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

Université A/Mira de Béjaia
Faculté des Sciences Exactes
Département d'Informatique

MÉMOIRE DE MASTER RECHERCHE

En Informatique

Option

Intelligence Artificielle

Thème

Analyse de sentiment sur twitter

Présenté par : Mr AKIF AMINE

Mlle BELHOCINE LILIA

Soutenu le 13 Septembre 2023 devant le jury composé de :

Présidente	**Mme ALOUI soraya**	**Maître de conf. A**	**U. A/Mira Béjaia.**
Encadrant	**Mr AKILAL Karim**	**Maître De conf B**	**U. A/Mira Béjaia.**
Examinatrice	**Mme TAHAKOURT Zineb**	**Maître de conf. B**	**U. A/Mira Béjaia.**

Béjaia, Septembre 2023.

Remerciements

Tout d'abord, nous aimerions exprimer notre profonde gratitude envers notre université pour nous avoir offert l'opportunité de poursuivre nos études. Les ressources mises à notre disposition ont été cruciales pour la réalisation de ce travail.

Nous aimerions également remercier Mr. Akilal pour sa disponibilité, sa réactivité et son expertise. Ses conseils éclairés ont été inestimables dans l'élaboration de notre travail. De plus, nous souhaitons exprimer notre reconnaissance envers les membres du jury pour avoir accepté d'évaluer notre mémoire. Vos commentaires constructifs et votre expertise sont très précieux pour nous.

Enfin nous exprimons notre profonde reconnaissance à tous les responsables et enseignants de l'université de Bejaia qui ont contribué à notre formation.

Dédicaces

Dédicaces Amine :

Ce mémoire est dédié à mes parents et à mes frères, qui ont toujours été ma source d'inspiration, de soutien et d'amour inconditionnel. Votre confiance en moi a été le moteur de ma persévérance.

Je tiens également à dédier ce travail à ma deuxième famille, ma tante Ouardhia et mes chers cousins et cousines Redouane, Nadira, Sassi et Nadjet. Vous avez été plus qu'une famille d'accueil pendant mes années d'études, vous avez été un pilier essentiel de mon épanouissement académique.

À mes amis dévoués, vous avez été mes compagnons de route dans cette aventure académique. Votre amitié, vos encouragements et votre présence m'ont donné la force de persévérer dans les moments difficiles.

À Mr Akilal, Mr Bouchebah, à l'université de Abderahman Mira, aux membres du jury, à nos collègues et camarades de classe, je vous adresse également mes remerciements sincères pour votre contribution à ma formation académique.

Merci à tous pour avoir été une part précieuse de mon voyage vers la réalisation de ce mémoire de master 2.

Dédicaces lilia :

Je dédie ce travail à ces êtres chers qui m'ont épaulé inlassablement durant ces 5 dernières. Ma famille, une source constante de soutien, ma mère, la pierre angulaire de ma vie, mon frère, le roc de mon quotidien, ma tante qui ma ouvert ses porte,et celle qui dynamise mes journée Maylisse.

Je ne peux oublier de remercier mes chères amies, compagnes de mon parcours : à Kamilia, Damya, Kahina, Wahiba, wissam,Hanane,Nadjet votre impact n'est pas des moindres

Sommaire

Dédicaces **II**

Sommaire **III**

Table des figures **1**

Liste des tableaux **1**

Introduction générale **2**

1 Analyse de sentiments **4**
1.1 Introduction 4
1.2 Les reseaux sociaux 4
1.3 Analyse des sentiments 7
1.4 Étapes d'analyse des sentiments 9
1.5 Défis d'analyse des sentiments 10
1.6 Applications d'analyse des sentiments 12
1.7 Conclusion 14

2 Apprentissage Automatique **15**
2.1 Introduction 15
2.2 Machine learning 16
2.2.1 Les Algorithmes de Machine Learning 16
2.3 Approches d'apprentissage automatique utilisées dans l'analyse des sentiments 18
2.3.1 Random Forest 18
2.3.2 La régression logistique 19
2.3.3 Support Vector Machine 19
2.3.4 Apprentissage Ensembliste 20

2.4 Conclusion 21

3 Etat de l'art **22**
3.1 Introduction 22
3.2 Aperçu des approches trouvées dans la littérature 22
3.2.1 Approches Machine Learning 22
3.2.2 Approaches hybrides 25
3.2.3 Deep Learning Approches 27
3.3 Tableau Comparatif 28
3.3.1 Comparaison des approches de la littérature 29
3.4 Conclusion 29

4 Approche proposée **35**
4.1 Introduction 35
4.2 Plateformes et outils de développement 36
4.2.1 Environnement de développement 36
4.2.2 Langage de programmation 36
4.2.3 Bibliothèques Python 36
4.3 Métriques d'évaluation 37
4.4 Dataset Utilisé 38
4.5 Prétraitement 39
4.5.1 Dataset Shuffling 39
4.5.2 Tokenization 40
4.5.3 Suppression de bruit 40
4.5.4 Lemmatization 40
4.5.5 Gestion des mots rares 41
4.6 Selection Des Fonctionnalités 41
4.6.1 Longueur du texte 41
4.6.2 Caractéristiques spécifiques au domaine 42
4.6.3 POS Taging features 42
4.7 Description de l'approche proposée 42
4.7.1 Importations et Configuration 44
4.7.2 Mélange de l'ensemble de données 44
4.7.3 Prétraitement des données 44
4.7.4 Architecture du modèle 45
4.7.5 l'entraînement du modèle 45
4.8 Algorithmes compares 45
4.9 Résultats expérimentaux 46
4.10 Discussion des résultats 47
4.11 Conclusion 47

Conclusion générale **49**

Abstract **59**

Table des figures

1.1 Réseaux sociaux [25] 6
1.2 Statistiques sur les réseaux sociaux dans le monde. [79] 6
1.3 Statistiques sur les réseaux sociaux en Algérie. [76] 7
1.4 Roue des émotions de Robert Plutchik. [104] 8
1.5 Sentiments. [2] 8
1.6 Étapes d'analyse des sentiments. 11

4.1 Modéle Proposé. 43

Liste des tableaux

3.1 Résumé des travaux de la littérature (1) 30
3.2 Résumé des travaux de la littérature (2) 31
3.3 Résumé des travaux de la littérature (3) 32
3.4 Résumé des travaux de la littérature (4) 33
3.5 Résumé des travaux de la littérature (5) 34

4.1 Résultats de l'étude comparative 46

Introduction générale

Depuis leur émergence, les réseaux sociaux ont révolutionné la façon dont les individus interagissent, partagent des informations et expriment leurs opinions.
Parmi ces plateformes, Twitter s'est imposé comme un espace privilégié pour la libre expression et la diffusion d'idées à l'échelle mondiale. Dans ce contexte, l'analyse de sentiments sur les réseaux sociaux est devenue une discipline de recherche et d'application cruciale, offrant des opportunités fascinantes pour comprendre les opinions et les émotions des utilisateurs à grande échelle.

À l'ère du numérique, les médias sociaux font désormais partie intégrante de nos vies, influençant et reflétant les sentiments de la société à grande échelle. L'analyse des sentiments sur les médias sociaux est vitale, car elle permet de connaître en temps réel l'opinion et les émotions du public.

Cette recherche explore l'importance et les applications de l'analyse des sentiments dans le paysage numérique contemporain. Comprendre les sentiments sur les médias sociaux est essentiel pour les entreprises, les gouvernements, et toute personne cherchant à comprendre et à s'engager efficacement dans le monde numérique. Cette étude se penche sur les nuances de l'analyse des sentiments, ses implications et son potentiel pour informer la prise de décision et façonner les expériences des utilisateurs.

Le présent mémoire se propose d'explorer en profondeur l'analyse de sentiments sur Twitter, en mettant l'accent sur les méthodes, les outils, et les défis associés à cette pratique. À travers cette étude, nous chercherons à comprendre le fonctionnement des méthodes d'analyse de sentiments qui sont appliquées sur les émotions, les attitudes, et les opinions exprimées par les utilisateurs sur cette plateforme, et comment l'analyse de ces sentiments peut être utilisée pour des applications concrètes.

Dans un premier temps, nous examinerons les fondements conceptuels de l'analyse de sentiments, en nous concentrant sur les théories et les approches couramment utilisées pour l'évaluation des sentiments. Nous aborderons également les principaux défis liés à l'analyse de sentiments ainsi que ses différentes applications.

Ensuite, nous présenterons un aperçu détaillé des différentes méthodes employées pour l'analyse de sentiments sur Twitter. Ceci inclura l'utilisation de techniques de traitement automatique du langage naturel et d'apprentissage automatique pour extraire et classifier les sentiments exprimés dans les tweets. Nous mettrons en lumière les avantages et les limites

de ces méthodes, tout en discutant leurs applications potentielles.

Dans la partie empirique de notre mémoire, nous procéderons à une étude approfondie de l'analyse de sentiments sur Twitter en utilisant un échantillon représentatif de tweets. Nous explorerons les différentes dimensions des sentiments exprimés sur le sujet spécifique abordé dans l'échantillon. Nous analyserons également les différents résultats obtenus par l'utilisation des différentes méthodes d'analyse sur l'échantillon. Et pour finir nous proposerons notre propre méthode d'analyse et nous fournirons des détails sur les performances de notre méthode.

Notre étude comparative a démontré que l'approche hybride(CNN-BiLSTM-BiGRU) que nous proposons est plus performante que les méthodes conventionnelles, ce qui en fait un candidat convaincant pour une mise en œuvre innovante. Notre recherche s'efforce de fournir une compréhension complète de l'analyse des sentiments sur Twitter, en englobant les théories, les méthodes, les applications et les défis de cette discipline. En explorant les opinions et les émotions exprimées sur cette plateforme, nous aspirons à contribuer à une compréhension plus profonde du sentiment public et à ouvrir de nouvelles voies pour l'application de l'analyse du sentiment dans divers domaines.

En synthèse, ce mémoire vise à fournir une compréhension approfondie de l'analyse de sentiments sur Twitter, en examinant les théories, les méthodes, les applications et les défis liés à cette discipline. En explorant les opinions et les émotions exprimées sur cette plateforme, nous espérons contribuer à une meilleure compréhension de l'opinion publique et à de nouvelles perspectives d'application de l'analyse de sentiments dans divers domaines.

Chapitre 1

Analyse de sentiments

1.1 Introduction

Dans notre monde de plus en plus connecté, les réseaux sociaux sont devenus un moyen incontournable pour les individus et les organisations de s'exprimer et de partager des informations en temps réel. Twitter est l'un des réseaux sociaux les plus populaires, et les utilisateurs y partagent une quantité incroyable d'opinions, d'émotions, et de pensées sur une variété de sujets.

L'analyse de sentiments sur Twitter est une méthode qui permet de comprendre ce qui se cache derrière ces données, en identifiant les opinions positives, négatives, ou neutres exprimées par les utilisateurs. Cette analyse peut être utile pour la prise de décisions et la planification stratégique.

1.2 Les reseaux sociaux

Les réseaux sociaux sont des plateformes numériques qui permettent aux utilisateurs de se connecter et d'interagir avec d'autres personnes, de partager du contenu, de discuter de sujets d'intérêt commun, et de créer des communautés en ligne[33].

Leur histoire remonte aux années 1970, lorsque les premiers systèmes de messagerie électronique ont été développés pour permettre aux utilisateurs de communiquer entre eux sur des réseaux informatiques [32].

Au cours des dernières décennies, les réseaux sociaux ont évolué et ont connu une croissance exponentielle. Voici une liste de quelques-uns des réseaux sociaux les plus populaires (Voir La figure 1.1 1.2 1.3) [29, 64, 84] :

Facebook : avec plus de 2,8 milliards d'utilisateurs actifs mensuels, Facebook est le plus grand réseau social au monde. Il permet aux utilisateurs de se connecter avec des amis et de la famille, de rejoindre des groupes et de suivre des pages d'entreprises [33, 101].

Instagram : cette plateforme de partage de photos et de vidéos compte plus de 1 milliard d'utilisateurs actifs mensuels. Elle permet aux utilisateurs de partager des images et

des vidéos, de suivre d'autres utilisateurs et de découvrir du contenu à travers des hashtags et des tendances [32].

Twitter : ce réseau social permet aux utilisateurs d'envoyer des messages courts appelés "tweets" et de suivre d'autres utilisateurs. Il compte plus de 330 millions d'utilisateurs actifs mensuels [32].

LinkedIn : cette plateforme est conçue pour les professionnels qui cherchent à se connecter avec d'autres professionnels et à trouver des opportunités d'emploi. Elle compte plus de 740 millions de membres dans le monde [32].

