République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université A. Mira de Béjaïa
Faculté des Sciences Exactes
Département d'Informatique

MEMOIRE DE MASTER RECHERCHE
En Informatique

Option : Systèmes d'Informations Avancés

Thème

Analyse des sentiments sur l'éducation en Algérie sur You-Tube : approche hybride

Réalisé par:

LAIDOUI Imad

DJEMADI Rafik

Soutenu le 03 Juillet devant le jury composé de :

Présidente :	*M^me* MECHIOURI Sarah	M.C.B Université de Béjaïa.
Examinatrice :	*M^me* HOUHA Amel	M.A.A Université de Béjaïa.
Encadrante :	*M^me* BOUADEM Nassima	M.C.B Université de Béjaïa.

Promotion 2023 - 2024.

Remerciements

Tout d'abord, nous exprimons notre gratitude envers Dieu tout-puissant, qui nous a donné la force de survivre et le courage de surmonter toutes les difficultés.

Nous remercions Madame Nassima Bouadem pour son encadrement et son aide.

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

Enfin, nous souhaitons exprimer notre gratitude à chacun des membres du jury pour l'intérêt qu'ils ont porté à ce travail et pour avoir accepté de l'évaluer.

Dédicaces

Je remercie chaleureusement mon père, dont la sagesse, les conseils et le dévouement ont été des piliers essentiels dans ma vie. Sa présence et son appui m'ont donné la force de surmonter les obstacles.

Je remercie également ma mère, sans qui rien de tout cela n'aurait été possible. Son soutien inconditionnel, ses sacrifices et son amour m'ont permis de persévérer et de réussir. Merci de m'avoir toujours encouragé et d'avoir cru en moi.

Je suis profondément reconnaissant envers mon frère, qui a toujours été un modèle de détermination et de persévérance. Son soutien et sa camaraderie m'ont inspiré à donner le meilleur de moi-même.

Enfin, je remercie tous mes amies qui ont toujours été là pour moi. Leur soutien inconditionnel et leurs encouragements ont été d’une grande aide.

À la mémoire de mes grands-pères et mon cousin, que Dieu les accueille dans son vaste paradis.

À tous ces intervenants, je présente mes remerciements, mon respect et ma gratitude

Imad

Dédicaces

Je remercie mon père Abdelhamid et ma mère et ma sœur Chahrazed, pour leurs encouragements.

Je remercie mes amis les plus proches Amine Ladrani, Youba Bouchala, Youba Boumenir , Ouazene Amazigh, Samir Ab Kehoul Oussama, Koussaila Kerrache , et à la fin Lyes sd.

De plus je remercie mon binôme Imad et ma encadrant, ça m'a fait plaisir de travailler avec eux.

Rafik

Table des matières

Table des figures

Liste des tableaux

Liste des équations

Liste des abréviations

ABSA - (**A**spect-**B**ased **S**entiment **A**nalysis).
API - **A**pplication **P**rogramming **I**nterface.
APL - **A** **P**rogramming **L**anguage.
AraBERT - **U**n modèle **B**ERT pré-entraîné sur le texte arabe.
AUC - **A**rea **U**nder the **C**urve.
AVG – **A**verage.
Bag-of-Words - **U**n modèle utilisé dans le **T**raitement du **L**angage **N**aturel.
BERT - **B**idirectional **E**ncoder **R**epresentations from **T**ransformers.
BJP - **B**haratiya **J**anata **P**arty (Parti politique indien).
CHI - **C**hi-squared **T**est.
COVID - **C**oronavirus **D**isease.
CRM - **C**ustomer **R**elationship **M**anagement.
CSV - **C**omma-**S**eparated **V**alues.
DBSCAN - **D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise.
EIM - **E**ffets **I**ndésirables des **M**édicaments (Adverse Drug Reactions).
E-Learning - **E**lectronic **L**earning.
EdX - **O**nline **L**earning **P**latform.
ESI - **É**cole **N**ationale d'**I**nformatique.
FN - **F**alse **N**egatives.
FP - **F**alse **P**ositives.
FPR - **F**alse **P**ositive **R**ate.
GMT - **G**reenwich **M**ean **T**ime.
GloVe - **G**lobal **V**ectors for **W**ord **R**epresentation.
GPU - **G**raphics **P**rocessing **U**nit.
GridSearchCV - **G**rid **S**earch with **C**ross-**V**alidation.
GRC - **G**estion de la **R**elation **C**lient (Customer Relationship Management).
HH – **H**ours.
HMM - **H**idden **M**arkov **M**odel.
HTML - **H**yper**T**ext **M**arkup **L**anguage.
IA - **I**ntelligence **A**rtificielle (Artificial Intelligence).
IE - **I**nformations **É**lectroniques (Electronic Information).
IG - **I**nformation **G**ain.
IPython - **I**nteractive **P**ython.
ItalWordNet - **U**ne base de données lexicale italienne similaire à **W**ord**N**et.
KNN - **K**-**N**earest **N**eighbors.
LDA - **L**atent **D**irichlet **A**llocation.
LSTM - **L**ong **S**hort-**T**erm **M**emory.
MI - **I**nformation **M**utuelle.
ML - **M**achine **L**earning.
MM – **M**inutes.

MOOC - **M**assive **O**pen **O**nline **C**ourse.
MongoDB - **N**o**SQL** database.
MySQL - **R**elational **D**atabase **M**anagement **S**ystem.
NB - **N**aive **B**ayes.
NFL - **N**ational **F**ootball **L**eague.
NLTK - **N**atural **L**anguage **T**oolkit.
NLP - **N**atural **L**anguage **P**rocessing.
NN - **N**eural **N**etwork.
NoSQL - **N**on-**R**elational **D**atabase.
PDF - **P**ortable **D**ocument **F**ormat.
PNL - **P**rogrammation **N**euro-**L**inguistique (Neuro-Linguistic Programming).
POS - **P**art-of-**S**peech.
POStagging - **P**art-of-**S**peech **T**agging.
PostgreSQL - **O**pen-**S**ource **R**elational **D**atabase.
RBF - **R**adial **B**asis **F**unction.
RE - **R**egular **E**xpression.
RNN - **R**ecurrent **N**eural **N**etwork.
ROC - **R**eceiver **O**perating **C**haracteristic.
SA - **S**entiment **A**nalysis.
SciPy - **S**cientific **P**ython.
SGBDR - **S**ystème de **G**estion de **B**ase de **D**onnées **R**elationnelle (Relational Database Ma.nagement System).
SML - **S**tandard **M**L.
SVM - **S**upport **V**ector **M**achine.
SVMlight - **I**mplementation of **S**upport **V**ector **M**achines.
TAL - **T**raitement **A**utomatique du **L**angage.
TALN - **T**raitement **A**utomatique du **L**angage **N**aturel (Natural Language Processing).
TF-IDF - **T**erm **F**requency-**I**nverse **D**ocument **F**requency.
TN - **T**rue **N**egatives.
TP - **T**rue **P**ositives.
TPR - **T**rue **P**ositive **R**ate.
TSV - **T**ab-**S**eparated **V**alues.
UCD - **U**nicode **C**haracter **D**atabase.
URL - **U**niform **R**esource **L**ocator.
UTC - **C**oordinated **U**niversal **T**ime.
Word2Vec - **W**ord to **V**ector.
WordNet - **U**ne base de données lexicale pour la langue anglaise.
XML - e**X**tensible **M**arkup **L**anguage.
YYYY-MM-DD - **Y**ear-**M**onth-**D**ay.

Résumé

YouTube est devenu essentiel pour l'éducation en ligne grâce à son vaste contenu éducatif. L'analyse des commentaires, vues et likes aide à comprendre la qualité du contenu et les opinions du public. L'éducation numérique, facilitée par Internet, permet un apprentissage dynamique. L'analyse des sentiments, grâce aux avancées technologiques, permet de suivre et analyser en temps réel les opinions sur YouTube. Les défis incluent la catégorisation des commentaires, la distinction entre avis authentiques et faux, et la prise en compte du langage et du sarcasme.

Les techniques de Machine Learning, comme Random Forest, SVM et Naïve Bayes, sont utilisées pour analyser les sentiments dans l'éducation en ligne. Ce mémoire se concentre sur le dialecte algérien, un défi en raison de sa complexité linguistique. Nous proposons une méthode pour analyser les commentaires en dialecte algérien sur YouTube, incluant la collecte de données, l'analyse avec les algorithmes mentionnés, et l'obtention de résultats précis sur la polarité et la fiabilité des avis.

Mots clés : Algorithmes, Analyse des sentiments, Commentaires, Dataset, Dialecte algérien, KNN, Machine Learning, NB, NLP, Opinion Mining, Polarité, Random Forest SVM, YouTube.

Abstract

YouTube has become essential for online education due to its vast educational content. Analyzing comments, views, and likes helps understand content quality and public opinions. Digital education, facilitated by the Internet, allows for dynamic learning. Sentiment analysis, enabled by technological advancements, allows for real-time tracking and analysis of opinions on YouTube. Challenges include categorizing comments, distinguishing between authentic and fake reviews, and considering language and sarcasm.

Machine Learning techniques like Random Forest, SVM, and Naïve Bayes are used to analyze sentiments in online education. This thesis focuses on the Algerian dialect, a challenge due to its linguistic complexity. We propose a method to analyze comments in the Algerian dialect on YouTube, including data collection, analysis with the mentioned algorithms, and obtaining accurate results on the polarity and reliability of reviews.

Keywords : Algorithms, Comment Analysis, Dataset, Dialect Algerian, KNN, Machine Learning, Naïve Bayes (NB), NLP, Opinion Mining, Random Forest, Sentiment Analysis, Support Vector Machine (SVM),Polarity, YouTube.

1 Chapitre 1

Introduction

Sommaire

1.1 Introduction

Depuis l'aube de l'humanité, l'homme s'efforce constamment d'améliorer sa qualité de vie. Cette quête incessante a engendré de nombreuses révolutions qui ont transformé le monde. Parmi celles-ci, la révolution de la technologie de l'information occupe une place prépondérante.

Avec l'avènement du Web 2.0, un nombre croissant de personnes expriment leurs opinions sur internet sur une multitude de sujets. De nombreuses plateformes facilitent cette interaction, telles que les blogs personnels, les commentaires sur des articles, les vidéos sur YouTube, et les critiques de produits sur divers sites de commerce en ligne. Ces plateformes génèrent ainsi d'énormes quantités de données potentiellement exploitables. En raison de leur volume, des techniques de traitement automatique sont nécessaires pour convertir ces données en connaissances utiles.

Avec toutes les autres opportunités qu'Internet offre aux gens, il est désormais devenu une mine d'or pour ceux qui veulent apprendre. L'une des principales raisons pour lesquelles les gens visitent les plateformes en ligne est la recherche d'informations [1]. L'enseignement et l'apprentissage ont largement évolué vers le paradigme en ligne. Les gens ne sont plus obligés d'aller à un endroit particulier ou d'étudier à une heure précise [2].

La littérature démontre que les vidéos éducatives en ligne sont utilisées comme un outil d'apprentissage efficace. Avec plus de 2 milliards d'utilisateurs actifs, YouTube est le site d'hébergement de vidéos le plus populaire [3]. Il est couramment utilisé pour regarder du contenu éducatif [4]. Une revue de la littérature existante a révélé que les vidéos éducatives sur YouTube aident les étudiants à améliorer leur apprentissage [5]. Des centaines d'universités et de collèges ont créé leurs propres chaînes YouTube éducatives. YouTube contient de nombreuses informations sur une grande variété de sujets qui sont accessibles au public sans aucun frais [6]. Les utilisateurs de YouTube peuvent exprimer leurs opinions par des likes, des

dislikes et des commentaires. Ces commentaires permettent d'analyser les avis des spectateurs sur le contenu.

Ainsi, les émotions exprimées dans les commentaires pourraient être utilisées comme indicateurs [7]. Le traitement du langage naturel est un ensemble de techniques utilisées pour analyser et représenter les formes textuelles des langues que les humains utilisent naturellement à des fins de communication.

Ces techniques entraînent les machines à traiter le langage comme les humains afin d'effectuer diverses tâches plus efficacement [8]. L'analyse des sentiments est le processus d'évaluation d'un texte écrit et d'en extraire des opinions ou d'autres informations [2]. L'analyse des sentiments peut être appliquée aux avis des clients, aux commentaires, aux blogs, aux articles de presse, et plus encore. Elle aide les entreprises à développer des stratégies efficaces pour leur croissance, permet aux créateurs de contenu d'améliorer leur production, et offre aux chercheurs et analystes une meilleure compréhension des données. Cette analyse peut être réalisée à l'aide de méthodes d'apprentissage automatique ainsi que de techniques basées sur le lexique.

Une approche hybride peut également être choisie, utilisant à la fois des techniques basées sur le lexique et des techniques d'apprentissage automatique [9]. Les approches d'apprentissage automatique requièrent des données étiquetées pour entraîner les classificateurs, qui sont ensuite utilisés pour analyser de nouvelles données.

À l'inverse, les techniques lexicales utilisent une collection prédéterminée de mots, où chaque mot est connecté à une émotion particulière [10].

La satisfaction est un critère essentiel souvent utilisé pour mesurer le succès et la fiabilité d'un système. Elle indique si le système est plaisant et s'il répond aux attentes des consommateurs. La satisfaction des consommateurs est devenue une priorité élevée pour les prestataires de services car elle est un gage de succès à long terme [11]. L'attitude des étudiants envers leurs services éducatifs, leur expérience et leur institut définit leur satisfaction [12]. La satisfaction des étudiants est directement liée aux scores de sentiment de leurs commentaires. Des scores de sentiment élevés prédisent une satisfaction élevée [13]. Les étudiants ressentent une grande satisfaction en apprenant en ligne, car ce mode d'apprentissage diffère de l'enseignement traditionnel. Il est amusant d'apprendre sur les sites Web car ils sont plus attrayants et interactifs [14]. Dans cette étude, des techniques d'apprentissage automatique ont été employées pour analyser les sentiments des commentaires sous les vidéos éducatives sur YouTube, dans le but de mesurer le niveau de satisfaction des étudiants. Au cours de la littérature, il a été constaté que divers chercheurs ont mesuré le niveau de satisfaction des consommateurs grâce à l'analyse des sentiments [15] [16] [17].

Pour ce travail, les sentiments positifs et négatifs exprimés par les téléspectateurs dans les commentaires seront comparés. Les résultats serviront à conclure si les étudiants sont satisfaits du contenu pédagogique disponible sur YouTube.

Le domaine de recherche présenté dans ce mémoire est l'analyse des sentiments. Le but est d'analyser automatiquement les sentiments des commentaires en dialecte algérien sur les vidéos éducatives de YouTube.

1.2 Problématique

Avec la popularité croissante des plateformes éducatives en ligne comme YouTube, les commentaires des utilisateurs deviennent essentiels pour évaluer la qualité et la pertinence des vidéos. En Algérie, où YouTube est une source importante de ressources éducatives, la diversité et l'ambiguïté des commentaires rendent difficile pour les spectateurs de faire des choix éclairés. Analyser les sentiments exprimés dans les commentaires, notamment en dialecte algérien, est crucial pour mesurer la perception des vidéos éducatives et aider les créateurs à améliorer leur contenu. Une telle analyse permettra de mieux comprendre les attentes des utilisateurs et d'améliorer la qualité des ressources éducatives disponibles.

1.3 Objectifs

L'objectif de ce mémoire est de développer un système d'analyse des sentiments pour classer et détecter automatiquement les sentiments des étudiants et leurs avis positifs, négatifs ou neutres exprimé dans les commentaires en dialecte algérien sur les vidéos éducatives sur YouTube, et d'évaluer les différentes techniques et approches proposées pour l'analyse des sentiments et leurs résultats. L'objectif est également d'évaluer l'impact général de l'éducation sur YouTube en Algérie.

1.4 Organisation du mémoire

Nous avons structuré notre travail pour atteindre notre objectif comme suit :

Introduction générale :

Nous débutons notre rapport par une introduction générale, suivie d'une présentation de la problématique qui permettra de bien encadrer notre étude. Nous préciserons également les objectifs que nous avons fixés dès le début de notre projet de fin d'études afin de clarifier notre but final.

Le reste du mémoire est structuré comme suit :

Chapitre 02 : Généralités sur l'analyse de sentiments :

Le deuxième chapitre est consacré aux généralités sur l'analyse des sentiments. On présentera l'importance de l'analyse des sentiments, quelques domaines d'application, les différents niveaux et types d'opinions, ainsi que la catégorisation des sentiments. Nous aborderons également les disciplines en relation avec l'analyse des sentiments, les défis de cette analyse, son application dans le domaine de l'éducation et le processus complet d'analyse des sentiments.

Chapitre 03 : Etat de l'art :

Dans le troisième chapitre, nous élaborerons l'état de l'art en représentant tous les travaux connexes que nous synthétiserons. Nous présenterons ces travaux dans un tableau contenant les grandes lignes de chaque document synthétisé, suivi de brefs paragraphes résumant chaque travail. Ensuite, nous procéderons à une analyse comparative entre les approches des documents connexes et notre approche.

Chapitre 04 : Analyse des sentiments sur les videos d'éducation :

Le quatrième chapitre porte sur l'analyse des sentiments sur les vidéos d'éducation. Il présente notre approche proposée, incluant la collecte des données, le prétraitement, l'annotation,

l'analyse du vocabulaire, la création d'un lexique des émotions, la visualisation des commentaires et leur classification en utilisant diverses méthodes telles que les forêts aléatoires, les machines à vecteurs de support et les modèles de Naive Bayes.

Chapitre 05 : Expérimentation :

Dans le cinquième chapitre, nous présenterons l'expérimentation de notre approche. Ce chapitre inclut la description du dataset, l'environnement de travail, les langages et bibliothèques de programmation utilisés, ainsi que la mise en service et l'évaluation des modèles avec divers métriques de performance telles que la précision, le rappel, le F1-score, l'accuracy, les moyennes macro et pondérée, ainsi que les courbes ROC (**R**eceiver **O**perating **C**haracteristic) et AUC (**A**rea **U**nder the **C**urve). Une comparaison des modèles sera également effectuée.

Chapitre 06 : Conclusion générale :

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

2 Chapitre 2

Généralités sur l'analyse des sentiments

Sommaire

2.1 Introduction

Pour presque toutes les activités humaines, les opinions constituent le noyau qui influence grandement notre comportement. En outre, l'évaluation du monde par les autres et la façon dont ils le voient jouent un rôle en influençant nos croyances et nos perceptions de la réalité, et donc les choix que nous faisons. Ceci explique le phénomène de recherche de l'avis des autres lorsque l'on veut prendre une décision. Ce phénomène s'applique non seulement aux individus mais aussi aux organisations.

De nos jours, le processus de publication et de partage d'expériences et d'opinions via Internet et les réseaux sociaux, et donc d'expression de sentiments, est devenu quelque chose que pratiquent la plupart des gens dans le monde. Habituellement, une énorme quantité de données résultera de cette action sur Internet. Ce qui rend ces données plus utiles la plupart du temps, c'est lorsqu'elles sont analysées, que ce soit par des chercheurs ou même par des entreprises, des gouvernements et des organisations. Par exemple, la plupart des entreprises industrielles et des campagnes électorales s'appuient sur les sites de réseaux sociaux pour savoir si les opinions des gens sont positives, négatives ou neutres. Gérer les réponses et feedback, par exemple sur les sites commerciaux pour connaître leur acceptation ou leur refus

d'un produit ; cela contribuera à améliorer les ventes de l'entreprise car il indique le choix d'un client, et c'est ce qu'est SA (**S**entiment **A**nalysis), une étude textuelle largement utilisée sur les avis et les enquêtes sur Internet et les réseaux sociaux. Ainsi, les opinions et leurs concepts associés tels que 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.

Les racines de l'analyse des sentiments se trouvent dans les études sur l'analyse de l'opinion publique au début du 20e siècle et dans l'analyse de la subjectivité des textes réalisée par la communauté de la linguistique informatique dans les années 1990. Cependant, l'essor de l'analyse informatique des sentiments n'a eu lieu qu'avec la disponibilité de textes subjectifs sur le Web [18].

L'analyse des sentiments est le processus de détermination du ton émotionnel derrière une série de mots. Cette analyse est utilisée pour mieux comprendre la perception, les opinions et les émotions exprimées dans une déclaration en ligne [19].

L'analyse des sentiments, également appelée exploration d'opinions, permet de classer les idées et opinions en positives, négatives ou neutres. Les deux termes désignent essentiellement le même domaine d'études. L'analyse des sentiments et l'exploration d'opinions sont des concepts étudiés depuis le début des années 2000, avec diverses recherches pionnières dans le domaine. Quoi qu'il en soit, l'analyse des sentiments et L'exploration d'opinions se concentre principalement sur les opinions qui expriment ou impliquent des sentiments positifs ou négatifs [20].

De nouvelles idées sont générées par les régimes, les politiciens, les psychologues, les industriels et les chercheurs pour analyser les différentes opinions qui ont explosé sur les réseaux sociaux et ainsi mettre en œuvre les meilleures décisions.

Dans ce deuxième chapitre, nous allons explorer les fondements théoriques et les concepts essentiels de l'analyse des sentiments. Nous commencerons par définir l'analyse des sentiments et son importance croissante dans divers domaines d'application. Ensuite, nous examinerons les différents niveaux et types d'opinions, ainsi que la catégorisation des sentiments. Ce chapitre abordera également les disciplines connexes et les défis spécifiques liés à l'analyse des sentiments.

Enfin, nous détaillerons le processus complet de l'analyse des sentiments, de la collecte des données à leur prétraitement, en passant par l'annotation, la représentation et la classification des sentiments. Nous comparerons les approches basées sur le lexique, l'apprentissage automatique et les approches hybrides, en mettant en lumière leurs avantages et leurs limites respectifs.

2.2 Analyse de sentiments

L'analyse des sentiments, également appelée analyse d'opinion, est le domaine d'étude qui analyse les opinions, les sentiments, les évaluations, les attitudes et les émotions des personnes envers des entités telles que des produits, des services, des organisations, des individus, des problèmes, des événements, des sujets et leurs attributs [20].

Il existe également de nombreux noms et tâches légèrement différentes, par exemple l'analyse des sentiments, l'exploration d'opinions, l'extraction d'opinions, l'exploration de sentiments, l'analyse de subjectivité, l'analyse d'affect, l'analyse d'émotions, l'exploration d'avis, etc. Cependant, elles sont désormais toutes regroupées sous l'égide de l'analyse des sentiments. Ou l'exploration d'opinions. Dans l'industrie, le terme analyse des sentiments est plus couramment utilisé, mais dans le monde universitaire, l'analyse des sentiments et l'exploration d'opinions sont fréquemment utilisées. Ils représentent essentiellement le même domaine d'études [20].

2.2.1 L'importance d'analyse de sentiments [20]

Depuis le début des années 2000, l'analyse des sentiments est devenue l'un des domaines de recherche les plus actifs dans le domaine du traitement du langage naturel.

Il est également largement étudié dans l'exploration de données, l'exploration de sites Web et l'exploration de textes.

En fait, elle s'est étendue de l'informatique aux sciences de gestion et aux sciences sociales en raison de son importance pour l'entreprise et la société dans son ensemble.

Ces dernières années, les activités industrielles autour de l'analyse du sentiment 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.

Donc, les différents domaines d'application d'analyse des sentiments lui confèrent une grande importance.

2.2.2 Quelque Domaines d'application d'analyse de sentiments

Étant donné la diversité des domaines d'application de l'analyse de sentiment, nous allons nous concentrer sur quelques exemples représentatifs pour illustrer son potentiel et son impact. Les six domaines d'application suivants sont illustrés dans la figure ci-jointe :

Figure 2.1 – Les domaines d'application d'analyse de sentiments

1. **Marketing et Publicité:** les entreprises peuvent désormais mieux ajuster leurs campagnes publicitaires, leurs messages et leurs stratégies de marque, améliorant ainsi les attentes des consommateurs, grâce aux précieuses sources d'avis clients, de forums, de blogs et d'avis en ligne qui ont été collectés et analysés, notamment avec la disponibilité d'un accès à ces informations dès qu'elles sont trouvées. Après avoir été publiées en ligne par les consommateurs eux-mêmes du monde entier, permettant ainsi des analyses à grande échelle, c'est après que la collecte des avis des consommateurs ait longtemps été ardue.

2. **Revue des produits(product review mining) :** l'analyse des sentiments permet de classer les avis en fonction des sentiments exprimés, qu'ils soient positifs, négatifs ou neutres, ce qui facilite la prise de décision lors de l'achat d'un produit. Ainsi, non seulement évaluez rapidement les opinions des autres utilisateurs sur un produit particulier, mais créez également des résumés d'avis, détectez le spam et les faux avis publiés par des agences ou des particuliers dans le but de manipuler les avis des consommateurs. Ces informations sont précieuses pour les consommateurs dans le processus d'achat. Ainsi que pour les entreprises qui souhaitent améliorer leurs produits et services. Par exemple, eBay utilise des outils d'analyse des sentiments pour mettre en évidence les meilleurs avis et les rendre plus visibles pour les utilisateurs, aidant ainsi les acheteurs à prendre des décisions éclairées.
3. **E-commerce et du CRM (Customer Relationship Management) ou GRC (Gestion de la Relation Client) en français :** les entreprises de e-commerce acquièrent une connaissance approfondie de leurs clients, ce qui les aide à personnaliser leurs offres pour mieux

les satisfaire, grâce à l'analyse des avis consommateurs, qui permet de fidéliser les clients après avoir recueilli leurs critiques et avis puis répondu à leurs commentaires. Cela améliore la relation client-fournisseur et augmente également le contrôle de la qualité des produits.

4. **Politique :** afin que les politologues puissent déterminer comment le public reçoit les publicités politiques, ils s'appuient sur l'analyse des sentiments. En 2012, l'administration Obama a appliqué l'analyse des sentiments pour évaluer les annonces politiques. De plus, cette forme d'analyse peut être utilisée pour étudier le nombre de mentions négatives concernant les candidats dans diverses sources d'information et médiatiques.

5. **Education :** les éducateurs et les établissements évaluent la satisfaction globale, l'engagement et l'expérience d'apprentissage des étudiants grâce au processus d'analyse des sentiments. Ce processus dans le domaine de l'éducation comprend l'examen des attitudes, des opinions et des sentiments exprimés par les étudiants dans différents contextes éducatifs.

6. **Santé :** aujourd'hui, les plateformes en ligne telles que les réseaux sociaux et les sites Web sont également utilisées pour exprimer des opinions sur des questions de santé par des médecins et même des patients. Les organismes de santé utilisent l'analyse des sentiments dans leur domaine pour pouvoir analyser les avis, les commentaires et les enquêtes des patients afin de mieux comprendre les émotions des patients. Elle peut également être appliquée pour détecter les signes avant-coureurs d'épidémies, suivre les préoccupations du public concernant les vaccins et anticiper les émergences. Les tendances en matière de santé peuvent également aider à identifier les effets indésirables des médicaments (EIM :) en analysant les expériences et les sentiments rapportés par les patients concernant des médicaments spécifiques.

Outre les applications réelles, de nombreux articles de recherche axés sur les applications ont également été publiés. Par exemple, dans (Liu et al., 2007), un modèle de sentiment a été proposé pour prédire les performances commerciales. Dans (McGlohon, Glance et Reiter, 2010), les avis ont été utilisés pour classer les produits et les commerçants. Dans (Hong et Skiena, 2010), les relations entre les lignes de paris de la NFL (**N**ational **F**ootball **L**eague) et les opinions publiques dans les blogs et sur Twitter ont été étudiées [21].

Il existe plusieurs autres articles et domaines, quel que soit le domaine d'application, le processus d'analyse des sentiments reste le même.

2.2.3 Différents niveaux d'analyse de sentiments

L'analyse peut être effectuée à différents niveaux qui sont illustrés dans la figure ci-jointe :

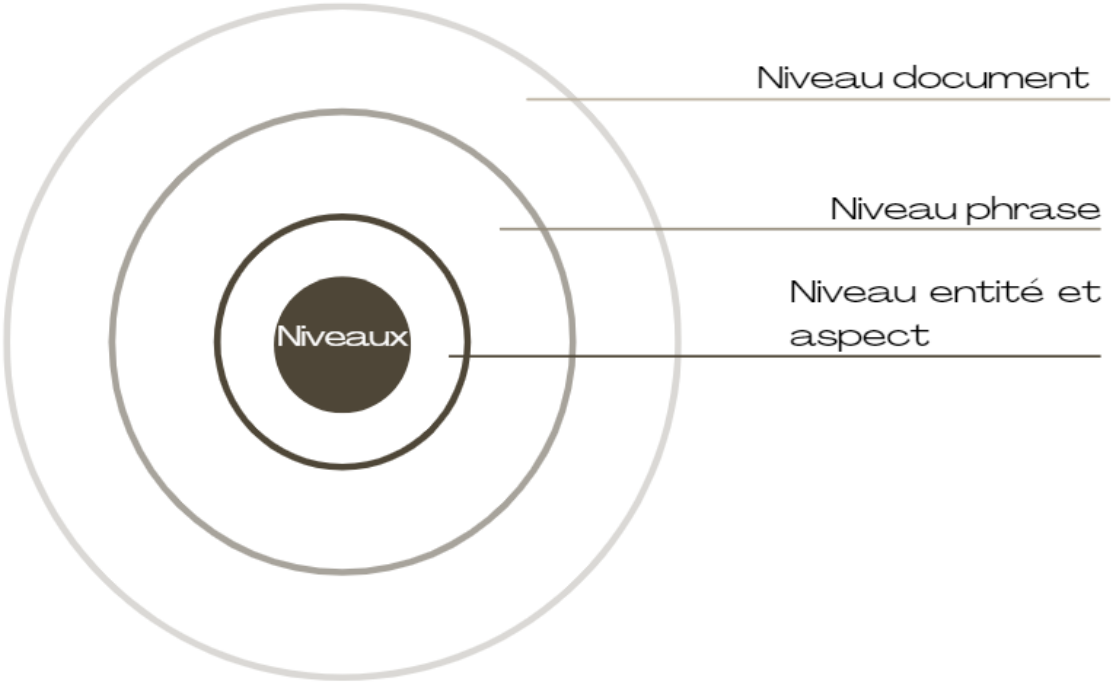

Figure 2.2 – Les niveaux d'analyse de sentiments

A. Niveau du document :

La tâche à ce niveau consiste à évaluer si un document d'opinion global exprime un sentiment positif ou négatif.

Par exemple, considérons un scénario dans lequel une université propose un cours en ligne de programmation informatique. À la fin du semestre, les étudiants sont invités à donner leur avis via une enquête ou un forum en ligne sur leur expérience globale du cours.

Un système d'analyse des sentiments au niveau du document pourrait analyser les réponses des étudiants pour déterminer si le sentiment général à l'égard du cours est positif ou négatif.

Ce niveau d'analyse suppose que chaque document exprime des opinions sur une seule entité (par exemple, cours de programmation informatique). Ainsi, il ne s'applique pas aux documents qui évaluent ou comparent plusieurs entités.

