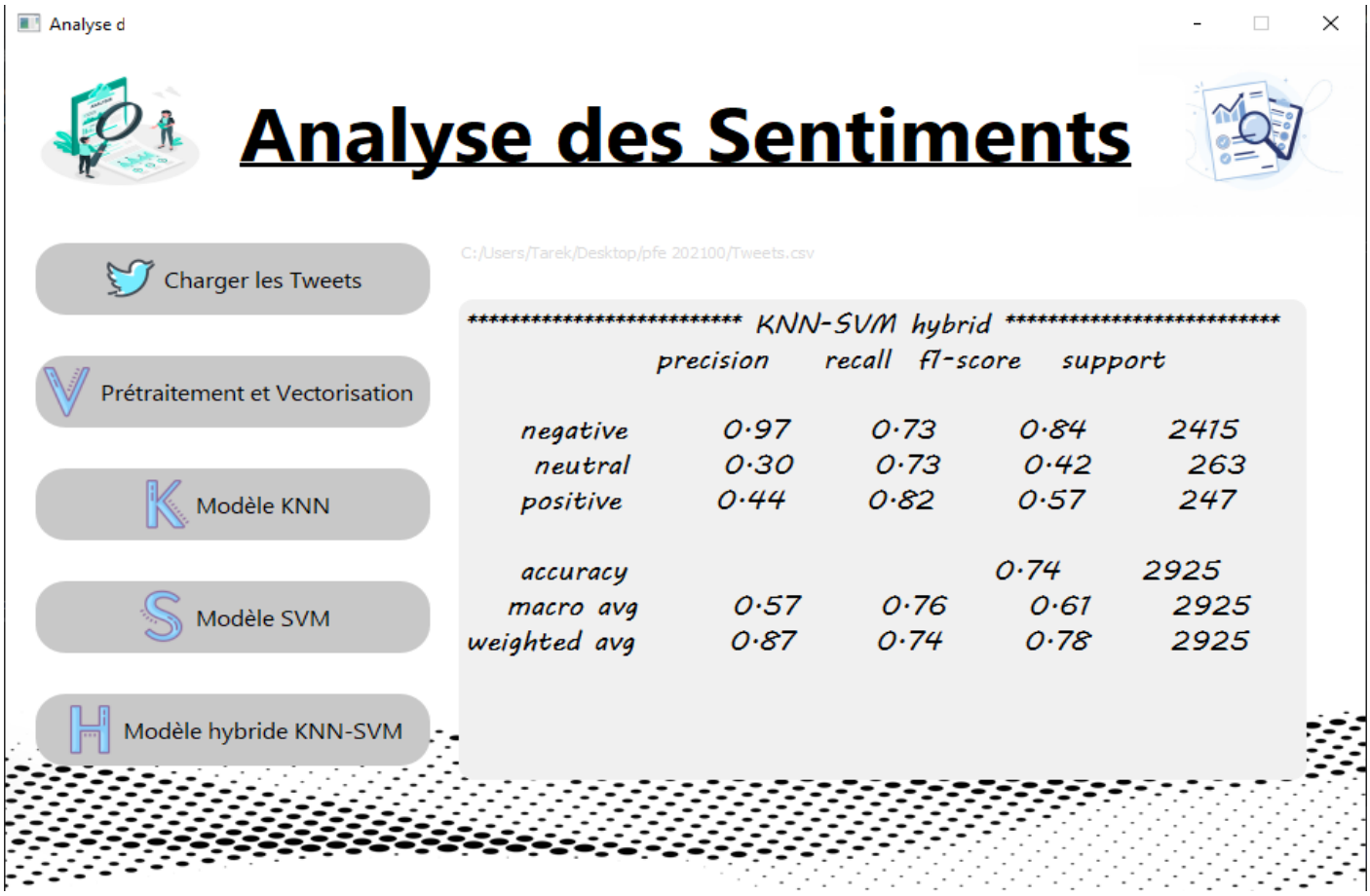

Figure 3-5: Precision KNN+SVM avec K=100 (Dataset US. Airlines Sentiment).

La Table 3.9 résume les résultats du notre modèle KNN+SVM sur le second ensemble de données avec différentes valeur de K.

	Mesure (%)	Positive	Neutral	Négative	Accuracy (%)
K = 5	Précision	**28**	**84**	**20**	**35**
	Rappel	**74**	**23**	**77**	
	F-mesure	**40**	**36**	**32**	
K = 20	Précision	**28**	**88**	**17**	**34**
	Rappel	**73**	**23**	**80**	
	F-mesure	**40**	**37**	**29**	
K = 50	Précision	**44**	**38**	**94**	**74**
	Rappel	**80**	**58**	**76**	
	F-mesure	**57**	**46**	**84**	
K = 100	Précision	**44**	**30**	**97**	**74**
	Rappel	**82**	**73**	**73**	
	F-mesure	**57**	**42**	**84**	

Table 3.9 : les résultats de modèles KNN+SVM (US. Airlines Sentiment)

La figure 3.2 montre une comparaison entre les différents modèles pour les deux ensembles de données en termes d'accuracy.