TikTok : cette application de partage de vidéos courtes est populaire auprès des jeunes. Elle permet aux utilisateurs de créer des vidéos en ajoutant de la musique et des effets spéciaux [32].

Snapchat : cette application de messagerie éphémère permet aux utilisateurs d'envoyer des messages qui disparaissent après avoir été vus. Elle est populaire auprès des jeunes utilisateurs [32].

Pinterest : cette plateforme de partage de photos permet aux utilisateurs de créer des tableaux d'images qui représentent leurs centres d'intérêt. Elle compte plus de 400 millions d'utilisateurs actifs mensuels [32].

Il existe de nombreux autres réseaux sociaux, chacun avec ses propres caractéristiques et son propre public [32].

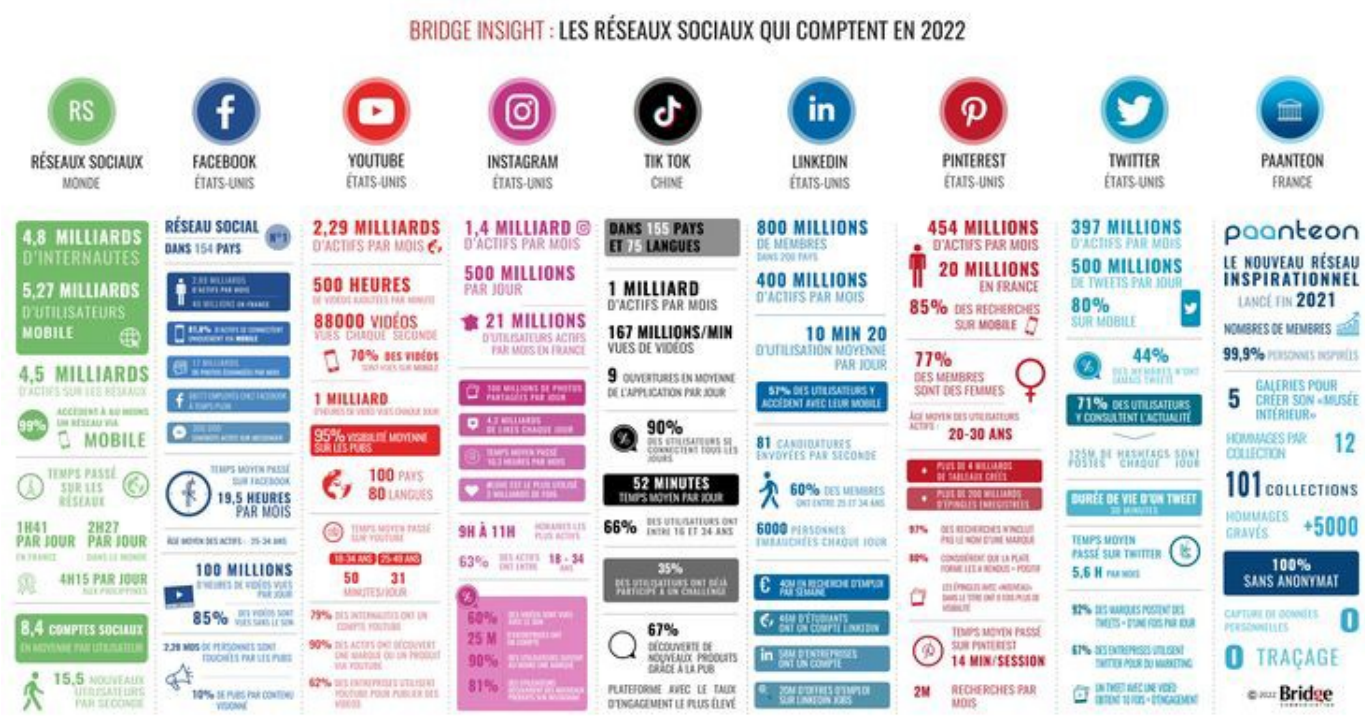

FIGURE 1.1 – Réseaux sociaux [25]

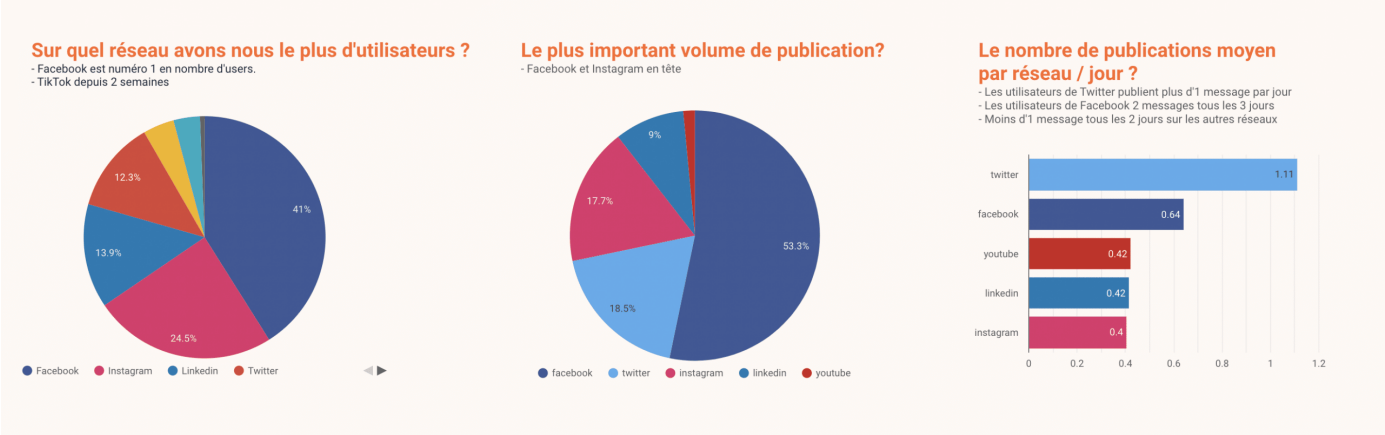

FIGURE 1.2 – Statistiques sur les réseaux sociaux dans le monde. [79]

Chiffres Clés de Facebook en Algérie

L'audience présente sur les réseaux sociaux

Réseau social	Nbre d'inscrits		
	24 000 000	63%	38%
	4 900 000	56%	44%
	3 650 000	40%	60%
	2 300 000	70%	30%
	846 500	70%	30%

FIGURE 1.3 – Statistiques sur les réseaux sociaux en Algérie. [76]

1.3 Analyse des sentiments

le terme *"analyse des sentiments"* fait référence au domaine d'étude et de recherche qui analyse les sensations, les émotions , les sentiments, les opinions, les points de vue, les jugements, les évaluations, les attitudes, et les comportements que les gens ont à l'égard des entités (comme les produits, les services, les organisations, les individus, les événements, les sujets) et de leurs attribut [21, 83].

De nos jours, cette analyse implique tous les médias possibles que les gens utilisent pour exprimer leur opinion, comme l'écriture, la parole, les mouvements, l'expression faciale, les actions, etc. Cependant, aujourd'hui encore, parmi tous les médias possibles, le principal champ d'application de l'AS est l'analyse du langage naturel[66]. Dans ce domaine, l'AS peut être considérée comme l'ensemble des activités visant à analyser et à extraire les opinions, les jugements, les émotions et les sentiments contenus dans un texte, écrit par un utilisateur au sujet d'une entité spécifique [83].

Les termes *"sentiment"*, *"émotion"* et *"opinion"* sont souvent utilisés de manière interchangeable, mais ils ont en réalité des significations distinctes.

Un sentiment est une réponse émotionnelle de base à une situation ou un stimulus, comme la joie, la tristesse, la peur, la colère, ou la surprise. Les sentiments sont généralement considérés comme universels et innés, et peuvent être exprimés de manière non verbale (Voir La figure 1.5) [67].

Une émotion, quant à elle, est une réponse psychologique plus complexe et subjective à une situation ou un stimulus. Les émotions sont souvent liées à des pensées et des croyances, et peuvent inclure des sentiments ainsi que des réponses physiologiques, telles que des changements dans le rythme cardiaque ou la respiration (Voir La figure1.4)[67].

Enfin, une opinion est une évaluation subjective et argumentée d'un sujet ou d'une question. Les opinions sont formées sur la base d'expériences personnelles, de croyances, de valeurs, et de connaissances individuelles [21, 53].

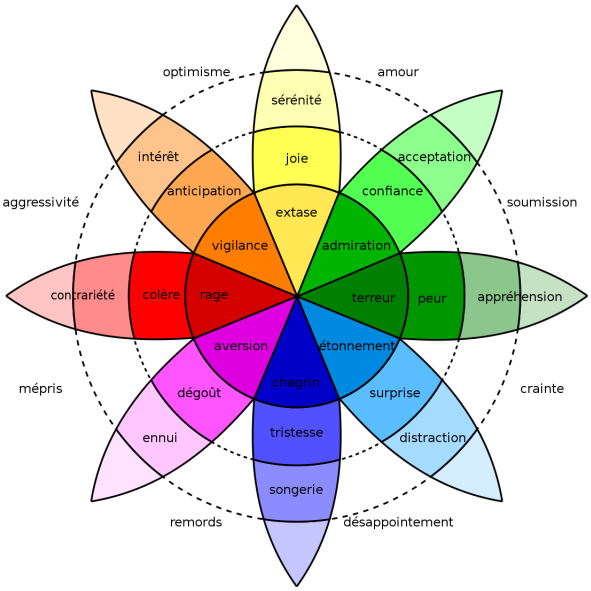

FIGURE 1.4 – Roue des émotions de Robert Plutchik. [104]

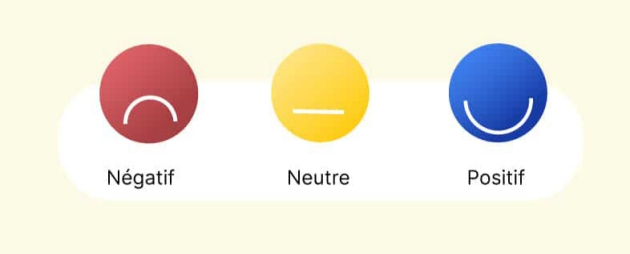

FIGURE 1.5 – Sentiments. [2]

1.4 Étapes d'analyse des sentiments

Les étapes d'analyse des sentiments sont les suivantes (Voir La figure1.6) :

Extraction de données Les données sont collectées à partir des requêtes de discussions d'utilisateurs sur des espaces publics tels que des blogs, des forums de discussions et des tableaux de critiques de produits, ainsi que sur des journaux privés via des sites de réseaux sociaux tels que Twitter et Facebook. Très souvent, le journal de données est volumineux, désorganisé et dispersé sur plusieurs portails stockées dans une base de données. Une fois les données extraites, elles seront ensuite préparées pour l'analyse [83] .

Prétraitement Le prétraitement est le processus de nettoyage des données préparant le texte pour la classification. Les textes en ligne contiennent généralement beaucoup de bruits et des éléments inutiles tels que des balises, des scripts. Le prétraitement des données réduit le bruit et contribue à améliorer les performances du classificateur. Le prétraitement accélère également le processus de classification, aidant ainsi en AS à temps réel.

La préparation du texte implique le nettoyage des données extraites avant que l'analyse ne soit effectuée. Habituellement, la préparation de texte implique l'identification et l'élimination du contenu non textuel de l'ensemble de données textuelles. En outre, tout autre contenu qui n'est pas jugé pertinent pour le domaine d'étude est également supprimé de l'ensemble de données textuelles, comme par exemple des mots vides ou des mots qui ne sont pas pertinents pour le cours de l'analyse [66].

Pour un système qui donne une AS des flux de données, la stratégie de prétraitement est la suivante :

a) Suppression des mots Certaines approches considèrent que certains mots très communs n'apportent aucune information utile pour l'analyse du texte. Dans ce cas, il est courant, d'utiliser une liste de ces mots pour filtrer le texte. La suppression de ces mots présente l'avantage de réduire la phrase à des mots pleins.

b) Structuration des phrases : Structurer une phrase permet de mieux appréhender la sémantique de chaque mot. Certaines techniques d'analyse de sentiments utilisent la structure des phrases afin d'identifier l'opinion.

c) Suppression des localisateurs de ressources uniformes (URL), hashtags, références, caractères spéciaux : Le nettoyage des données des hashtags, références, caractères spéciaux, aidera à réduire la plupart des bruits.

d) Traduction de mots d'argot - Pour cela, nous prenons l'aide du dictionnaire d'argot Internet et remplaçons les mots d'argot dans leur format significatif.

e) Suppression des lettres supplémentaires des mots : Les mots qui ont la même lettre plus de deux fois et qui ne sont pas présents dans le lexique sont réduits au mot avec la lettre répétitive n'apparaissant qu'une seule fois. Par exemple, le mot exagéré *"Happyyyyyy"* est réduit à *"Happy"*.

f) Enracinement : L'enracinement vient du mot racine, est se fait à l'aide de Natural Language Tool Kit (NLTK). Par exemple, des mots tels que *"waiting"*, *"waits"*, *"waited"* sont remplacés par le mot *"wait"*.[66] .