B. Niveau phrase :

La tâche à ce niveau concerne les phrases et détermine si chaque phrase contenue dans un texte exprime une opinion positive, négative ou neutre. L'hypothèse est que chaque phrase, dans un texte donné, désigne une seule opinion sur une seule entité.

C. Niveau Entité et Aspect :

Une analyse plus fine consiste à évaluer les sentiments au niveau des aspects spécifiques du document. Autrefois appelé niveau de fonctionnalité, ce type d'analyse explore et

synthétise les opinions basées sur différentes fonctionnalités. Au lieu d'examiner les constructions linguistiques (documents, paragraphes, phrases, clauses ou expressions), le niveau aspect s'intéresse directement à l'opinion elle-même. Il est basé sur l'idée qu'une opinion consiste en un sentiment (positif ou négatif) et une cible (d'opinion). Un avis sans que sa cible soit identifiée est d'une utilité limitée. Prendre conscience de l'importance des cibles d'opinion nous aide également à mieux comprendre le problème de l'analyse des sentiments.

Par exemple, considérons les commentaires fournis par un étudiant concernant un cours en ligne :

"Le matériel de cours est complet, mais les forums de discussion ne sont pas très engageants."

Dans cet exemple, l'avis de l'étudiant évalue deux aspects du cours en ligne : le support de cours et les forums de discussion. Le sentiment à l'égard du matériel de cours est positif (complet), tandis que le sentiment à l'égard des forums de discussion est négatif (pas très engageant). Le matériel de cours et les forums de discussion servent de cibles d'opinion.

2.2.4 Différents types d'opinions

I. Opinion réguliers et comparatifs

A. Opinion régulière : Une opinion régulière est souvent simplement appelée une opinion dans la littérature et elle comporte deux sous-types principaux :

- **Opinion directe :** fait référence à une opinion exprimée directement sur une entité ou un aspect de l'entité, par exemple : « Le contenu du cours est engageant» [20].
- **Opinion indirecte :** est exprimée indirectement sur une entité ou un aspect d'une entité en fonction de ses effets sur certaines autres entités [20]. Par exemple, l'expression « Après avoir assisté à la conférence, je me suis senti plus confus à propos du sujet » décrit un effet indirect de la conférence sur la compréhension de l'orateur, exprimant ainsi indirectement une opinion ou un sentiment négatif à propos de la conférence. Dans ce cas, l'entité est le cours magistral et l'aspect est son effet sur la compréhension du locuteur.

Une grande partie des recherches actuelles se concentrent sur les opinions directes. Ils sont plus simples à manipuler. Les opinions indirectes sont souvent plus difficiles à gérer.

B. Opinion comparatif : exprime une relation de similitudes ou de différences entre deux ou plusieurs entités et/ou une préférence de l'auteur de l'opinion basée sur certains aspects partagés des entités. Par exemple, dans un contexte éducatif :"Je trouve les cours en ligne plus intéressants que les cours traditionnels en classe.""Etudier en groupe est pour moi plus efficace que d'étudier seul."Dans ces exemples, les opinions comparatives expriment des préférences ou des comparaisons entre différentes méthodes ou approches pédagogiques.

II. Opinions explicites et implicites

A. Opinion explicite : une déclaration subjective qui donne une opinion régulière ou comparative [22]. Par exemple, « "Le support de cours est très instructif." » Et «" Je pense que les projets de groupe sont plus bénéfiques que les missions individuelles" ».

B. Opinion implicite : est une déclaration objective qui implique un commun régulier ou comparatif. Une telle déclaration objective exprime généralement un fait souhaitable ou indésirable [22].Par exemple :"Les cours en ligne sont souvent interrompus par des problèmes techniques." (Implique un avis négatif sur la fiabilité de la plateforme de cours en ligne)

"Le manuel contient de nombreuses erreurs qui doivent être corrigées." (Implique une opinion négative sur l'exactitude du manuel).

Les opinions explicites sont plus faciles à détecter et à classer que les opinions implicites. Une grande partie des recherches actuelles se sont concentrées sur des opinions explicites.

2.2.5 Catégorisation des sentiments

I. Une phrase objective

Présente des informations factuelles sur le monde [22].

Un exemple de phrase objective est "L'examen final aura lieu vendredi", autre exemple : "Le manuel a été écrit par le professeur Smith."

II. Une phrase subjective

Exprime des sentiments, des opinions ou des croyances personnels, un exemple de phrase subjective est : "Je trouve le style d'enseignement du professeur très engageant.", autre exemple : "Selon moi, le projet de groupe a été la partie la plus enrichissante du cours."

Les expressions subjectives peuvent se manifester sous diverses formes, telles que les opinions, les allégations, les désirs, les croyances, les soupçons et les spéculations.

Ici, nous devons noter ce qui suit :

Une phrase subjective ne peut exprimer aucun sentiment, par exemple : « Je crois que l'examen final couvrira plusieurs chapitres », autre exemple : "À mon avis, le manuel offre une couverture complète du sujet."

Les phrases objectives peuvent impliquer des opinions ou des sentiments en raison de faits souhaitables ou indésirables. Par exemple, les deux phrases suivantes qui exposent certains faits impliquent clairement des sentiments négatifs (qui sont des opinions implicites)

"La plateforme de cours en ligne plante fréquemment lors des sessions en direct."

"Les lectures assignées pour ce cours sont obsolètes et sans rapport avec le sujet."

2.2.6 Disciplines en relation avec l'analyse des sentiments

L'analyse des sentiments est un domaine multidisciplinaire qui a plusieurs disciplines qui lui sont directement ou indirectement liées pour détecter, analyser et interpréter les sentiments exprimés dans les données textuelles :

A. Intelligence Artificielle (IA) :

Les fondements théoriques et les outils pratiques nécessaires au développement de modèles prédictifs sont fournis par l'intelligence artificielle dans le domaine de l'analyse des sentiments. Elle fournit également des systèmes d'analyse des sentiments basés sur l'apprentissage automatique et d'autres technologies capables d'analyser, de comprendre et de générer des données textuelles.

B. Text Mining et Data Mining :

Le Text Mining (ou fouille de texte) sont des disciplines qui englobes de nouvelles méthodes de recherche et des outils logiciels et se concentrent sur l'extraction d'informations de haute à partir de données textuelles dans le milieu universitaire ainsi que par des entreprises et des organismes gouvernementaux, Le Text Mining a connu une augmentation significative de la demande au cours des dernières années.

Les processus de fouille de textes comprennent généralement non seulement l'extraction d'informations (IE : Informations Électroniques) qui visent principalement à identifier les objets en extrayant les informations pertinentes des fragments, puis en plaçant toutes les pièces extraites dans un cadre mais aussi le traitement du langage naturel (NLP), Le NLP fait partie du domaine de l'intelligence artificielle et tente d'aider à transformer des messages imprécis et ambigus en messages sans ambiguïté et précis. Text Mining comprennent aussi la Classification de texte.

C. Traitement Automatique du Langage Naturel (TALN) :

Le TALN est une branche de l'IA qui permet aux ordinateurs de comprendre la structure grammaticale des phrases et même la reconnaissance vocale et la génération de réponses à des questions. Il fournit les outils et les techniques nécessaires pour analyser, traiter et interpréter les données textuelles dans le but de détecter les opinions, les sentiments et les émotions.

2.3 Les défis de l'analyse des sentiments [23]

Le chercheur dans le domaine de l'analyse des sentiments est confronté à plusieurs contraintes et défis allant du coût de calcul à l'écriture informelle et à la présence de variations dans les langues et les phrases ou mots vagues difficiles à identifier.

Ces contraintes constituent un obstacle à l'analyse de données ciblées et peuvent conduire à un résultat peu fiable. Il existe différents types d'obstacles qui constituent un défi pour l'analyse des sentiments à l'aide de l'un des questionnaires connus, des questionnaires simples ou des questionnaires basés sur les rôles. Mais les événements clairs sont hautement acceptés et

candidats à l'obtention de bons commentaires et de haute qualité. Nous illustrerons certains types de défis répertoriés ci-dessous :

A. Style d'écriture informel :

Le style d'écriture informel constitue le plus grand défi pour toutes les tâches de PNL (**P**rogrammation **N**euro-**L**inguistique), y compris l'analyse des sentiments. Les gens sont très décontractés lorsqu'il s'agit d'écrire des critiques ou des textes ; ils ont tendance à utiliser des acronymes, des emojis, des raccourcis dans leur texte, ce qui est très difficile à comprendre. Les acronymes peuvent être traités s'ils sont universels. Il existe de nombreux acronymes régionaux qui changent et grandissent de jour en jour.

B. Adaptations du langage :

Les langues changent à mesure qu'elles se déplacent vers différentes régions et lieux ; bien que la langue de base reste la même, de nombreux facteurs influencent la langue, tels que la proéminence de la langue, la prononciation, le taux d'alphabétisation, etc. Par exemple, considérons la langue anglaise, qui est largement parlée dans le monde entier, mais on constate que de nombreuses variétés anglaises sont parlées. Dans le monde en fonction des régions comme l'Inde, l'Amérique… etc. De nombreux mots sont utilisés différemment selon la région dans laquelle ils sont utilisés.

Par exemple, en Australie, le mot «'pumps' » désigne des machines qui soulèvent, transfèrent, délivrent ou compriment des fluides, tandis qu'aux États-Unis, «pumps» désigne des chaussures pour femmes. En Australie, ils ont appellent les chaussures pour femmes «high heels».

Cela créera des doublons et pourrait affecter la précision et le coût de calcul du modèle. La barrière de la langue est le plus difficile des défis de la PNL (**P**rogrammation **N**euro-**L**inguistique). Il existe des milliers de langues parlées dans le monde, bien que les techniques de PNL soient difficilement disponibles pour 5 à 10 langues et que les ressources soient largement disponibles pour l'anglais.

C. Données de code mixtes :

Le mélange de codes est l'utilisation du vocabulaire et de la grammaire de différentes langues dans une même phrase.

Le Code Mixing est un phénomène linguistique qui peut se produire dans une situation multilingue où les locuteurs parlent plusieurs langues.

Par exemple dans notre dialecte algérien qui est mélangé avec de l'arabe, du français parfois de l'anglais et d'autres mots qui ne sont spéciaux que dans notre dialecte.

Un examen des publications Facebook créées par des utilisateurs hindi-anglais a révélé un niveau élevé de mélange de codes dans les publications.

L'absence de grammaire formelle pour les phrases à code mixte rend difficile l'identification de la sémantique compositionnelle, ce qui est essentiel pour mener une analyse des sentiments à l'aide de techniques basées sur des règles et sur l'apprentissage automatique.

En conséquence, afin de mener une analyse des sentiments sur des données mixtes, de nouveaux modèles de langage doivent être développés.

D. Erreurs grammaticales :

Les erreurs grammaticales sont très courantes dans les textes informels et peuvent être corrigées, mais seulement dans une certaine mesure ; les fautes d'orthographe peuvent également être corrigées de manière limitée. Il est très difficile de détecter à chaque fois les fautes d'orthographe des utilisateurs. La précision de l'analyse des sentiments et des tâches de PNL peut être améliorée si ces erreurs peuvent être traitées et corrigées.

E. Coût de calcul :

Pour obtenir une meilleure précision, nous devons augmenter la taille des données de formation et compliquer le modèle, ce qui augmentera de manière exponentielle le coût de calcul du modèle de formation ; Un GPU (**G**raphics **P**rocessing **U**nit) haut de gamme peut être nécessaire pour entraîner un modèle avec un énorme corpus. Les modèles comme SVM, NB ne sont pas coûteux en calcul, mais les réseaux de neurones et les modèles d'attention ont montré qu'ils le sont.

F. Disponibilité des données :

La PNL et l'analyse des sentiments étant une technologie en plein essor, la disponibilité des données peut également constituer un défi dans certains cas. Bien que des données soient disponibles sur Twitter pour l'analyse des sentiments, les données de formation de haute qualité constituent un défi pour les algorithmes d'apprentissage supervisé. Les données de formation pour l'ABSA (Aspect-Based Sentiment Analysis) sont difficiles à trouver en ligne et doivent donc être préparées manuellement. Les données de formation d'un domaine peuvent ne pas être applicables et utiles à d'autres domaines. Par exemple, un modèle formé sur un ensemble de données d'évaluation d'hôtels n'est pas utile pour prédire les sentiments d'un ensemble de données d'actions ou de fonds communs de placement et vice versa.

G. Détection des avis abusifs et des avis contrefaits :

Sur le Web, il existe de nombreuses informations de spam contenant du spam et des avis abusifs destinés à la classification des sentiments ; il est inacceptable de traiter des données contenant de fausses données car cela réduit la fiabilité des résultats ; nous devons d'abord identifier les messages indésirables et les supprimer, puis procéder au traitement.

H. Le focus sur un domaine :

Ce défi constitue un obstacle majeur à l'analyse des sentiments car il dépend principalement du caractère limité du mot analyse des sentiments ; cela peut conduire à se concentrer sur un seul sujet. Par exemple, nous pouvons trouver dans un domaine plusieurs fonctionnalités et de bonnes performances ; en même temps, ces fonctionnalités peuvent être très mauvaises dans certains autres domaines [24].

Remarque :

La performance de pointe de l'analyse des sentiments : en moyenne, la précision est d'environ 80 % et ne descend pas en dessous de 70 % [25].

2.4 Analyse des sentiments en éducation

Notre société vit une transformation impressionnante, peut-être la plus importante de ces dernières années, qui, à travers la forte diffusion des nouvelles technologies, modifie radicalement la nature des relations entre les pays, les marchés, les personnes et les cultures. Cette révolution technologique a clairement facilité le processus de mondialisation et l'échange d'informations. De cette manière, l'information peut être considérée comme un bien économique dont la valeur est étroitement liée à la quantité de connaissances qu'elle peut apporter à ses utilisateurs. L'acquisition de nouvelles connaissances, compétences ou aptitudes a déterminé la nécessité d'une mise à jour continue de la part des acteurs de la chaîne d'approvisionnement de la nouvelle économie.

Dans ce scénario, un rôle clé est joué par l'apprentissage tout au long de la vie, qui se poursuit tout au long de la vie et vise à améliorer l'épanouissement des personnes, tant au niveau personnel que social. Dans la société apprenante, être continuellement à jour est la condition essentielle pour y vivre et suivre les changements.

La transition radicale vers l'éducation numérique depuis une vingtaine d'années, accélérée par les récents événements mondiaux, a fermement établi l'apprentissage en ligne comme la norme contemporaine et l'une des approches de formation les plus largement utilisées. Ces dernières années, les universités et les établissements de formation ont adopté cette approche pour dispenser cours ou accompagner les étudiants dans leur processus de formation.

Grâce à l'utilisation d'Internet et de ses services, les démarches d'accompagnement et de suivi de l'utilisateur peuvent être facilement intégrées des aspects pédagogiques et technologiques pour un apprentissage dynamique.

L'apprentissage mixte est une approche utile pour accompagner les étudiants et mieux comprendre leurs problématiques d'apprentissage. La possibilité d'utiliser des outils collaboratifs et d'interagir avec d'autres étudiants permet à l'étudiant de partager ses doutes sur certains sujets.

Cependant, l'enseignant reste souvent en dehors de cette dynamique et ne comprend pas les problèmes d'apprentissage qui caractérisent la classe. Une solution possible est l'analyse des sentiments.

Ce changement monumental a généré une énorme quantité de données, englobant les dialogues élèves-enseignants, les discussions entre pairs et les commentaires évaluatifs. Malgré leur potentiel, une grande partie de ces données restent inexplorées.

Plonger dans ce vaste domaine constitue une occasion en or d'évaluer les sentiments des étudiants, de mettre en lumière leurs perceptions et leurs attitudes à l'égard de leur parcours d'apprentissage. Ceci, à son tour, peut jouer un rôle central dans l'amélioration de la qualité des cours et l'optimisation des résultats d'apprentissage.

L'analyse des sentiments, une branche spécialisée du traitement du langage naturel, a déjà démontré ses prouesses dans divers domaines, allant du monde du commerce aux coulisses de

la politique, capturant efficacement le pouls du sentiment public. Pourtant, étonnamment, son intégration dans le domaine de l'éducation en ligne n'est pas largement étudiée.

2.5 Le processus de l'analyse de sentiments

Dans cette partie nous allons présenter brièvement les différentes étapes de l'analyse de données qui sont illustrés dans la figure ci-jointe :

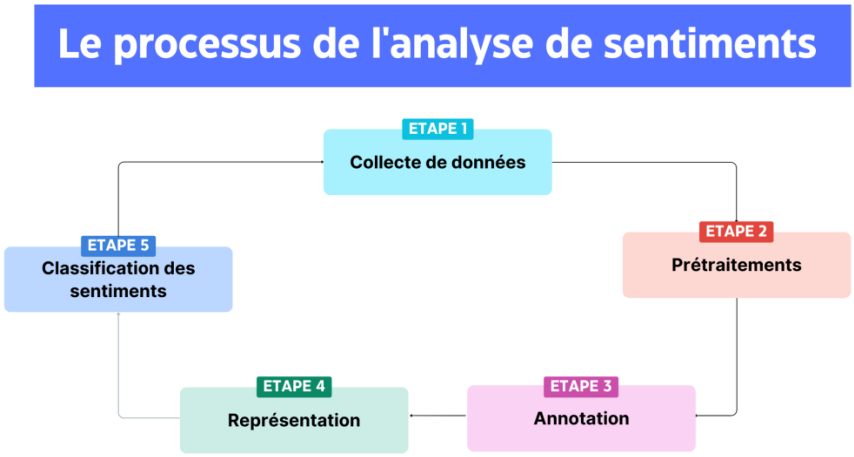

Figure 2.3 – Le processus de l'analyse de sentiment

2.5.1 Collecte de données

La collecte de données est la première étape de l'analyse des sentiments. Afin de prendre des décisions basées sur ces données collectées à des fins de recherche, nous devons nous assurer que ces données sont informatives et fiables. Elles peuvent être collectées à partir de sources telles que des groupes d'utilisateurs, Twitter, Facebook, des blogs, des forums et des sites Web commerciaux tels qu'Amazon.com, alibaba.com, etc.

2.5.2 Prétraitement

Examiner les données avant de les analyser, c'est les préparer.

Des mots inappropriés et offensants sont trouvés dans certains avis et conversations sur les sites de médias sociaux. Pour une analyse plus fiable, ils doivent être examinés et préparés.

Le contenu qui n'est pas pertinent pour l'analyse, le spam ou les avis inappropriés, incorrects, incomplets, non pertinents, en double ou mal formatés doivent être identifiés puis supprimés avant d'être envoyés pour une analyse automatisée. C'est l'objectif du processus.

Dans ce cas, ce prétraitement implique plusieurs étapes, notamment le nettoyage, la transformation par négation, la transformation emoji, la casse, l'encodage, le filtrage des mots vides et la suppression.

2.5.3 Annotation

Pour permet au classifieur de s'entraîner dans la construction du modèle de classification pour l'étape suivante, il faut d'abord L'attribution de labels positifs, négatifs ou neutres à chaque message.

2.5.4 Représentation

C'est La conversion des données prétraitées en caractéristiques. Selon l'objectif, différentes techniques peuvent être utilisées, telles que le TF-IDF (**T**erm **F**requency-**I**nverse **D**ocument **F**requency) et les Word Embeddings.

2.5.5 Classification des sentiments

La classification des sentiments est une tâche d'extraction et de classification du texte visant à déterminer la polarité de l'opinion qu'il contient, par exemple, positif ou négatif, bon ou mauvais, aimer ou ne pas aimer.

La classification des sentiments contient plusieurs techniques et est classée en trois techniques principales, à savoir l'approche d'apprentissage automatique, l'approche des techniques hybrides et l'approche basée sur le lexique.

Actuellement, la technique Naive Bayes et les machines à vecteurs de support (SVM) sont plus populaires et utilisées pour la classification des sentiments.

Dans la figure 2, nous illustrerons les approches actuellement utilisées dans la classification des sentiments [24].

La figure ci-dessous représente les approches de l'analyse de sentiments:

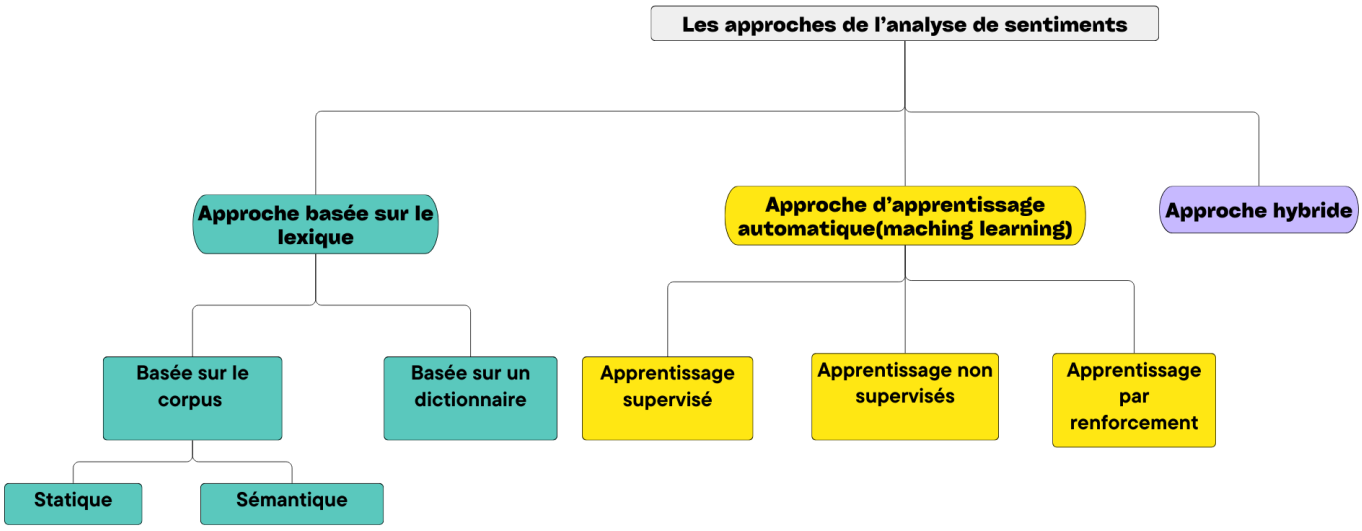

Figure 2.4 – Les approches de l'analyse de sentiment

I. Approche basée sur le lexique

Les lexiques sont une collection de termes où chaque terme est associé à un score prédéfini indiquant la polarité neutre, positive ou négative du texte. Un score est attribué aux jetons en fonction de la polarité telle que + 1, 0, − 1 pour positif, neutre, négatif ou le score peut être attribué en fonction de l'intensité de la polarité et ses valeurs varient de [+ 1, − 1] où + 1 représente un résultat hautement positif et − 1 représente un résultat hautement négatif [26].

Pour une critique ou un texte donné, l'agrégation des scores de chaque jeton est effectuée, c'est-à-dire que les scores positifs, négatifs et neutres sont additionnés séparément. Lors de la dernière étape, la polarité globale est attribuée au texte en fonction de la valeur la plus élevée des scores individuels. Ainsi, le document est d'abord divisé en jetons de mots simples, après quoi la polarité de chaque jeton est calculée et finalement agrégée [26].

Par exemple si:

g(t)>0 la polarité est positif.

g(t)<0 la polarité est négatif.

g(t)=0 la polarité est neutre.

Si on a :

C'est très bon explication (+1), mais la qualité d'image est médiocre (-1), et si vous pouvez partagés le lien de PDF (0), merci (+1). Donc :

g(t)=1-1+0+1=1 donc la polarité est positif.

L'approche basée sur le lexique ne requiert pas de données de formation, ce qui la classe comme une méthode non supervisée par certains experts. Le principal inconvénient de l'approche basée sur le lexique est qu'elle est fortement orientée domaine et que les mots appartenant à un domaine ne peuvent pas être utilisés dans un autre domaine [26].

Il existe deux méthodes selon l'approche basée sur le lexique. La première est une approche basée sur le corpus et la seconde est une approche basée sur un dictionnaire.

A. Approche basée sur le corpus

L'approche utilise des modèles sémantiques et syntaxiques pour déterminer l'émotion de la phrase. Cette approche commence par un ensemble prédéfini de termes de sentiment et leur orientation, puis étudie des modèles syntaxiques ou similaires pour découvrir des jetons de sentiment et leur orientation dans un vaste corpus. Il s'agit d'une méthode spécifique à une situation qui nécessite une quantité importante de données étiquetées pour être entraînée. Cependant, cela aide à résoudre le problème des mots d'opinion avec des orientations dépendant du contexte [26].

L'approche basée sur le corpus comprend les types d'approches suivants : approche statistique et approche sémantique comme expliqué ci-dessous.

B. Approche statistique

Les mots d'opinion de départ ou les modèles de cooccurrence peuvent être trouvés à l'aide d'une approche statistique. L'idée générale derrière cette approche est que si elle apparaît plus dans les textes positifs que dans les textes négatifs, alors elle est plus susceptible d'être positive ou vice versa. Le principe clé de cette approche est que si des jetons de sentiment comparables sont fréquemment observés dans le même environnement, ils auront probablement la même orientation. De ce fait, l'orientation du nouveau jeton est déterminée par la fréquence à laquelle il apparaît aux côtés d'autres jetons détectés dans un contexte similaire [26].

C. Approche sémantique

Dans cette approche, le score de similarité est calculé entre les jetons utilisés pour l'analyse des sentiments. Les antonymes et les synonymes peuvent être facilement trouvés en utilisant cette approche, car des mots similaires ont un score positif ou une valeur plus élevée [26].

D. Approche basée sur un dictionnaire

L'approche basée sur un dictionnaire utilise une liste de mots d'opinion prédéfinis collectés manuellement. L'hypothèse principale derrière cette approche est que les synonymes ont la même polarité que le mot de base, tandis que les antonymes ont une polarité opposée.

Les grands corpus comme le thésaurus ou WordNet (Une base de données lexicale pour la langue anglaise) sont recherchés pour les antonymes et les synonymes, après quoi ils sont ajoutés à un groupe ou à une liste de départ préparée plus tôt.

Dans la première étape, un ensemble initial de mots est collecté manuellement avec leur polarité respective. Par la suite, la liste est enrichie en explorant les antonymes et synonymes dans les ressources lexicales disponibles. Ensuite, les mots sont ajoutés de manière itérative à la

liste, et la liste est développée. Une évaluation ou une correction manuelle peut être effectuée dans la dernière étape pour en garantir la qualité. Stefano et Andrea ont créé SentiWordNet à trois voies à l'aide d'annotations automatiques de WordNet 3 [26].

II. Approche d'apprentissage automatique

L'approche d'apprentissage automatique est utilisée pour résoudre les problèmes liés à la classification de textes contenant des caractéristiques syntaxiques ou linguistiques. Bien que l'approche basée sur le lexique soit utilisée pour extraire les sentiments d'un texte, elle dépend d'un lexique de sentiments ; la collection de termes de sentiments connus et précompilés dans les algorithmes d'apprentissage automatique divisés en apprentissage par renforcement, apprentissage non supervisé et apprentissage supervisé [24].

A. Apprentissage supervisé

Il s'agit d'un type d'approche d'apprentissage automatique qui utilise un ensemble de données appelé ensemble de données d'entraînement pour faire des prédictions. Ces ensembles de données contiennent des données d'entrée ainsi que des valeurs de réponse. Dans les méthodes d'apprentissage supervisé, il fait appel à un grand nombre de documents de formation variés [24].

En apprentissage supervisé, il existe deux types d'algorithmes :

A. Algorithmes de régression, qui cherchent à prédire une valeur continue, une quantité.
B. Algorithmes de classification, qui cherchent à prédire une classe/catégorie.

Pour créer un modèle d'apprentissage supervisé, on peut recourir à différents algorithmes, on peut citer en guise d'exemple la régression linéaire et logistique, l'arbre de choix avec différentes variables de sortie, le Naive Bayes, Random Forest, SVM (**S**upport **V**ector **M**achine) et KNN.

L'apprentissage supervisé consiste à utiliser un ensemble d'apprentissage

$Dn = \{(x_1, y_1), \dots , (x_n, y_n)\backslash\} \backslash)$ pour inférer la relation entre les caractéristiques x et les étiquettes y . Cela peut également être appelé "discrimination" ou "reconnaissance de formes". Les termes associés comprennent "caractéristique", "feature" ou "variable explicative", notée x_i.

B. Apprentissage non supervisés

Il s'agit d'un type unique d'algorithme d'apprentissage automatique et est utilisé dans la plupart des cas pour tirer diverses inférences de données ; ces groupes de données sont constitués de données d'entrée sans aucune réponse étiquetée. Il est utilisé lorsqu'il est impossible d'obtenir des documents de formation labellisés [24].

Les modèles d'apprentissage non supervisé sont notamment utilisés pour :

A. Le classement des données.
B. Le calcul approximatif de la densité de distribution.
C. La réduction des dimensions.

L'apprentissage non supervisé implique de partitionner un ensemble d'apprentissage

$D_n = \{x_1,\ldots, x_n\} \subseteq X$ en classes pertinentes sans l'utilisation d'étiquettes de classe préalablement définies.

C. Apprentissage par renforcement

Son intégralité indique comment prendre une décision optimale, une technique importante qui diffère relativement de son homologue d'apprentissage non supervisé. Cette technique vise fortement à améliorer l'efficacité de la classification des textes pour montrer que la technique d'apprentissage par renforcement est importante et proéminente [24].

III. Approche hybride

L'approche hybride combine l'apprentissage automatique et les approches basées sur le lexique. Hybride est un terme qui fait référence à la combinaison de techniques d'apprentissage automatique et de lexique pour l'analyse des sentiments. La technique hybride combine les deux et est extrêmement populaire, les lexiques de sentiments jouant un rôle important dans la majorité des systèmes [26].

2.6 Conclusion

Dans ce chapitre, nous avons exploré les généralités sur l'analyse des sentiments, en commençant par une introduction qui définit cette discipline et en mettant en évidence son importance croissante dans divers domaines d'application. Nous avons détaillé les différents niveaux et types d'analyse de sentiment, y compris les opinions régulières, comparatives, explicites et implicites, ainsi que la catégorisation des sentiments en phrases objectives et subjectives.

Nous avons également examiné les défis spécifiques auxquels fait face l'analyse des sentiments, en particulier dans le contexte éducatif. Le processus de l'analyse de sentiments a été décortiqué en plusieurs étapes, allant de la collecte des données à la classification des sentiments, en passant par les prétraitements, l'annotation, la représentation et la classification.

Différentes approches pour la classification des sentiments ont été présentées, y compris les méthodes basées sur le lexique, l'apprentissage automatique (supervisé, non supervisé et par renforcement), et les approches hybrides, chacune ayant ses propres avantages et inconvénients.

Ce chapitre servira de base théorique pour comprendre les techniques et les méthodes utilisées dans l'analyse des sentiments, préparant ainsi le terrain pour les discussions plus techniques et spécifiques à venir dans les chapitres suivants.