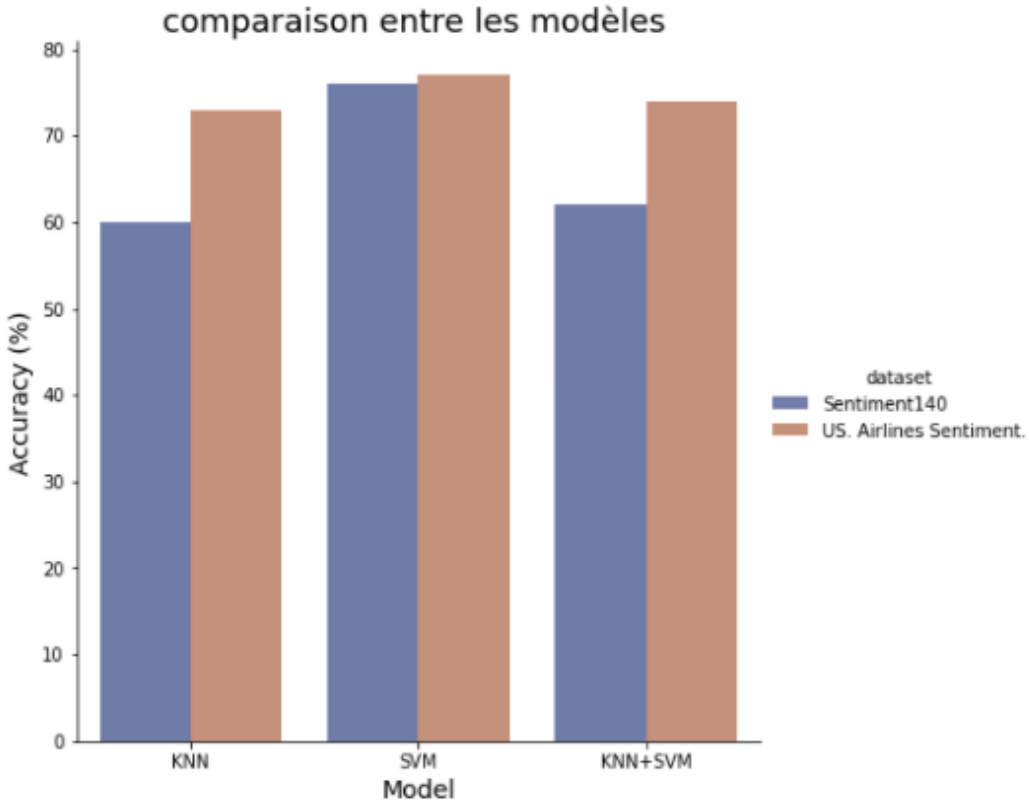

Figure 3-2 Comparaison des modèles pour une meilleure précision.

La figure 3.3 montre une comparaison entre les différents modèles pour les deux ensembles de données en termes de temps d'exécution.