Détection des sentiments La troisième étape est la détection des sentiments. La détection des sentiments nécessite d'évaluer et d'extraire des critiques et des opinions à partir de l'ensemble de données textuelles grâce à l'utilisation de tâches de calcul. Chaque phrase est examinée pour la subjectivité. Seules les phrases avec des expressions subjectives sont conservées dans l'ensemble de données. Les phrases qui véhiculent des faits et une communication objective sont écartées de toute analyse ultérieure.

La détection des sentiments se fait à différents niveaux, soit un seul terme, des phrases, des phrases complètes ou un document complet avec des techniques couramment utilisées.[55].

Sélection des fonctionnalités le but principal de la sélection de fonctionnalités est de diminuer la dimensionnalité de l'espace de fonction. Un espace de fonctionnalités réduit le coût de calcul.

En tant que deuxième objectif, la sélection des fonctionnalités réduira également la sur-adaptation du schéma d'apprentissage aux données d'apprentissage. Au cours de ce processus, il est également important de trouver un bon compromis entre les richesse des fonctionnalités et des contraintes de calcul impliquées lors de la résolution de la tâche de catégorisation.[87]

Classification des sentiments La cinquième étape est la classification de la polarité qui classe chaque phrase subjective de l'ensemble de données textuelles en groupes de classification. Habituellement, ces groupes sont représentés sur deux points extrêmes d'un continuum (positif, négatif, bon, mauvais, ect).

Les techniques de classification des sentiments (SC) peuvent être divisées en deux parties, à savoir l'approche d'apprentissage automatique (Machine Learning) et l'approche basée sur le lexique.

Les approches ML sont basées sur la formation d'un algorithme, principalement la classification, dépend sur un ensemble de fonctionnalités sélectionnées pour une mission spécifique, puis testées sur un autre ensemble s'il est capable de détecter les bonnes fonctionnalités et de donner les bonnes classifications.

Les approches ML utilisent des algorithmes de ML se basant sur des attributs linguistiques tandis que l'approche basée sur un lexique s'inspire du *"lexique des sentiments"* [67, 6].

1.5 Défis d'analyse des sentiments

L'analyse des sentiments (AS) concerne principalement le traitement des avis, les commentaires sur différentes personnes, et leur traitement pour en obtenir des informations significatives.

FIGURE 1.6 – Étapes d'analyse des sentiments.

Différents facteurs affectent le processus de AS et doivent être traités correctement pour obtenir le rapport final de classification ou de regroupement. Quelques-uns de ces défis sont abordés ci- dessous :

Résolution de coréférence Ce problème se réfère principalement à savoir ce qu'indique un pronom ? [50]Par exemple, dans la phrase *"Après avoir regardé le film, nous sommes partis manger ; c'était bien"*. À quoi se réfère le mot *"c'était"* ; que ce soit le film ou la nourriture ? Ainsi, lorsque l'analyse du film est en cours, si la phrase concerne le film ou la nourriture ? C'est une préoccupation pour l'analyste. Ce type de problème se produit principalement dans le cas d'une AS orientée aspect.

Association avec une période le moment de la collecte d'avis est une question importante dans le cas de l'AS. Le même utilisateur ou groupe d'utilisateurs peut donner une réponse positive pour un produit à un moment donné, et il peut y avoir un cas où il peut donner une réponse négative. C'est donc un défi pour l'analyseur de sentiment à un autre moment. Ce type de problème survient principalement dans l'AS comparative.

Gestion du sarcasme l'utilisation de mots qui signifient le contraire de ce qu'ils informent sont surtout connus comme des mots de sarcasme [41]. Par exemple, la phrase *"Quel bon batteur il est, il marque zéro dans toutes les autres manches"*. Dans ce cas, le mot positif *"bon"* a un sens négatif. Ces phrases sont difficiles à trouver et par conséquent, elles affectent l'analyse du sentiment.

Indépendance de domaine Dans l'AS, les mots sont principalement utilisés comme fonction d'analyse. Mais, le sens des mots n'est pas fixé de bout en bout[103]. Il y a peu de mots dont la signification change d'un domaine à l'autre. En dehors de cela, il existe des mots qui ont une signification opposée dans différentes situations connues sous le nom de contronyme. Ainsi, il est difficile de connaître le contexte pour lequel le mot est utilisé, car il affecte l'analyse du texte et finalement le résultat.

Négations Les mots négatifs présents dans un texte peuvent totalement changer le sens de la phrase dans laquelle il est présent[75].A insi, lors de l'analyse des critiques, ces mots doivent être pris en compte. Par exemple, les phrases *"Ceci est un bon livre"*. Et *"Ce n'est pas un bon livre"*. Ont une signification opposée, mais lorsque l'analyse est effectuée en utilisant un seul mot à la fois, le résultat peut être différent. Pour gérer ce type de situations, une analyse en n-gramme est préférable.

Détection de spam le Web contient à la fois des contenus authentiques et du spam. Pour une classification efficace des sentiments, ce contenu de spam doit être éliminé avant le traitement [61]. Cela peut être fait en identifiant les doublons, en détectant les valeurs aberrantes, et compte tenu de la réputation du critique.

Mots orthographiques les gens utilisent des mots orthographiques pour exprimer leur excitation, leur bonheur, par exemple : le mot *Sooo*, *Sweeettt*, *Haappy* ou s'ils sont pressés, ils insistent sur les mots par exemple : *comeeeee*, *fasssssssst* , *waittttnggg*...etc [6]

Manière d'exprimer son sentiment les gens n'expriment pas toujours leurs sentiments de la même manière. Le sentiment de chaque individu est différent la façon de penser, la manière d'exprimer varie d'une personne à l'autre.

Asymétrie dans la disponibilité des logiciels d'extraction d'opinion le logiciel d'extraction d'opinion est très coûteux et actuellement abordable uniquement pour les grandes organisations et les gouvernements[98]. C'est au-delà des attentes des citoyens ordinaires. Cela doit être accessible à tous, afin que chacun en profite.

1.6 Applications d'analyse des sentiments

La gestion de contenu axée sur l'opinion possède de nombreuses applications importantes comme la détermination des opinions des critiques concernant un certain produit via la classification des critiques de produits en ligne, ou l'enregistrement de l'évolution des attitudes du public à l'égard d'un parti politique via l'extraction de sites d'actualités en ligne ou de contenu de blogs [82, 66].

Alors que les applications basées sur l'opinion ou sur le feedback sont plus populaires, le domaine du traitement du langage naturel s'intéresse actuellement aux analyses de sentiments ainsi qu'aux systèmes d'exploration d'opinion [21, 67].

Les principales applications de des analyses de sentiments et de l'extraction d'opinions sont données ci-dessous :

Achat de produits ou de services lorsque vous décidez d'acheter des produits ou des services, prendre des décisions précises n'est plus un travail difficile. Grâce à cette

méthode, les individus peuvent évaluer les opinions et les expériences des autres concernant tous les produits et services et comparer les marques concurrentes [82].

Amélioration de la qualité du produit ou du service grâce à l'exploration d'opinions et à l'analyse des sentiments, les fabricants peuvent recueillir les opinions des critiques ainsi que les critiques positives concernant leurs produits, ou services, et ainsi améliorer leur qualité [82, 65].

Recherche marketing les résultats des analyses de sentiment peuvent être utilisés à des fins d'étude de marché. Grâce à des méthodes d'analyse des sentiments, les tendances récentes des clients concernant des produits ou services particuliers peuvent être examinées. De même, les attitudes actuelles du public à l'égard des nouvelles politiques de l'État peuvent également être facilement examinées [65, 22].

Systèmes de recommandation Grâce à la classification des opinions des individus comme positives ou négatives, le système peut déterminer laquelle est recommandée et laquelle ne l'est pas [65].

Détection de flamme la supervision des sites d'information, des articles de blogs ainsi que des réseaux sociaux est simplifiée grâce à des analyses de sentiment. L'exploration d'opinions et l'analyses des sentiments sont capables de détecter les mots arrogants, surchauffés, incitant à la haine, jurons utilisés dans les e-mails, les blogs ou les sites d'informations de manière automatisée [39, 96].

Détection d'opinion Spam Comme Internet est également accessible à tout le monde, n'importe qui peut télécharger n'importe quoi sur Internet. Cela signifie la probabilité que le contenu soit du spam augmente de jour en jour. Les particuliers pourraient télécharger du contenu de spam dans le but d'induire les gens en erreur. L'exploration d'opinions ainsi que les analyses de sentiment sont capables de classer le contenu Internet en spam et en non spam [21].

Élaboration de politiques en utilisant des analyses de sentiment, les décideurs politiques sont en mesure de prendre en considération les perspectives des citoyens concernant certaines politiques et ces connaissances peuvent être utilisées pour créer de nouvelles politiques en faveur des citoyens [65].

Prise de décision les opinions et les expériences du public sont des facteurs très utiles lors de la prise de décisions. Une analyse OM (opinion mining) ainsi qu'une analyse des sentiments fournissent une analyse des opinions du public qui peuvent être utilisées efficacement lors de la prise de décisions. Les sites de réseaux sociaux tels que Twitter ou Facebook ne sont pas des sources de données publiques à grande échelle à traiter légèrement. Le public les utilise pour révéler ses opinions et ses sentiments sur plusieurs sujets. L'utilisation de l'analyse de sentiments sur les avis et leur classification automatique en classes positives, négatives, ou neutres peut offrir des données cruciales aux entreprises sous la forme d'études de marché [81, 65].

1.7 Conclusion

L'analyse des sentiments est une méthode pour déterminer les opinions et les émotions exprimées dans les données textuelles comme les tweets sur Twitter. Elle est devenue une méthode importante pour comprendre l'opinion publique et les tendances dans divers domaines, y compris la politique, la santé, l'économie, et la culture.

Bien que l'analyse des sentiments sur Twitter présente des défis tels que la difficulté à interpréter les tweets ambigus et ironiques, elle reste une méthode utile pour les entreprises, les organisations, et les décideurs politiques. Elle peut leur permettre de comprendre les opinions et les émotions des utilisateurs de Twitter sur un sujet donné.

En outre, avec l'essor de l'intelligence artificielle et de l'apprentissage automatique, l'analyse des sentiments devient de plus en plus sophistiquée et précise. Les chercheurs et les praticiens travaillent ensemble pour développer de nouvelles techniques et méthodes pour améliorer l'analyse des sentiments sur les plateformes de médias sociaux telles que Twitter, et ainsi aider les entreprises et les organisations à prendre des décisions éclairées en fonction de l'opinion publique.

Chapitre 2

Apprentissage Automatique

2.1 Introduction

En raison de la croissance rapide des médias sociaux, une grande quantité de données générées par les utilisateurs est maintenant disponible. Analyser les sentiments et la classification précise de cette quantité gigantesque de données est une tâche très difficile. La plupart des données disponibles sur Internet sont sous forme textuelle car c'est la forme la plus naturelle et la plus lisible pour présenter les pensées et les opinions aux utilisateurs. Dans cette étude, une analyse sommaire des algorithmes et des techniques d'apprentissage automatique utilisés pour l'analyse des sentiments est réalisée. Ces algorithmes sont plus adaptés aux entrées changeantes[55].

Les unigrammes (mot unique), les bigrammes (deux mots) et les n-grammes (plusieurs mots) sont utilisés par différents algorithmes pour l'étiquetage et le traitement des données. Les techniques d'apprentissage automatique sont généralement utilisées pour la classification binaire et les prédictions de sentiments comme étant soit positifs, soit négatifs.

2.2 Machine learning

L'apprentissage automatique est un domaine d'étude axé sur le développement d'algorithmes et de modèles qui permettent aux systèmes informatiques d'apprendre à partir de données et d'améliorer leur performance sur une tâche spécifique sans être explicitement programmés [11].

L'objectif de l'apprentissage automatique est de permettre aux systèmes d'apprendre et de s'adapter à de nouveaux entrées et situations, améliorant ainsi leur capacité à faire des prédictions et des décisions [72]. Au cœur de l'apprentissage automatique, se trouve le concept de données d'entraînement. Les algorithmes d'apprentissage automatique sont conçus pour analyser de grands ensembles de données et identifier des modèles et des relations qui peuvent être utilisés pour faire des prédictions ou des décisions. Ces modèles et relations sont ensuite codés dans un modèle qui peut être utilisé pour faire des prédictions sur de nouvelles données [44].