3 Chapitre 3

Etat de l'art

Sommaire

3.1 Introduction

L'analyse des sentiments, également connue sous le nom d'analyse d'opinions, est l'une des applications PNL les plus largement utilisées pour identifier les intentions humaines à partir de leurs avis.

Les techniques d'analyse des sentiments ont été largement utilisées pour extraire les opinions des utilisateurs sur les produits et services à partir de leurs avis et créer des connaissances exploitables pour une entité [27]. Cela permet aux entreprises d'améliorer leurs stratégies et d'obtenir des informations sur les commentaires des clients sur leurs produits et services [27].

L'analyse des sentiments est un domaine multidisciplinaire qui inclut l'apprentissage automatique, la PNL, la sociologie et la psychologie pour détecter les opinions sous-jacentes des clients ou des utilisateurs. Les sites de médias sociaux comme Twitter, Facebook et Instagram sont d'importantes sources d'opinions des utilisateurs[27]. L'analyse de ces sources pour extraire les sentiments ou les opinions des utilisateurs a commencé il y a dix ans [27].

Les progrès technologiques ont transformé des domaines tels que les soins de santé et l'éducation en adoptant des techniques d'IA et de PNL [27].

Dans le secteur de l'éducation, afin d'améliorer pédagogiquement les pratiques d'apprentissage-enseignement des étudiants, les techniques d'analyse des sentiments sont utilisée pour recueillir les opinions des étudiants, et étiqueter leur commentaires avec leur orientation sentimentale grâce aux progrès des techniques d'annotation des sentiments et des méthodologies d'IA et sans trop d'intervention humaine. Les établissements d'enseignement ont largement investi dans la création d'outils d'analyse des sentiments et dans le traitement des commentaires de leurs étudiants afin de recueillir leurs opinions et leurs idées [27]. Il existe plusieurs différentes applications du domaine éducatif telles que les limites de l'infrastructure éducative existante, la compréhension de la pédagogie de l'engagement des étudiants, la prise de décision sur les politiques éducatives, l'évaluation des cours et de l'enseignement...etc. et le mécanisme d'analyse des sentiments peut être s'adapter à eux.

L'annotation manuelle ou l'étiquetage de l'orientation sentimentale prend du temps [28] et nécessite de nombreuses ressources avec une compréhension pédagogique en éducation. Ce défi a été relevé en développant différentes approches d'annotation de sentiments utilisant des lexiques et des corpus. Ces techniques agissent comme des techniques non supervisées pour une compréhension initiale des commentaires des étudiants.

Le rôle de l'IA dans l'analyse des sentiments est inévitable car elle aide à traiter et à analyser un grand nombre de commentaires d'étudiants [29]. Les méthodologies d'IA telles que l'apprentissage automatique, l'apprentissage profond et les transformateurs [30] sont capables d'apprendre les opinions des étudiants grâce à des mécanismes d'attention et de classer ou de prédire leurs émotions pour les commentaires non étiquetés des étudiants [31]. Les techniques d'annotation de sentiments non supervisées et les méthodologies d'IA surmontent dans une certaine mesure le défi de l'étiquetage manuel.

Dans ce chapitre, nous allons explorer l'état de l'art en matière d'analyse des sentiments, en examinant les travaux connexes et les contributions significatives dans ce domaine. Cette revue de la littérature nous permettra de situer notre recherche dans le contexte des études existantes et de comprendre les approches et méthodes les plus récentes et efficaces utilisées par les chercheurs et les praticiens.

3.2 Traveaux connex

L'analyse des émotions et des opinions des individus est devenue l'un des sujets les plus critiques du monde de la recherche et de ses évolutions dans toutes les entreprises, en particulier dans le secteur de l'éducation. Ces dernières années, de nombreuses recherches ont été menées pour créer des méthodes permettant d'évaluer et de documenter le processus d'analyse des sentiments dans de nombreuses langues [32]. L'application de l'analyse des sentiments dans les environnements d'apprentissage en ligne nécessite une approche soigneusement structurée et méthodologiquement solide [33].

L'analyse des sentiments est généralement effectuée selon un format en deux phases, où dans la première phase, les données pertinentes sont collectées, et dans la deuxième phase, l'extraction du sentiment a lieu [34]. Les travaux rapportés peuvent être globalement classés en trois approches principales : (a) basée sur l'apprentissage automatique, (b) basée sur le lexique et (c) hybride [35]. Dans la littérature et travaux connexes, il existe de nombreuses approches liées à l'analyse des sentiments.

D'après Bensba Amal et al (2022) [36], ils ont proposé une solution au problème de détection des émotions des étudiants dans un contexte d'apprentissage en ligne pendant la pandémie Covid 19 (Coronavirus Disease), les auteurs ont adopté une approche hybride combinant des techniques d'apprentissage automatique (machine learning) supervisé et non supervisé. Cette détection s'appuie sur les avis des étudiants en langue française extraits de formulaires envoyés aux étudiants de l'École nationale d'informatique ESI (l'**É**cole **N**ationale d'**I**nformatique) d'Algérie , concernant différents aspects des cours en ligne suivis pendant la pandémie Covid19 (Contenu, Structure, Présentation, Communication, Design et Général),donc ils obtiennent un Dataset composé de plus de 13 000 avis d'étudiants ,Ces données ont été filtrées manuellement, annotées dans différentes classes pour chaque niveau puis cer-

taines étapes de prétraitement ont été appliquées. (Suppression des URL (**U**niform **R**esource **L**ocator), des Hashtags, suppression des symboles de ponctuation et mise en minuscules de tout), aussi la traduction des avis qui n'étaient pas en français et supprimé les mots vides et la normalisation du texte.

Pour l'apprentissage supervisé, ils ont utilisé trois algorithmes classiques : SVM, KNN (K-Nearest Neighbors) et RNN (**R**ecurrent **N**eural **N**etwork) avec une représentation TF-IDF pour entraîner les modèles et les tester avec des mesures de précision et de rappel. Cela leur a permis de classer les commentaires d'étudiants selon 4 niveaux : polarité (positif/négatif/neutre), émotion (joie, mécontentement, confusion, colère, ennui, anxiété, neutre), attitude (amicale/hostile/neutre) et aspect (contenu, structure, présentation, Communication, Design et Général).

Pour l'apprentissage non supervisé, ils ont utilisé les algorithmes K-means et K-modes pour identifier différents profils/groupes d'étudiants, ainsi que l'algorithme Apriori pour trouver des règles d'association intéressantes entre les différentes variables.

Leur approche a donné de bons résultats pour détecter la polarité avec une précision de 91%, même pour les aspects des commentaires ont une précision autour de 80%, les deux sont avec SVM, mais le manque de données a limité les performances sur les émotions/attitudes selon les auteurs.

D'après Qaqish Evon et al (2023) [32], l'article traite la problématique d'évaluer les opinions et les expériences des Jordaniens concernant le passage de l'éducation en présentiel à l'éducation hybride (en ligne et en présentiel) dans le contexte post-COVID .À l'aide de l'API (**A**pplication **P**rogramming **I**nterface) Twitter via la programmation Python,4 000 tweets sur les réflexions des gens sur l'apprentissage hybride en Jordanie à partir de Twitter du 20 octobre 2021 à février 2022 ont été obtenu ce qui a abouti un dataseet au format *.csv.

Pour le Prétraitement du texte arabe ils ont utilisé :

DeNosie qui est un processus de suppression de tout bruit dans le texte tels que les mots anglais, les caractères spéciaux comme les signes diacritiques qui sont des symboles similaires aux voyelles de la langue anglaise qui apparaissent au-dessus ou en dessous des lettres de texte arabe, les emojies, supprimer tous les caractères anglais...etc

Tokénisation qui est le processus de dissection du texte en morceaux significatifs, appelés jetons.

Suppression des mots vides, normalisation et Morphological disambiguation qui est un processus consistant à fournir la signification morphologique la plus probable pour un mot particulier dans son contexte.

Les auteurs proposés une approche basée sur l'apprentissage profond (deep learning) pour la détection d'émotions et l'analyse de sentiments à partir des tweets en arabe en utilisant d'un modèle Deep Learning LSTM (Long Short-Term Memory) qui est un type de réseau de neurones récurrent, pour la classification des émotions des tweets en 5 classes : colère, haine, tristesse, joie et neutre , avec Plongement de mots (word embedding) avec GloVe (**G**lobal **V**ectors for **W**ord **R**epresentation) pour représenter les mots sous forme de vecteurs numériques. GloVe a été préféré à Word2Vec (**W**ord to **V**ector) dans cette étude.

Cette approche d'apprentissage profond proposée a permis d'obtenir des résultats satisfaisants avec F1-score global du modèle : 0,85.

D’après Ameur Mohamed Seghir Hadj et al (2023) [37], ils ont proposé une solution au problème de détection du sarcasme et de comprendre les sentiments des gens sur les réseaux sociaux par rapport au COVID-19,cela peut aider les autorités et les organismes de santé à mieux gérer la crise, à diffuser des informations plus précises et à prendre de meilleures décisions, La première étape qu'ils ont suivie pour créer le dataset « AraCOVID19-SSD » a été de préparer un ensemble de mots-clés, puis ils ont récupéré les tweets en fonction de ces mots-clés. Les mots-clés qu'ils ont utilisés ont été faits pour récupérer le plus grand nombre possible de tweets liés au COVID 19, donc après l’étape de filtrage, ils ont retrouvés avec un total de 300 000 tweets arabes uniques liés à la pandémie de COVID-19 dans la période allant du 15 décembre 2019 au 15 décembre 2020. L'annotation des données leur a permis d'annoter un total de 5 162 tweets arabes sur les 300 000 tweets collectés. Ils ont appliqué un prétraitement de base à tous les tweets arabes collectés, qui comprend :

- La suppression des signes diacritiques.
- La suppression des caractères allongés et répétés.
- Normalisation des caractères arabes.
- La suppression des liens et des références des utilisateurs (notifications des utilisateurs).
- Tokenisation des tweets dans laquelle la ponctuation, les mots et les chiffres sont séparés.

Les auteurs ont adopté une approche d'apprentissage automatique supervisé, ils ont utilisé plusieurs modèles :

Des modèles classiques de sacs de mots Bag-of-Words(**A m**odel used in **n**atural **l**anguage **p**rocessing) models comme les machines à vecteurs de support (SVMs), les forêts aléatoires (Random Forests model) et la régression logistique (Logistic Regression) pour Fournir des références de base pour évaluer les performances sur les tâches et explorer les capacités de ces modèles classiques largement utilisés.

Des modèles de transformeurs pré-entraînés (comme AraBERT(Arabic **B**idirectional **E**ncoder **R**epresentations from **T**ransformers), BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) multilingue et XLM-RoBERTa) pour Tirer parti des dernières avancées en transfert d'apprentissage pour les tâches de NLP et évaluer les performances de pointe de ces grands modèles neuronaux pré-entraînés.

Les différents modèles testés ont atteint des performances élevées sur les deux tâches de détection du sarcasme et d'analyse des sentiments, avec des scores F1 dépassant 0,89 pour tous les modèles.

Les bons résultats sont attribués principalement à la richesse de l'ensemble de données et aussi avec un grand nombre d'instances pour chaque classe.

D’après Colace Francesco (2019) [35], ils ont proposé une solution au problème d’analyser les sentiments des commentaires d'étudiants avec des questions de type descriptif des étudiants et non seulement le type objectif à choix multiples comme le font la plupart des systèmes de feedback existants.

Pour le dataset, c'est des commentaires des étudiants qui sont collectés via un portail étudiant en ligne. Où un étudiant a un identifiant séparé. Ensuite, un étudiant peut donner un seul commentaire par identifiant de connexion.

Les auteurs ont adopté une approche hybride combinant l'utilisation d'un lexique de sentiments (approche lexicale) et des méthodes d'apprentissage automatique supervisé. Tous d'abord, ces données sont passées par le prétraitement de texte qui est divisé en trois sous-catégories :Tokenisation, Suppression des mots vides, L'étiquetage des parties du discours (Part-of-Speech Tagging ou POS Tagging en anglais) qui est une technique de traitement du langage naturel qui consiste à assigner une étiquette de catégorie grammaticale à chaque mot ou jeton dans une phrase. Ensuite, l'extraction de caractéristiques, notamment les mots à sentiment positif et négatif à l'aide de SentiWordNet (lexique).Après, ils ont Entraînés des modèles d'apprentissage automatique supervisé comme les forêts aléatoires (Random Forest) et les machines à vecteurs de support (SVM) en utilisant les caractéristiques extraites, y compris celles du lexique. Enfin ils ont Appliqué du modèle entraîné sur de nouvelles données pour la classification des sentiments.

Donc, leur méthode hybride a permis d'obtenir une précision de classification des sentiments de 90%, ce qui est un très bon résultat selon les auteurs.

D'après Lin Fan-gyuan (2023) [33], ils ont proposé une solution pour rationaliser le processus d'évaluation automatique des sentiments et d'extraction d'opinions de la vaste mer de contenu produit par les apprenants lors de leurs interactions en ligne, les auteurs ont adopté une approche hybride combinant des methodes d'apprentissage automatiqe supervisé et non supervisé . Ils obtiennent un Dataset qui est structuré sous forme de fichier CSV (**C**omma-**S**eparated **V**alues) et englobe diverses formes de communication, notamment les publications sur les forums de discussion, les commentaires sur les devoirs, les e-mails et les transcriptions de discussions sur plusieurs cours sur des plateformes telles que Moodle, MOOC (**M**assive **O**pen **O**nline **C**ourse), Blackboard, Coursera et EdX(, cette datest a été réalisée à l'aide de scripts Python avec plusieurs outils et techniques y compris Beautiful Soup pour l'analyse du HTML(**H**yper**T**ext **M**arkup **L**anguage)/XML(**e**Xtensible **M**arkup **L**anguage) et la création d'arbres pour l'extraction de données. Selenium pour automatiser les actions de navigation, soumission de formulaires, etc.

Pour le prétraitement du texte, ils ont utilisé : La tokenisation pour décomposer le texte en jetons, la suppression des mots vides et la vectorisation (Bag of Words, TF-IDF), ainsi que les bibliothèques Python NLTK (**N**atural **L**anguage **T**oolkit), scikit-learn et gensim ont été utilisées. Une association entre sentiments et sujets a été réalisée et visualisée via des cartes thermiques.

Pour l'apprentissage supervisé, les algorithmes (Naive Bayes, SVM, RNN) ont permis de classer les commentaires selon leur sentiment (positif/négatif/neutre), guidée par des principes de psychologie éducative, et l'apprentissage non supervisé (LDA : **L**atent **D**irichlet **A**llocation) pour l'extraction des opinions/sujets abordés par les apprenants.

Pour l'apprentissage non supervisé, LDA a identifié les principaux sujets/opinions abordés par les apprenants. Une association sentiments-sujets a ensuite été réalisée, visualisée

via des cartes thermiques, permettant d'identifier les aspects spécifiques des cours générant des sentiments positifs ou négatifs chez les apprenants.
Enfin, cette approche a permis d'atteindre une précision de 89,6% pour la classification des sentiments et d'identifier les aspects spécifiques des cours générant des sentiments positifs ou négatifs chez les apprenants.

D'après Kechaou Zied et al (2011) [38], les auteurs ont proposé une approche pour l'analyse automatique des sentiments exprimés dans les blogs et forums d'e-learning (**E**lectronic **L**earning) afin d'aider les développeurs à améliorer leurs systèmes. Ils ont adopté une approche hybride car elle combine deux techniques d'apprentissage automatique différentes(supervisé avec non supervisé).Un corpus de 2000 revues d'e-learning (1000 positives, 1000 négatives) extrait de blogs et forums a été constitué.
Pour le prétraitement du texte, ils ont utilisé la suppression des mots vides, la racinisation (stemming) et la vectorisation TF-IDF. Les bibliothèques Python NLTK et SVMlight (**I**mplementation of **S**upport **V**ector **M**achines) ont été utilisées. Trois méthodes de sélection de caractéristiques qui est une étape non supervisée visant à identifier les termes les plus discriminants pour la tâche de classification de sentiments : (Information Mutuelle (MI), Gain d'Information (IG), Statistiques du Chi (Chi-squared Test) ont été évaluées.

Pour l'apprentissage supervisé, les algorithmes de Markov cachés HMM (**H**idden **M**arkov **M**odel) et Machines à vecteurs de support SVM ont été utilisés et combinés avec différentes règles de combinaison pour classer les commentaires selon leur sentiment positif/négatif.
Les performances en termes de précision, rappel et F-mesure n'étaient pas très élevées (entre 0.72 et 0.82 environ), ce que les auteurs attribuent à la nature complexe et bruitée des blogs d'e-learning.

Pour la sélection de caractéristiques, la méthode du gain d'information (IG) a donné les meilleurs résultats pour la sélection des termes exprimant des sentiments et pour la classification des sentiments, devant l'information mutuelle (MI) et les statistiques du Chi (CHI).

D'après Colace Francesco (2014) [39], ils ont proposé une solution au problème de détecter les émotions et sentiments des étudiants dans un environnement d'apprentissage en ligne (e-learning) pour permettre aux enseignants d'adapter leur approche pédagogique en fonction de l'état émotionnel perçu des étudiants, L'objectif principal de cet article est de montrer comment mGT (Mixed Graphs of Terms) peut être appliqué efficacement pour l'exploration de sentiments à partir de textes : la méthode proposée peut être utilisée pour construire un détecteur de sentiments capable d'étiqueter un document en fonction de son sentiment, les auteurs ont adopté une approche hybride qui combine une méthode non supervisée , l'utilisation d'un lexique de sentiments annoté manuellement et une petite partie supervisée uniquement pour l'entraînement initial des graphes de termes représentatifs des sentiments positifs et négatifs, en utilisant des documents étiquetés manuellement. Ils ont utilisé Dataset standard de critiques de films, L'objectif principal de cette expérimentation était d'évaluer les performances de la méthode et de faire une comparaison avec les autres approches bien connues dans la littérature (section travaux connexes),après ils ont utilisé une dataset réelles d'un

cours en ligne (forums, chats) de la populaire plateforme d'apprentissage en ligne Moodle, environ 75 étudiants ont assisté aux cours et ont utilisé Moodle pour partager des commentaires. Ces données ont passe par la phase de prétraitement qui implique la tokenisation, le filtrage et la radicalisation des mots vides, une matrice de document à terme est construite pour alimenter l'allocation latente de Dirichlet (LDA). Un graphe mixte de termes est ensuite construit à partir de plusieurs clusters. Ce module construit un graphe mixte de termes à partir d'un ensemble de documents appartenant à un domaine de connaissances bien défini et préalablement étiquetés en fonction du sentiment qui y est exprimé. De cette façon, le graphe mixte de termes obtenu contient des informations sur les mots et leurs cooccurrences représentant ainsi un certain sentiment dans un domaine de connaissance bien défini. Ils utilisent le module Sentiment Mining : ce module extrait le sentiment d'un document grâce à l'utilisation du Mixed Graph of Term comme filtre de sentiment. L'extraction des sentiments est obtenue par une comparaison entre le document et le graphique mixte à l'aide d'un algorithme proposé, qui nécessite l'utilisation d'un lexique annoté, comme par exemple WordNet ou ItalWordNet (Italian lexical database similar to WordNet), pour la récupération des synonymes des mots contenus dans le document et non inclus dans la référence mGT. L'approche proposée est efficace dans une classification asynchrone des sentiments, mais peut également fonctionner de manière synchrone. Deux nouveaux modules ont été introduits : Capture de documents, ce module vise à collecter des documents provenant de sources web (réseaux sociaux, blogs, etc.) et classification des sentiments des documents, les nouveaux documents insérés dans l'ensemble de formation doivent être classés avec l'appui d'un expert.

Sur le dataset de critiques de films, leur approche a obtenu une précision de 88,5%, supérieure à d'autres méthodes comparées.

D'après Clarizia Fabio et al (2018) [40], Le problème abordé était de permettre aux enseignants d'avoir une visibilité sur les sentiments exprimés par les étudiants dans les forums, chats et autres outils collaboratifs des plateformes e-learning, afin d'ajuster leur approche pédagogique en conséquence.

Les auteurs ont proposé une approche hybride combinant l'allocation de Dirichlet latente (LDA), une technique d'apprentissage non supervisé, avec des règles définies par les auteurs, pour le calcul final des scores de sentiments.

Leur méthode consistait à construire des "Mixed Graphs of Terms" (mGT) spécifiques aux sentiments positifs et négatifs à partir d'un corpus de documents annotés, en capturant les mots et co-occurrences de mots représentatifs de chaque sentiment. Pour classer un nouveau document, ils calculaient des scores de correspondance avec les mGT positifs et négatifs, et classaient le document comme positif, négatif ou neutre selon les scores obtenus. Ces règles de calcul de scores constituent l'aspect "supervisé" ou du moins défini manuellement par les auteurs, et non appris automatiquement à partir des données.

Différents datasets ont été utilisés pour évaluer l'approche, notamment un dataset standard de critiques de films, ainsi que des tweets, des commentaires Facebook et des posts de forums issus de la plateforme Moodle.

Les expériences ont montré de bonnes performances de cette approche, en particulier sur les tweets (courts textes). Les auteurs attribuent ces bons résultats à la capacité de leur

méthode à bien capturer les sentiments exprimés dans des textes courts.ils ont obtient plus de 82% pour Accuracy , plus de 78% pour Recall et 0,83 pou F score.

D'après Nandal Neha (2020) [41], ils ont proposé une solution au problème d'analyse de sentiments au niveau des aspects pour les produits Amazon. Cela peut aider les commerçants et détaillants à mieux comprendre les forces et faiblesses perçues de leurs produits par les clients. La première étape qu'ils ont suivie a été de collecter des données d'avis de produits à grande échelle en développant un crawler web basé sur Scrapy. Après avoir filtré les données, ils ont obtenu un total de 300 000 avis uniques sur différents produits Amazon entre 2019 et 2020. L'annotation manuelle des données leur a permis d'identifier 5 162 avis annotés au niveau des aspects. Ils ont appliqué un prétraitement de base à tous les avis, qui comprend: Vectorisation, POStagging, Stop word removal, La lemmatisation/radicalisation des mots : contribue à réduire la spatialité des mots. Par exemple, les mots comme « brillant », « plus brillant » et « éclaircissant » sont considérés comme un seul mot « brillant ».

Les auteurs ont adopté une approche d'apprentissage automatique supervisé. Ils ont utilisé plusieurs modèles :

Des modèles classiques de type sac de mots comme les machines à vecteurs de support (SVM), les forêts aléatoires et la régression logistique pour fournir des références de base.

Des modèles de transformeurs pré-entraînés comme RBF (**R**adial **B**asis **F**unction) et polynomial pour tirer parti des dernières avancées et évaluer les performances de ces grands modèles neuronaux.

Les différents modèles testés ont atteint des performances élevées, avec le modèle SVM avec noyau RBF donnant les meilleurs résultats (taux d'apprentissage de 97%). Les bons résultats sont attribués à la richesse de l'ensemble de données et à la prise en compte des mots bipolaires dont la polarité change selon le contexte.

D'après Ansari Mohd Zeeshan (2020) [34], les auteurs ont proposé une approche d'apprentissage supervisé pour analyser les orientations des sentiments politiques sur Twitter avant les élections générales indiennes de 2019. Cela pouvait aider à comprendre les tendances et opinions politiques sur les réseaux sociaux, ce qui est utile pour les campagnes électorales. La première étape a été d'extraire 3896 tweets liés aux partis politiques indiens en utilisant des mots-clés pertinents. Après un prétraitement de base comme la suppression des données redondantes et la conversion en minuscules, suppression de mentions, hashtags..etc, les tweets ont été annotés manuellement dans 8 catégories de sentiments comme favorable au parti BJP (**B**haratiya **J**anata **P**arty), Congrès, autres partis, ou des combinaisons.

Différentes représentations de texte ont été explorées comme caractéristiques d'entrée, notamment la fréquence des termes (TF) et la TF-IDF pour les uni-grammes, bi-grammes et tri-grammes. Les auteurs ont adopté plusieurs approches d'apprentissage supervisé, à la fois des modèles classiques comme les machines à vecteurs de support (SVM), les arbres de décision, la régression logistique, les forêts aléatoires ; et des modèles de deep learning comme les réseaux de neurones LSTM. Les différents modèles ont été entraînés et évalués sur le corpus annoté.

Les résultats ont montré que les réseaux LSTM avec les tri-grammes TF-IDF ont atteint la meilleure précision de 0,76 et un score F1 de 0,74, suivis de près par les forêts aléatoires avec une précision de 0,77 pour les uni-grammes TF-IDF. Cependant, les LSTM étaient beaucoup plus lents lors de l'entraînement par rapport aux autres méthodes. Les performances des modèles classiques comme les SVM étaient inférieures. L'ensemble de données relativement petit et le déséquilibre de classe sont cités comme des limites potentielles.

3.3 Étude comparative et analyse

Les tableaux ci-dessous synthétisent les principales caractéristiques des approches mentionnées précédemment. Ils comportent neuf (09) colonnes correspondant chacune à un critère de comparaison :

• **La colonne Auteur et Référence** mentionne le nom de l'auteur et la source de l'article.

• **La colonne Catégorie de l'approche** indique le type d'approche utilisée dans l'étude, qu'il s'agisse d'une approche hybride, d'apprentissage supervisé, d'apprentissage non supervisé ou lexicale.

• **La colonne Méthodes** décrit les différentes étapes et techniques employées dans l'approche, telles que la collecte et le prétraitement des données, les algorithmes d'apprentissage automatique utilisés, etc.

• **La colonne Output** précise le résultat final ou la production de l'approche.

• **La colonne Domaine** indique le domaine d'application de l'approche, comme l'éducation, etc.

• **La colonne Dataset** décrit les sources de données utilisées dans l'approche.

• **La colonne Outil supporté** mentionne si un outil logiciel spécifique a été utilisé pour mettre en œuvre l'approche.

• **La colonne Évaluation** présente les principaux résultats d'évaluation de l'approche, tels que la précision, le rappel, le F-score, etc.

Auteur et Référence **Catégorie de l'approche**	**Méthodes**	**Outils supportés**	**Année**	**Domaine**	**Dataset**	**Output**	**Évaluation**
Bensba Amal et al (2022) [36] Approche hybride (apprentissage supervisé et non supervisé)	-Collecte et annotation manuelle des données -Prétraitement des données (suppression d'URL, hashtags, ponctuation, mise en minuscule, suppression des doublons, normalisation du texte) -Représentation vectorielle des données avec TF-IDF -Clustering avec les algorithmes K-means et K-modes -Règles d'association avec l'algorithme Apriori -Classification avec les algorithmes SVM, RNN et KNN -Évaluation avec les métriques de précision, rappel et F-score	Scikit-learn pour les algorithmes de machine learning	2022	Education	13 902 avis d'étudiants avis des étudiants extraits de formulaires envoyés aux (ESI) d'Algérie	Détection des émotions des étudiants dans un contexte d'apprentissage en ligne pendant la pandémie Covid 19	-Bons résultats pour détecter la polarité avec une précision de 91% et les aspects avec une précision autour de 80% -Le manque de données a limité les performances sur les émotions/attitudes.
Qaqish Evon et al (2023) [32] Apprentissage profond (deep learning)	-Collecte de données -Prétraitement des données(Suppression du bruit (caractères anglais, émojis, diacritiques), tokenisation, suppression des mots vides, normalisation (réduction de l'ambiguïté orthographique, dérivation),désambiguïsation morphologique -Plongement de mots (Word Embedding) :Word2Vec (modèle CBOW),GloVe -Développement du modèle :Réseaux de neurones récurrents (RNN) - Long Short-Term Memory (LSTM) -Évaluation du modèle :Matrice de score F1,Précision et rappel pour chaque classe	Python, bibliothèques Keras, Tensorflow, Gensim (pour GloVe)	2023	Education	4 000 tweets post-COVID en Jordanie à partir de Twitter á l'aide de l'API Twitter	Evaluer les opinions des Jordaniens concernant le passage de l'éducation en présentiel à l'éducation hybride (en ligne et en présentiel) dans le contexte post-COVID	Des résultats satisfaisants et bonnes performances pour la détection d'émotions avec F1-score global du modèle : 0,85
Ameur Mohamed Seghir Hadj et al (2023) [37] Approche d'apprentissage automatique supervisé	-Collecte de données -Prétraitement des données :(Suppression des diacritiques, Suppression des ,caractères répétés et allongés, Normalisation des caractères arabes, Suppression des liens et des références aux utilisateurs, tokenisation des tweets) -Annotation des données :Annotation manuelle des tweets pour la détection du sarcasme (Oui/Non) et pour l'analyse de sentiments (Positif/Négatif/Neutre) -Modèles d'apprentissage automatique utilisés : --Modèles Bag-of-Words :Machines à vecteurs de support (SVM),Forêts aléatoires (Random Forests),Régression logistique --Modèles de transformeurs pré-entraînés:AraBERT,BERT multilingue (mBERT), XLM-RoBERTa Évaluation des modèles :Validation croisée stratifiée à 5 plis, Calcul des métriques : Précision, Rappel et F1-score	-Scikit-learn -Flair -Hugging Face Transformers -PyTorch	2023	La santé	le dataset « Ara-COVID19-SSD », contenant 5162 tweets arabes	Comprendre les sentiments des gens sur les réseaux sociaux par rapport au COVID-19	Des performances élevées sur les deux tâches de détection du sarcasme et d'analyse des sentiments avec des scores F1 dépassant 0,89 pour tous les modèles.
Colace Francesco (2019) [35]	-Collecte de données Prétraitement des données :(Tokenisation Suppression des mots vides (stopwords), Étiquetage des parties du discours (Part-of-Speech Tagging))	Senti-WordNet Stanford POS Tagger Scikit-learn	2019	Education	Des commentaires des étudiants sont collectés via	analyser les questions de type descriptif des étudiants pour	Des performances élevées avec précision de 90 %.