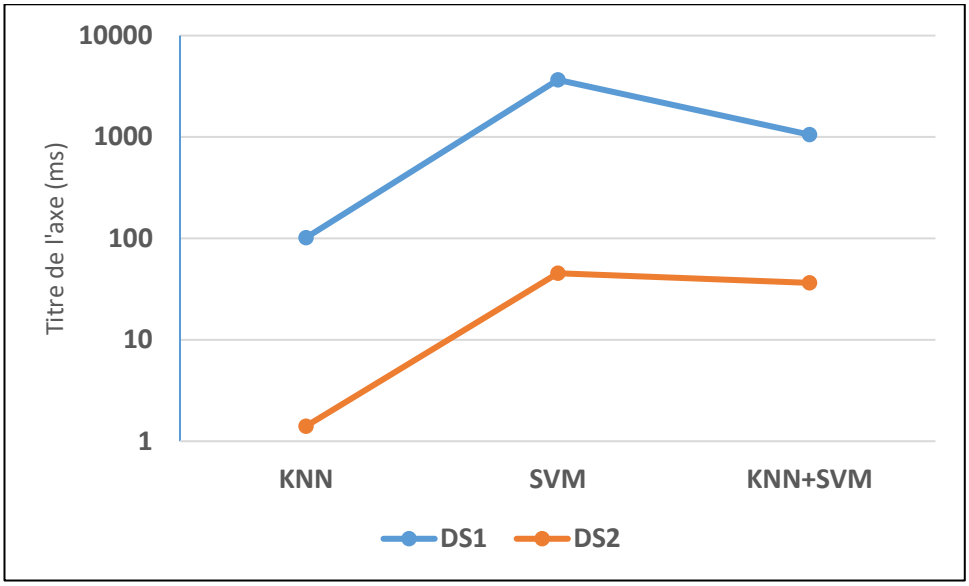

Figure 3-6 comparaisons entre les modèles en termes de temps d'exécution

En termes de classification, les premiers résultats obtenus montrent que notre modèle hybride basée sur la combinaison des algorithmes SVM et KNN (KNN+SVM) améliore encore les performances de classification par rapport aux cas où le modèle KNN qui est utilisé individuellement. La méthode proposée consomme moins d'espace mémoire et prend moins de temps de calcul par rapport au modèle SVM qui a de bonnes performances mais qu'elle souffre d'une forte consommation de temps d'exécution et d'une utilisation importante de la mémoire physique, en raison de ses processus d'apprentissage et de classification alambiqués, surtout lorsque la dimensionnalité des données est élevée.

3-3 Conclusion

Dans ce chapitre, nous avons vu notre part de contribution au problème de l'analyse du sentiment, représentant les outils et les jeux de données utilisés, ainsi que les étapes que nous avons suivies pour obtenir les résultats que nous montrons également pour différents modèles dont le but de faire la comparaison pour deux différents benchmarks.

Conclusion générale :

L'analyse des sentiments se réfère à l'extraction automatique de texte évaluative, qui aide à produire des résultats prédictifs. Dans ce mémoire nous avons étudié les différents approches d'analyse des sentiments en particulier celles appliquées sur les données Twitter.

Nous avons implémenté notre modèle de classification où nous avons utilisons une méthode hybride basée sur la combinaison des algorithmes SVM et KNN. Notre objectif est d'optimiser l'algorithme SVM et de simplifier le processus d'apprentissage, en intégrant l'algorithme KNN dans le SVM. Dans la littérature, nous avons constaté que l'approche de classification SVM a de bonnes performances mais qu'elle souffre d'une forte consommation de temps d'exécution et d'une utilisation importante de la mémoire physique, en raison de ses processus d'apprentissage et de classification alambiqués, surtout lorsque la dimensionnalité des données est élevée, en ajoutant que lorsque le nombre de données pour l'apprentissage est moindre par rapport aux données de test. En outre, l'approche de classification KNN est remarquable par sa simplicité et son processus de formation à faible coût. L'idée principale du classificateur hybride basé sur la combinaison de KNN et de SVM (KNN-SVM) est de trouver les voisins les plus proches d'un échantillon de requête par l'algorithme KNN, et de former un SVM local qui préserve la fonction de distance sur l'ensemble des voisins.

Ceci étant dit, il faut noter que l'environnement matériel utilisé est relativement limité. Par conséquent nous n'avons pas pu conduire nous expérimentations sur des corpus de tailles Importantes. En plus, cette limite nous a privés d'utiliser des méthodes de validation plus sophistiquées de l'approche implémentée.

Les premières études expérimentales basées sur deux ensembles de données disponibles publiquement prouvent l'efficacité de la méthode proposée en ce qui concerne la précision et ont une complexité de calcul raisonnable à la fois en apprentissage et en temps d'exécution.

Reference web ET bibliographies:

[01] Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites, Matthew A. Russell, "O'Reilly Media, Inc.", Jan 14, 2011.

[2] An Introduction to Sentiment Analysis. Ashish Karyekar AVP, Big Data Analytics.

[3] Sentiment Analysis and Opinion Mining Bing Liu

[4] Sentiment Analysis by Professor Dan Jurafsky

[5] Méthodes de calcul est le terme introduit par Donald Knuth pour séparer les algorithmes validés mathématiquement et les méthodes empiriques fréquemment utilisées dans la pratique

[6] M. Neff, B. Blaser, J.M. Lange, H. Lehmann, get it where you can: Acquiring and maintaining bilingual lexicons for machine translation, in: Proceedings of AAAI Spring Symposium on Building Lexicons for Machine Translation, Stanford, California, 1993

[7] Metaphor and Corpus Linguistics by Alice Dignam

[8] Corpus Linguistics: Method, Theory and Practice by Tony McEnery, Andrew Hardie.

[9] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis." Foundations and Trends in Information Retrieval 2(1-2), pp. 1–135, 2008

[10] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin, "Learning Subjective Language," Computational Linguistics, vol. 30, pp. 277–308, September 2004

[11] B. Liu. Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010.