L'apprentissage automatique est largement utilisé dans une variété d'applications, notamment le traitement du langage naturel, la vision par ordinateur, et l'analyse prédictive. Certaines des techniques d'apprentissage automatique les plus courantes comprennent les arbres de décision, les réseaux de neurones, et les machines à vecteurs de support.

Dans l'ensemble, l'objectif de l'apprentissage automatique est de développer des algorithmes et des modèles qui permettent aux systèmes informatiques d'apprendre à partir de données et d'améliorer leur performance au fil du temps, ce qui leur permet de résoudre des problèmes complexes et de faire des prédictions qui n'étaient pas possibles auparavant [49].

2.2.1 Les Algorithmes de Machine Learning

Les algorithmes d'apprentissage automatique sont généralement classés en fonction du type de tâche pour laquelle ils sont conçus. Certains des types d'algorithmes d'apprentissage automatique les plus courants comprennent l'apprentissage supervisé, l'apprentissage non supervisé, et l'apprentissage par renforcement.

Supervisé Dans ces algorithmes, un ensemble de données d'entraînement avec des classes préétiquetées est donné, et sur la base de cet ensemble de données entraîné, les entrées sont étiquetées avec la classe/résultat de sortie. Ces algorithmes classifient l'ensemble de données d'entrée à l'aide d'un classificateur entraîné. Les données d'entraînement sont composées d'un ensemble d'exemples d'entraînement, chacun étant constitué d'un objet d'entrée et de résultats de sortie souhaités. Une fonction déduite est créée en analysant les données d'entraînement par des méthodes d'apprentissage supervisé qui peuvent être utilisées ultérieurement pour mapper de nouvelles données entrantes, également appelées données de test. La plupart des techniques d'apprentissage automatique utilisent l'approche supervisée. Elle peut être catégorisée davantage en deux méthodologies, c'est-à-dire la classification et la régression. Les exemples les plus courants d'algorithmes d'apprentissage automatique supervisé sont la régression linéaire, la forêt aléatoire, et les machines à vecteurs de support [9].

Non supervisé Ces types d'algorithmes d'apprentissage automatique prennent les données d'entrée non étiquetées et, avec l'aide de différents algorithmes, découvrent la structure/correspondance cachée. Contrairement à l'apprentissage supervisé, cette technique n'utilise pas de données préétiquetées pour entraîner le classificateur. L'apprentissage automatique non supervisé peut être subdivisé en clustering et en association. Les exemples les plus courants d'algorithmes d'apprentissage automatique non supervisé étant K-Means et l'algorithme d'Apriori [9].

Semi-Supervisé Ces types d'algorithmes traitent à la fois les ensembles de données étiquetées et non étiquetées. Ils examinent différents outils et techniques basés sur le lexique et mentionnent la comparaison entre les caractéristiques et les résultats de précision des différentes techniques lexicales. En allant plus loin, différentes techniques et algorithmes d'apprentissage automatique sont étudiés et analysés dans cette recherche. Une analyse exhaustive est également formulée entre différentes techniques et précisions [9].

Apprentissage par renforcement L'apprentissage par renforcement est un type d'algorithme d'apprentissage automatique qui permet à un agent d'apprendre par des interactions de type essai-erreur avec son environnement. Dans l'apprentissage par renforcement, l'agent reçoit des commentaires sous forme de récompenses ou de pénalités pour les actions qu'il entreprend, et le but est d'apprendre une politique qui maximise la récompense cumulative attendue au fil du temps. L'agent apprend en explorant l'environnement et en mettant à jour sa politique en fonction des commentaires qu'il reçoit. L'apprentissage par renforcement a été appliqué avec succès dans un large éventail d'applications, y compris les jeux, la robotique, et les véhicules autonomes [73].

Nous présentons ci-dessous les approches les plus courrament entreprises par les chercheurs dans ce domaine, tout en tâchant d'expliquer le principe de fonctionnement. Le lecteur retrouvera dans le chapitre suivant une revue de l'état de l'art détaillant les efforts les plus récents.

2.3 Approches d'apprentissage automatique utilisées dans l'analyse des sentiments

Parmi les approches basées sur le ML, nous pouvons citer les approches suivantes :

2.3.1 Random Forest

Random Forest est une méthode d'apprentissage en ensemble qui vise à améliorer et à stocker les arbres de classification. Les prédicteurs d'arbre sont organisés de telle manière que chaque arbre dépende de valeurs de vecteurs aléatoires indépendamment structurées et que tous les arbres soient uniformément répartis dans la forêt [68]. Tel que défini, la forêt aléatoire est un classificateur composé de classificateurs structurés en arbre $h(x, Ok)$, $k = 1$, ayant Ok comme vecteurs aléatoires distribués de manière indépendante et identiquement répartis, et chaque arbre émet un vote unitaire pour la classe la plus populaire pour l'entrée x. Les forêts aléatoires ont été efficacement appliquées à de nombreux problèmes complexes en épidémiologie génétique, et en microbiologie au cours des dernières années [71]. En Et aussitôt, la forêt aléatoire est devenue, parmi d'autres méthodes standard, un outil majeur d'analyse de données [97].

Random Forest est un algorithme souvent utilisé pour l'analyse de sentiments. Il consiste à identifier le sentiment (positif, négatif, ou neutre) exprimé dans un texte. L'algorithme peut être entraîné sur un grand corpus de données textuelles étiquetées pour apprendre les motifs et les caractéristiques qui sont indicatifs de chaque sentiment.

Les forêts aléatoires présentent plusieurs avantages pour l'analyse de sentiments. Elles peuvent gérer des espaces de caractéristiques à haute dimensionnalité, ce qui est important pour l'analyse de texte où il y a souvent de nombreuses caractéristiques à considérer. Elles sont également robuste aux caractéristiques bruyantes ou non pertinentes, qui peuvent être courantes dans les données textuelles. De plus, la nature d'ensemble de l'algorithme contribue à améliorer la précision et à réduire le surajustement, ce qui peut être important dans les tâches d'analyse de sentiments.

Plusieurs études ont montré que Random Forest peut atteindre une grande précision dans les tâches d'analyse de sentiments. Par exemple, dans une étude comparant plusieurs algorithmes d'apprentissage automatique pour l'analyse de sentiments des données Twitter, la forêt aléatoire a atteint la plus haute précision [3]. Dans une autre étude, la forêt aléatoire a été utilisée pour classifier les avis sur les restaurants comme positifs ou négatifs, et a atteint une précision de 89,1 % [109].

En résumé, la forêt aléatoire est un algorithme puissant qui peut être utilisé pour les tâches d'analyse de sentiments. Sa capacité à gérer des espaces de caractéristiques à haute dimensionnalité, sa robustesse aux caractéristiques bruyantes ou non pertinentes, et sa nature d'ensemble en font un choix populaire pour cette application.

2.3.2 La régression logistique

La régression logistique (Logistic Regression) est un type de modèle linéaire généralisé qui est utilisé pour prédire la probabilité d'un résultat binaire. Elle est couramment utilisée pour les problèmes de classification où la variable dépendante est binaire (par exemple, oui/non, vrai/faux, 1/0). Les variables indépendantes peuvent être soit continues, soit catégorielles [51].

En régression logistique, la relation entre les variables indépendantes et la variable dépendante est modélisée à l'aide de la fonction logistique. La fonction logistique est une courbe en forme de S qui transforme tout nombre réel en une valeur entre 0 et 1. La sortie de la fonction logistique peut être interprétée comme la probabilité de la classe positive (par exemple, oui/vrai/1) [7].

L'objectif de la régression logistique est de trouver le meilleur modèle qui décrit la relation entre les variables indépendantes et la variable dépendante. Cela se fait en estimant les coefficients de l'équation de régression logistique à l'aide d'une technique appelée estimation du maximum de vraisemblance (EMV). L'EMV trouve les valeurs des coefficients qui maximisent la vraisemblance d'observer les données étant donné le modèle [7].

Une fois que les coefficients ont été estimés, ils peuvent être utilisés pour faire des prédictions sur de nouvelles données. La probabilité prédite de la classe positive peut être calculée en introduisant les valeurs des variables indépendantes dans l'équation de régression logistique et en appliquant la fonction logistique [51]. Une façon courante de faire une classification binaire est de choisir une valeur de seuil (par exemple, 0,5) et de classer les entrées avec une probabilité prédite supérieure au seuil comme une classe et en dessous du seuil comme l'autre [10].

2.3.3 Support Vector Machine

Les Machines à Vecteurs de Support (SVM) sont une classe de modèles d'apprentissage supervisé largement utilisés pour l'analyse de classification et de régression [26]. Les SVM sont particulièrement utiles lorsqu'il s'agit de données multidimensionnelles, car ils peuvent apprendre des limites de décision complexes qui séparent les points de données de différentes classes.

Dans le contexte de l'analyse de sentiment, les SVM sont couramment utilisés pour classifier les données textuelles comme étant positives, négatives, ou neutres [82]. Cela implique de former le modèle SVM sur un ensemble de données étiquetées de documents textuels, où chaque document est annoté avec son sentiment correspondant [82]. Au cours du processus de formation, les SVM apprennent à classer de nouveaux documents textuels en fonction des caractéristiques extraites des données d'entraînement.

Le processus d'utilisation des SVM dans l'analyse de sentiment implique généralement les étapes suivantes :

Prétraitement des données cela implique le nettoyage des données, la suppression des mots vides, et la conversion des données textuelles en un format approprié pour les SVM [106]. Cette étape est cruciale pour garantir que les SVM puissent apprendre des

motifs significatifs à partir des données.

Extraction de caractéristiques Il s'agit là de l'extraction de caractéristiques pertinentes des données textuelles, telles que la fréquence des mots, les n-grammes, et la fréquence du terme-inverse de document (TF-IDF) [57]. L'extraction de caractéristiques est importante dans l'analyse de sentiment car elle capture le sentiment sous-jacent du texte.

Formation des SVM cela implique l'utilisation d'un ensemble de données étiquetées pour former le modèle SVM. Au cours du processus de formation, les SVM apprennent à identifier les motifs dans les données qui sont associés à un sentiment positif, négatif, ou neutre [65].

Test et évaluation Ceci consiste en l'utilisation d'un ensemble de données de test distinct pour évaluer les performances du modèle SVM dans la classification du sentiment. Les performances des SVM sont généralement évaluées à l'aide de mesures telles que l'exactitude, la précision, le rappel, et le score F1 [93].

Les SVM ont prouvé leur efficacité dans l'analyse de sentiment car ils peuvent traiter des données multidimensionnelles et sont robustes au bruit et aux valeurs aberrantes [26]. De plus, les SVM peuvent apprendre des limites de décision complexes et ont une bonne performance de généralisation, ce qui est important dans l'analyse de sentiment où le modèle doit classer avec précision des données inconnues [82, 65].

Ces dernières années, les SVM ont été étendus pour traiter des tâches plus complexes dans l'analyse de sentiment, telles que l'identification de l'intensité du sentiment et la détection du sarcasme et de l'ironie [82]. Ces extensions ont encore augmenté l'efficacité et l'applicabilité des SVM dans l'analyse de sentiment.

2.3.4 Apprentissage Ensembliste

Les méthodes d'ensemble (Ensembled Methods) sont des techniques d'apprentissage automatique qui combinent plusieurs modèles pour améliorer leur pouvoir prédictif [31]. En d'autres termes, au lieu de s'appuyer sur un seul modèle pour faire des prédictions, les méthodes d'ensemble utilisent la connaissance collective de plusieurs modèles pour produire des résultats plus précis et plus robustes [31].

Il existe plusieurs types de méthodes d'ensemble, mais les plus courantes sont :

Bagging Cette méthode consiste à former plusieurs modèles indépendants sur des sous-ensembles aléatoires des données et à combiner leurs prédictions par appariement ou par moyennage [20].

Boosting Cette méthode entraîne une séquence de modèles, chaque modèle suivant les résultats de autres modèles ciblant des échantillons mal classés par les modèles précédents. Ensuite, la prédiction finale est faite en combinant les prédictions de tous les modèles [35].

Stacking Cette méthode consiste à entraîner plusieurs modèles de types différents et à combiner leurs prédictions à l'aide d'un méta-modèle [105].

Les méthodes d'ensemble sont couramment utilisées dans l'analyse des sentiments, qui identifie et extrait des informations subjectives à partir de données textuelles telles que les opinions, les attitudes, et les émotions [65]. En effet, l'analyse des sentiments implique souvent de traiter des données bruyantes et ambiguës, ce qui rend difficile l'obtention d'une grande précision avec un seul modèle [31].