Approche hybride combinant une approche et une approche d'apprentissage automatique supervisé	-Extraction des caractéristiques (features) :Mots à sentiment positif,Mots à sentiment négatif -Réduction des caractéristiques : Utilisation des techniques d'Information Gain et Gain Ratio. Entraînement du modèle :Random Forest, Support Vector Machines (SVM) -Classification des phrases en sentiments positifs, négatifs ou neutres à l'aide des modèles entraînés. -Affichage des résultats finaux sous forme graphique pour faciliter la compréhension des polarités.	APIs Java			un portail étudiant en ligne.	augmenter la précision du système de rétroaction.	
Lin Fangyuan (2023) [33] Approche hybride (apprentissage automatique supervisé et non supervisé)	-Collecte des données -Nettoyage des données (Suppression des éléments non textuels (balises HTML, URL, émoticônes) avec NLTK et les expressions régulières) -Prétraitement du texte (Tokenisation, Radicalisation (stemming) avec PorterStemmer-Lemmatisation avec WordNet Lemmatizer, Suppression des mots vides (stop words), Vectorisation avec Bag of Words ou TF-IDF) -Calcul des sentiments(Entraînement de modèles d'apprentissage automatique (Naïve Bayes, SVM, RNN), Classification en sentiments positifs, neutres ou négatifs et Prise en compte des principes de psychologie de l'éducation -Extraction des opinions (Modélisation des sujets avec LDA (Latent Dirichlet Allocation), Interprétation contextuelle des sujets guidée par la psychologie de l'éducation, Association sentiment-sujet) -Visualisation des résultats (Visualisation interactive des sujets avec pyLDAvis, Visualisation des associations sentiment-sujet avec des cartes thermiques) -Évaluation (Métriques comme la précision, le rappel, le F1-score, l'exactitude, l'AUC-ROC)	Bibliothèques Python comme NLTK, scikit-learn, gensim, seaborn	2023	Éducation	Contenu de communication en ligne provenant d'apprenants de multiples cours sur des plateformes comme Moodle, Coursera, EdX, etc.	analyse nuancée des sentiments et opinions des apprenants dans les environnements d'éducation en ligne	Bons résultats avec une précision de 89,6% pour la classification des sentiments et d'identifier les aspects spécifiques des cours

Kechaou Zied et al (2011) [38] Approche hybride combinant apprentissage machine supervisé et non supervisé	-Prétraitement du texte (Suppression des mots vides (stop words), Racinisation (stemming) avec l'algorithme de Porter -Sélection des caractéristiques (features) (Calcul du TF-IDF (Term Frequency-Inverse Document Frequency), Sélection des termes les plus discriminants avec 3 méthodes : Information Gain (IG), Mutual Information (MI), Chi-Square (CHI) -Apprentissage supervisé (Modèles de Markov cachés (HMM), Machines à vecteurs de support (SVM)) -Combinaison des classificateurs HMM et SVM avec différentes règles	Toolbox de David G. Stork et Elad Yom Tov pour les HMM, SVMlight pour les SVM	2011	Education	Corpus de 2000 revues d'e-learning (1000 positives, 1000 négatives) extraites de blogs et forums	analyse automatique des sentiments exprimés dans les blogs et forums d'e-learning	Les performances en termes de précision, rappel et F-mesure n'étaient pas très élevées (entre 0.72 et 0.82 environ)
Colace Francesco (2014) [39] Approche hybride combinant apprentissage non supervisé, supervisé et techniques basées sur lexique.	-Prétraitement (tokenisation, filtrage des mots vides, stemming) -Construction d'une matrice terme-document -Allocation de Dirichlet Latente (LDA) pour obtenir les probabilités mot-sujet et document-sujet -Calcul des probabilités d'occurrence de mots et paires de mots -Sélection des mots racines (roots) et mots agrégés -Construction d'un graphe mixte de termes (mGT) -Détection du sentiment à l'aide du graphe mGT : -Construction de graphes mGT séparés pour les sentiments positifs et négatifs à partir d'un ensemble d'entraînement étiqueté -Algorithme de classification du sentiment basé sur la correspondance des mots du document avec les mots racines et agrégés des graphes mGT positifs et négatifs -Utilisation d'une ressource lexicale (ex: WordNet) pour récupérer les synonymes	Plateforme e-learning Moodle WordNet comme lexique annoté	2014	Education	Dataset standard de critiques de films Dataset réelles d'un cours en ligne (forums, chats)	détecter les émotions et sentiments des étudiants dans un environnement d'apprentissage en ligne (e-learning)	Obtenu une précision de 88,5%, supérieure à d'autres méthodes comparées.

Clarizia Fabio et al (2018) [40] Approche hybride combinant machine learning non supervisé (LDA) et règles définies	-Prétraitement des données (tokenisation, filtrage des mots vides (stop words) et radicalisation (stemming)) -Construction d'une matrice terme-document -Utilisation de l'allocation de Dirichlet latente (LDA) -Calcul des probabilités d'occurrence des mots, probabilités conditionnelles entre paires de mots et probabilités jointes entre paires de mots -Sélection des racines agrégées (mots les plus impliqués par l'occurrence d'autres mots) -Sélection des mots agrégés liés aux racines agrégées selon les poids probabilistes les plus élevés -Construction d'un graphe de termes mixtes (mixed graph of terms, mGT) contenant les racines agrégées liées aux mots agrégés -Optimisation du mGT -Utilisation d'un lexique -Comparaison du document d'entrée avec les mGT d'orientation positive et négative -Classification du sentiment en positif, négatif ou neutre	Plateformes e-learning comme Moodle	2018	Education	-Dataset standard de critiques de films -Tweets -Commentaires Facebook -Posts de forums e-learning (Moodle)	Detection les sentiments exprimés par les étudiants dans les forums, chats et autres outils collaboratifs des plateformes e-learning,permettre à l'enseignant d'avoir un aperçu de l' "ambiance" de la classe	bonnes performances avec : plus de 82% pour Accuracy , plus de 78% pour Recall et 0,83 pou F score
Nandal Neha (2020) [41] Apprentissage automatique supervisé	-Développement d'un crawler web basé sur Scrapy pour extraire les avis de clients d'Amazon -Identification manuelle des termes d'aspect liés aux caractéristiques du produit, Agrégation des aspects (regroupement des termes synonymes) -Prétraitement des données (Vectorisation, Étiquetage des parties du discours (Part-of-Speech tagging) ,Stemming et lemmatisation, Suppression des mots vides) -Identification des mots bipolaires -Classification et évaluation (Utilisation des machines à vecteurs de support (SVM) comme classificateur et évaluation avec différentes métriques : taux d'apprentissage, erreur quadratique moyenne (MSE), précision, rappel, matrice de confusion, courbes ROC -Approche d'ensemble pour améliorer la précision et l'efficacité, Mappage des sentiments aux notes	Python, Scrapy, Matlab	2020	e-commerce	Avis d'utilisateurs Amazon collectés pour différents produits	Analyse de sentiments au niveau des aspects spécifiques d'un produit mentionnés dans les avis	Atteint des performances élevées taux d'apprentissage de 97%).
Ansari Mohd Zeeshan (2020) [34] Apprentissage supervisé (machine learning et deep learning)	-Collecte des données -Pré-traitement (Suppression des mentions, hashtags, émoticons et ponctuation non conventionnelle, Élimination des tweets en double et des retweets, Suppression des mots vides (stopwords) Mise en minuscule (case folding)) -Annotation manuelle des tweets dans 8 classes de sentiments politiques par 3 annotateurs, Calcul de l'accord inter-annotateurs -Extraction des caractéristiques (TF-IDF pour les unigrammes, bigrammes et trigrammes) -Machines à vecteurs de support (SVM), Arbre de décision (Decision Tree), Régression logistique, Forêt aléatoire (Random Forest), Réseau de neurones récurrent à mémoire à long et court terme (LSTM) -Évaluation : Validation croisée stratifiée à 10 plis Mesures de performance : précision, rappel, f1-score, exactitude -Analyse des résultats	Non spécifié	2020	Politique	3896 tweets liés aux élections générales indiennes de 2019	capturer les orientations des sentiments politiques sur les réseaux sociaux en vue des élections	une précision de 0,76 et un score F1 de 0,74 avec les réseaux LSTM une précision de 0,77 avec les forêts aléatoires Les performances des modèles classiques comme les SVM étaient inférieures

Table 3.1 – État de l'art des travaux connexes

Les travaux étudiés concernent majoritairement le domaine d'éducation, mais il existe d'autres domaines tels que la santé, e-commerce et poliqtique, ils tentent d'analyser les données d'avis d'étudiants, de tweets et de commentaires en ligne, et d'obtenir de meilleurs résultats en termes de précision, de rappel et d'exactitude pour la classification des sentiments et la détection des émotions. Plusieurs approches ont été utilisées, comme l'apprentissage automatique supervisé avec des algorithmes comme les SVM, les forêts aléatoires, la régression logistique et les réseaux de neurones (RNN, LSTM), ainsi que l'apprentissage non supervisé comme le clustering K-means et l'allocation latente de Dirichlet (LDA). Des approches hybrides combinant ces deux types d'apprentissage ont également été explorées. Des techniques de prétraitement comme la tokenisation, la suppression des mots vides, la lemmatisation et la vectorisation (TF-IDF, Word Embedding) ont également été appliquées.

Les avantages majeurs des approches supervisées comprennent leur capacité à classer efficacement les sentiments grâce à des algorithmes performants comme les forêts aléatoires et SVM. Les modèles avancés comme les LSTM excellent à capturer les nuances du langage et les dépendances à longue distance dans le texte. Quant aux approches non supervisées, elles permettent de découvrir des motifs et des thèmes cachés dans les données sans annotations préalables, comme l'ont montré les techniques de clustering. Les approches hybrides tirent parti des forces combinées du supervisé et du non supervisé.

Cependant, ces approches présentent plusieurs inconvénients. L'apprentissage supervisé dépend de la disponibilité de grands ensembles de données annotées, ce qui est coûteux. Les modèles complexes comme les réseaux de neurones profonds souffrent d'un manque d'interprétabilité, étant qualifiés de "boîtes noires". Certaines techniques comme les SVM font l'hypothèse que les caractéristiques d'entrée sont indépendantes, ce qui n'est pas toujours valide dans le langage naturel. Enfin, les approches hybrides ajoutent de la complexité due à la combinaison de plusieurs techniques à pondérer correctement.

3.4 Conclusion

Dans ce chapitre, nous avons exploré l'état de l'art en matière d'analyse des sentiments, en commençant par une introduction qui a posé les bases de notre discussion. Nous avons passé en revue divers travaux connexes, mettant en lumière les recherches et les approches précédemment entreprises dans ce domaine. Ces travaux nous ont permis de comprendre les différentes méthodes et techniques utilisées par d'autres chercheurs pour aborder l'analyse des sentiments.

Nous avons également mené une étude comparative et une analyse approfondie des différentes approches. Cette analyse nous a permis d'identifier les avantages et les limites de chaque méthode, ainsi que les tendances actuelles et les défis persistants dans le domaine.

Ce chapitre fournira ainsi une base solide pour situer notre propre contribution et justifier les choix méthodologiques adoptés dans notre étude.

4 Chapitre 4

Analyse des sentiments sur les videos d'éducation

Sommaire

4.1 Introduction

Avec des millions de vues, YouTube est l'un des sites de partage de vidéos les plus utilisés. Avec la popularité toujours croissante des vidéos en ligne et la croissance exponentielle du contenu généré par les utilisateurs, comprendre la qualité et la pertinence du contenu est devenu crucial pour les téléspectateurs en examinant manuellement les commentaires, le nombre de vues et le nombre de likes. Les commentaires et les débats de la plateforme fournissent une mine d'informations qui peuvent être utilisées pour étudier le sentiment et l'opinion du public sur une série de sujets. Cela fait des données YouTube une ressource inestimable pour un large éventail de domaines universitaires. De plus, l'énorme bibliothèque de contenu de YouTube, qui comprend tout, des actualités et vlogs personnels aux divertissements et tutoriels, en fait un outil inestimable pour les universitaires, les entreprises, les éducateurs et les spécialistes du marketing souhaitant exploiter la richesse des données produites par les utilisateurs et leurs interactions avec la plateforme.

Récemment, l'utilisation de YouTube comme outil éducatif a augmenté en raison de la pandémie de covid-19. La qualité des vidéos pédagogiques est cruciale dans le processus d'apprentissage. Les commentaires des utilisateurs sur les vidéos éducatives peuvent aider à déterminer la qualité des vidéos et aident les universitaires et les institutions à évaluer l'efficacité des tactiques pédagogiques et de l'apprentissage en ligne. Ces commentaires peuvent être utilisés à l'aide d'une technique de traitement du langage naturel appelée analyse des sentiments.

Dans ce chapitre, nous nous concentrerons sur l'application pratique de l'analyse des sentiments aux vidéos éducatives. Nous présenterons tout d'abord l'approche proposée pour cette analyse, en détaillant les différentes étapes du processus. Cela inclut la collection des

données, leur prétraitement, et l'annotation des commentaires pour en extraire les sentiments exprimés.

Ensuite, nous décrirons les étapes supplémentaires, telles que l'analyse du vocabulaire utilisé dans les commentaires, la création d'un lexique des émotions spécifique au contexte éducatif, et la visualisation des données collectées. Ces étapes visent à enrichir l'analyse et à fournir des insights plus détaillés sur les sentiments exprimés par les utilisateurs.

Enfin, nous aborderons la classification des sentiments à l'aide de plusieurs modèles d'apprentissage automatique. Nous comparerons l'efficacité des forêts aléatoires (Random Forest), des machines à vecteurs de support (Support Vector Machines, SVM) et des modèles de Naive Bayes pour déterminer les sentiments des commentaires. Cette comparaison nous permettra d'identifier le modèle le plus performant pour notre analyse spécifique.

4.2 Approche proposée

Notre projet consiste en l'analyse des sentiments sur les commentaires des vidéos YouTube d’éducation en Algérie. Pour cela, plusieurs étapes doivent être effectuées pour obtenir de meilleurs résultats.

Voici une figure qui donne un aperçu de l'approche proposée et des différentes étapes qui la composent : collecte des données, annotation, Classification des sentiments: utiliser des classifieurs tels que Random Forest, SVM et Naive Bayes pour prédire les sentiments des commentaires et en utilisant un lexique spécifique à l'algérien mots et expressions avec deux types de sentiments :positif et negatif, Évaluation avec plusieurs métriques :

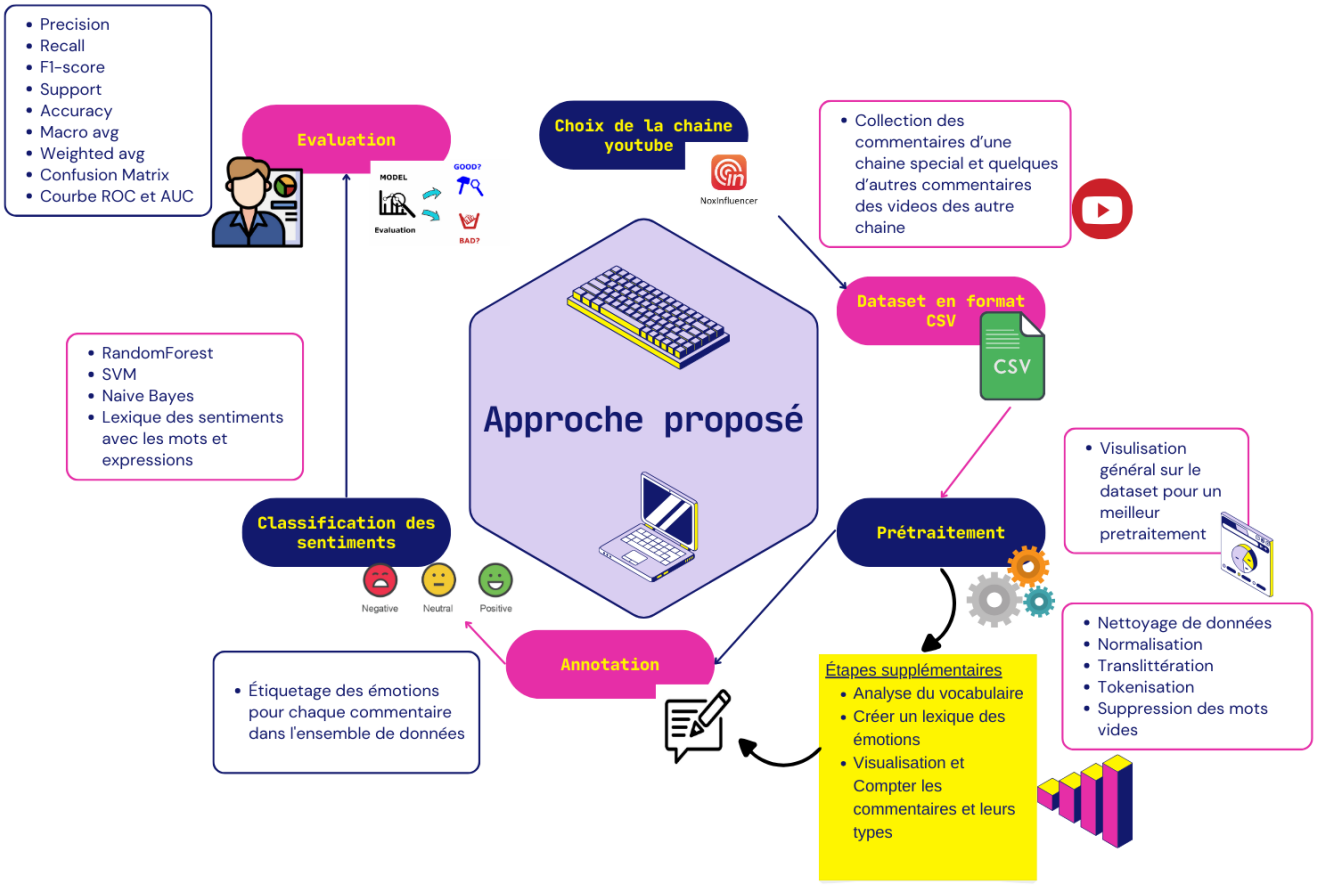

Figure 4.1 – Schéma globale de l'approche proposé

Nous détaillons ci-dessous chaque étape comme suit :

4.2.1 Collection des données

La première étape que nous avons suivie pour créer l'ensemble de données consistait à rechercher la chaine d'enseignement scolaire algérienne la plus populaire sur youtube (en particulier enseignement secondaire), pour cela nous avons utilisé un site Web appelé noxinfluencer. NoxInfluencer est un logiciel complet de marketing d'influence pour la découverte d'influenceurs, l'exécution de campagnes marketing, la gestion des relations avec les influenceurs et l'analyse des médias sociaux. Il couvre plus de 20 millions de ressources d'influenceurs dans le monde sur YouTube, TikTok et Instagram [42].

Il nous aide à déterminer la chaîne d'éducation la plus populaire en Algérie en termes de :

- **Sorting by NoxScore (Tri par NoxScore):** Noxscore est la façon dont NoxInfluencer Data System considère les influenceurs selon cinq facteurs différents et donne 0 à 5 étoiles en utilisant des calculs avec ces 5 facteurs, pour plus d'informations sur les méthodes de calcul utilisées, vous pouvez consulter les lien [43].

- **Sorting by Subscribed(Trier par Abonné) :** Triés du grand nombre d'abonnements au moins.
- **Sorting by AVG.Views(Triés par spectateurs moyens):** Prendre sa durée totale de visionnage et la diviser par le nombre de lectures vidéo.
- **Sorting by Growth(Tri par développement):** Plus la chaîne se développe rapidement au cours des 30 derniers jours, plus le score est élevé.
- **Sorting by unsubscribed(Tri par désabonnement):** Trié du petit nombre d'abonnements au plus grand.
- **Sorting by Monthly Views(Tri par vues mensuelles):** C'est le nombre de vues par mois

Nous avons pris en compte uniquement : les vues mensuelles, l'abonnement, NoxScore, spectateurs moyens, car le reste est uniquement destiné pour intéresser à les chaînes non populaires.

Après avoir choisi le pays d'Algérie et la catégorie d'éducation et l'éducation pour la catégorie Noxscore, nous avons constaté que la chaine " الاستاذ نورالدين " s'est plutôt bien classée par rapport aux autres chaînes, voici des captures d'écran du classement de cette chaîne sur le site NoxInfluencer :

- La chaine s'est classée sixième par Noxscore :

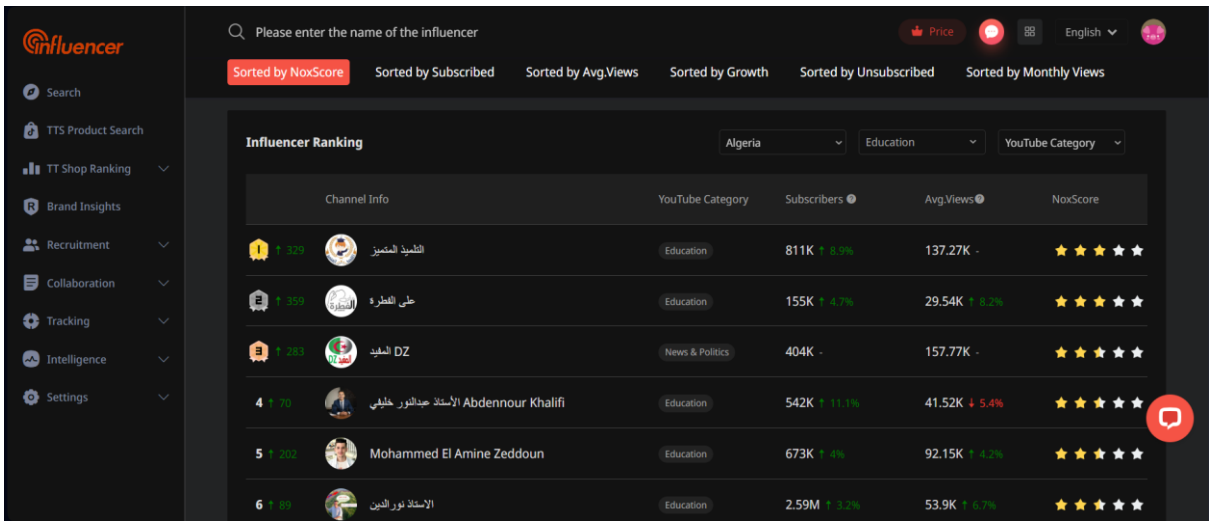

Figure 4.2 – Capture d'écran du site NoxInfluencer pour Sorted by NoxScore

- Elle s'est classée cinquième par spectateurs moyens **:**

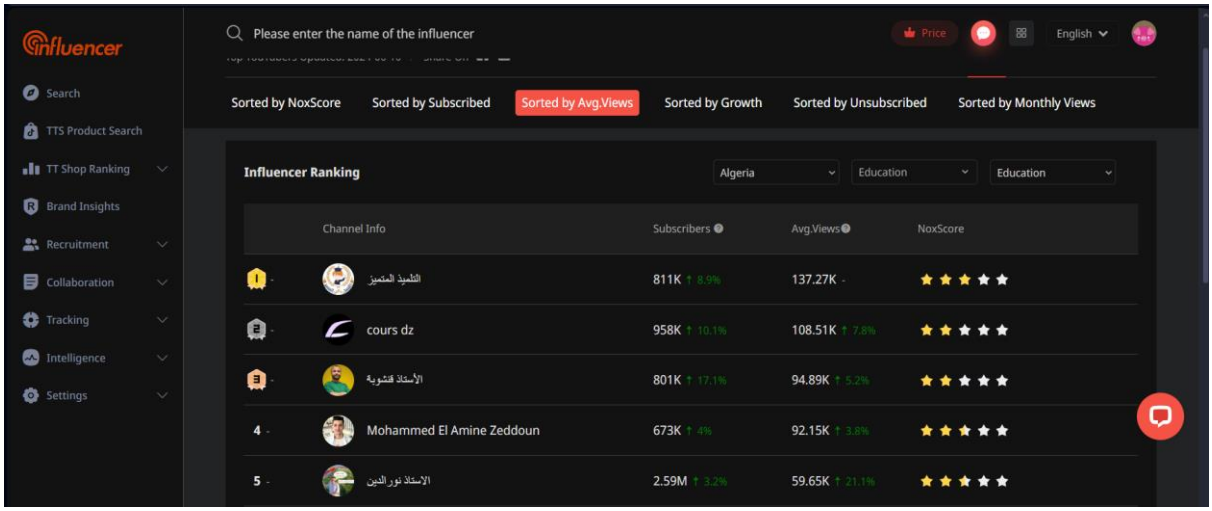

Figure 4.3 – Capture d'écran du site NoxInfluencer pour Sorted by AVG.Views

- Elle s'est classée deuxième par vues mensuelles :

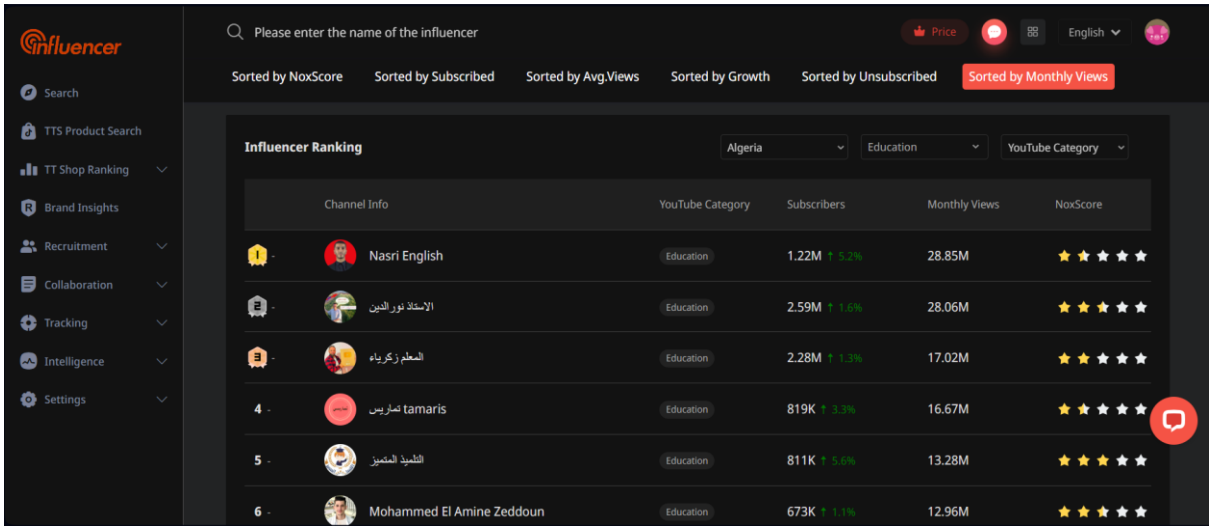

Figure 4.4 – Capture d'écran du site NoxInfluencer pour Sorted by Monthly Views

- Elle s’est classée première par les abonnés:

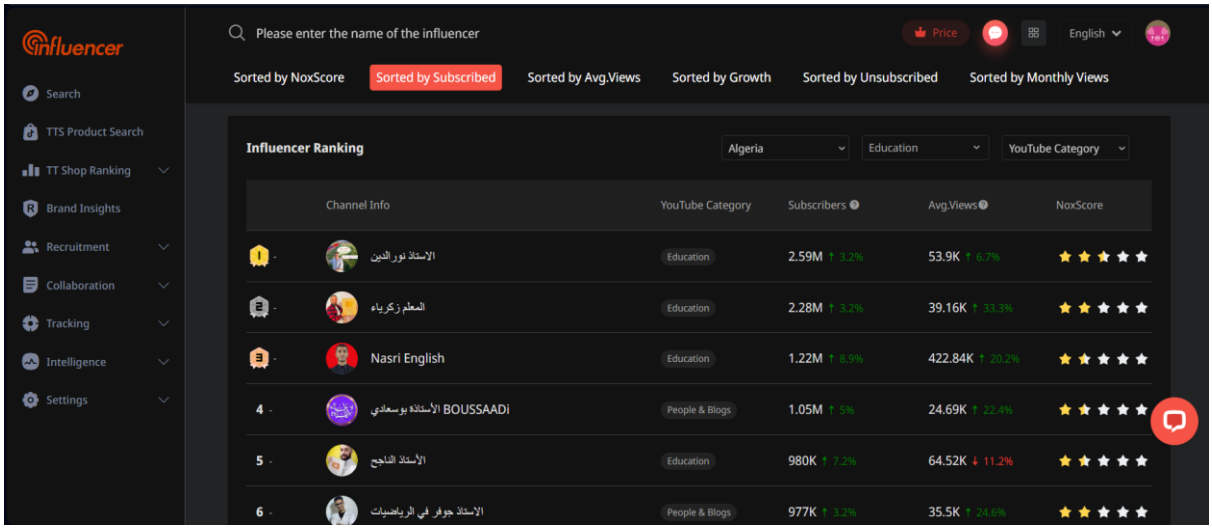

Figure 4.5 – Capture d’écran du site NoxInfluencer pour Sorted by Subscribed

Nous avons donc choisi cette chaine pour collecter notre ensemble de données à l'aide de l'API Youtube via la programmation Python, ce qui a donné un ensemble de données au format *.csv.

L'API YouTube est un ensemble d'outils fournis par YouTube qui permettent aux développeurs d'accéder et d'interagir avec diverses fonctionnalités de YouTube par programmation, telles que la récupération de détails sur les vidéos (tels que les titres, les descriptions, le nombre de vues et les mesures d'engagement), l'accès aux listes de lecture et à leur contenu, et recueillir des informations sur les chaînes YouTube. Cet accès permet la création d'applications diverses, depuis les agrégateurs vidéo et les moteurs de recommandation de contenu jusqu'aux outils d'analyse pour les créateurs de contenu et les spécialistes du marketing [38].

Il existe des limites d'utilisation et des quotas pour garantir une utilisation équitable, comme un quota par défaut de 10 000 unités par jour, chaque requête API consommant un certain nombre d'unités en fonction de sa complexité.

Après avoir collecté les commentaires et afin de garantir que les informations des données soient riches, nous avons effectué une analyse superficielle manuellement, et nous avons remarqué la grande présence de commentaires positifs et neutres par rapport aux négatifs, ce qui pourrait nous poser un problème dans le étapes suivantes, nous avons donc recherché des vidéos spécifiques (peuvent être consultées [45]) provenant de différentes sources (A partir d'autres chaînes éducatives algériennes) pour enrichir les ensembles de données avec plus de commentaires négatifs, puis nous avons combiné les deux ensembles de données en utilisant le code Python.

4.2.2 Prétraitement

Puisque notre ensemble de données est en dialecte algérien, cela signifie qu'il y a des commentaires avec des alphabets en arabe ou de latin alphabet qui est connu sous le nom d'Arabizi, aussi le même commentaire peut avoir les deux : des arabes alphabets et des latins.

Arabizi est un système d'écriture informel qui utilise des caractères et des chiffres latins pour représenter l'écriture arabe, souvent utilisée dans la communication numérique telle que les SMS et les réseaux sociaux. Il est apparu à la fin des années 1990 et au début des années 2000, facilité par la diffusion d'une technologie qui, initialement, ne prenait pas en charge l'écriture arabe. Arabizi combine des éléments d'arabe et d'anglais (le français en Algérie) et est particulièrement populaire auprès des jeunes générations dans des contextes informels.

En employant Arabizi, les lettres arabes sont transcrites en lettres ou chiffres latins en fonction de leurs phonèmes et formes correspondants. Par exemple, la lettre « ع », qui n'existe pas phonétiquement en anglais, est transcrite sous la forme du chiffre « 3 » en raison de sa forme similaire. Les sons comme « س » et « ش » sont respectivement transcrits en « s » et « sh », car ils se prononcent phonétiquement de la même manière [46].

Cette forme d'arabe romanisé n'existait pas avant la création et les progrès d'Internet. Vers la fin du XXe siècle, les services de messages courts (SMS), le courrier électronique et les plateformes de chat instantané se sont répandus dans le monde arabe. Cependant, comme de nombreux premiers systèmes électroniques ne pouvaient pas prendre en charge les écritures non latines [46].