En combinant plusieurs modèles avec différentes forces et faiblesses, les méthodes d'ensemble peuvent capturer efficacement la nature complexe et diversifiée du sentiment dans les données textuelles [65].

Par exemple, un ensemble d'arbres de décision d'ensachage (Bootstrap Aggregation) peut être utilisé pour classer les humeurs en fonction de diverses caractéristiques lexicales et syntaxiques, tandis qu'un ensemble de renforcement de réseau neuronal profond peut être utilisé pour capturer des modèles sémantiques plus complexes. De même, un ensemble d'empilement de différents modèles peut être utilisé pour combiner les points forts de chaque modèle et obtenir une plus grande précision [65].

2.4 Conclusion

En somme, l'apprentissage automatique est un domaine informatique en constante évolution qui implique de former des machines à apprendre des données et à prendre des décisions ou des prédictions. La technologie a le potentiel de révolutionner diverses industries et présente des avantages tels que l'évolutivité et la capacité de reconnaître des modèles dans les données [31].

Cependant, l'apprentissage automatique s'accompagne également de défis, notamment le besoin de grandes quantités de données de haute qualité et la difficulté d'interpréter les modèles. Malgré ces défis, l'apprentissage automatique évolue rapidement et devient un outil de plus en plus important pour les entreprises et les organisations. Il est important pour les chercheurs et les praticiens de se tenir au courant des progrès de l'apprentissage automatique.

Chapitre 3

Etat de l'art

3.1 Introduction

Nous avons étudié, dans les chapitres précédents, ce qu'est l'analyse de sentiments à partir de données textuelles en utilisant des techniques d'apprentissage automatique. Notre objectif est de comprendre comment ces méthodes peuvent être mises en œuvre pour analyser les grandes quantités de données générer par les utilisateurs sur les réseaux sociaux. Nous avons, alors, commencé par présenter les différentes étapes de l'analyse de sentiments, les défis que l'on peut rencontrer, ainsi que les techniques et outils utilisés pour nettoyer et préparer les données textuelles utilisée dans l'analyse. Nous avons également exploré la sélection des fonctionnalitées pour obtenir un rapport final de classification ou de regroupement, ainsi que l'utilisation des SVM pour l'analyse de sentiments.

Dans le présent chapitre, nous nous penchons sur les efforts des chercheurs dans le domaine, et tâcherons de passer en revue les travaux les plus significatifs en la matière.

3.2 Aperçu des approches trouvées dans la littérature

De nos recherches, nous avons retenu trois familles d'approches : Machine learning, deep learning, hybride que nous allons développer ci-dessous.

3.2.1 Approches Machine Learning

K. Gajbhiye et N.Guptaont [36] ont utilisé une approche basée sur l'apprentissage automatique, et plus précisément la classification probabiliste Naïve Bayes, afin de classer des commentaires extraits en temps réel depuis le site social twitter. Pour ce, ils ont procédé comme suit : tout d'abord, ils ont commencé par collecter et stocker des commentaires sur Twitter 'a l'aide de l'outil d'apprentissage automatique `R` et plus précisément les packages `twitter` et `ROAuth`. Ensuite, ils ont prétraité les tweets collectés en supprimant les données indésirables telles que les urls, les caractères spéciaux, les

nombres, les mots d'arrêt, etc. du texte, transformé le texte en minuscules et cela grâce au package `tm` utilisé qui utilise la technique de traitement du langage naturel (NLP).

Enfin, après avoir prétraité les tweets collectés, ils ont appliqué le classificateur bayes naïf pour les classifier en différentes émotions et polarités de sentiment. Au terme de la procédure, ils ont obtenu un résultat avec un pourcentage de précision de 81%.

De leurs coté, Maryum Bibi et al [17], présentent un nouveau cadre d'ensemble non supervisé utilisant des méthodes linguistiques basées sur des concepts et l'apprentissage automatique pour l'analyse des sentiments sur Twitter. Afin de classifier les commentaires, ils ont procédé comme suit :

Tout d'abord, ils ont prétraité les ensembles de données en supprimant les éléments inutiles, notamment les signes de ponctuation, les URL, les nombres, les émoticônes. . .etc, ont converti tous les tweets sont convertis en lettres minuscules, et ont supprimé les retweets.

Ensuite, ils ont mis en place des méthodes différentes de représentation des caractéristiques pour l'analyse des sentiments sur Twitter : la méthode booléenne et la méthode TF-IDF (Term frequency-inverse document frequency). Ils ont également expérimenté des classificateurs bien connus (Naive Bayes, réseau neuronal) pour une comparaison équitable.

La mesure de la précision (proportion de prédictions correctes) est utilisée pour évaluer les performances des techniques étudiées. Il est démontré empiriquement que les performances des techniques d'apprentissage non supervisé sont comparables à celles des techniques d'apprentissage supervisé.

En effet, ils ont obtenu une précision de classification allant jusqu'à 80% pour certains ensembles de données. Ces résultats sont supérieurs à ceux obtenus avec les méthodes classiques basées sur la méthode booléenne. Cependant, la performance du modèle dépend fortement de la qualité des données d'entrée et de la sélection des paramètres.

Dans le but d'analyser l'influence des compagnes de promotion des villes intelligentes sur la perception de ces projet par le public Robert N.A et al [12] ont effectué une analyse des sentiments des tweets concernant quatre projet de villes intelligentes en Afrique en utilisant le lexique des émotions du NRC pour annoter les mots avec des émotions telles que la colère peur, l'anticipation, la confiance, la surprise, la tristesse, la joie, et le dégoût, après avoir lemmatise les tweets ils ont utilisé des scriptes python pour faire correspondre les mots utilises dans les tweets avec leurs émotions les plus proches. Les résultats indiquent que les stratégies de promotion des villes intelligentes créent potentiellement un sentiment illusoire parmi le public qui détourne leur attention sur les réalités urbaines existantes y compris les promesses non tenues.

Dans leur article, Samuels et Mcgonical [90] ont exploré comment les technologies de l'information ont changé la façon dont nous recevons et interprétons les nouvelles. Cet recherche se concentre sur l'analyse des sentiments à l'égard de l'actualité, en utilisant un lexique pour analyser les articles de presse.

L'approche proposée dans cet article est basée sur un lexique pour l'analyse des sentiments dans les articles de presse. Cette approche consiste à utiliser un ensemble de mots pré-définis pour évaluer le ton général d'un article, en attribuant des scores positifs, négatifs, ou neutres à chaque mot. Ces scores sont agrégés pour donner une mesure globale

du sentiment de l'article. Cette approche est considérée comme efficace car elle ne nécessite pas de données d'apprentissage et peut être appliquée à un grand nombre d'articles en peu de temps. Cependant, elle a également des limites car elle ne prend pas en compte le contexte et la subjectivité des opinions exprimées dans les articles.

Siddhaling Urologin [99] présente une nouvelle approche pour l'analyse et la catégorisation des articles de presse en combinant des techniques de résumé, d'analyse des sentiments, de visualisation et de classification.

Diverses techniques de visualisation, telles que les nuages de mots et les cartes thermiques de sentiment, sont utilisées pour fournir des représentations intuitives du sentiment de l'article. Des algorithmes d'apprentissage automatiquetsont appliqués pour la classification, ce qui permet de classer les articles d'actualité en fonction de leur contenu émotionnel. Afin de classifier les commentaires, ils ont procédé comme suit : Une approche novatrice est proposée pour l'analyse de sentiment, la visualisation et la classification d'articles de presse résumés.

L'approche combine la technique de résumé de texte avec l'analyse de sentiment pour extraire et représenter efficacement les données à partir d'un grand volume de texte. La méthode utilise un algorithme de remplacement des pronoms par des noms propres pour le résumé du texte, ainsi que l'analyseur de sentiment VADER pour déterminer les informations sur le sentiment. Des schémas de visualisation en 3D sont ensuite utilisés pour représenter les informations sur le sentiment. Enfin, une classification est effectuée sur les articles originaux ainsi que sur les articles résumés à l'aide de classificateurs tels que la régression logistique, la forêt aléatoire et Adaboost.

Les résultats de classification montrent également que les articles résumés peuvent être classés avec une précision similaire à celle des articles originaux. Enfin , ils ont obtenu un résultat avec un pourcentage de précision de 83 %.

Indra Irawanto et al [54] se sont concentré sur l'application de techniques d'analyse des sentiments aux données des reseaux sociaux concernant les incendies de forêt en Indonésie. Les chercheurs ont collecté un ensemble de données de messages de médias sociaux liés aux incendies de forêt et les ont annotés avec des étiquettes de sentiment.

Ils ont utilisé des algorithmes d'apprentissage automatique tels que SVM, Naive Bayes, et Random Forest, ainsi que des méthodes d'extraction de caractéristiques telles que le sac de mots et les enchâssements de mots pour classer les sentiments exprimés dans les messages. L'étude a montré que les trois algorithmes ont bien fonctionné dans la classification des sentiments positifs, négatifs et neutres à l'égard des incendies de forêt.

Modak et Mondal [74] présentent une méthode d'analyse des sentiments des données Twitter en utilisant la technique de clustering. Pour l'analyse du langage naturel, les mots inutiles tels que les prépositions et les articles, appelés mots vides, sont supprimés car ils ne sont pas utiles pour l'analyse des sentiments. Ensuite, un score de sentiment est calculé pour chaque tweet en classant le sentiment en trois catégories : positif, négatif, et neutre. Des techniques non supervisées telles que le clustering sont appliquées aux scores de sentiment pour regrouper naturellement les tweets similaires en fonction de leur similarité. Cela permet d'identifier rapidement les tweets positifs ou négatifs.

Asha. R et al[88], proposent l'utilisation de la régression logistique pour l'analyse des

sentiments et la détection des rumeurs dans les données des réseaux sociaux. La régression logistique est une technique d'apprentissage automatique qui prédit la probabilité d'un résultat binaire. Dans ce cas, l'analyse des sentiments consiste à classer un texte comme positif ou négatif, tandis que la détection des rumeurs détermine la véracité des informations.

Leur approche comprend la collecte de données sur les plateformes de réseaux sociaux, suivie d'étapes de prétraitement telles que le nettoyage du texte et l'extraction de caractéristiques telles que la fréquence des mots et les lexiques de sentiment. Ces caractéristiques sont utilisées pour entraîner le modèle de régression logistique. Ce modèle a ensuite été appliqué à de nouvelles données pour l'analyse des sentiments et la détection des rumeurs.

Young Gyo Jung et al [58], se sont concentrés sur l'amélioration des performances de l'analyse des sentiments à l'aide d'un classificateur Naive Bayes amélioré mis en œuvre dans le cadre de Spark. Leur étude vise à répondre au besoin d'analyse des sentiments en temps réel, qui nécessite un traitement efficace de grands volumes de données. Ce dernier souligne l'importance de l'analyse des sentiments dans divers domaines, tels que la surveillance des médias sociaux, l'analyse des commentaires des clients et la gestion de la réputation des marques. L'analyse des sentiments en temps réel permet de prendre des décisions en temps opportun et de réagir aux tendances ou aux sentiments émergents.

L'approche proposée améliore le classificateur Naive Bayes traditionnel en incorporant des techniques de sélection des caractéristiques et en tirant parti des capacités de traitement distribué du cadre Spark. En sélectionnant les caractéristiques les plus pertinentes à partir des données d'entrée, le classificateur amélioré vise à améliorer la précision des résultats de l'analyse des sentiments.

L'article mentionne que l'évaluation expérimentale du classificateur Naive Bayes amélioré démontre sa performance améliorée par rapport aux approches traditionnelles. Le classificateur amélioré atteint une plus grande précision et efficacité dans les tâches d'analyse des sentiments en temps réel, ce qui en fait un outil précieux pour les applications qui nécessitent une analyse rapide des sentiments à partir de grands flux de données.

Par ailleurs, Vadivukarassi et al. [1] ont exploré différentes approches pour l'analyse du sentiment et la classification des tweets relatifs aux compagnies aériennes afin de mieux comprendre le sentiment du public à l'égard des compagnies aériennes et pouvoir identifier l'approche la plus efficace en comparant différents modèles de classification.

Les auteurs soulignent l'importance des techniques de prétraitement telles que la normalisation du texte, la tokenisation, et l'extraction de caractéristiques pour préparer les données des tweets à l'analyse. Ces techniques permettent de traiter les données textuelles bruyantes et non structurées et d'améliorer la précision de la classification des sentiments.

3.2.2 Approaches hybrides

Ravinder et Sharma [8] ont présenté une approche efficace pour la détection des sarcasmes en combinant le BERTweet, les GRU bidirectionnelles, et les mécanismes d'attention. Tout d'abord ils ont proposé une approche qui utilise BERTweet, un modèle de

langage pré-entraîné spécifiquement formé sur les données de Twitter, pour extraire des représentations contextualisées des mots. Ensuite, des GRU bidirectionnelles ont été incorporées pour capturer les informations séquentielles et les dépendances entre les mots. Un mécanisme d'attention est utilisé pour se concentrer sur les parties pertinentes du texte.

Les auteurs ont obtenu des résultats expérimentaux qui démontrent l'efficacité de l'approche proposée, le modèle atteignant une grande précision dans la détection des sarcasmes à travers divers ensembles de données. Le modèle B^2GRUA proposé surpasse les méthodes existantes de détection des sarcasmes avec un pourcentage de précision de 85 %.

De leur coté, Kian Long Tan et al. [95], ont presenté leur approche *"RoBERTa-LSTM"* qui est un modèle hybride pour l'analyse de sentiments avec transformateur et réseau neuronal récurrent. Plus précisement, cette approche hybride combine le modèle de transformateur RoBERTa et la LSTM pour l'analyse de sentiments.

Leur modèle hybride a atteint des performances supérieures en capturant à la fois les informations contextuelles et les dépendances séquentielles. Et bien que le modèle hybride puisse augmenter la complexité de calcul, ses avantages résident dans l'amélioration de la précision et la capacité à traiter des modèles nuancés dans l'analyse des sentiments.

Les résultats expérimentaux montrent que le modèle hybride RoBERTa-LSTM est plus performant que les modèles individuels et les approches traditionnelles d'apprentissage automatique dans l'analyse des sentiments. La combinaison du transformateur et de la LSTM permet au modèle de capturer efficacement les informations contextuelles locales et globales, ce qui améliore la précision de la classification des sentiments avec un pourcentage de précision de 91%.

Danday et Murthy [27] présentent une étude sur l'analyse des données Twitter pendant les catastrophes. Leur approche est basée sur l'utilisation de Distill BERT et du réseau neuronal de convolution basé sur le graphe pour analyser les données Twitter pendant les catastrophes. Cette approche permet d'extraire des informations précieuses en temps réel, telles que les besoins des victimes, les zones touchées, et les ressources disponibles. Les auteurs soulignent également l'importance de l'utilisation de techniques plus avancées telles que le traitement du langage naturel et l'apprentissage automatique pour améliorer la précision de l'analyse à l'avenir.

Banerjeea et al. [14] ont exploré l'analyse de sentiment dans des données codées en tamoul-anglais et malayalam-anglais. Les données codées sont difficiles à traiter car elles contiennent des mots et des phrases provenant de plusieurs langues différentes. Les auteurs ont utilisé un modèle XLNet auto-régressif pour effectuer l'analyse de sentiment. Ce modèle a été entraîné sur des ensembles de données tamouls-anglais et malayalam-anglais codés. Les auteurs ont également utilisé une technique appelée *"fine-tuning"* pour améliorer les performances du modèle.

Leurs résultats ont montré que le modèle XLNet auto-régressif était efficace pour l'analyse de sentiment dans les données codées en tamoul-anglais et malayalam-anglais. Les performances du modèle ont été améliorées grâce à la technique de *"fine-tuning"*. Les auteurs ont également constaté que les tendances en matière d'opinions variaient selon les régions géographiques.

3.2.3 Deep Learning Approches

Sunithaa et al [94], se sont penché sur l'utilisation d'un modèle d'apprentissage profond basé sur un ensemble pour l'analyse du sentiment des tweets liés au COVID-19 provenant de l'Inde et des pays européens. Les chercheurs ont utilisé des réseaux neuronaux convolutifs (CNN) et des réseaux de mémoire à long terme (LSTM) dans une approche combinée. Les modèles sont entraînés sur un vaste ensemble de données de tweets liés au COVID-19 et visent à capturer des modèles complexes et des informations contextuelles dans les données.

Les résultats expérimentaux de cette étude ont montré que le modèle d'apprentissage en profondeur basé sur un ensemble a obtenu une précision de prédiction élevée. Les résultats ont également montré que la majorité des tweets étaient positifs à l'égard des vaccins COVID-19, avec seulement 8% de tweets négatifs et 9% de tweets neutres en provenance des Philippines.

Dans leur article, Dholpuria et al. [30] soulignent l'importance de l'analyse des sentiments dans l'industrie cinématographique, car elle permet aux cinéastes, aux producteurs, et aux spécialistes du marketing de comprendre l'accueil réservé à leurs films et de prendre des décisions éclairées en fonction des réactions du public. L'approche proposée utilise des algorithmes d'apprentissage profond, qui sont capables d'apprendre automatiquement des modèles et des représentations complexes à partir de données textuelles. En entraînant le modèle d'apprentissage profond sur un vaste ensemble de données de critiques de films, les chercheurs visent à permettre une classification précise des sentiments, en distinguant les sentiments positifs, négatifs, et neutres exprimés dans les critiques. Les auteurs soulignent l'importance des techniques de prétraitement telles que la tokenisation, la suppression des mots vides et le stemming, qui aident à préparer les données textuelles pour l'analyse et améliorent la performance du modèle d'apprentissage profond.

Leurs résultats démontrent l'efficacité de l'approche d'apprentissage profond dans l'analyse des sentiments des critiques de films. Le modèle atteint une précision prometteuse en classifiant correctement le sentiment exprimé dans les critiques, ce qui permet aux cinéastes et aux professionnels de l'industrie d'évaluer les réactions du public à l'égard de leurs films.

Hossen et al. [52] ont exploré l'utilisation d'une approche d'apprentissage profond pour analyser les critiques d'hôtels et prédire leur impact sur les affaires. L'étude vise à tirer parti de la puissance des algorithmes d'apprentissage profond pour extraire des informations précieuses des critiques d'hôtels et prévoir leur influence sur les performances de l'entreprise. L'approche proposée utilise des algorithmes d'apprentissage profond qui permet l'extraction automatique de caractéristiques et l'apprentissage de modèles complexes à partir de données textuelles, ce qui améliore la précision de l'analyse des sentiments et des tâches de prédiction. Les résultats de l'étude démontrent l'efficacité de l'approche d'apprentissage profond dans l'analyse des critiques d'hôtels et la prédiction de leur influence sur les affaires.

Garg et Kaliyar [38] ont utilisé des techniques d'apprentissage en profondeur pour analyser les sentiments politiques sur les plateformes de réseaux sociaux tels que Twitter et Facebook. Les auteurs ont proposé une approche basée sur l'apprentissage en profondeur pour classer les tweets en sentiments positifs, négatifs, ou neutres concernant deux dirigeants politiques.

Leur approche a fournie des résultats de pointe pour l'analyse des sentiments politiques. Les auteurs ont également comparé les résultats obtenus avec d'autres modèles de machine learning tels que Bayes, SVM, Decision Tree, Random Forest et Logistic Regression, et leurs résultats ont montré que l'approche basée sur l'apprentissage en profondeur a fourni des résultats supérieurs à ces autres modèles.

Rhanoui et al. [89] ont présenté une approche de Deep Learning pour l'analyse de sentiment au niveau des documents. Les auteurs proposent un modèle CNN-BiLSTM avec Doc2vec embedding, qui est comparé à d'autres modèles couramment utilisés pour l'analyse de sentiment. Le modèle proposé a été testé sur un ensemble de données d'articles de presse français et a obtenu une précision de 90,66%, ce qui est supérieur aux autres modèles testés. Les résultats montrent que la combinaison de CNN et BiLSTM avec Doc2vec embedding est une approche efficace pour l'analyse de sentiment au niveau des documents.

De leur coté, Fang et al. [34], ont proposé une méthode efficace pour analyser le sentiment des avis sur les attractions touristiques en utilisant un pipeline basé sur ELECTRA. Les avis en ligne sont une source d'information précieuse pour les voyageurs, mais ils sont souvent informels et bruyants. Pour résoudre ce problème, l'article propose une approche de prétraitement des données et un modèle de classification basé sur ELECTRA.

Le pipeline proposé dans l'article comprend deux étapes principales : le prétraitement des données et la classification du sentiment. Le prétraitement des données consiste à supprimer les mots vides, les caractères spéciaux, les doublons, et à remplacer les négations pour réduire le bruit dans les avis. La classification du sentiment est réalisée à l'aide d'ELECTRA. Les résultats expérimentaux montrent que le pipeline proposé est plus performant que plusieurs modèles de classification de texte profond représentatifs. En outre, le pipeline peut être appliqué à d'autres domaines tels que la finance et la santé pour analyser le sentiment des commentaires en ligne.

3.3 Tableau Comparatif

Dans les tableaux ci-dessous nous effectuerons une étude comparatives des approches proposées ci-dessus selon les 4 facteurs suivants :

Titre : Le titre du papier.

Approche : désigne l'approche du papier.

Avantages : avantages de l'approche abordée.

Inconvénients : inconvénients de l'approche abordée.

3.3.1 Comparaison des approches de la littérature

Les tableaux 3.1, 3.2, 3.3, 3.4, et 3.5 fournissent une analyse comparative des diverses approches d'analyse de sentiments décrites dans les articles étudiés dans le présent chapitre, et cela en se basant sur cinq critères préétablis.

Le tableau examine l'approche utilisée pour chacune des méthodes mentionné dans les différents articles. Les différentes approches d'analyse de sentiments incluent des techniques d'apprentissage automatique, telles que la classification probabiliste Naïve Bayes, ou des techniques de Deep Learning, telles que le TF-IDF et le biLSTM, et aussi des techniques hybrides telle que Robera-LSTM.

Par ailleurs, le tableau passe en revue les ensembles de données utilisés pour chaque méthode. Les données utilisées varient généralement selon les approches, allant des tweets sur les évaluations de produits à des évaluations de films sur IMDb ou encore des commentaires sur des hôtels.

En outre, le tableau liste également les avantages et les inconvénients de chaque approche. Les atouts identifiés comportent des éléments tels la rapidité d'exécution, la précision, et la classification efficace des sentiments pour les commentaires politiques, par exemple. Parmi les points faibles, on note la complexité de certaines techniques, la nécessité de grands ensembles de données étiquetées, ou encore les préjugés éventuels dans les données d'apprentissage.

3.4 Conclusion

En conclusion, le troisième chapitre a mis en évidence les différentes tendances et approches de l'analyse des sentiments. Plusieurs articles scientifiques ont été examinés, classés en trois catégories d'approches différentes (apprentissage automatique, apprentissage en profondeur, et hybrides) et qui ont été appliqués à diverses domaines tels que les réseaux sociaux, les critiques de films, et les commentaires de touristes.

La comparaison exhaustive des différentes approches a montré qu'il n'y avait pas de méthode unique pour aborder l'analyse des sentiments, soulignant ainsi l'importance de choisir l'approche la plus appropriée pour les données et les objectifs de l'analyse. Cela a également permis de mieux comprendre les forces et les faiblesses de chaque méthode.

Ces résultats démontrent l'importance de suivre les dernières évolutions dans ce domaine pour les chercheurs et les professionnels afin de fournir des résultats précis et d'aider les entreprises à prendre des décisions éclairées pour améliorer la satisfaction des clients et la qualité de leurs produits.

TABLE 3.1 – Résumé des travaux de la littérature (1)

Titre	**approches**	**Avantages**	**Inconvénients**
Real Time Twitter Sentiment Analysis for Product Reviews Using Naive Bayes Classifier(K. Gajbhiye et N. Guptaont)	Naive Bayes	La technique peut traiter de grandes quantités de données non structurées, peut être utilisé pour identifier des modèles et des tendances dans les données qui peuvent ne pas être immédiatement apparentes, elle peut traiter rapidement et efficacement de grandes quantités de données, Il peut être utilisé pour classer le texte en diverses émotions et polarités avec un haut degré de précision,Il est capable de traiter un grand volume de données en peu de temps, ce qui le rend adapté à l'analyse en temps réel	La précision de la technique dépend de la qualité des données analysées, La technique s'appuie fortement sur des modèles statistiques, La technique peut ne pas être capable de gérer le sarcasme ou l'ironie, La technique nécessite une quantité importante de ressources de calcul
A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis(Maryum Bibi, Wajid Arshad Abbasi a , Wajid Aziz a , Sundus Khalil a , Mueen Uddinb , Celestine Iwendi c , Thippa Reddy Gadekallud)	TF-IDF,Naïf Bayes	Il utilise des méthodes linguistiques basées sur des concepts et l'apprentissage automatique, permet une analyse plus fine et précise des sentiments exprimés dans les tweets. Il fonctionne de manière non supervisée	Il nécessite un prétraitement des données pour éliminer les bruits et les erreurs, Il est basé sur des méthodes linguistiques, Les résultats expérimentaux ont montré que la performance du cadre dépend fortement de la qualité des données d'entrée et de la sélection des paramètres.