Nous avons effectué un prétraitement spécifique afin de distinguer les émotions de manière appropriée et fiable en utilisant des techniques de programmation en Python :

Nettoyage de données (Text Cleaning):

Text cleaning, également appelé nettoyage de données ou nettoyage de données, est le processus de préparation de données textuelles brutes pour un traitement et une analyse ultérieurs. C'est une étape cruciale car elle impacte directement les performances du modèle. Plus les données sont propres et structurées, plus le modèle peut en tirer des leçons [47].

Il nous aide à améliorer la qualité de nos données textuelles en :

1. Supprime les signes diacritiques du latin et de l'arabe : (Dédiacritisation), Les signes diacritiques sont des symboles qui apparaissent au-dessus ou en dessous des lettres. Ces signes diacritiques sont supprimés afin de réduire la rareté des données [32]. Par exemple : "école" devient "ecole".
2. Éliminer les informations non pertinentes telles que les balises HTML, les hashtags, les URL et autres caractères spéciaux comme les emojis.
3. Supprime les espaces supplémentaires qui incluent les espaces, les tabulations, les nouvelles lignes et autres caractères similaires. Il garantit que tout le texte est uniformément espacé, réduisant ainsi le bruit dans l'ensemble de données, ce qui contribue à améliorer la tokenisation et facilite son traitement et son analyse.
4. Gère les caractères répétés et les lettres simples isolées pour réduire le bruit.

Normalisation(Normalization):

La normalisation du texte consiste à réduire les variations des formes de mots à une forme standard lorsque ces variations ont la même signification. Il n'y a pas de liste fixe de

tâches incluses dans la normalisation du texte, car elles varient en fonction des besoins de l'application.

La normalisation du texte contribue à améliorer la qualité de nos données textuelles en :

1. Minuscules : les données textuelles peuvent apparaître dans différents cas (par exemple, "Bien", "bien", "BIEN"). La mise en minuscules convertit tous les caractères en casse uniforme (généralement en minuscules), garantissant ainsi la cohérence dans l'ensemble de données, et réduit également la taille du vocabulaire. Par exemple, « Bien » et « bien » seront considérés comme le même mot plutôt que comme deux mots distincts et cela aide à faire correspondre les mots quelle que soit leur casse d'origine dans le texte.
2. Standardiser certaines lettres pour réduire la variabilité, car différentes formes de la même lettre peuvent augmenter inutilement la complexité et la taille du vocabulaire et une représentation cohérente des lettres garantit que les mots sont symbolisés correctement et de manière cohérente. Par exemple : 'أ ' et 'آ ' devient 'ا'.
3. Remplace les chiffres spécifiques couramment utilisés dans le texte arabe en ligne (comme dans le dialecte algérien) par les lettres latines correspondantes basées sur un mappage prédéfini. Ces chiffres représentent souvent des sons arabes qui n'ont pas d'équivalents latins directs, ils sont donc remplacés par des caractères latins qui se rapprochent phonétiquement des sons arabes.

Translittération :

Translittère le texte arabe en écriture latine : uniformité de la représentation du texte pour le rendre plus facile à manipuler.

Tokenisation :

La tokenisation est une étape de prétraitement cruciale dans l'analyse des sentiments au sein du traitement du langage naturel (NLP) et du NLTK (Natural Language Toolkit). Cela implique de décomposer le texte en unités plus petites, appelées jetons, qui peuvent être des mots, des expressions ou des phrases. La tokenisation aide à comprendre et à analyser le texte en simplifiant la structure et en facilitant le traitement des algorithmes.

Modèles de tokenisation :

- Tokenisation Unigram : Divise le texte en mots individuels.

  Exemple : "L'éducation est essentielle" ➔ ["L'éducation", "est", "essentielle"]

  Ce type de modèle est facile à mettre en œuvre et à comprendre, il permet une analyse individuelle au niveau des mots, ce qui est utile pour l'analyse de base des sentiments.

- Tokenisation Bigramme: Divise le texte en paires de mots consécutifs.

  Exemple : "L'éducation est essentielle" ➔ [("L'éducation", "est"), ("est", "essentielle")]

  Il capture le contexte entre les paires de mots, améliorant ainsi la compréhension des phrases, afin d'améliorer la précision des modèles en prenant en compte les paires de mots.

- Tokenisation Trigramme : Divise le texte en triplets de mots consécutifs.

  Exemple : "L'éducation est essentielle" ➔ [("L'éducation", "est", "essentielle")]

  Il capture plus de contexte en considérant des triplets de mots, utiles pour des relations plus complexes, et améliore les prédictions en comprenant des phrases plus étendues.

- Tokenisation N-gram : Divise le texte en n mots consécutifs.

Exemple pour n=4 : "L'éducation est essentielle pour tous" ➔ [("L'éducation", "est", "essentielle", "pour"), ("est", "essentielle", "pour", "à nous")]
Il peut être ajusté pour capturer autant de contexte que nécessaire. Il est également utile pour une analyse très détaillée lorsque les dépendances à long terme sont importantes.

Dans notre cas, la tokenisation unigramme est utilisée, car elle est bénéfique pour l'analyse de base des sentiments en fournissant un moyen simple d'analyser le sentiment de chaque mot, pour l'extraction de fonctionnalités en permettant l'extraction de mots individuels en tant que fonctionnalités pour les modèles d'apprentissage automatique et pour plus de simplicité, plus rapide à calculer et nécessite moins de ressources de calcul par rapport aux n-grammes d'ordre supérieur.

Suppression des mots vides (Stopword Removal):

Il s'agit du processus de suppression de termes qui apparaissent souvent dans tous les textes du corpus. Ces mots sont abondants dans toutes les langues humaines. En les supprimant, nous éliminons toutes les données de bas niveau de notre texte, permettant ainsi de nous concentrer davantage sur les informations cruciales [32].

Dans cette étude, nous avons créé nos propres documents de mots vides pour garantir que les mots susceptibles de provoquer une ambiguïté dans la compréhension du sens sémantique et réel du commentaire ne soient pas supprimés.

Afin de garantir que l'ensemble de données contient le plus grand nombre possible de mots importants et de supprimer les mots qui ne sont pas importants et qui peuvent causer des problèmes lors du processus de classification des sentiments, nous avons suivi les étapes suivantes :

1. Téléchargé un document prêt à l'emploi qui contient des mots vides avec des lettres arabes pour: le dialecte algérien et la langue arabe [48].
2. Translittération de ces mots vides en lettres latines pour maintenir la cohérence dans la représentation du texte.
3. Téléchargé un document prêt à l'emploi contient des mots vides en français [49].
4. Combinaison les deux documents de mots vides.
5. Analysé et vérifié manuellement pour garantir qu'aucun mot important pertinent pour notre étude n'a été inclus et supprimer les redondantes.
6. Ajout de certains des mots les plus fréquemment utilisés mais sans importance de dataset aux documents des mots vides.

4.2.3 L'annotation

L'annotation de sentiment consiste à attribuer un label de sentiment (positif, négatif, neutre) à chaque commentaire du dataset. Cette étape est essentielle pour construire et évaluer des modèles d'analyse de sentiment. Par exemple, un commentaire comme "J'adore cette vidéo éducative" serait annoté comme positif, tandis que "Je déteste cette leçon" serait négatif. Les commentaires neutres, comme "Cette vidéo traite de l'éducation", ne contiennent pas d'émotion marquée. Cette annotation peut être faite manuellement ou à l'aide de modèles automatiques, et elle permet de mieux comprendre les opinions des utilisateurs et d'améliorer les services en conséquence.

4.2.4 Étapes supplémentaires

I. Analyse du vocabulaire

L'analyse du vocabulaire est cruciale pour comprendre les données textuelles, identifier le bruit et préparer un prétraitement efficace. Les étapes impliquées comprennent :

1. Compte la fréquence et les occurrences de chaque mot dans le texte, dans le but d'identifier les mots les plus courants et les plus rares dans l'ensemble de données. Il aide à construire des listes de mots vides (mots courants qui peuvent ne pas ajouter de valeur à l'analyse) et des dictionnaires de sentiments (listes de mots positifs et négatifs). Cette étape est essentielle pour la sélection des fonctionnalités et la réduction de la dimensionnalité dans l'analyse de texte.
2. Calcule la taille du vocabulaire avant le prétraitement, dans le but de fournir une mesure de base de la diversité et du bruit dans les données textuelles brutes.
3. Calcule la taille du vocabulaire après prétraitement, dans le but de mesurer l'impact des étapes de prétraitement sur le vocabulaire.

   En comparant la taille du vocabulaire avant et après le prétraitement, on peut mesurer l'impact des étapes de prétraitement. Cette comparaison permet d'évaluer l'efficacité de ces étapes pour réduire le bruit et améliorer la qualité des données.

II. Créer un lexique des émotions (dictionnaire des émotions)

Un lexique des émotions, également appelé dictionnaire des émotions, est un ensemble de mots ou d'expressions associés à des émotions spécifiques. Chaque entrée du lexique est étiquetée avec une catégorie d'émotion (par exemple positive, négative, joie, colère, etc.) pour aider à la classification et à l'analyse des données textuelles basées sur le contenu émotionnel.

À partir de notre ensemble de données, à l'aide de la fonction qui calcule les mots les plus fréquemment utilisés, nous sélectionnons manuellement les mots positifs et négatifs, puis nous créons un fichier texte contenant ces mots catégorisés selon leur sentiment. Ce dictionnaire de sentiments est utilisé dans les tâches d'analyse des sentiments pour classer le texte en fonction de la présence de mots positifs et négatifs. Le but de cette étape est :

- Aider le modèle à découvrir, lors de son processus d'apprentissage sur l'ensemble de données d'entraînement, les mots qui ont rendu la phrase positive ou négative.
- Augmentez la quantité d'exemples appris manuellement sans affecter l'ensemble de données d'origine.
- Augmenter les chances de prédire correctement le sentiment des commentaires.

III. Visualisation et comptage les commentaires et leurs types

Notre script compte le nombre de commentaires dans chaque catégorie (latin, arabe, mixte, contenant des emojis, uniquement des emojis, pas d'emojis et nombre général des commentaires).

Cette étape a pour objectif de comprendre la distribution des différents types de commentaires et d'aider à adapter les stratégies de prétraitement et de modélisation.

Nous avons utilisé un diagramme circulaire pour représenter visuellement la répartition des commentaires selon leurs caractéristiques afin de permettre de comprendre rapidement la composition de l'ensemble de données et d'identifier les catégories dominantes.

4.2.5 Classification des sentiments

Les deux figures ci-dessous représentent les résultats l'exécution du script Python pour l'entraînement du modèle :

```
Sentiment distribution before resampling:
positif    200
neutral    113
negatif     37
Name: Sentiment, dtype: int64
Positive words: 69
Negative words: 12
Positive phrases: 0
Negative phrases: 48
Modified dataset saved to 'modified_comments_data.txt'
Total positive and negative words saved to 'total_positive_words.txt' and 'total_negative_words.txt'
Resampled Sentiment Distribution:
positif    200
neutral    200
negatif    200
Name: Sentiment, dtype: int64
```

Figure 4.6 – Résultat de l'exécution du script Python pour l'entraînement du modèle

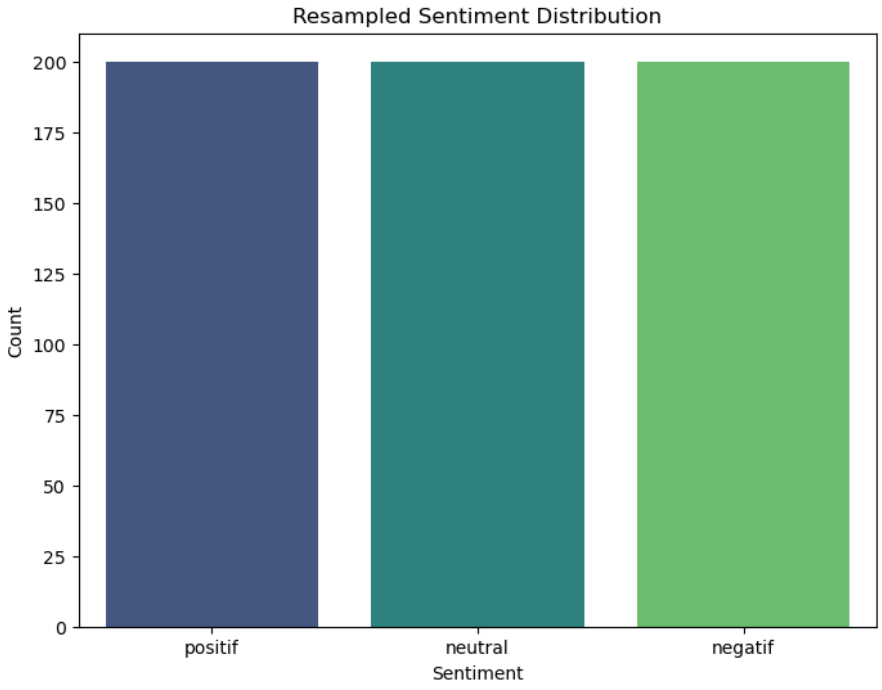

Figure 4.7 – Histogramme de distribution du sentiment rééchantillonnage

Lorsque l'on travaille avec des ensembles de données déséquilibrés, une ou plusieurs classes peuvent être largement surreprésentées par rapport aux autres. Par exemple, dans notre ensemble de données initial, la distribution des sentiments est la suivante :

- Positif : 200
- Neutre : 113
- Négatif : 37

Cela signifie que les classes "neutre" et "négatif" sont sous-représentées par rapport à la classe "positif". Ce déséquilibre peut poser des problèmes pour les algorithmes d'apprentissage automatique, car ils peuvent avoir tendance à privilégier la classe majoritaire, entraînant des biais et une mauvaise performance sur les classes minoritaires.

Après avoir appliqué des techniques de rééchantillonnage, la distribution des sentiments devient :

- Positif : 200
- Neutre : 200
- Négatif : 200

Cela indique que des méthodes telles que le sur-échantillonnage des classes minoritaires ou le sous-échantillonnage de la classe majoritaire ont été utilisées pour équilibrer les classes.

En outre, des statistiques sur les mots et les phrases positifs et négatifs dans le texte montrent:

- Mots positifs : 69
- Mots négatifs : 12
- Phrases positives : 0
- Phrases négatives : 48

Ces statistiques peuvent être importantes pour analyser la polarité du texte et comprendre les sentiments exprimés.

I. Les forêts aléatoires (Random Forest)

En machine learning, les forêts aléatoires (Random Forest) sont des modèles d'apprentissage supervisé qui utilisent une combinaison d'arbres de décision pour améliorer la précision des prédictions et réduire le surapprentissage (overfitting). Cet algorithme est utilisé pour la classification et la régression, et il fonctionne en construisant de multiples arbres de décision lors de l'entraînement et en sortant la classe qui est le mode des classes (classification) ou la moyenne des prédictions (régression) des arbres individuels.

Dans l'algorithme de Random Forest, chaque élément de données est représenté par ses caractéristiques qui servent d'entrées pour les arbres de décision. Pour traiter les textes, ceux ci sont d'abord vectorisés en utilisant le TF-IDF Vectorizer, transformant chaque document en un vecteur de poids TF-IDF. De plus, des caractéristiques supplémentaires basées sur les mots positifs et négatifs identifiés dans les textes sont ajoutées aux vecteurs de caractéristiques.

Le Random Forest fonctionne en construisant de nombreux arbres de décision à partir de sous-échantillons du jeu de données et en utilisant la moyenne pour augmenter la précision prédictive et contrôler le surapprentissage. Chaque arbre de décision est construit en sélectionnant un sous-ensemble aléatoire des caractéristiques à chaque division. La combinaison de ces arbres de décision réduit la variance et conduit à un modèle plus robuste.

Un modèle Random Forest est une collection d'arbres de décision où chaque arbre vote pour une classe donnée et la classe finale attribuée à un nouvel échantillon est celle qui reçoit le plus de votes parmi tous les arbres. Dans notre code, après vectorisation des textes et ajout des nouvelles caractéristiques, les données sont sur-échantillonnées pour équilibrer les classes avant d'entraîner le modèle. Le modèle est ensuite optimisé à l'aide de la validation croisée et de la recherche en grille GridSearchCV (**G**rid **S**earch with **C**ross-**V**alidation) pour trouver les meilleurs hyperparamètres.

Une fois le modèle Random Forest entraîné, il peut être utilisé pour prédire les sentiments des nouveaux exemples de texte. Dans ce projet, les prédictions sont faites non seulement en se basant sur le modèle appris, mais aussi en utilisant des règles basées sur le nombre de mots positifs et négatifs présents dans chaque texte. Les résultats finaux sont ensuite sauvegardés, et le modèle et le vectoriseur sont stockés pour une utilisation future. Les performances du modèle sont évaluées à l'aide de rapports de classification et de matrices de confusion.

II. Les machines à vecteurs de support (Support Vector Machines, SVM)

Les machines à vecteurs de support (Support Vector Machines, SVM) sont des modèles d'apprentissage supervisé utilisés pour la classification et la régression. Elles fonctionnent en trouvant l'hyperplan qui maximise la marge entre les différentes classes dans l'espace des caractéristiques. Les SVM peuvent être utilisés avec différents noyaux (kernel) pour traiter des données non linéaires en les projetant dans un espace de dimensions supérieures.

Dans notre modèle, les textes sont représentés par des vecteurs de caractéristiques utilisant la technique TF-IDF (Term Frequency-Inverse Document Frequency). Cette technique transforme chaque document en un vecteur pondéré basé sur la fréquence des termes, permettant de capturer l'importance relative des mots dans les documents. Des caractéristiques supplémentaires, basées sur le nombre de mots positifs et négatifs dans chaque texte, sont également ajoutées pour enrichir les vecteurs de caractéristiques.

Le SVM fonctionne en trouvant l'hyperplan optimal qui sépare les différentes classes avec la plus grande marge possible. L'algorithme utilise des noyaux (kernels) comme le noyau linéaire ou le noyau radial de base (RBF) pour transformer les données et rendre possible la séparation non linéaire. Le SVM peut également ajuster des hyperparamètres tels que le coefficient de régularisation (C) et le paramètre de noyau (gamma) pour améliorer les performances du modèle.

Le modèle SVM est utilisé pour la classification des sentiments dans les textes. Après la vectorisation des textes et l'ajout des nouvelles caractéristiques, les données sont suréchantillonnées pour équilibrer les classes avant d'entraîner le modèle. Le modèle est optimisé en utilisant GridSearchCV pour trouver les meilleurs hyperparamètres (C, gamma, et kernel). Une fois le modèle entraîné, il peut prédire les sentiments des nouveaux textes en utilisant les caractéristiques textuelles transformées et les nouvelles caractéristiques ajoutées.

Une fois le modèle SVM optimisé et entraîné, il est utilisé pour prédire les sentiments des textes. Les prédictions sont faites en utilisant les vecteurs de caractéristiques textuelles et les nouvelles caractéristiques ajoutées. Le modèle et le vectoriseur sont sauvegardés pour une utilisation future, et les performances du modèle sont évaluées à l'aide de rapports de classification et de matrices de confusion. Les résultats finaux sont sauvegardés dans des fichiers pour une analyse ultérieure.

III. Les modèles de Naive Bayes

Les modèles de Naive Bayes sont des classificateurs probabilistes basés sur l'application du théorème de Bayes avec une forte (naïve) hypothèse d'indépendance entre les caractéristiques. Ils sont particulièrement efficaces pour les problèmes de classification de texte, comme la classification des sentiments, où chaque mot est traité comme une caractéristique indépendante.

Dans notre modèle, les textes sont représentés par des vecteurs de caractéristiques utilisant la technique TF-IDF (Term Frequency-Inverse Document Frequency). Cette technique transforme chaque document en un vecteur pondéré basé sur la fréquence des termes, permettant de capturer l'importance relative des mots dans les documents. Des caractéristiques supplémentaires, basées sur le nombre de mots positifs et négatifs dans chaque texte, sont également ajoutées pour enrichir les vecteurs de caractéristiques.

Le modèle Naive Bayes, et plus particulièrement le Multinomial Naive Bayes, est utilisé ici. Il est bien adapté pour les données de comptage, comme les fréquences de mots dans des documents textuels. L'algorithme calcule la probabilité qu'un document appartienne à chaque classe et choisit la classe avec la probabilité la plus élevée. Les hyperparamètres, comme le paramètre alpha (utilisé pour le lissage de Laplace), peuvent être ajustés pour améliorer les performances.

Le modèle Naive Bayes est utilisé pour la classification des sentiments dans les textes. Après la vectorisation des textes et l'ajout des nouvelles caractéristiques, les données sont suréchantillonnées pour équilibrer les classes avant d'entraîner le modèle. Le modèle est optimisé en utilisant GridSearchCV pour trouver les meilleurs hyperparamètres (comme alpha). Une fois le modèle entraîné, il peut prédire les sentiments des nouveaux textes en utilisant les caractéristiques textuelles transformées et les nouvelles caractéristiques ajoutées.

Une fois le modèle Naive Bayes optimisé et entraîné, il est utilisé pour prédire les sentiments des textes. Les prédictions sont faites en utilisant les vecteurs de caractéristiques textuelles et les nouvelles caractéristiques ajoutées. Le modèle et le vectoriseur sont sauvegardés pour une utilisation future, et les performances du modèle sont évaluées à l'aide de rapports de classification et de matrices de confusion. Les résultats finaux sont sauvegardés dans des fichiers pour une analyse ultérieure.

4.3 Conclusion

Dans ce chapitre, nous avons présenté notre approche pour l'analyse des sentiments sur les vidéos d'éducation et détaillé notre méthodologie. Cette approche structurée et méthodique nous permet d'extraire des informations précieuses sur les émotions exprimées dans ces commentaires, et de classifier efficacement les sentiments. Les techniques et les étapes décrites ici constituent une base solide pour la réalisation de notre projet et pour les analyses futures.

5 Chapitre 5

Expérimentation

Sommaire

5.1 Introduction

Dans le cadre de notre recherche, nous avons réalisé une analyse des sentiments des commentaires en dialecte algérien publiés sous les vidéos éducatives sur YouTube. Ce traitement permet de classer les opinions exprimées en trois catégories : négatif, neutre et positif. Les données d'entrée utilisées pour cette étude sont les commentaires extraits d'une chaîne et de quelques autres vidéos, servant à entraîner et tester notre approche en temps réel.

Ce chapitre est consacré à la description détaillée de notre cadre expérimental pour l'analyse des sentiments appliquée aux vidéos éducatives. Nous commencerons par présenter le dataset utilisé, en expliquant sa composition et ses caractéristiques essentielles. Ensuite, nous décrirons l'environnement de travail et les outils logiciels employés pour mener à bien cette expérimentation.

Nous aborderons les différentes bibliothèques Python utilisées, en détaillant leur rôle et leur importance dans le traitement des données, la construction des modèles et la visualisation des résultats. Cela inclut des bibliothèques telles que Numpy, Pandas, Scikit-learn, et bien d'autres, qui sont cruciales pour la manipulation des données, l'implémentation des algorithmes de machine learning, et l'évaluation des modèles.

Par la suite, nous décrirons le processus de mise en service, y compris l'évaluation des modèles de classification des sentiments. Nous expliquerons les différentes métriques utilisées pour évaluer la performance des modèles, telles que la précision, le rappel, le F1-score, l'exactitude, ainsi que les courbes ROC et AUC. Ces métriques nous permettront d'obtenir une évaluation complète et précise des performances des modèles développés.

Enfin, nous présenterons les résultats obtenus et fournirons une analyse critique de ces résultats. Ce chapitre se conclura par un résumé des principaux enseignements tirés de cette expérimentation et des pistes de réflexion pour les travaux futurs.

5.2 Description du Dataset

Un dataset, ou jeu de données, est un ensemble structuré de données utilisé dans le Machine Learning et l'analyse des données. Il est constitué de plusieurs éléments clés. **Les lignes** (ou instances) représentent des enregistrements individuels, tandis que **les colonnes** (ou variables/features) représentent les différentes caractéristiques de ces enregistrements. Chaque variable contient des données spécifiques telles que des nombres, des textes ou des catégories. Un dataset peut également inclure des **étiquettes** (labels) dans des problèmes supervisés, qui sont les valeurs que l'algorithme doit prédire. Enfin, les **métadonnées** fournissent des informations contextuelles sur la structure et la source des données.

Les datasets peuvent être disponibles dans divers formats, chacun ayant ses propres avantages en fonction de l'application. Les formats courants incluent **CSV** (Comma-Separated Values) et **TSV** (Tab-Separated Values), qui sont largement utilisés en raison de leur simplicité et de leur compatibilité avec de nombreux outils d'analyse. Les fichiers **Excel** sont également populaires pour leur interface utilisateur conviviale. Les **SGBDR** (Systèmes

de Gestion de Bases de Données Relationnelles) comme MySQL (**R**elational **d**atabase **m**anagement **s**ystem) ou PostgreSQL (Open-source relational database) offrent des fonctionnalités robustes pour le stockage et la gestion des données. Les bases de données **NoSQL (N**on-**r**elational **d**atabase**)** comme MongoDB (NoSQL database) sont adaptées pour les données non structurées. Enfin, les **services Web** et les API fournissent un accès en temps réel à des datasets dynamiques provenant de diverses sources.

Dans le domaine du machine learning, la qualité et la composition des datasets jouent un rôle crucial dans la performance des modèles prédictifs. Deux problèmes courants liés à la qualité de l'ajustement des modèles sont l'overfitting et l'underfitting :

L'overfitting se produit lorsqu'un modèle de machine learning s'adapte trop étroitement aux données d'entraînement, capturant le bruit et les anomalies au lieu de discerner les tendances générales. Cela entraîne une excellente performance sur le jeu de données d'entraînement, mais une faible capacité à généraliser sur des données nouvelles et non vues auparavant. L'overfitting est souvent le résultat d'un modèle trop complexe par rapport à la quantité de données disponibles, ou d'un trop grand nombre de caractéristiques par rapport aux observations.

L'underfitting se produit lorsque le modèle est trop simple pour capturer les tendances présentes dans les données d'entraînement, entraînant une performance médiocre à la fois sur les données d'entraînement et sur les nouvelles données. Cela se produit généralement lorsque le modèle est incapable de s'adapter aux structures sous-jacentes des données en raison de sa trop faible complexité ou d'un manque de caractéristiques pertinentes.

Le dataset utilisé est un ensemble des commentaires extrait de la chaîne YouTube de "الأستاذ نور الدين" ainsi que d'autres vidéos éducatives de différentes chaînes. Les liens vers ces vidéos peuvent être consultés. Il est au format CSV, car ce format est plus pratique pour Python dans le domaine de l'analyse des sentiments. La taille du dataset est de 117KB et il comporte 424 avis sur des leçons vidéo en Algérie sur YouTube.

Le Dataset est composé de six (6) colonnes :

- **Username :** cette colonne contient les noms d'utilisateur des personnes qui ont posté les commentaires sur les videos en YouTube. Chaque entrée représente un utilisateur unique qui a interagi avec la vidéo en laissant un commentaire.

  Exemple : @SARAyhi, @mehdidjerboua10

- **Updated_at :** cette colonne contient les horodatages indiquant la dernière mise à jour du commentaire. Cela inclut la date et l'heure de la mise à jour, ce qui permet de comprendre la récence et la pertinence du commentaire.

  Format : Typiquement au format YYYY-MM-DD HH:MM:SS (Year-Month-Day Hour :Minute :Second).

  Exemple : 2022-05-16T21:45:54Z

T : Séparateur indiquant le début de la composante temporelle.

Z : indique que l'heure est en UTC (temps universel coordonné).

Le Temps Universel Coordonné (UTC) est la norme de temps principale utilisée dans le monde pour réguler les horloges et le temps. Il est essentiellement équivalent au Temps

Moyen de Greenwich GMT (**G**reenwich **M**ean **T**ime), mais il est plus précis grâce à la mesure du temps atomique.

- **Video_id :** cette colonne contient des identifiants uniques pour les vidéos YouTube auxquelles les commentaires sont associés. Chaque video_id correspond à une vidéo spécifique, permettant l'association des commentaires avec un contenu vidéo particulier.
  Exemple : sYDBM8657yE, dlUYsXqVX7k

- **Original Text :** cette colonne contient le texte réel des commentaires laissés par les utilisateurs sur les vidéos YouTube. Les commentaires sont en dialecte algérien et fournissent des avis et opinions précieux des spectateurs.
  Exemple :'lah ybarak mrc'

La figure ci-dessous représente une partie (9 lignes) de notre Data frame :

	Username	updated_at	video_id	Original Text
0	@SARAyhi	2024-05-10T20:50:31Z	sYDBM8657yE	شكرا استاذ
1	@bouallegzaineb4602	2024-04-23T18:33:53Z	sYDBM8657yE	Les types de grammaire?
2	@ala_clach9	2024-03-29T09:30:34Z	sYDBM8657yE	مع كل احتراماتي ليك بصح حا...
3	@user-po6gc2wg8h	2023-03-30T09:24:59Z	sYDBM8657yE	salam ostad win nal9a pdf ...
4	@magraouihassen1730q	2022-02-26 08:16:29+00:00	sYDBM8657yEh	صديقي شكرا على الفيديو و ا...
5	@inesabd7343	2022-05-24T16:40:37Z	sYDBM8657yE	استاذ وين نقدر نتواصل معاك...
6	@music4you335m	2020-10-19 10:18:22+00:00	idY_LcHRNkol	ممكن تشرحلي كيف تسجل شاشة ...
7	@mehdidjerboua10	2022-05-16T21:45:54Z	sYDBM8657yE	baraka allaho fik 👌
8	@saraanwr6442	2022-02-15T17:16:40Z	sYDBM8657yE	حضرتك بتشرح إيه ؟
9	@benamarabouchra6368	2021-11-24T18:24:31Z	sYDBM8657yE	merciiii bieen bon explica...

Figure 5.1 – Une partie de Dataframe

5.3 Environnement de travail

5.3.1 Anaconda

Anaconda est une distribution open source des langages de programmation Python et R pour la science des données qui vise à simplifier la gestion et le déploiement des packages. Les versions de packages dans Anaconda sont gérées par le système de gestion de packages, conda, qui analyse l'environnement actuel avant d'exécuter une installation pour éviter de perturber d'autres frameworks et packages [50].

5.3.2 Jupyter Notebook

Jupyter Notebook (anciennement IPython (**I**nteractive **P**ython) Notebook) est une application Web interactive permettant de créer et de partager des documents informatiques. Le projet a d'abord été nommé IPython, puis renommé Jupyter en 2014. Il s'agit d'un produit

entièrement open source et les utilisateurs peuvent utiliser gratuitement toutes les fonctionnalités disponibles. Il prend en charge plus de 40 langages, dont Python, R et Scala [51].