TABLE 3.2 – Résumé des travaux de la littérature (2)

Titre	approches	Avantages	Inconvénients
B2GRUA : BERTweet Bi-Directional Gated Recurrent Unit with Attention Model for Sarcasm Detection (RAVINDER AHUJA AND S. C. SHARMA)	Bi-Directionnel GRU, BERTweet Base	capable à capturer des informations contextuelles, Cette combinaison permet au modèle d'identifier et de comprendre efficacement les modèles complexes associés au sarcasme.	Elle repose sur la disponibilité de grands ensembles de données annotées pour l'entraînement, la complexité informatique du modèle peut être supérieure à celle d'approches plus simples
RoBERTa-LSTM : A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network(Tan, K. L., Lee, C. P., Anbananthen, K. S. M., Lim, K. M)	Robuste Optimized BERTa, LSTM	la capacité à exploiter les points forts des modèles de transformateur et de LSTM, capturer les informations contextuelles et les dépendances séquentielles, gère efficacement les dépendances à long terme et les modèles nuancés.	sa complexité informatique accrue par rapport à l'utilisation d'un seul modèle.
Sentiment Analysis, Visualization and Classification of Summarized News Articles : A Novel Approach (Siddhaling Urologin)	AdaBoost Classifier, Random Forest, Logistic Regression	offrant ainsi une solution holistique pour l'analyse des articles d'actualité. permet de comprendre clairement le ton émotionnel exprimé dans les articles.	Dépendance à l'égard du résumé, Granularité limitée des sentiments.
Sentiment analysis and classification of Forest Fires in Indonesia (Hasanah,A. Widodo,C. Irawanto,I. Kusumah,P. A.D.Kusrini, K.Kusnawi,K)	SVM, Bayes Naif, Random Forest, Vader Classifier	Prise de décision éclairée.	Subjectivité dans l'étiquetage des sentiments : Le processus d'annotation des messages des médias sociaux avec des étiquettes de sentiment implique de la subjectivité, car différents annotateurs peuvent interpréter le sentiment différemment, ce qui peut avoir un impact sur la fiabilité de l'ensemble de données étiquetées.

TABLE 3.3 – Résumé des travaux de la littérature (3)

Titre	approches	Avantages	Inconvénients
Clustering and Sentiment Analysis on Twitter Data (S.Ahuja, G.Dubey)	Afinn, Textblob, Clustering, (K-means ,Fuzzy C-means)	capable de traiter un grand nombre de tweets en peu de temps. Elle permet de classer rapidement les tweets positifs ou négatifs, Elle utilise des techniques non supervisées.	La méthode peut être moins précise que les méthodes supervisées, Elle peut être influencée par la qualité des données, notamment la présence de bruit ou d'erreurs dans les tweets, Elle ne prend pas en compte le ton ou l' ironie dans les tweets.
Twitter sentiment analysis using ensemble based deep learning model towards COVID-19 in India and European countries(D.Sunithaa,R.Kumar Patra,N.V.Babu,A.Suresh,S.Chand Gupta)	SVM, LSTM, Naïve Bayes, logistic regression , CNN,KNN, Text blob, GloVe Embedding	Capacités d'apprentissage en profondeur, Précision et robustesse accrues, fournit des informations précieuses sur les perceptions, les attitudes et les émotions du public	déficient au Les biais, le bruit ou une couverture insuffisante des donnée.
Sentimental Analysis and Detection of Rumours for Social Media Data using Logistic Regression(Asha.R,Rahul.J,Gourav.A,Pranjnay.B)	SVM, Lexicon-Based, LOGISTIC REGRESSION	traite également bien les données de haute dimension et peut traiter efficacement de grands ensembles de données, sa simplicité et sa facilité d'interprétation	nécessiter un effort manuel, suppose une relation linéaire entre les caractéristiques et le résultat, ce qui peut ne pas être réaliste dans de nombreux cas, es observations aberrantes ou influentes peuvent fausser les résultats de la régression logistique.
Enhanced Naive Bayes Classifier for Real-time Sentiment Analysis with SparkR	Binarized Naif Bayes Classifier, Laplce Smoothing, SparkR	la capacité de traiter des données volumineuses, effectuer des analyses de sentiments en temps réel.	la nécessité de mises à jour continues du modèle pour maintenir la précision.

TABLE 3.4 – Résumé des travaux de la littérature (4)

Titre	approches	Avantages	Inconvénients
A Sentiment analysis approach through deep learning for a movie review	CountVectorizer,CNN	la capacité à apprendre automatiquement les caractéristiques et les représentations, capturer les nuances subtiles dans l'expression des sentiments.	prendre en compte les défis potentiels tels que les biais dans les données d'entraînement, la nécessité de disposer de grands ensembles de données étiquetées.
An Exploration of Airline Sentimental Tweets with Different Classification Model	SVM,Logistic Regression, AdaBoost Classifier, Random Forest Classifier, KNeighbors Classifier, Decision Tree Classifier,GaussianNB Clasifier	Explorer diverses techniques d'apprentissage automatique ou de traitement du langage naturel, Identifier l'approche la plus appropriée pour l'analyse des sentiments	la gestion du sarcasme ou de l'ironie dans les tweets.
Hotel review analysis for the prediction of business using deep learning approach	LSTM,GRU	gèrer la complexité du langage naturel, capturant les nuances subtiles dans le sentiment et l'opinion, éliminer la nécessité d'une ingénierie manuelle des caractéristiques, réduisant ainsi les efforts humains, L'amélioration de la précision et de l'efficacité permet aux entreprises de gagner du temps et d'économiser des ressources.	nécessite de grands ensembles de données étiquetées pour une formation efficace, L'interprétation peut s'avérer difficile en raison de leur complexité, Les biais potentiels dans les données d'apprentissage peuvent influencer les prédictions du modèle.
PSent20 : An Effective Political Sentiment Analysis with Deep Learning using Real-time Social Media Tweets	TF-IDF,RNN(BiLSTM)	-entraîné sur des tweets de médias sociaux en temps réel, en incorporant des informations contextuelles et temporelles, PSent20 utilise des algorithmes d'apprentissage profond spécialement conçus pour l'analyse des sentiments politiques	nécessite des mises à jour continues du modèle, nécessite des ressources informatiques importantes et une grande quantité de données pour entraîner les modèles

TABLE 3.5 – Résumé des travaux de la littérature (5)

Titre	approches	Avantages	Inconvénients
A CNN-BiLSTM Model for Document-Level Sentiment Analysis	CNN, RNN(BiLSTM), Word2Vec,CBOW	Utilise une combinaison de réseaux de neurones convolutionnels et récurrents bidirectionnels pour capturer les relations entre les mots et les phrases dans le document, Utilise une technique d'embedding appelée Doc2vec pour représenter les mots dans un espace vectoriel continue,	ces modèles nécessitent de grandes quantités de données et de ressources informatiques, La précision des modèles peut être affectée par des facteurs tels que les biais dans les données d'entraînement et la complexité du langage utilisé dans le texte analysé.
Twitter Data Analysis using Distill BERT and Graph Based Convolution Neural Network during Disaster	Distilling BERT	Permet d'obtenir des informations précieuses en temps réel, Utilise des techniques avancées telles que Distill BERT et le réseau neuronal de convolution basé sur le graphe pour améliorer la précision de l'analyse.	Difficulté à distinguer les tweets pertinents des tweets non pertinents. -Nécessité de traiter un grand volume de données en temps réel.
NUIG-Shubhanker@Dravidian-CodeMix-FIRE2020 : Sentiment Analysis of Code-Mixed Dravidian text using XLNet	XLNet, fine-tuning	n'a pas besoin de pré-traitement	Le modèle est spécifique aux données codées en tamoul-anglais et malayalam-anglais.
An Effective ELECTRA-Based Pipeline for Sentiment Analysis of Tourist Attraction Reviews	ELECTRA	améliore la précision de la classification du sentiment par rapport à d'autres modèles de classification de texte profond, Le prétraitement des données permet de réduire le bruit dans les avis en supprimant les mots vides, les caractères spéciaux, les doublons et en remplaçant les négations, peut être appliquée à d'autres domaines tels que la finance et la santé pour analyser le sentiment des commentaires en ligne.	le prétraitement des données peut entraîner une perte d'informations importantes dans les avis, ce qui peut affecter la précision de l'analyse du sentiment.

Chapitre 4

Approche proposée

4.1 Introduction

Dans l'étude que nous avons faite, nous avons entrepris une analyse approfondie des travaux les plus récents portant sur l'analyse des sentiments. Notre objectif principal était d'examiner les différentes approches utilisées pour résoudre ce problème en utilisant des techniques de deep learning, telles que le traitement du langage naturel (NLP) et les modèles de réseaux de neurones. Nous avons passé en revue les chapitres précédents, mettant en évidence les méthodes les plus prometteuses pour améliorer la précision et la performance de l'analyse des sentiments.

Nous avons étudié comment ces travaux ont abordé la collecte et la préparation des données, ainsi que les architectures de réseaux neuroniques utilisées pour extraire les informations émotionnelles des messages sur les réseaux sociaux. Au cours de notre analyse, nous avons examiné les divers facteurs qui influencent les performances de l'analyse des sentiments, tels que la nature volatile et informelle du langage sur les réseaux sociaux, la présence de sarcasme, de doubles sens et d'émoticônes, ainsi que les défis liés à l'identification des sentiments dans des contextes multilingues et multiculturels. Nous avons également évalué les avantages et les inconvénients de ces approches, en tenant compte de critères tels que la facilité d'utilisation, la nécessité de grandes quantités de données annotées, la complexité des modèles, la capacité à gérer les variations linguistiques et culturelles, ainsi que les implications éthiques liées à la confidentialité et à la protection des données des utilisateurs.

Cette étude comparative nous a permis d'avoir une vision approfondie de l'état actuel de la recherche sur l'analyse des sentiments sur Twitter et les réseaux sociaux en general. Elle a également fourni des orientations pour la conception de futures approches de deep learning visant à améliorer la précision et l'efficacité de cette analyse. En effectuant cette étude, nous avons pu obtenir les sources d'informations nécessaires qui nous ont aidés et facilités dans la mise en oeuvre de notre approche que nous presentons dans ce chapitre.

4.2 Plateformes et outils de développement

4.2.1 Environnement de développement

Anaconda Anaconda est une distribution libre et open source des langages de programmation Python et R, utilisée pour le développement d'applications dédiées à la science des données et à l'apprentissage automatique. Elle vise à simplifier la gestion des paquets et du déploiement [78].

Jupyter notebook Jupyter notebook est le dernier environnement de développement interactif basé sur le Web pour les blocs-notes, le code, et les données. Son interface flexible permet aux utilisateurs de configurer et d'organiser des flux de travail en science des données, en informatique scientifique, en journalisme informatique, et en apprentissage automatique. Une conception modulaire permet aux extensions étendre et d'enrichir ses fonctionnalités [86].

Kaggle Kaggle, une filiale de Google LLC, est une communauté en ligne regroupant des scientifiques des données et des experts en apprentissage automatique. Les utilisateurs peuvent utiliser Kaggle pour rechercher et partager des ensembles de données, explorer et construire des modèles dans un environnement basé sur le web dédié à la science des données, ainsi que collaborer avec d'autres professionnels de la science des données et de l'apprentissage automatique [43].

Google Colab Est un environnement de bloc-notes Jupyter basé sur le cloud fourni par Google, et qui permet aux utilisateurs d'écrire et d'exécuter du code Python de manière collaborative dans un navigateur sans installation locale. Il s'intègre à Google Drive, donne accès à des GPU et TPU gratuits pour accélérer le calcul, prend en charge la collaboration en temps réel, est livré avec des bibliothèques préinstallées, et offre une interface conviviale pour organiser le code, le texte, et les médias. Dans l'ensemble, il s'agit d'un outil puissant, accessible et collaboratif pour coder et expérimenter avec Python dans le nuage [37].

4.2.2 Langage de programmation

Python est un langage de programmation puissant et facile à apprendre. Il possède des structures de données de haut niveau efficaces et une approche simple mais efficace de la programmation orientée objet. Sa syntaxe élégante et son typage dynamique, ainsi que sa nature interprétée, en font un langage idéal pour l'écriture de scripts et le développement rapide d'applications dans de nombreux domaines sur la plupart des plates-formes existantes.