5.4 Langage de programmation:

5.4.1 Python

Python est un langage de programmation largement utilisé dans les applications Web, le développement de logiciels, la science des données et l'apprentissage automatique ML (**M**achine **L**earning). Les développeurs utilisent Python car il est efficace et facile à apprendre et peut fonctionner sur de nombreuses plates-formes différentes. Le logiciel Python est téléchargeable gratuitement, s'intègre bien à tous les types de systèmes et augmente la vitesse de développement [52].

5.5 Bibliothèques de Python

5.5.1 Numpy (pour les opérations numériques)

NumPy est le package fondamental pour le calcul scientifique en Python. Il s'agit d'une bibliothèque Python qui fournit un objet tableau multidimensionnel, divers objets dérivés (tels que des tableaux et matrices masqués) et un assortiment de routines pour des opérations rapides sur les tableaux, notamment mathématiques, logiques, manipulation de forme, tri, sélection, E/S. , transformées de Fourier discrètes, algèbre linéaire de base, opérations statistiques de base, simulation aléatoire et bien plus encore [53].

5.5.2 Pandas (pour la manipulation et l'analyse des données)

Pandas est une bibliothèque Python utilisée pour travailler avec des ensembles de données. Il dispose de fonctions d'analyse, de nettoyage, d'exploration et de manipulation des données. Le nom « Pandas » fait référence à la fois à « Panel Data » et à « Python Data Analysis » et a été créé par Wes McKinney en 2008.

Pandas nous permet d'analyser le Big Data et de tirer des conclusions basées sur des théories statistiques. Peut nettoyer des ensembles de données désordonnés et les rendre lisibles et pertinents. Les données pertinentes sont très importantes en science des données [54].

5.5.3 Scikit-learn (pour les algorithmes de machine learning, la vectorisation de texte, la modélisation et l'évaluation des modèles)

Scikit-learn, également connu sous le nom de sklearn, est une bibliothèque open source d'apprentissage automatique et de modélisation de données pour Python. Il propose divers algorithmes de classification, de régression et de clustering, notamment des machines à vecteurs de support, des forêts aléatoires, l'augmentation de gradient, les k-means et DBSCAN (**D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise), et est conçu pour interopérer avec les bibliothèques Python, NumPy et SciPy (**S**cientific **P**ython) [55].

5.5.4 Imbalanced-learn (pour le suréchantillonnage des données déséquilibrées)

Imbalanced-learn est une boîte à outils Python open-source visant à fournir un large éventail de méthodes pour faire face au problème des ensembles de données déséquilibrés, fréquemment rencontré en apprentissage automatique et en reconnaissance de formes. Les

méthodes de pointe implémentées peuvent être catégorisées en 4 groupes : (i) sous-échantillonnage, (ii) sur-échantillonnage, (iii) combinaison de sur-échantillonnage et de sous-échantillonnage, et (iv) méthodes d'apprentissage en ensemble. La boîte à outils proposée ne dépend que de numpy, scipy et scikit-learn et est distribuée sous la licence MIT (**M**assachusetts **I**nstitute of **T**echnology). De plus, elle est entièrement compatible avec scikit-learn et fait partie du projet soutenu par scikit-learn-contrib [56].

5.5.5 Matplotlib (pour la création de graphiques et de visualisations)

Matplotlib est une bibliothèque complète pour créer des visualisations statiques, animées et interactives en Python. Matplotlib rend les choses faciles simples et les choses difficiles possibles [57].

5.5.6 Joblib (pour la sauvegarde des modèles)

Joblib est une bibliothèque Python pour exécuter des tâches intensives en calcul en parallèle. Elle fournit un ensemble de fonctions pour effectuer des opérations en parallèle sur de grands ensembles de données et pour mettre en cache les résultats de fonctions coûteuses en calcul. Joblib est particulièrement utile pour les modèles d'apprentissage automatique car elle permet de sauvegarder l'état de vos calculs et de reprendre votre travail plus tard ou sur une autre machine [58].

5.5.7 Pickle (pour la sérialisation et la désérialisation des objets Python)

Pickle en Python est principalement utilisé pour sérialiser et désérialise r une structure d'objet Python. En d'autres termes, c'est le processus de conversion d'un objet Python en un flux d'octets pour le stocker dans un fichier ou une base de données, maintenir l'état du programme entre les sessions, ou transporter des données sur le réseau [59].

5.5.8 Re ‘Regular Expression Syntax ‘ (pour les opérations sur les expressions régulières)

Une expression régulière (ou RE) spécifie un ensemble de chaînes qui lui correspondent ; les fonctions de ce module vous permettent de vérifier si une chaîne particulière correspond à une expression régulière donnée (ou si une expression régulière donnée correspond à une chaîne particulière, ce qui revient au même) [60].

5.5.9 Itertools (pour les outils d'itération)

C'est un « module [qui] implémente un certain nombre de blocs de construction d'itérateurs inspirés de constructions provenant de APL (**A P**rogramming **L**anguage), Haskell et SML (Standard ML)… Ensemble, ils forment une ‘algèbre d'itérateurs’ rendant possible la construction d'outils spécialisés de manière succincte et efficace en Python pur [61].

5.5.10 Scipy (pour les fonctions et algorithmes scientifiques)

SciPy est une bibliothèque de calcul scientifique qui utilise NumPy en sous-jacent. SciPy signifie Scientific Python. Elle fournit davantage de fonctions utilitaires pour l'optimisation, les statistiques et le traitement du signal. Comme NumPy, SciPy est open source, donc

nous pouvons l'utiliser librement. SciPy a été créée par Travis Oliphant, le créateur de NumPy [62].

5.5.11 Os (pour les opérations système)

Python possède un module intégré nommé os avec des méthodes pour interagir avec le système d'exploitation, comme la création de fichiers et de répertoires, la gestion des fichiers et des répertoires, les entrées et sorties, les variables d'environnement, la gestion des processus…etc [63].

5.5.12 Googleapiclient (pour l'accès à l'API de Google)

La bibliothèque cliente Google API pour Python est conçue pour les développeurs d'applications clientes en Python. Elle offre un accès simple et flexible à de nombreuses API Google [64].

5.5.13 Emoji (pour la manipulation des emojis)

La bibliothèque emoji en Python est un utilitaire qui fournit des outils pour gérer les emojis dans les chaînes de texte. Les emojis sont des symboles graphiques utilisés pour exprimer des émotions, des idées ou des concepts dans la communication numérique, souvent intégrés dans les messages texte, les publications sur les réseaux sociaux et d'autres contenus en ligne. Cette bibliothèque permet aux développeurs de trouver, remplacer et manipuler facilement les emojis dans les chaînes Python.

5.5.14 Unicodedata (pour la manipulation des données Unicode)

Ce module permet d'accéder à la base de données des caractères Unicode UCD (**U**nicode **C**haracter **D**atabase) qui définit les propriétés des caractères pour tous les caractères Unicode [65].

5.5.15 Pyarabic (pour la manipulation des textes en arabe)

Une bibliothèque spécifique pour la langue arabe en Python, qui fournit des fonctions de base pour manipuler les lettres et le texte arabe, comme la détection des lettres arabes, les groupes et les caractéristiques des lettres arabes, la suppression des diacritiques…etc [66].

5.5.16 Aaransia ou 3aransia (pour la translittération et les erreurs liées aux langues sources)

Est un terme désignant la translittération de l'arabe avec l'utilisation des chiffres pour représenter certains sons arabes spécifiques qui n'ont pas d'équivalent direct en lettres latines, couramment utilisée dans les communications informelles sur Internet. La translittération des langues et des dialectes est rapide et fiable, utilisant des variables par défaut pour accéder aux données, et offre des fonctionnalités telles que la translittération en vrac, une API disponible, la translittération multilingue et la prise en charge de 70 langues et dialectes [67].

5.5.17 Nltk (pour le traitement du langage naturel)

Le Natural Language Toolkit, ou plus communément NLTK, est un ensemble de bibliothèques et de programmes pour le traitement automatique du langage naturel (TAL) sym-

bolique et statistique pour l'anglais, écrit en langage de programmation Python. Il prend en charge les fonctionnalités de classification, de tokenisation, de stemming, d'étiquetage, d'analyse syntaxique et de raisonnement sémantique [68].

5.5.18 Collections (pour les conteneurs de données)

Le module Collections en Python fournit différents types de conteneurs. Un conteneur est un objet utilisé pour stocker différents objets et offrir un moyen d'accéder aux objets contenus et de les parcourir. Certains des conteneurs intégrés sont Tuple, List et Dictionary [69].

5.5.19 Ast ' Abstract Syntax Trees ' (pour la manipulation des arbres syntaxiques abstraits)

Le module ast aide les applications Python à traiter les arbres de la grammaire syntaxique abstraite de Python. La syntaxe abstraite elle-même peut changer avec chaque version de Python ; ce module permet de découvrir de manière programmatique à quoi ressemble la grammaire actuelle [70].

5.6 Mise en service

Nous allons donner une orientation pour les étapes d'exécution des cellules et scriptes python avec une figure qui génère ces étapes :

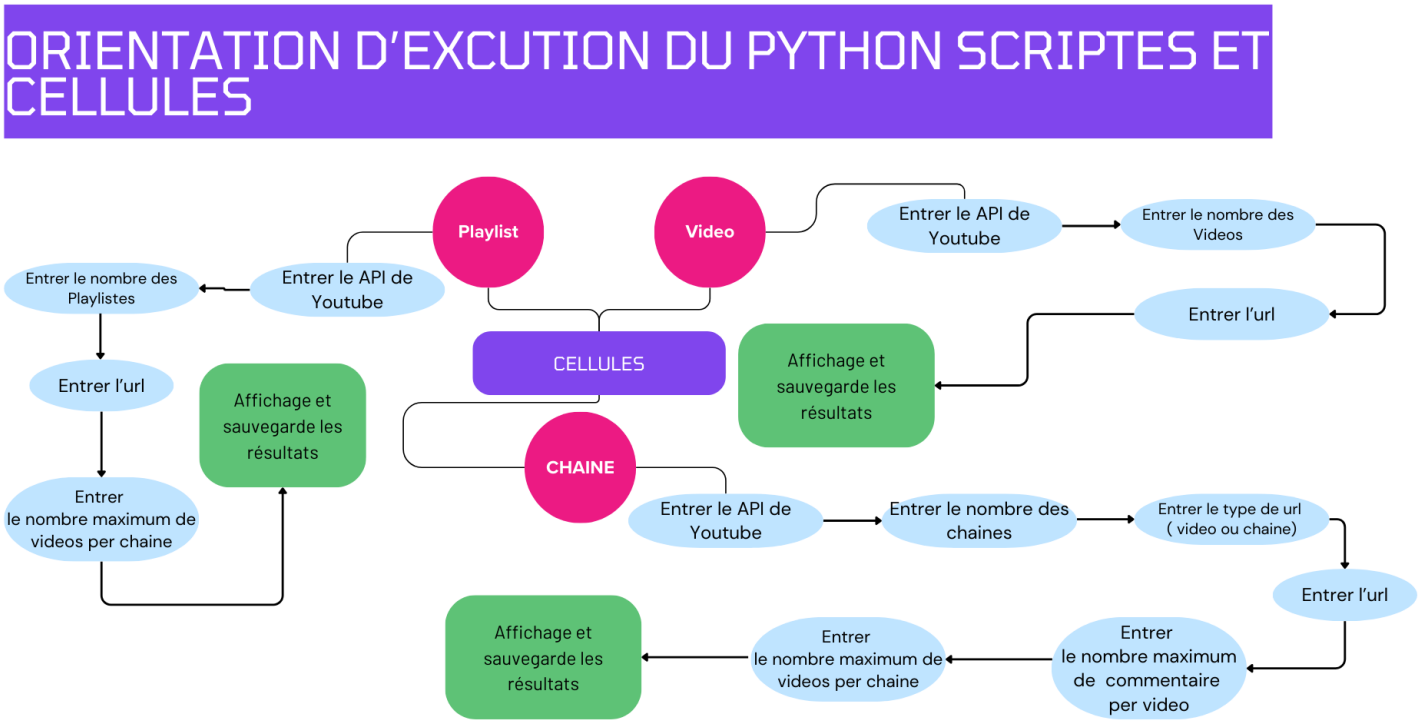

Figure 5.2 – Les étapes d'exécution du python scriptes et cellules

A. Pour la cellule chaine :

Tout d'abord, Le scripte Python vous demande d'entrer votre clé API YouTube, qu'il validera ensuite en faisant une requête de test à l'API YouTube Data.

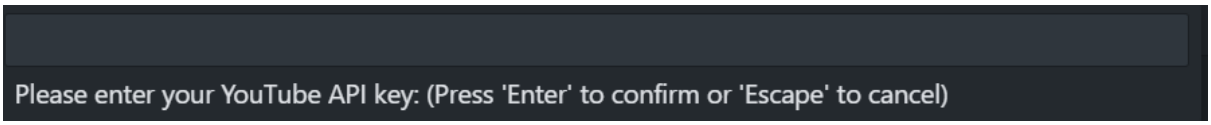

Figure 5.3 – Capture d'écran pour input API YouTube

Si la clé est valide, elle sera enregistrée dans un fichier nommé "api_key.txt". Si la clé est invalide, le script vous demandera de la saisir à nouveau jusqu'à ce qu'une clé valide soit fournie. Le résultat sera un message de confirmation indiquant que la clé API est valide et a été enregistrée.

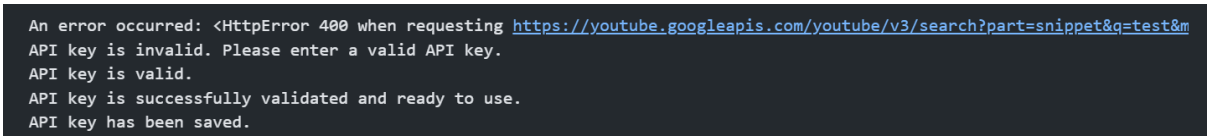

Figure 5.4 – Capture d'écran d'output après saisir API YouTube

Ensuite, il vous demande de saisir le nombre de chaînes que vous souhaitez traiter.

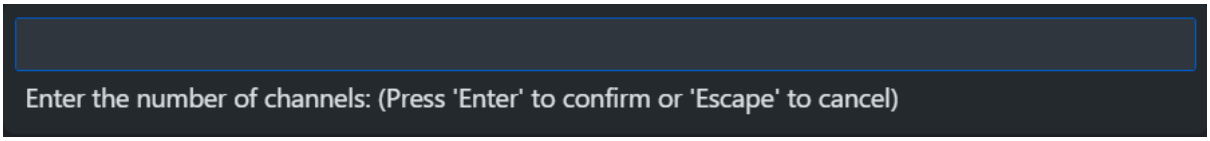

Figure 5.5 – Capture d'écran pour input nombre des chaines

Puis, Pour chaque chaîne, vous choisissez si vous voulez entrer une URL de vidéo ou de chaîne.

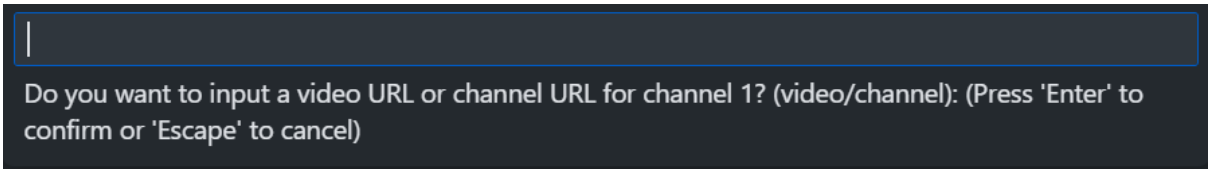

Figure 5.6 – Capture d'écran pour input type de l'url (video ou chaine)

Après, vous devez saisir l'url pour chaque chaine.

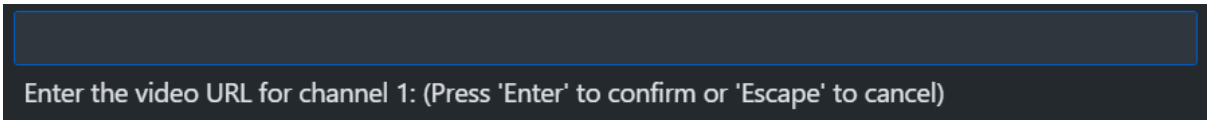

Figure 5.7 – Capture d'écran pour input l'Url

Par la suite, le script extrait les IDs des vidéos et des chaînes à partir des URLs fournies. En sortie, il vous affiche une liste de dictionnaires contenant les IDs des vidéos et des chaînes correspondantes, ainsi qu'une liste de tous les IDs de chaînes extraits.

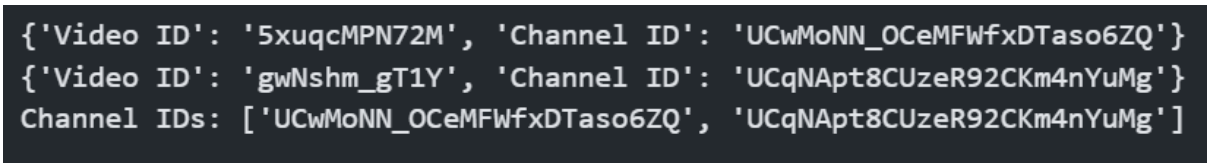

Figure 5.8 – Capture d'écran d'output après saisir l'Url

De plus, il vous demande d'entrer le nombre maximaum de commentaires par vidéo.

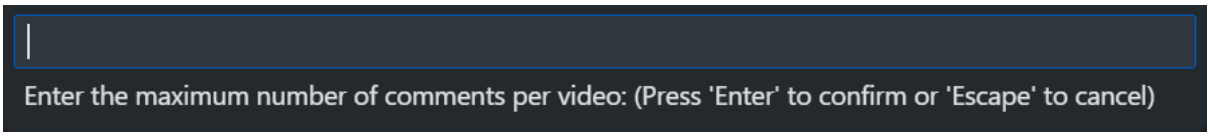

Figure 5.9 – Capture d'écran pour input nombre maximum des commentaires par video

En outre, le nombre maximal de vidéos par chaîne.

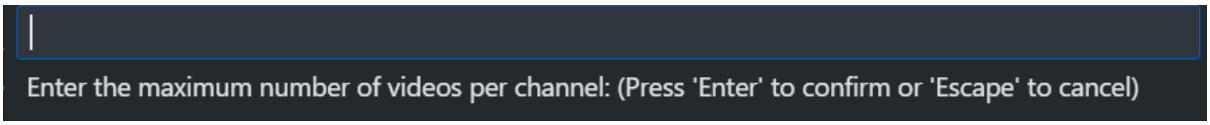

Figure 5.10 – Capture d'écran pour input nombre maximum de videos par chaine

Alors, il récupère les vidéos de chaque chaîne YouTube, extrait les commentaires de ces vidéos, et stocke les informations dans un fichier CSV et un fichier texte. En sortie, vous obtiendrez un fichier CSV et un fichier texte contenant les détails des commentaires pour toutes les chaînes spécifiées.

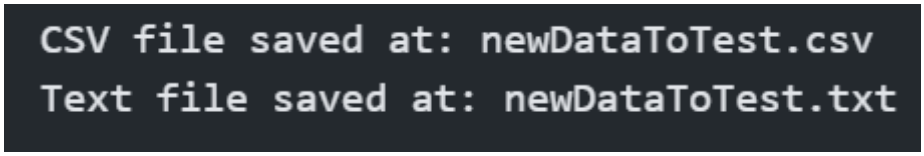

Figure 5.11 – Capture d'écran d'output après l'extraction des commentaires

En suivant, Ce script analyse et visualise les commentaires extraits précédemment d'une chaîne YouTube. Il identifie et compte les commentaires contenant des alphabets latins, arabes, mixtes (latin et arabe), des emojis, uniquement des emojis, et sans emojis. Ensuite, il génère un diagramme circulaire pour visualiser la distribution de ces catégories de commentaires. En outre, il effectue une analyse de vocabulaire en tokenisant le texte des commentaires et en comptant la fréquence des mots avant le prétraitement. Les résultats, y compris la taille du vocabulaire, sont enregistrés dans des fichiers CSV et TXT.

```
Total number of comments: 50
Number of comments with Latin alphabet: 5
Number of comments with Arabic alphabet: 43
Number of comments with mixed alphabet (Latin + Arabic): 1
Number of comments with only emojis: 2
Number of comments with emojis: 21
Number of comments without emojis: 29
```

Figure 5.12 – Capture d'écran d'output la visualisation des commentaires

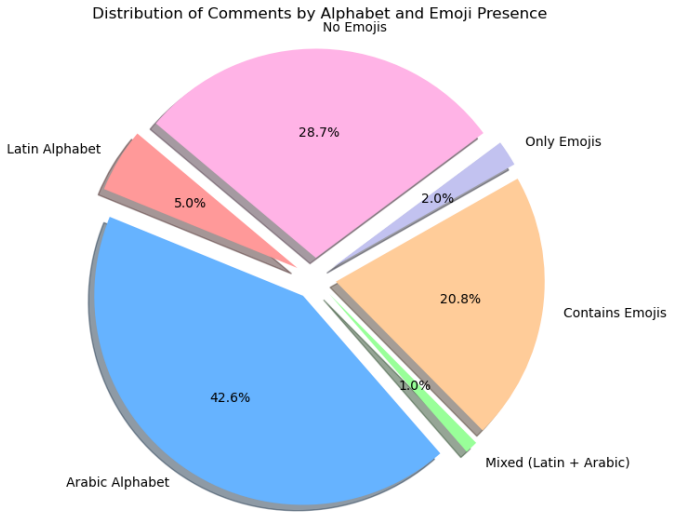

Figure 5.13 – Diagramme circulaire de distribution des commentaires

```
Vocabulary Size Before Preprocessing: 339
```

Figure 5.14 – Capture d'écran d'output la taille de vocabulaire avant le prétraitement

En continuant, Ce script traite et nettoie les commentaires YouTube extraits précédemment. Il enlève les diacritiques, translittère l'arabe en alphabet latin, remplace les chiffres par des lettres, supprime les hashtags, les URL, les caractères spéciaux, et normalise le texte en minuscules. Ensuite, il tokenise le texte prétraité, compte la taille du vocabulaire et enregistre les résultats dans des fichiers CSV et TXT. En sortie, vous obtiendrez les commentaires prétraités, ainsi que la taille du vocabulaire après le prétraitement.

```
Vocabulary Size After Preprocessing: 300
```

Figure 5.15 – Capture d'écran d'output la taille de vocabulaire après le prétraitement

Pour continuer, ce script va utiliser un fichier texte contenant les stop words translittérés (transliterated_unique_stop_words.txt) et enlève les stop words des textes, tokenise les textes restants, compte la taille du vocabulaire après suppression des stop words et enregistre les résultats dans des fichiers CSV et TXT. En sortie, vous obtiendrez un fichier CSV avec le vocabulaire et sa fréquence, un fichier TXT avec la taille du vocabulaire et les détails de chaque texte traité, ainsi qu'un nouveau DataFrame sans stop words.

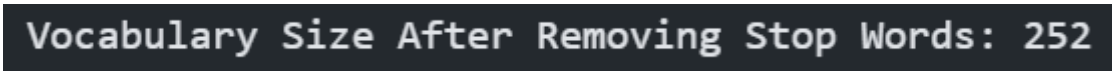

Figure 5.16 – Capture d'écran d'output la taille de vocabulaire après la suppression des mots vides

À ce moment-là, le script va utiliser un fichier CSV contenant les données prétraitées, ainsi qu'un modèle et un vectoriseur enregistrés. Le script charge ces fichiers, applique des fonctions pour extraire et compter les mots et phrases de sentiment, et utilise ces informations pour prédire les sentiments des nouveaux commentaires. En sortie, vous obtiendrez des fichiers CSV et TXT contenant les résultats des prédictions, avec des détails sur chaque commentaire et des statistiques sur la répartition des sentiments.

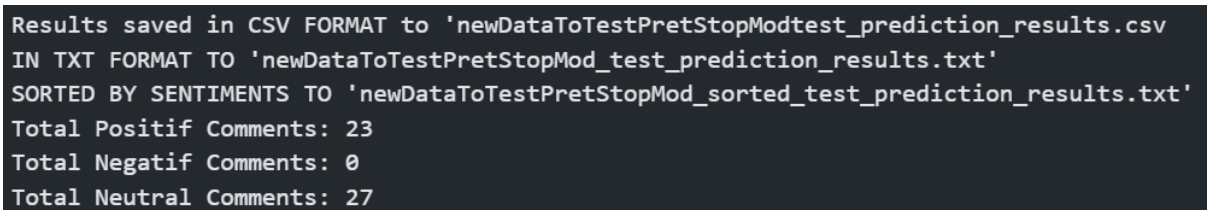

Figure 5.17 – Capture d'écran d'output les résultats du modèle Random Forest

Parallèlement, ce script va utiliser un fichier CSV contenant les données prétraitées, ainsi qu'un modèle SVM et un vectoriseur enregistrés. Le script charge ces fichiers, applique des fonctions pour extraire et compter les mots et phrases de sentiment, et utilise ces informations pour prédire les sentiments des nouveaux commentaires. En sortie, vous obtiendrez des fichiers CSV et TXT contenant les résultats des prédictions, avec des détails sur chaque commentaire et des statistiques sur la répartition des sentiments.

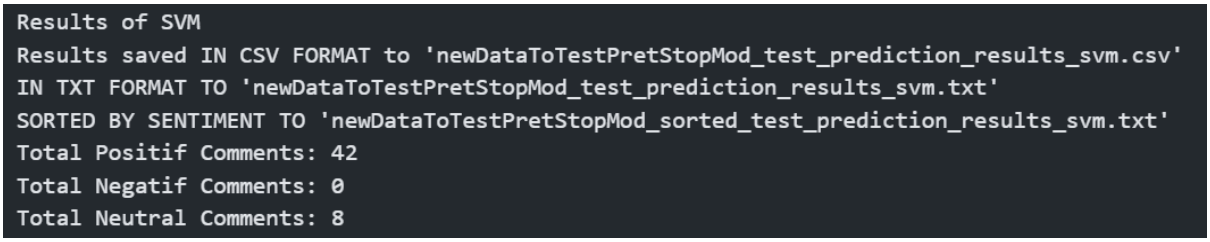

Figure 5.18 – Capture d'écran d'output les résultats du modèle SVM

Enfin, Le script chargera également un modèle Naive Bayes pré-entraîné et un vectoriseur TF-IDF depuis des fichiers `.pkl`. Ensuite, il effectuera une analyse des sentiments sur les données en utilisant des mots et des phrases de sentiment définis dans un fichier, et produira des prédictions de sentiments. Les résultats seront sauvegardés dans un fichier CSV et deux fichiers TXT, l'un avec les résultats bruts et l'autre avec les résultats triés par sentiment.

Vous obtiendrez également un compte-rendu du nombre de commentaires positifs, négatifs et neutres.

```
Results of NAIVE
Results saved IN CSV FORMAT to 'newDataToTestPretStopMod_naive_test_prediction_results.csv'
IN TXT FORMAT TO 'newDataToTestPretStopMod_naive_test_prediction_results.txt'
SORTED BY SENTIMENT TO 'newDataToTestPretStopMod_naive_sorted_test_prediction_results.txt'
Total Positif Comments: 46
Total Negatif Comments: 0
Total Neutral Comments: 4
```

Figure 5.19 – Capture d'écran d'output les résultatS du modèle Naive Bayes

B. Pour les cellules videos et playlistes :

Si vous souhaitez analyser les sentiments à partir des vidéos ou des playlists, c'est le même processus que pour les chaînes. Après avoir saisir l'API de YouTube, la seule différence réside dans les autres entrées : pour les vidéos, vous devez fournir le nombre de vidéos et bien sûr leurs URL. Pour les playlists, la différence concerne le nombre de playlists, le nombre de commentaires par playlist et toujours les URLs.

5.6.1 Évaluation des modèles

L'évaluation d'un modèle en apprentissage automatique est une étape cruciale pour déterminer sa performance et son utilité.

Lorsque l'on évalue un modèle de classification, plusieurs métriques clés permettent de mesurer ses performances comme la précision, rappel, F1-score, et d'autres termes importants utilisés dans le contexte des rapports de classification et des matrices de confusion :

I. Précision (Precision)

La précision mesure la proportion de prédictions positives correctes parmi toutes les prédictions positives faites par le modèle. Une précision élevée signifie moins de faux positifs.

$$(5.1): \text{Précision} = \frac{\text{Vrais Positifs (TP)}}{\text{Vrais Positifs (TP)} + \text{Faux Positifs (FP)}}$$

Exemple :

$$\text{Précision} = \frac{45}{45 + 1} = 0.98$$

II. Rappel (Recall)

Le rappel, également appelé sensibilité ou taux de vrais positifs, mesure la proportion de vrais positifs correctement identifiés parmi tous les échantillons réellement positifs. Un rappel élevé signifie moins de faux négatifs.

$$(5.2): \text{Rappel} = \frac{\text{Vrais Positifs (TP)}}{\text{Vrais Positifs (TP)} + \text{Faux Négatifs (FN)}}$$

Exemple:

$$\text{Rappel} = \frac{45}{45 + 0} = 1.00$$

III. F1-score

Le F1-score est la moyenne harmonique de la précision et du rappel. Il est utile lorsque vous avez besoin d'un équilibre entre précision et rappel.

$$(5.3): \text{F1 score} = 2 \times \frac{\textbf{Précision} \times \textbf{Rappel}}{\textbf{Précision} + \textbf{Rappel}}$$

Exemple:

$$\text{F1 score} = 2 \times \frac{0.98 \times 1.00}{0.98 + 1.00} = 0.99$$

IV. Exactitude (Accuracy)

L'exactitude mesure la proportion de prédictions correctes parmi l'ensemble des prédictions.

$$(5.4): \text{Exactitude} = \frac{\text{Prédictions Correctes}}{\text{Nombre Total d'Échantillons}}$$

Exemple:

$$\text{Exactitude} = \frac{34 + 45 + 31}{120} = 0.92$$

V. Moyenne macro (Macro Average)

La moyenne macro est la moyenne arithmétique des précisions, rappels et F1-scores de chaque classe, donnant à chaque classe le même poids.

Exemple :

$$\text{Macro Précision} = \frac{1.00 + 0.83 + 0.97}{3} = 0.93$$

$$\text{Macro Rappel} = \frac{11.00 + 0.98 + 0.78}{3} = 0.92$$

$$\text{Macro F1 score} = \frac{1.00 + 0.90 + 0.86}{3} = 0.92$$

VI. Moyenne pondérée (Weighted Average)

La moyenne pondérée tient compte du support (le nombre d'échantillons) de chaque classe lors du calcul des précisions, rappels et F1-scores, donnant plus de poids aux classes avec plus d'échantillons.

Exemple :

$$\text{Weighted Précision} = \frac{34 \times 1.00 + 46 \times 0.83 + 40 \times 0.97}{120} = 0.93$$

$$\text{Weighted Rappel} = \frac{34 \times 1.00 + 46 \times 0.98 + 40 \times 0.78}{120} = 0.92$$

$$\text{Weighted F1 score} = \frac{34 \times 1.00 + 46 \times 0.90 + 40 \times 0.86}{120} = 0.92$$

VII. Courbe ROC et AUC

La courbe ROC (Receiver Operating Characteristic) est un outil graphique utilisé pour évaluer la performance d'un modèle de classification binaire en visualisant la sensibilité (ou rappel) par rapport au taux de faux positifs (FPR). Chaque point sur la courbe ROC représente un seuil de classification différent.

A. Axes de la courbe ROC :

- **Axe des Y (True Positive Rate, TPR)** : il s'agit du rappel, calculé comme le nombre de vrais positifs divisé par la somme des vrais positifs et des faux négatifs.

$$(5.5): \text{TPR} = \frac{\text{Vrais Positifs (TP)}}{\text{Vrais Positifs (TP)} + \text{Faux Négatifs (FN)}}$$

- **Axe des X (False Positive Rate, FPR)** : il s'agit du taux de faux positifs, calculé comme le nombre de faux positifs divisé par la somme des faux positifs et des vrais négatifs.

$$(5.6): \text{FPR} = \frac{\text{Faux Positifs (FP)}}{\text{Faux Positifs (FP)} + \text{Vrai Négatifs (TN)}}$$

B. AUC (Area Under the Curve)

L'aire sous la courbe ROC (AUC) est une mesure de la capacité du modèle à distinguer entre les classes positives et négatives. Une AUC (**A**rea **U**nder the **C**urve) de 1.0 représente un modèle parfait, tandis qu'une AUC de 0.5 représente un modèle qui fait des prédictions au hasard.

5.6.2 Les résultat pour les 3 modèles

I. Random Forest :

Voici les figures ci-dessous qui représente l'évaluation de modèle Random Forest :

```
Classification Report:
              precision    recall  f1-score   support

     negatif       1.00      1.00      1.00        34
     neutral       0.83      0.98      0.90        46
     positif       0.97      0.78      0.86        40

    accuracy                           0.92       120
   macro avg       0.93      0.92      0.92       120
weighted avg       0.93      0.92      0.92       120

Confusion Matrix:
[[34  0  0]
 [ 0 45  1]
 [ 0  9 31]]
```

Figure 5.20 – Capture d'écran d'output l'évaluation du modèle Random Forest

- **Les mesures de précision, rappel et F1-score pour chaque classe:**

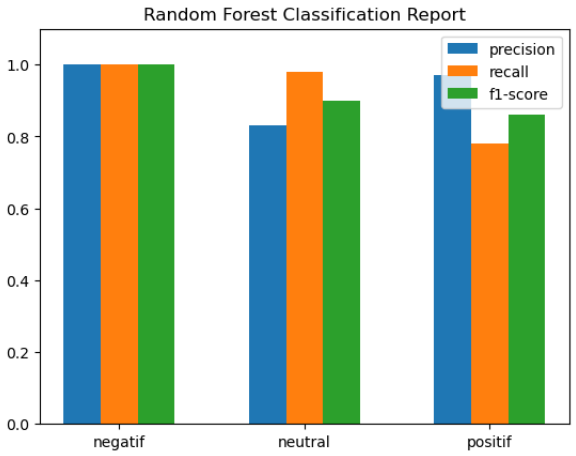

Figure 5.22 – Rapport de Classification avec la Forêt d'Arbres Décisionnels (Random Forest)

- La précision est excellente pour la classe négative (100%) et positive (97%), ce qui signifie que presque toutes les prédictions pour ces classes sont correctes.
- La précision pour la classe neutre est légèrement inférieure (83%), indiquant qu'il y a quelques erreurs de classification dans cette classe.
- Le rappel est parfait pour la classe négative (100%) et très élevé pour la classe neutre (98%), ce qui montre que le modèle est capable de capturer presque tous les exemples de ces classes.
- Le rappel est un peu plus faible pour la classe positive (78%), suggérant que le modèle manque quelques exemples de cette classe.

- Le f1-score, qui combine la précision et le rappel, est parfait pour la classe négative (100%) et très bon pour les classes neutre (90%) et positive (86%).

➢ **Le support,accuracy,macro moyenne et moyenne pondérée :**

- Le support indique le nombre de vrais exemples présents dans chaque classe. La classe neutre a le plus grand support (46), suivie de la classe positive (40) et négative (34).
- Une accuracy de 0.92 signifie que le modèle classifie correctement 92% des exemples.
- La macro moyenne prend la moyenne des mesures pour chaque classe sans tenir compte du support, ce qui donne une vue équilibrée de la performance du modèle.
- La moyenne pondérée prend en compte le support de chaque classe, offrant une mesure plus représentative de la performance globale du modèle.

➢ **Pour le Matrice de Confusion :**

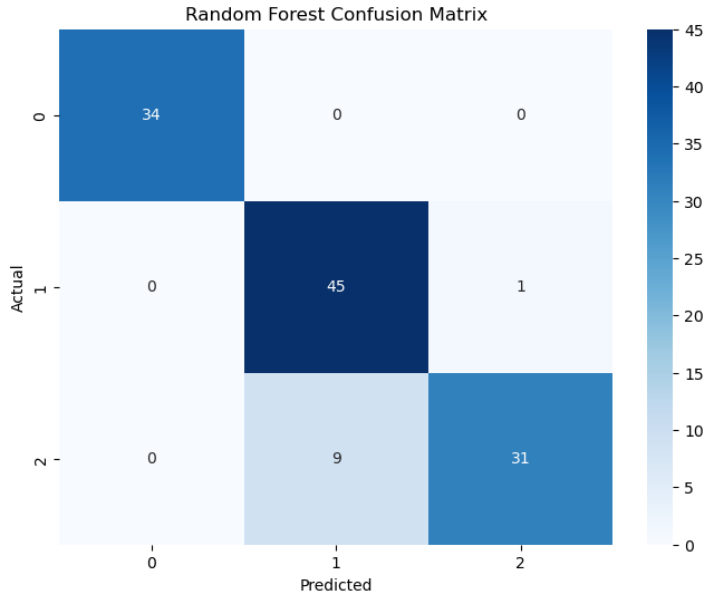

Figure 5.23 – Le matrice de confusion de la Forêt d'Arbres Décisionnels (Random Forest)

- Tous les exemples négatifs (34 exemples) sont correctement classifiés (aucune fausse prédiction).
- Un exemple neutre est mal classé comme positif parmi 46 exemples.
- Neuf exemples positifs parmi 40 exemples sont mal classés comme neutres.

➢ **Pour le Courbes ROC :**

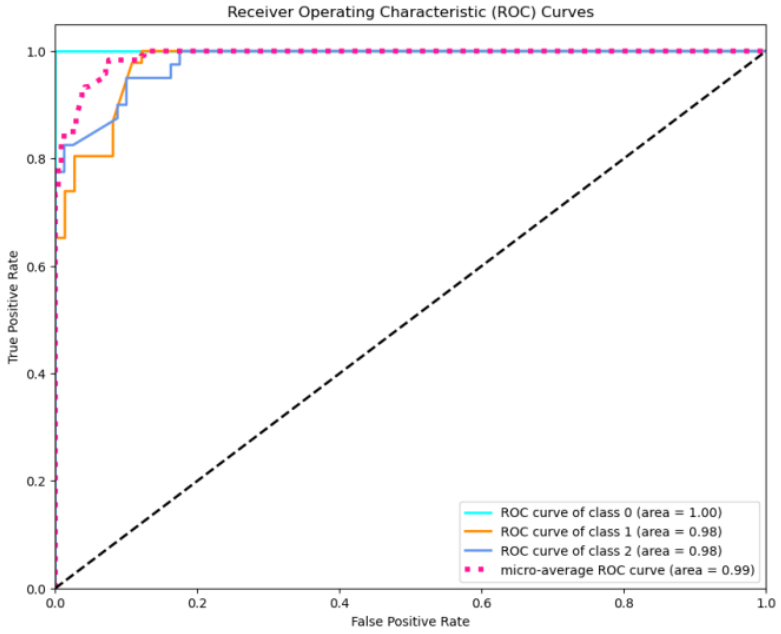

Figure 5.24 – Courbe ROC (Receiver Operating Characteristic) de Random Forest

- **Classe 0 (négatif):**
  - **Aire sous la courbe (AUC)**: 1.00
  - **Interprétation**: Une AUC de 1.00 indique une performance parfaite du modèle pour distinguer la classe négative des autres classes. Cela signifie que le modèle classe tous les exemples négatifs correctement sans aucune erreur.
- **Classe 1 (neutre):**
  - **Aire sous la courbe (AUC)**: 0.97
  - **Interprétation**: Une AUC de 0.97 montre que le modèle a une excellente capacité à distinguer la classe neutre des autres classes. Bien que ce ne soit pas parfait, le modèle est très proche de la performance idéale.
- **Classe 2 (positif):**
  - **Aire sous la courbe (AUC)**: 0.98
  - **Interprétation**: Avec une AUC de 0.98, le modèle démontre une capacité presque parfaite à distinguer la classe positive des autres classes. Cela montre que le modèle est très efficace pour classer les exemples positifs correctement.
- **Micro-moyenne:**
  - **Aire sous la courbe (AUC)**: 0.98
  - **Interprétation**: L'AUC de la micro-moyenne, qui agrège les performances sur toutes les classes, est également très élevée à 0.98. Cela signifie que, globalement, le modèle a une performance très forte et est capable de classer correctement la grande majorité des exemples, indépendamment de leur classe.

- Donc, globalement les résultats sont très bons parce que :
  1. **Courbes ROC**: Indiquent une excellente discrimination entre les classes.
  2. **Précision et Rappel**: Très bons pour les classes négatif et neutre, légèrement moins pour la classe positive.

3. **Matrice de Confusion**: Montre que les erreurs de classification sont rares et se produisent principalement entre les classes neutre et positif.

Le modèle de forêt aléatoire semble donc bien ajusté et performant pour ce jeu de données, avec une très bonne capacité à distinguer les classes, en particulier les classes négatives et neutres. Les paramètres optimisés (max_depth=None, min_samples_leaf=1, min_samples_split=10, n_estimators=300) semblent bien adaptés. Toutefois, des améliorations pourraient être explorées pour réduire les erreurs entre les classes neutre et positif.

II. SVM :

Voici les figures ci-dessous qui représente l'évaluation de modèle SVM :

```
Classification Report:
              precision    recall  f1-score   support

     negatif       1.00      1.00      1.00        34
     neutral       0.89      0.85      0.87        46
     positif       0.83      0.88      0.85        40

    accuracy                           0.90       120
   macro avg       0.91      0.91      0.91       120
weighted avg       0.90      0.90      0.90       120

Confusion Matrix:
[[34  0  0]
 [ 0 39  7]
 [ 0  5 35]]
```

Figure 5.25 – Capture d'écran d'output l'évaluation du modèle SVM

➢ **Les mesures de précision, rappel et F1-score pour chaque classe:**

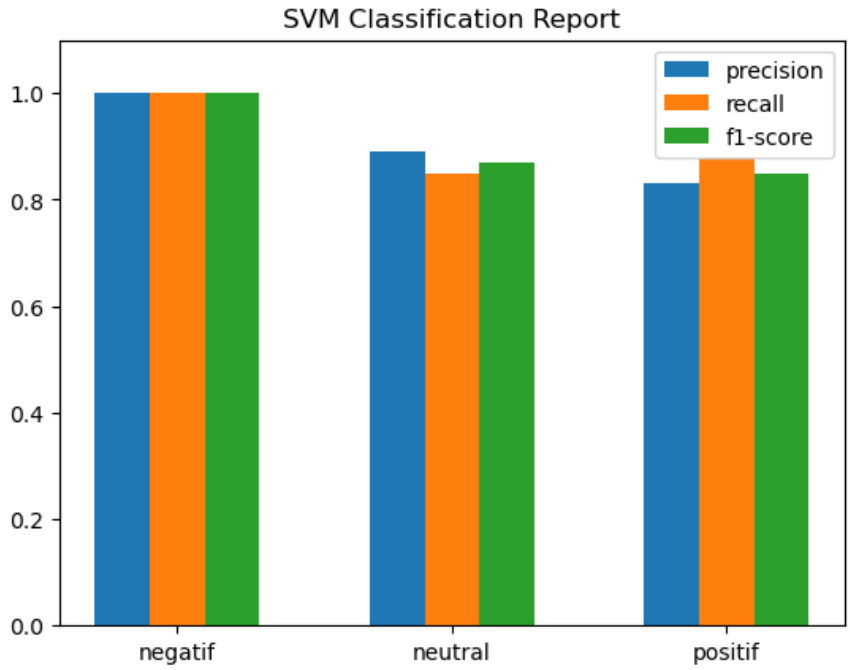

Figure 5.27 – Rapport de Classification avec SVM

- La précision est excellente pour la classe négative (100%), ce qui signifie que toutes les prédictions pour cette classe sont correctes.
- La précision pour la classe neutre est très bonne (89%), mais il y a quelques erreurs de classification.
- La précision pour la classe positive est également élevée (83%), bien qu'il y ait quelques erreurs.
- Le rappel est parfait pour la classe négative (100%), indiquant que le modèle capture tous les exemples de cette classe.
- Le rappel pour la classe neutre est légèrement inférieur (85%), suggérant que certains exemples neutres sont mal classés.
- Le rappel pour la classe positive est bon (88%), bien que quelques exemples soient classés dans d'autres classes.
- Le f1-score, qui combine la précision et le rappel, est parfait pour la classe négative (100%).
- Le f1-score pour la classe neutre est élevé (87%), indiquant un bon équilibre entre précision et rappel.
- Le f1-score pour la classe positive est bon (85%), mais légèrement inférieur aux autres classes.

➢ **Le support,accuracy,macro moyenne et moyenne pondérée :**

- Le support indique le nombre de vrais exemples présents dans chaque classe. La classe neutre a le plus grand support (46), suivie de la classe positive (40) et négative (34).
- Une accuracy de 0.90 signifie que le modèle classifie correctement 90% des exemples.
- La macro moyenne prend la moyenne des mesures pour chaque classe sans tenir compte du support, ce qui donne une vue équilibrée de la performance du modèle.

- La moyenne pondérée prend en compte le support de chaque classe, offrant une mesure plus représentative de la performance globale du modèle.

➢ **Pour le Matrice de Confusion :**

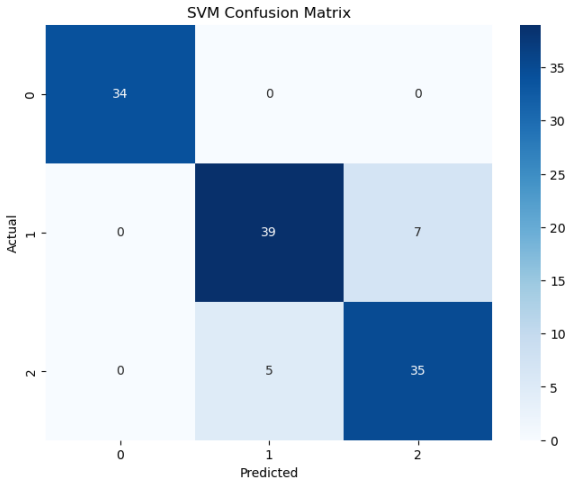

Figure 5.28 – Le matrice de confusion de SVM

- Tous les exemples négatifs (34 exemples) sont correctement classifiés (aucune fausse prédiction).
- Sept exemples neutres sont mal classés comme positifs parmi 46 exemples.
- Cinq exemples positifs parmi 40 exemples sont mal classés comme neutres.

➢ **Pour le Courbes ROC :**

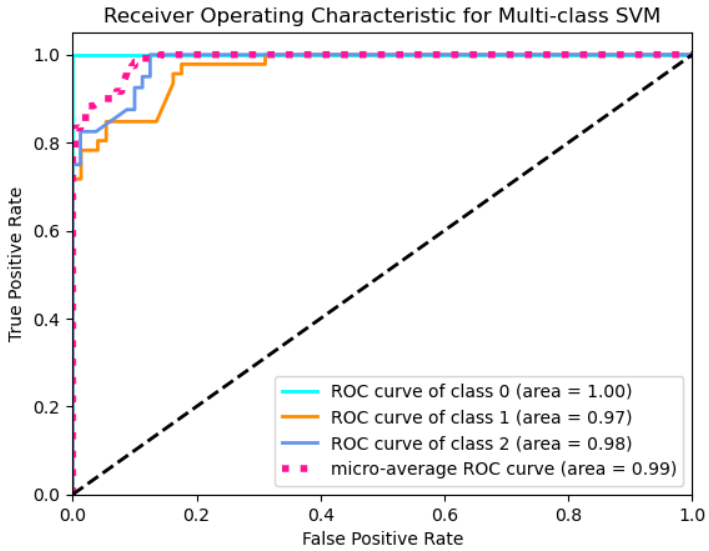

Figure 5.29 – Courbe ROC (Receiver Operating Characteristic) de SVM

- **Classe 0 (négatif):**
  - **AUC**: 1.00
  - **Interprétation**: Une AUC de 1.00 indique une performance parfaite du modèle pour distinguer la classe négative des autres classes. Le modèle classifie tous les exemples négatifs correctement sans aucune erreur.
- **Classe 1 (neutre):**
  - **AUC**: 0.97
  - **Interprétation**: Une AUC de 0.97 montre que le modèle a une excellente capacité à distinguer la classe neutre des autres classes. Bien que ce ne soit pas parfait, le modèle est très proche de la performance idéale.
- **Classe 2 (positif):**
  - **AUC**: 0.98
  - **Interprétation**: Avec une AUC de 0.98, le modèle démontre une capacité presque parfaite à distinguer la classe positive des autres classes. Cela montre que le modèle est très efficace pour classer les exemples positifs correctement.
- **Micro-moyenne:**
  - **AUC**: 0.99
  - **Interprétation**: L'AUC de la micro-moyenne, qui agrège les performances sur toutes les classes, est également très élevée à 0.99. Cela signifie que, globalement, le modèle a une performance très forte et est capable de classer correctement la grande majorité des exemples, indépendamment de leur classe.
- Globalement, les résultats sont également très bons pour le SVM parce que :
  1. **Courbes ROC** : Indiquent une excellente discrimination entre les classes.
  2. **Précision et Rappel** : Très bons pour la classe négatif, légèrement inférieurs pour les classes neutre et positif par rapport à la forêt aléatoire.
  3. **Matrice de Confusion** : Montre que les erreurs de classification se produisent principalement entre les classes neutre et positif.

Le modèle SVM semble bien ajusté et performant pour ce jeu de données, avec une très bonne capacité à distinguer les classes, en particulier la classe négative. Les paramètres optimisés (C=1, gamma=1, kernel='linear') sont adaptés. Toutefois, comme pour la forêt aléatoire, des améliorations pourraient être explorées pour réduire les erreurs entre les classes neutre et positif.

III. Naive Bayes :

Voici les figures ci-dessous qui représente l'évaluation de modèle Naive Bayes :

```
Classification Report:
              precision    recall  f1-score   support

     negatif       0.89      1.00      0.94        34
     neutral       1.00      0.72      0.84        46
     positif       0.82      1.00      0.90        40

    accuracy                           0.89       120
   macro avg       0.90      0.91      0.89       120
weighted avg       0.91      0.89      0.89       120

Confusion Matrix:
[[34  0  0]
 [ 4 33  9]
 [ 0  0 40]]
```

Figure 5.30 – Capture d'écran d'output l'évaluation du modèle Naive Bayes

- **Les mesures de précision, rappel et F1-score pour chaque classe:**

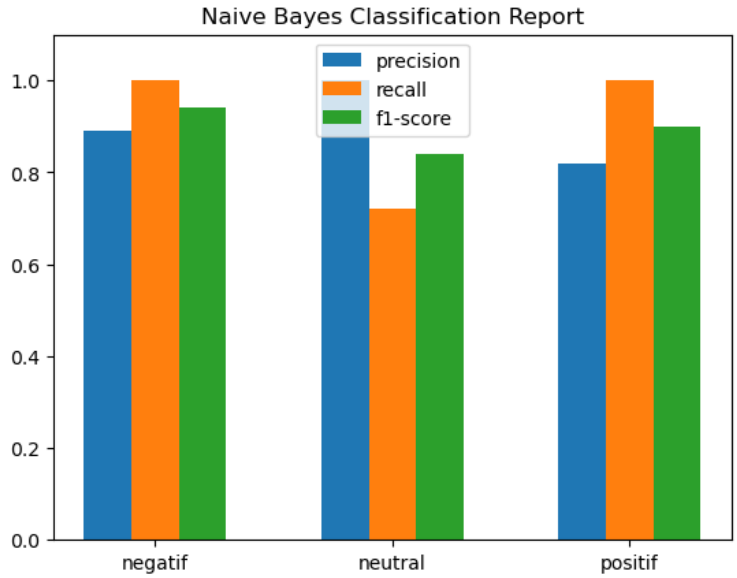

Figure 5.32 – Rapport de Classification avec Naive Bayes

- La précision est élevée pour la classe négative (0.89) et la classe neutre (1.00), mais légèrement inférieure pour la classe positive (0.82). Cela signifie que les prédictions pour les classes négative et neutre sont très fiables, tandis que la classe positive a quelques erreurs.
- Le rappel est parfait pour les classes négative (1.00) et positive (1.00), ce qui indique que le modèle capture tous les exemples de ces classes. En revanche, le rappel pour la

classe neutre est plus bas (0.72), ce qui signifie que certains exemples neutres sont mal classés.

- Le f1-score, qui combine la précision et le rappel, est très bon pour la classe négative (0.94) et la classe positive (0.90), mais légèrement inférieur pour la classe neutre (0.84). Cela montre que le modèle équilibre bien précision et rappel pour les classes négative et positive, mais qu'il y a une marge d'amélioration pour la classe neutre.

➢ **Le support,accuracy,macro moyenne et moyenne pondérée :**

- Le support indique le nombre de vrais exemples présents dans chaque classe. La classe neutre a le plus grand support (46), suivie de la classe positive (40) et négative (34).
- Une accuracy de 0.89 signifie que le modèle classifie correctement 89% des exemples.
- La macro moyenne prend la moyenne des mesures pour chaque classe sans tenir compte du support, ce qui donne une vue équilibrée de la performance du modèle.
- La moyenne pondérée prend en compte le support de chaque classe, offrant une mesure plus représentative de la performance globale du modèle.

➢ **Pour le Matrice de Confusion :**

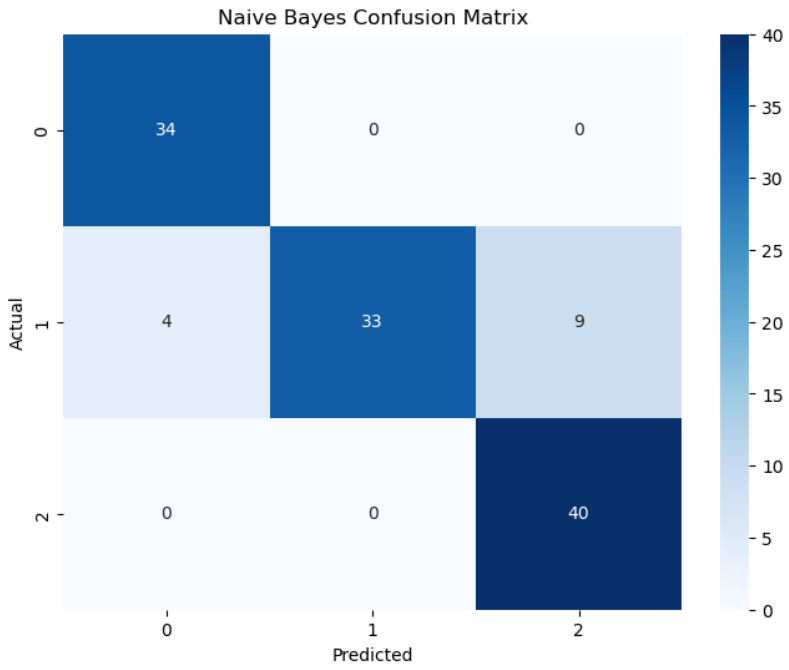

Figure 5.33 – Le matrice de confusion de Naive Bayes

- Tous les exemples négatifs (34 exemples) sont correctement classifiés (aucune fausse prédiction).
- Quatre exemples neutres sont mal classés comme négatifs parmi 46 exemples.
- Neuf exemples neutres sont mal classés comme positifs parmi 46 exemples.
- Tous les exemples positifs (40 exemples) sont correctement classifiés.

➢ **Pour le Courbes ROC**

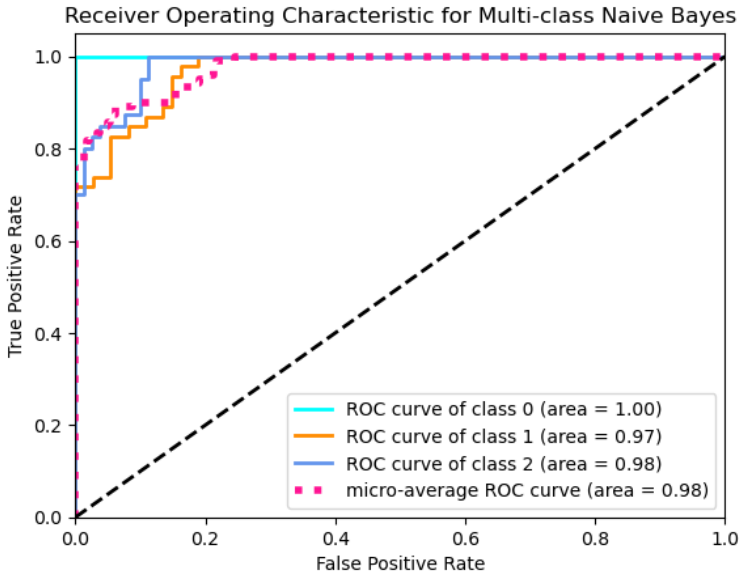

Figure 5.34 – Courbe ROC (Receiver Operating Characteristic) de Naive Bayes

Les courbes ROC (Receiver Operating Characteristic) montrent la performance de notre modèle Naive Bayes pour chaque classe. Voici les détails des courbes ROC :

➢ **Classe 0 (négatif):**
- **AUC**: 1.00
- **Interprétation**: Une AUC de 1.00 indique une performance parfaite du modèle pour distinguer la classe négative des autres classes. Le modèle classifie tous les exemples négatifs correctement sans aucune erreur.

➢ **Classe 1 (neutre):**
- **AUC**: 0.97
- **Interprétation**: Une AUC de 0.97 montre que le modèle a une excellente capacité à distinguer la classe neutre des autres classes. Bien que ce ne soit pas parfait, le modèle est très proche de la performance idéale.

➢ **Classe 2 (positif):**
- **AUC**: 0.98
- **Interprétation**: Avec une AUC de 0.98, le modèle démontre une capacité presque parfaite à distinguer la classe positive des autres classes. Cela montre que le modèle est très efficace pour classer les exemples positifs correctement.

➢ **Micro-moyenne:**
- **AUC**: 0.98
- **Interprétation**: L'AUC de la micro-moyenne, qui agrège les performances sur toutes les classes, est également très élevée à 0.98. Cela signifie que, globalement, le modèle a une performance très forte et est capable de classer correctement la grande majorité des exemples, indépendamment de leur classe.

➢ Globalement, les résultats sont bons pour Naive Bayes :

1. **Précision et Rappel** : Très bons pour les classes négatif et positif. Le rappel pour la classe neutre est légèrement plus faible.

2. **Matrice de Confusion** : Montre que les erreurs de classification se produisent principalement dans la classe neutre.
3. **Courbes ROC**: Indiquent une excellente discrimination entre les classes.

Le modèle Naive Bayes avec les paramètres optimisés (alpha=0.5) est performant pour ce jeu de données, surtout pour les classes négative et positive. Cependant, il pourrait bénéficier de quelques améliorations pour mieux distinguer les exemples de la classe neutre.

5.7 Comparison

Voici un tableau récapitulatif des performances des trois modèles (Random Forest, SVM, et Naive Bayes) basé sur les résultats fournis :

Classe	Métrique	Random Forest	SVM	Naive Bayes
Négatif	Précision	1.00	1.00	0.89
	Rappel	1.00	1.00	1.00
	F1-score	1.00	1.00	0.94
Neutre	Précision	0.83	0.89	1.00
	Rappel	0.98	0.85	0.72
	F1-score	0.90	0.87	0.84
Positif	Précision	0.97	0.83	0.82
	Rappel	0.78	0.88	1.00
	F1-score	0.86	0.85	0.90
Global	Précision	0.93	0.91	0.90
	Rappel	0.92	0.91	0.91
	F1-score	0.92	0.91	0.89
	AUC (moy)	0.99	0.99	0.98

Table 5.1 – Table de comparaison des performances des trois modèles.

Ci-dessous, des graphiques comparant les performances des trois modèles (Random Forest, SVM, et Naive Bayes) basés sur les résultats obtenus :

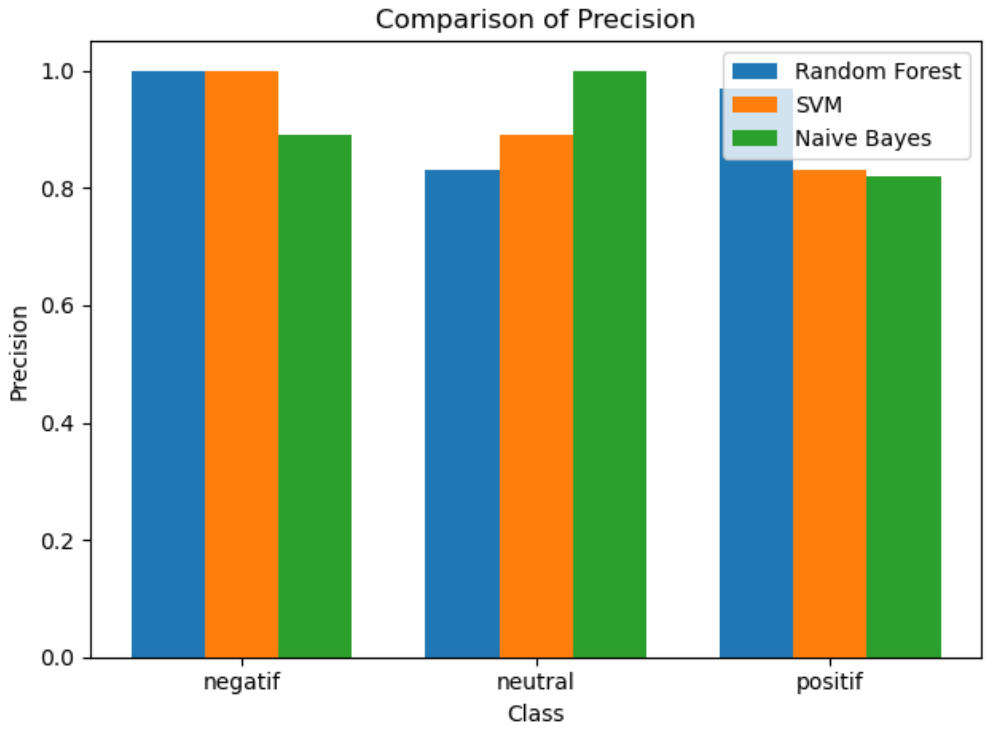

Figure 5.35 – Comparaison de la Précision entre Différents Modèles de Classification

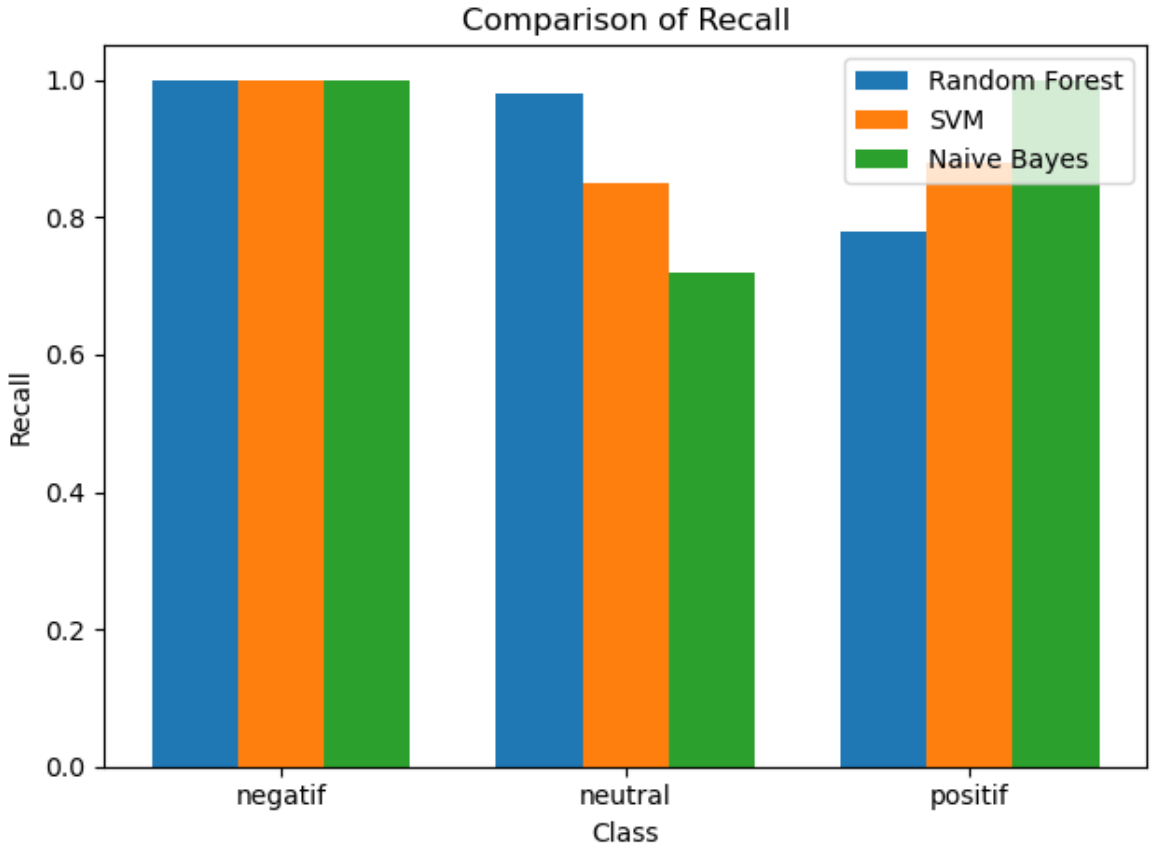

Figure 5.36 – Comparaison de 'Recall' entre Différents Modèles de Classification

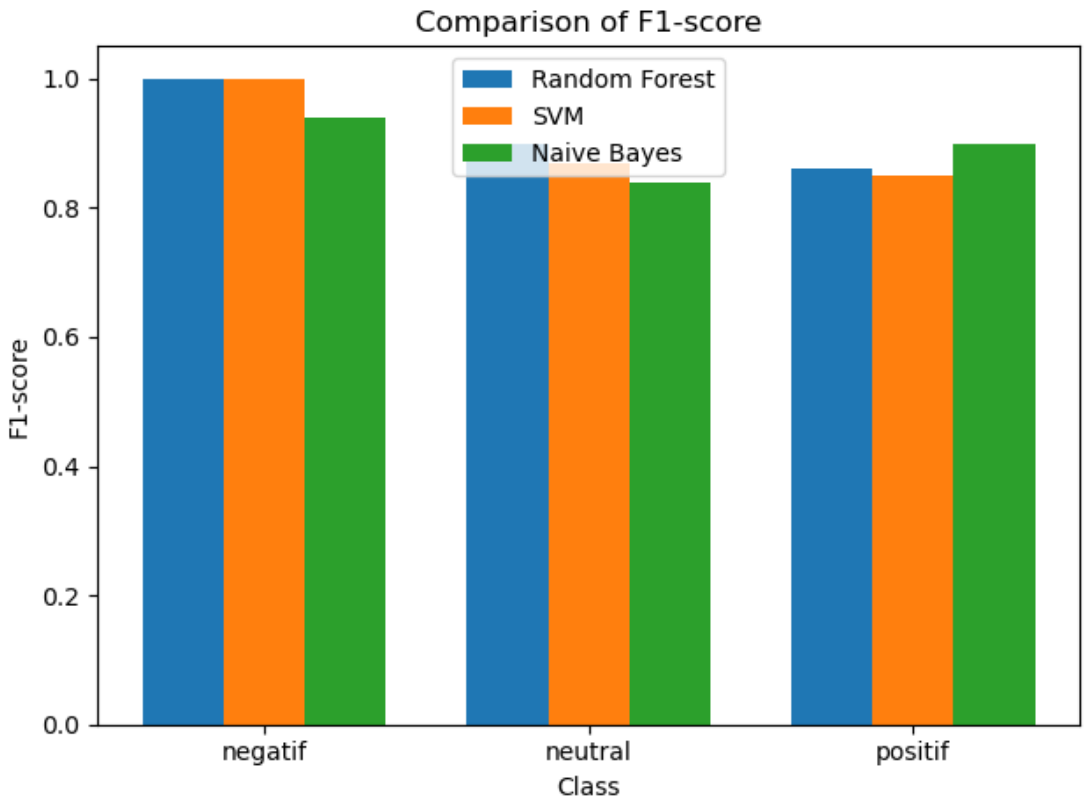

Figure 5.37 – Comparaison de 'F1-score' entre Différents Modèles de Classification

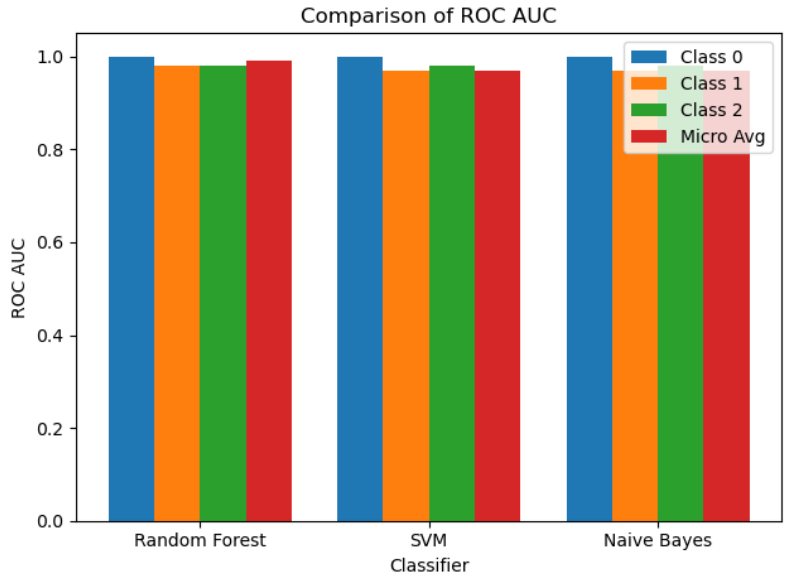

Figure 5.38 – Comparaison de 'ROC AUC' entre Différents Modèles de Classification

Remarques

1. Random Forest : Excellente performance globale avec des scores F1 équilibrés pour toutes les classes.

2. SVM: Très bonne performance globale, en particulier pour la classe négative, mais quelques erreurs entre les classes neutre et positif.

3. Naive Bayes : Très bon pour les classes négative et positive, mais le rappel pour la classe neutre est plus faible.

Les courbes ROC montrent une performance globale similaire en termes de discrimination pour Random Forest et SVM avec un AUC micro-moyenne de 0.99 et même pour Naive Bayes avec un AUC micro-moyenne de 0.98.

Donc, Random Forest est le modèle le plus performant pour ce jeu de données, suivi de SVM et Naive Bayes. Naive Bayes pourrait nécessiter des ajustements supplémentaires pour améliorer la classification des exemples neutres.

5.8 Conclusion

Ce chapitre a fourni une vue complète de notre cadre expérimental, allant de la description des données et des outils utilisés, jusqu'à l'évaluation détaillée des performances des modèles. Ces expérimentations et leurs résultats nous permettent de mieux comprendre l'efficacité de notre approche et de proposer des améliorations pour les travaux futurs. La rigueur de ce processus garantit la robustesse et la fiabilité des conclusions que nous tirons de notre étude.

6 Chapitre 6

Conclusion générale

Ce travail a été réalisé dans le cadre de notre projet de fin de cycle Master en informatique option System d'information avancé (SIA). Il consiste à proposer une approche hybride, combinant des techniques de machine learning et des méthodes basées sur un lexique pour l'analyse des sentiments des commentaires en dialecte algérienne des spectateurs des videos d'éducation sur Youtube algérien.

Ce travail concerne donc l'analyse de sentiment aussi appelée opinion mining, il existe également de nombreux noms : extraction d'opinion, sentiment mining... etc, pour simplifier la présentation, tout au long de ce travail nous avons utilisé le terme analyse de sentiment. Cependant, ces concepts ne sont pas équivalents [20].

Le Dataset qu'on a utilisé a été pris d'une chaine de youtube d'éducation nommé "الاستاذ نور الدين", il contient les commentaires des spectateurs en dialecte algérienne. L'analyse des sentiments a été effectuée sur chaque commentaire et ensuite classifier en utilisant une approche hybride.

L'approche hybride combinant lexique et Machine Learning est particulièrement efficace pour l'analyse des sentiments. Elle permet de passer en revue des milliers de commentaires en utilisant un modèle qui intègre à la fois des règles basées sur un lexique et l'apprentissage automatique.

Dans cette travail, les sentiments des consommateurs ont été analysés en utilisant un modèle de Random Forest,SVM et Naive bayes tout en exploitant un lexique de mots et phrases positifs et négatifs. Cette méthode permet non seulement de capturer des indices sentimentaux explicites, mais aussi d'apprendre des motifs plus subtils dans les données textuelles, améliorant ainsi la précision et la robustesse de la classification sentimentale.

Dans ce mémoire, nous avons exploré et analysé en profondeur le domaine de l'analyse des sentiments appliquée aux vidéos éducatives. Il est constitué de six (6) chapitres organisés comme suit :

Chapitre 1 (Introduction) a présenté le contexte de notre étude, en décrivant la problématique et les objectifs poursuivis. Il a également fourni un aperçu de l'organisation du mémoire, mettant en avant les principaux aspects que nous allions explorer.

Chapitre 2 (Généralités sur l'analyse des sentiments), nous avons introduit les concepts fondamentaux de l'analyse des sentiments, expliquant son importance et ses différents domaines d'application. Nous avons également abordé les divers niveaux d'analyse, les types d'opinions, ainsi que les défis associés à cette discipline.

Chapitre 3 (Etat de l'art) a passé en revue les travaux de recherche existants dans le domaine de l'analyse des sentiments. Une étude comparative et une analyse des différentes

approches utilisées par d'autres chercheurs ont été effectuées, permettant de situer notre travail dans le contexte plus large de la recherche.

Chapitre 4 (Analyse des sentiments sur les vidéos d'éducation), nous avons détaillé notre approche proposée pour analyser les sentiments dans les commentaires des vidéos éducatives. Cela inclut la collecte des données, le prétraitement, l'annotation, et les différentes techniques de classification utilisées, telles que les forêts aléatoires, les machines à vecteurs de support, et les modèles de Naive Bayes.

Chapitre 5 (Expérimentation), nous avons décrit le cadre expérimental de notre étude. Nous avons présenté le dataset utilisé, l'environnement de travail, les bibliothèques Python employées, ainsi que le processus de mise en service et d'évaluation des modèles. Les résultats obtenus ont été analysés et comparés pour mesurer l'efficacité de notre approche.

Nous avons développé le projet au maximum de nos capacités actuelles, mais de nombreuses étapes restent à franchir. Nous prévoyons de créer un site web pour notre modèle, une tâche déjà entamée avec l'extraction et le prétraitement des commentaires. Nous envisageons également d'enrichir notre dataset. Une difficulté rencontrée est la faible quantité de commentaires négatifs, car les vidéos éducatives ne génèrent pas beaucoup de ce type de commentaires donc le modèle manque d'exemples pour bien apprendre à les détecter à l'avenir.

Ce projet a été très enrichissant à bien des égards. Il nous a permis d'acquérir de nouvelles compétences et de mettre en pratique les connaissances théoriques acquises au cours de notre cursus universitaire.

Bibliographie

[1] Kim, Rosemary / Olfman, Lorne / Ryan, Terry / Eryilmaz, Evren, Leveraging a persona-lized system to improve self-directed learning in online educational environments, 2014-01, Computers & Education , Vol. 70, Elsevier BV, p. 150-160.

[2] Lee, Chei Sian / Osop, Hamzah / Goh, Dion Hoe-Lian / Kelni, Gani, Making sense of comments on YouTube educational videos: a self-directed learning perspective, 2017-09 , Online Information Review , Vol. 41, No. 5, Emerald , p. 611-625.

[3] Nehama, Marchal / Hubert, Au / Philip, N Howard, Coronavirus news and information on YouTube: A content analysis of popular search terms, 2020.

[4] Saurabh, Samant / Gautam, Sanjana, Modelling and statistical analysis of YouTube's educational videos: A channel Owner's perspective, 2019-01, Computers & Education , Vol. 128, Elsevier BV, p. 145-158.

[5]Chtouki, Yousra / Harroud, Hamid / Khalidi, Mohammed / Bennani, Samir, The impact of YouTube videos on the student's learning , 2012-06, 2012 International Conference on In-formation Technology Based Higher Education and Training (ITHET), IEEE.

[6] Chareen Snelson, YouTube across the disciplines: A review of the literature, 2011.

[7]Chelaru, Sergiu / Orellana-Rodriguez, Claudia / Altingovde, Ismail Sengor , How useful is social feedback for learning to rank YouTube videos?, 2013-11 , World Wide Web , Vol. 17, No. 5, Springer Science and Business Media LLC, p. 997-1025.

[8] Elizabeth, DuRoss Liddy , Natural Language Processing, in Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc. 2001.

[9] Mohamad Sham, Nabila / Mohamed, Azlinah, Climate Change Sentiment Analysis Using Lexicon, Machine Learning and Hybrid Approaches, 2022-04, Sustainability , Vol. 14, No. 8, MDPI AG, p. 4723.

[10] Gonçalves, Pollyanna / Araújo, Matheus / Benevenuto, Fabrício / Cha, Meeyoung , Comparing and combining sentiment analysis methods, 2013-10, Proceedings of the first ACM conference on Online social networks , COSN'13, ACM.

[l l] Jasper, Feine / Stefan, Morana / Ulrich, Gnewuch, Measuring Service Encounter Satisfaction with Customer Service Chatbots using Sentiment Analysis, in Proceedings of the 14th Interna tional Conference on Wirtschaftsinformatik (WI2019), Siegen, Germany, February 24–27. 2019.

[12] Hoang, Thai Son / Ngo, Thuy Ha / Pham, Thi Minh Khuyen, Measuring students satisfaction with higher education service-An experimental study at Thainguyen University, 2018.

[13] Asghar, Muhammad Zubair / Ullah, Ikram / Shamshirband, Shahab / Kundi, Fazal Ma-sud / Habib, Ammara, Fuzzy-Based Sentiment Analysis System for Analyzing Student Feed-back and Satisfaction, 2019-07, MDPI AG.

[14] Ohliati, Jenny / Abbas, Bahtiar Saleh, Measuring Students Satisfaction in Using Learn-ing Management System, 2019-02 , International Journal of Emerging Technologies in Learning (iJET) , Vol. 14, No. 04, International Association of Online Engineering (IAOE) , p. 180.

[15] Al-Otaibi, Shaha / Alnassar, Allulo / Alshahrani, Asma / Al-Mubarak, Amany / Albugami, Sara / Almutiri, Nada / Albugami, Aisha, Customer Satisfaction Measurement using Sentiment Analysis, 2018, International Journal of Advanced Computer Science and Applications , Vol. 9, No. 2, The Science and Information Organization.

[16] Kang, Daekook / Park, Yongtae, Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach, 2014-03, Expert Systems with Applications , Vol. 41, No. 4, Elsevier BV, p. 1041-1050.

[17] Gang, Zhou / Chenglin, Liao, Dynamic Measurement and Evaluation of Hotel Customer Satisfaction Through Sentiment Analysis on Online Reviews, 2021-10, Journal of Organiza-tional and End User Computing , Vol. 33, No. 6, IGI Global, p. 1-27.

[18] Mäntylä, Mika V / Graziotin, Daniel / Kuutila, Miikka, The evolution of sentiment analysis—A review of research topics, venues, and top cited papers, 2018-02, Computer Science Review , Vol. 27, Elsevier BV, p. 16-32.

[19]Alaoui, Imane El / Gahi, Youssef / Messoussi, Rochdi, Full Consideration of Big Data Characteristics in Sentiment Analysis Context, 2019-04, 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), IEEE.

[20] Liu, Bing, Sentiment Analysis and Opinion Mining, 2012, Synthesis Lectures on Human Language Technologies , Springer International Publishing.

[21] Liu, Bing, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions , 2015-06, Cambridge University Press.

[22] Gundla, Apurva V, A Review on Sentiment Analysis and Visualization of Customer Reviews, 2016-08, International Journal Of Engineering And Computer Science, Valley In-ternational.

[23] Wankhade, Mayur / Rao, Annavarapu Chandra Sekhara / Kulkarni, Chaitanya, A survey on sentiment analysis methods, applications, and challenges
2022-02, Artificial Intelligence Review , Vol. 55, No. 7, Springer Science and Business Me-dia LLC, p. 5731-5780.

[24] Aqlan, Ameen Abdullah Qaid / Manjula, Bairam / Lakshman Naik, R, A Study of Sentiment Analysis: Concepts, Techniques, and Challenges, 2019, Lecture Notes on Data Engineering and Communications Technologies, Springer Singapore, p. 147-162.

[25] Zunic, Anastazia / Corcoran, Padraig / Spasic, Irena, Sentiment Analysis in Health and Well-Being: Systematic Review, 2020-01, JMIR Medical Informatics , Vol. 8, No. 1, JMIR Publications Inc. p. e16023.

[26]BV,Pranay,Kumar/M,Sadanandam, A Comprehensive Review of Approaches, Methods, and Challenges and Applications in Sentiment Analysis , 2021, Turkish Journal of Computer and Mathematics Education ,Vol.12 No.14, p.6136-6161.

[27] Shaik, Than-veer / Tao, Xiaohui / Dann, Christopher / Xie, Haoran / Li, Yan / Galligan, Linda, Sentiment analysis and opinion mining on educational data: A survey, 2023-03, Natural Language Processing Journal , Vol. 2, Elsevier BV, p. 100003.

[28] Liu, Zhi / Yang, Chongyang / Rüdian, Sylvio / Liu, Sannyuya / Zhao, Liang / Wang, Tai, Temporal emotion-aspect modeling for discovering what students are concerned about in online course forums, 2019-04, Interactive Learning Environments , Vol. 27, No. 5–6, Informa UK Limited, p. 598-627.

[29] Zhu, John Jianjun / Chang, Yung-Chun / Ku, Chih-Hao / Li, Stella Yiyan / Chen, Chi-Jen, Online critical review classification in response strategy and service provider

rating: Algorithms from heuristic processing, sentiment analysis to deep learning, 2021-05, Journal of Business Research , Vol. 129 , Elsevier BV, p. 860-877.

[30] Acheampong, Francisca Adoma / Nunoo-Mensah, Henry / Chen, Wenyu, Transformer models for text-based emotion detection: a review of BERT-based approaches, 2021-02, Artificial Intelligence Review , Vol. 54, No. 8, Springer Science and Business Media LLC, p. 5789-5829.

[31] Kuleto, Valentin / Ilić, Milena / Dumangiu, Mihail / Ranković, Marko / Martins, Oliva M. D. / Păun, Dan / Mihoreanu, Larisa, Exploring Opportunities and Challenges of Artificial Intelligence and Machine Learning in Higher Education Institutions, 2021-09, Sustainability , Vol. 13, No. 18, MDPI AG, p. 10424.

[32] Qaqish, Evon / Aranki, Aseel / Etaiwi, Wael, Sentiment analysis and emotion detection of post-COVID educational Tweets: Jordan case, 2023-03, Social Network Analysis and Mining , Vol. 13, No. 1, Sprin-ger Science and Business Media LLC.

[33] Lin, Fangyuan, Sentiment analysis in online education: An analytical approach and application, 2024-02, Applied and Computational Engineering , Vol. 33, No. 1, EWA Publishing , p. 9-17.

[34] Ansari, Mohd Zeeshan / Aziz, M. B. / Siddiqui, M. O. / Mehra, H. / Singh, K. P. Analysis of Political Sentiment Orientations on Twitter, 2020, Procedia Computer Science , Vol. 167, Elsevier BV, p. 1821-1828.

[35] Nasim, Zarmeen / Rajput, Qu-ratulain / Haider, Sajjad, Sentiment analysis of student feedback using machine learning and lexicon based approaches, 2017-07, 2017 International Conference on Research and Innova-tion in Information Systems (ICRIIS), IEEE.

[36] Bensba, Amal / Ahmim, Naima / Zakaria, Chahnez / Bousbia, Nabila, Analysis of Students' Emotions in an Online Learning Environment, 2022-09, 2022 International Conference on Advanced Aspects of Software Engineering (ICAASE), IEEE

[37] Ameur, Mohamed Seghir Hadj / Aliane, Hassina, AraCOVID19-SSD: Arabic COVID-19 Sentiment and Sarcasm Detection Dataset, 2021, arXiv.

[38] Kechaou, Zied / Ben Ammar, Mohamed / Alimi, Adel. M. Improving e-learning with senti-ment analysis of users' opinions, 2011-04, 2011 IEEE Global Engineering Education Confe-rence (EDUCON), IEEE.

[39] Colace, Francesco / De Santo, Massimo / Greco, Luca, SAFE: A Sentiment Analysis Framework for E-Learning, 2014-12, International Journal of Emerging Technologies in Learning (iJET) , Vol. 9, No. 6, International Association of On-line Engineering (IAOE), p. 37.

[40] Clarizia, Fabio / Colace, Francesco / De Santo, Massimo / Lombardi, Marco / Pascale, Francesco / Pietrosanto, Antonio, E-learning and sentiment analysis: a case study, 2018-01 ,Proceedings of the 6th International Conference on Information and Education Technology ICIET '18, ACM.

[41] Nandal, Neha / Tanwar, Rohit / Pruthi, Jyoti, Machine learning based aspect level sentiment analysis for Amazon products, 2020-02,Spatial Information Research , Vol. 28, No. 5, Springer Science and Business Media LLC, p. 601-607

[42] 'NoxInfluencer' ,[Online], disponible: https://www.capterra.com/p/187786/NoxInfluencer/, consulté le : 05/07/2024.

[43]'Standard of Grading about Noxscore', [Online], disponible: https://www.noxinfluencer.com/youtube/score/UCrwAK 5m3hatI0Ftexh0fxQ, consulté le : 05/07/2024.

[44] Pranayteja, 'YouTube Comments Sentiment Analysis using YouTube Data API v3', [Online],disponible:https://medium.com/@pranayteja270/youtube-comments-sentiment-analysis-using-youtube-data-api-v3-bf4a2a041144] , consulté le : 05/07/2024.

[45] Liste des chaines youtube : consulté le : 05/07/2024.

1. chaine : Dz Brain,'Théorie des langages : Chapitre N° 3 Définition formelle d'une grammaire', [Online], disponible: https://www.youtube.com/watch?v=sYDBM8657yE&ab_channel=DzBrain.
2. chaine : Dz Brain, 'Théorie des langages : Chapitre N° 3 : Classification des grammaires :', [Online], disponible: https://www.youtube.com/watch?v=dlUYsXqVX7k&ab_channel=DzBrain.
3. chaine : Dz Brain, 'Théorie des langages : Chapitre N° 3 : Classification des grammaires :', [Online], disponible: https://www.youtube.com/watch?v=yikm0NvjenA&ab_channel=DzBrain.
4. chaine : Dz Brain, 'ALGEBRE 01 : résumé (partie 01)', [Online], disponible: https://www.youtube.com/watch?v=0wxsIh8GyKU&ab_channel=DzBrain.

5. chaine : Math Info DZ, ‘theorie des langages Rappel mathematique partie 1’, [Online], disponible: https://www.youtube.com/watch?v=gsNhs3R95Po&list=PLz8b0eOENotoKJhBwiBX6MGN5qhWCiUN&ab_channel=MathInfoDZ.
6. chaine : Amar Moulai, ‘2 قذيفة علوم تجريبية الموضوع الثاني الجزء 2016 باك’, [Online], disponible: https://www.youtube.com/watch?v=cjMeWft7jtM&ab_channel=AmarMoulai.
7. chaine : Math Info DZ,[Online], disponible: https://www.youtube.com/watch?v=IV0l26m9qE&list=PLz8b0eOENotoKJhBwiBX6MGN5qhWCiUN&index=6&ab_channel=MathInfoDZ.

[46] Ruofei Shang, ‘Modernity or Colonialism? The Use of ‘Arabizi’ and Its Controversy ‘, [en ligne],disponible:https://www.irreview.org/articles/the-use-of-arabizi-and-its-controversy, consulté le : 05/07/2024.

[47] Rajiv Chandra ,‘Text Cleaning in Python: Effective Data Cleaning Tutorial ‘, [en ligne],disponible: https://ecoagi.ai/topics/Python/text-cleaning-python, consulté le : 05/07/2024.

[48] Damazouz ,‘algerian_arabic_stopwords.txt ‘, [en ligne],disponible: https://github.com/Damazouz/Algerian-Arabic-stop-words , consulté le : 05/07/2024.

[49]genediazjr, ‘stopwords-fr.txt ‘, [en ligne],disponible:https://github.com/stopwords-iso/stopwords-fr/blob/master/stopwords-fr.txt, consulté le : 05/07/2024.

[50]‘Anaconda ‘, [en ligne],disponible:https://domino.ai/data-science-dictionary/anaconda, consulté le : 05/07/2024.

[51]‘Jupyter Notebook ‘, [en ligne],disponible: https://domino.ai/data-science-dictionary/jupyter-notebook, consulté le : 05/07/2024.

[52] ‘What is Python? ‘, [en ligne],disponible: https://aws.amazon.com/what-is/python/, consulté le : 05/07/2024.

[53] ‘What is NumPy? ‘, [en ligne], disponible: https://numpy.org/doc/stable/user/whatisnumpy.html, consulté le : 05/07/2024.

[54] ‘Pandas Introduction ‘, [en ligne],disponible:

https://www.w3schools.com/python/pandas/pandas_intro.asp, consulté le : 05/07/2024.

[55] ‘sklearn ‘, [en ligne],disponible:https://domino.ai/data-science-dictionary/sklearn, consulté le : 05/07/2024.

[56] ‘Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning ‘, [en ligne],disponible: https://jmlr.org/papers/v18/16-365.html#:~:text=imbalanced-learn%20is%20an%20open,machine%20learning%20and%20pattern%20recognition, consulté le : 05/07/2024.

[57] ‘Matplotlib: Visualization with Python ‘, [en ligne],disponible: https://matplotlib.org/, consulté le : 05/07/2024.

[58] Pallav Sharma,‘How to Save and Load Machine Learning Models in Python Using Joblib Library? ‘, [en ligne], disponible: https://www.analyticsvidhya.com/blog/2023/02/how-to-save-and-load-machine-learning-models-in-python-using-joblib-library/#:~:text=Joblib%20is%20a%20Python%20library,results%20of%20computationally%20expensive%20functions, consulté le : 05/07/2024.

[59] Ashutosh Agrawal ,‘Understanding Python pickling and how to use it securely ‘, [en ligne],disponible: https://www.synopsys.com/blogs/software-security/python-pickling.html#:~:text=Pickle%20in%20Python%20is%20primarily,transport%20data%20over%20the%20network, consulté le : 05/07/2024.

[60] ‘re — Regular expression operations ‘, [en ligne], disponible: https://docs.python.org/3/library/re.html, consulté le : 05/07/2024.

[61] David Amos, ‘Itertools in Python 3, By Example ‘, [en ligne], disponible: https://realpython.com/python-itertools/#:~:text=According%20to%20the%20itertools%20docs,and%20efficiently%20in%20pure%20Python, consulté le : 05/07/2024.

[62] ‘, [en ligne], disponible: https://www.w3schools.com/python/scipy/scipy_intro.php#:~:text=SciPy%20is%20a%20scientific%20computation,by%20NumPy%27s%20creator%20Travis%20Ollipha nt, consulté le : 05/07/2024.

[63] ‘Python os Module ‘, [en ligne],disponible:
https://www.w3schools.com/python/module_os.asp, consulté le : 05/07/2024.

[64] ‘, [en ligne], disponible:
https://googleapis.github.io/google-api-python-client/docs/#:~:text=The%20Google%20API%20Client%20Library,access%20to%20many%20Google%20APIs, consulté le : 05/07/2024.

[65] ‘unicodedata — Unicode Database ‘, [en ligne],disponible:
https://docs.python.org/3/library/unicodedata.html, consulté le : 05/07/2024.

[66] linuxscout ,‘PyArabic 0.6.15 ‘, [en ligne],disponible:https://pypi.org/project/PyArabic/, consulté le : 05/07/2024.

[67] 3aransia ,‘aaransia 1.1 ‘, [en ligne],disponible:
https://pypi.org/project/aaransia/, consulté le : 05/07/2024.

[68] ‘Natural Language Toolkit ‘, [en ligne],disponible:
https://en.wikipedia.org/wiki/Natural_Language_Toolkit#:~:text=The%20Natural%20Lan-guage%20Toolkit%2C%20or,parsing%2C%20and%20semantic%20reasoning%20functionalities, consulté le : 05/07/2024.

[69] ‘Python Collections Module ‘, [en ligne],disponible:
https://www.geeksforgeeks.org/python-collections-module/, consulté le : 05/07/2024.

[70] ‘ast — Abstract Syntax Trees ‘, [en ligne],disponible:
https://docs.python.org/3/library/ast.html, consulté le : 05/07/2024.