4.2.3 Bibliothèques Python

Pandas utilisée pour la manipulation des données, en particulier pour lire les données à partir d'un fichier CSV ou Excel, et effectuer des opérations sur les tableaux de données.

Numpy une bibliothèque utilisée pour les calculs numériques et les opérations sur les tableaux multidimensionnels.

Matplotlib une bibliothèque de visualisation utilisée pour tracer des graphiques et des figures.

Scikit-learn (Sklearn) Scikit-learn est une bibliothèque Python qui fournit une interface standard pour la mise en œuvre d'algorithmes d'apprentissage automatique. Elle comprend également d'autres fonctions auxiliaires qui font partie intégrante du pipeline d'apprentissage automatique, telles que les étapes de prétraitement des données, les techniques de rééchantillonnage des données, les paramètres d'évaluation, et les interfaces de recherche pour ajuster/optimiser les performances d'un algorithme.

TensorFlow est une bibliothèque logicielle open-source développée par Google pour le calcul numérique et l'apprentissage automatique. Elle offre un cadre flexible pour construire et déployer des modèles d'apprentissage automatique, en particulier des réseaux neuronaux profonds. TensorFlow traite efficacement les données à grande échelle et les graphes de calcul complexes, en utilisant à la fois les CPU et les GPU. Il fournit une API de haut niveau qui permet aux utilisateurs d'exprimer les calculs sous forme de graphes de flux de données, ce qui la rend populaire pour diverses applications telles que la reconnaissance d'images, le traitement de la parole, et la compréhension du langage naturel. L'évolutivité de TensorFlow, son outillage, et sa prise en charge de l'informatique distribuée lui ont permis d'être largement adopté par la recherche et l'industrie, et de s'imposer comme une technologie fondamentale dans le domaine de l'apprentissage profond.

4.3 Métriques d'évaluation

Les mesures d'évaluation des modèles d'apprentissage profond dépendent de la tâche et du problème spécifiques sur lesquels vous travaillez. Différentes tâches, telles que la classification, la régression, la détection d'objets et le traitement du langage naturel, peuvent nécessiter des mesures d'évaluation différentes. Voici quelques mesures d'évaluation couramment utilisées pour l'apprentissage profond :

Accuracy est une mesure d'évaluation fondamentale qui évalue l'exactitude globale des prédictions dans les tâches de classification. Elle mesure la proportion d'instances correctement classées par rapport au nombre total d'instances dans un ensemble de données. La précision est généralement utilisée lorsque les classes de l'ensemble de données sont équilibrées, c'est-à-dire qu'elles sont représentées de manière à peu près égale. Elle fournit une indication directe de la performance d'un modèle en termes de prédictions correctes.[92, 19]

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Predictions}}$$

Precision se concentre sur la précision des prédictions positives. Elle quantifie la proportion de vraies prédictions positives (instances positives correctement prédites) sur l'ensemble des prédictions positives, y compris les vrais positifs et les faux positifs. La précision est particulièrement importante lorsque la minimisation des faux positifs est une priorité, comme dans le cas du diagnostic médical ou de la détection des spams. Un score de précision élevé signifie un haut niveau de confiance dans les prédictions positives.[92, 19]

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall également connu sous le nom de sensibilité ou de taux de vrais positifs, mesure la capacité du modèle à identifier toutes les instances positives réelles. Il calcule la proportion de prédictions vraiment positives sur l'ensemble des instances positives réelles, y compris les vrais positifs et les faux négatifs. Le rappel est essentiel lorsque des instances positives manquantes (faux négatifs) peuvent avoir des conséquences importantes, comme c'est le cas dans des applications telles que la détection des maladies. Un rappel élevé indique l'efficacité d'un modèle à capturer des instances positives.[92, 19]

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

F1-score combine la précision et le rappel en une seule mesure, en prenant la moyenne harmonique des deux. Cette mesure équilibrée est particulièrement utile en cas de distribution inégale des classes ou de coûts de classification différents pour les faux positifs et les faux négatifs. Un score F1 élevé est obtenu lorsque la précision et le rappel sont tous deux élevés, ce qui fournit une mesure complète des performances globales d'un modèle qui tient compte à la fois de la précision des prédictions positives et de la capacité à capturer efficacement les instances positives.[92, 19, 47]

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

En résumé, ces mesures d'évaluation jouent un rôle crucial dans l'évaluation des performances des modèles de classification. L'exactitude mesure la justesse globale, la précision se concentre sur l'exactitude des prédictions positives, le rappel évalue la capacité du modèle à identifier les instances positives, et le score F1 trouve un équilibre entre la précision et le rappel, ce qui le rend particulièrement utile dans les situations où la distribution des classes est déséquilibrée où les coûts de classification erronée varient.

4.4 Dataset Utilisé

Le dataset *Sentiment140*, également connu sous le nom de *"Training Data for Sentiment Analysis"* ou *"Sentiment140 dataset"*, est un ensemble de données largement utilisé pour

l'analyse des sentiments. Il a été créé par Alec Go, Richa Bhayani et Lei Huang de l'Université de Stanford. [91]

Ce dataset contient environ 1,6 million de tweets en anglais, qui ont été annotés avec des étiquettes de sentiment (positif ou négatif). Les tweets ont été collectés en utilisant des mots-clés spécifiques liés à différentes émotions et sentiments.

Le *Sentiment140* dataset est souvent utilisé pour entraîner et évaluer des modèles d'analyse de sentiments, notamment des modèles basés sur l'apprentissage automatique et le traitement du langage naturel. Il a été utilisé dans de nombreuses études et compétitions académiques pour développer des systèmes capables de classifier automatiquement le sentiment exprimé dans un texte court, tel qu'un tweet.

Grâce à sa taille importante et à sa variété de sentiments annotés, le dataset *Sentiment140* est devenu une référence dans le domaine de l'analyse des sentiments et a contribué à l'avancement de la recherche dans ce domaine.

4.5 Prétraitement

Le prétraitement du texte consiste en un ensemble de techniques et de méthodes appliquées aux données textuelles brutes afin de les transformer dans un format adapté aux tâches d'analyse et d'apprentissage automatique. Il s'agit d'une série d'étapes visant à nettoyer, normaliser, et transformer les données textuelles afin d'en améliorer la qualité, de réduire le bruit, et d'améliorer les performances des tâches de traitement du langage naturel (NLP) en aval.

Le prétraitement du texte joue un rôle essentiel dans l'amélioration de la qualité et de la précision des tâches d'analyse de texte ultérieures, telles que l'analyse des sentiments, la modélisation des sujets, la recherche d'informations, et la traduction automatique. En nettoyant, et en transformant les données textuelles brutes, le prétraitement facilite l'extraction efficace des caractéristiques, réduit le bruit, et améliore les performances de divers algorithmes et modèles de NLP. [3]

4.5.1 Dataset Shuffling

L'expression *"tweets dataset shuffling"* fait référence au processus de randomisation de l'ordre des tweets au sein d'un ensemble de données [24, 85, 77]. Dans le contexte de l'analyse ou du traitement des données Twitter, un ensemble de données consiste souvent en une collection de tweets rassemblés dans un but spécifique, comme l'analyse des sentiments, la modélisation des sujets, ou les tâches d'apprentissage automatique [108, 46]. Mélanger l'ensemble de données de tweets consiste à réorganiser l'ordre des tweets de manière aléatoire, sans modifier le contenu des tweets individuels.

L'objectif du brassage d'un ensemble de données de tweets est d'introduire du hasard et d'éliminer tout biais ou modèle potentiel qui pourrait être présent en raison de la manière dont les tweets ont été collectés ou stockés [24, 85]. En mélangeant l'ensemble de données, l'ordre des tweets devient indépendant des facteurs externes, ce qui permet une analyse ou une formation plus précise et impartiale des modèles [77]. Cela permet de s'assurer que

l'ordre des tweets n'influence pas les résultats, et permet une meilleure généralisation et un échantillonnage représentatif des données [108, 46].

Le brassage d'un ensemble de données de tweets est une étape de prétraitement courante dans les tâches d'analyse de données ou d'apprentissage automatique impliquant des données issues de Twitter, car il permet d'améliorer la robustesse et la fiabilité des analyses ou des modèles ultérieurs en supprimant tout effet d'ordre inhérent à l'ensemble de données.

4.5.2 Tokenization

La technique de tokenisation dans l'analyse du sentiment des tweets fait référence au processus de décomposition d'un tweet en unités individuelles appelées *"tokens"* [80]. Les tokens sont généralement des mots, mais ils peuvent également être des phrases, des symboles, ou d'autres éléments significatifs. La tokenisation est une étape fondamentale de l'analyse des sentiments qui permet d'analyser les sentiments à un niveau plus granulaire en traitant chaque token comme une entité distincte [18]. Elle permet d'extraire les éléments clés d'un tweet, de supprimer la ponctuation, et les caractères spéciaux [70], et de préparer le texte pour une analyse plus approfondie. Les techniques de tokenisation jouent un rôle essentiel dans la compréhension des sentiments exprimés dans un tweet en permettant aux modèles d'analyse des sentiments de se concentrer sur les unités individuelles et leurs relations contextuelles dans le texte [4].

4.5.3 Suppression de bruit

La technique de suppression du bruit (*"ou noise removal"*) dans l'analyse du sentiment des tweets fait référence au processus de filtrage des informations non pertinentes ou non désirées des données des tweets afin d'améliorer la précision et la fiabilité de l'analyse du sentiment. Les tweets contiennent souvent du bruit sous la forme de caractères spéciaux, d'URL, de hashtags, de ponctuation, d'émoticônes, et d'erreurs grammaticales.

Les techniques de suppression du bruit comprennent des étapes telles que la suppression des caractères spéciaux, des URL, et des hashtags, ainsi que l'application de techniques de normalisation du texte telles que le stemming ou la lemmatisation pour réduire les mots à leur forme de base. En éliminant le bruit des tweets, les modèles d'analyse des sentiments peuvent se concentrer sur le contenu essentiel et les mots porteurs de sentiments, ce qui permet d'améliorer la classification des sentiments et l'interprétation des sentiments des tweets.

4.5.4 Lemmatization

La technique de lemmatisation dans l'analyse du sentiment des tweets se réfère au processus de réduction des mots à leur forme de base (ou dictionnaire), connue sous le nom de lemmes [70]. Il s'agit de transformer les mots en leur forme canonique, en tenant compte de leurs propriétés grammaticales et de leur signification sémantique [59]. Dans l'analyse des sentiments, la lemmatisation permet de normaliser les différentes formes infléchies des mots (telles que les pluriels, les conjugaisons de verbes et les différents temps) en les ramenant à

leur lemme commun [45]. En réduisant les mots à leur forme de base, la lemmatisation réduit la dimension du vocabulaire et améliore la précision des modèles d'analyse des sentiments en capturant le sens principal et le sentiment associé au mot, indépendamment de son inflexion ou de sa variation spécifique [18]. La lemmatisation est souvent utilisée comme étape de prétraitement dans l'analyse du sentiment des tweets afin de normaliser le texte et d'améliorer les performances des modèles de classification du sentiment [5].

4.5.5 Gestion des mots rares

La technique de traitement des mots rares dans l'analyse des sentiments des tweets se réfère aux stratégies et aux approches employées pour traiter la présence de mots peu fréquents ou rares dans l'ensemble de données des tweets [69]. Les mots rares sont ceux qui apparaissent avec une faible fréquence dans l'ensemble de données et peuvent ne pas avoir suffisamment d'occurrences pour capturer des modèles de sentiments fiables. Le traitement des mots rares implique différentes méthodes telles que le remplacement des mots rares par un jeton spécial ou un jeton "inconnu", leur suppression de l'ensemble de données, ou leur mise en correspondance avec un mot similaire plus fréquent [62]. Ces techniques permettent d'éviter l'ajustement excessif et garantissent que le modèle d'analyse des sentiments peut être généralisé à des données inédites [111]. En traitant efficacement les mots rares, les modèles d'analyse des sentiments peuvent se concentrer sur des mots plus représentatifs et plus informatifs, ce qui permet d'améliorer les performances de la classification des sentiments dans les tweets [100].

4.6 Selection Des Fonctionnalités

L'extraction de caractéristiques pour l'analyse des sentiments sur Twitter est un processus fondamental dans lequel les données textuelles brutes des tweets sont transformées en représentations numériques ou catégorielles qui conviennent aux modèles d'apprentissage automatique ou d'apprentissage profond [42, 81].

Les données Twitter posent des défis uniques pour l'analyse des sentiments en raison de leur format concis, de l'utilisation d'un langage informel, et de conventions spécifiques telles que les hashtags et les expressions propres à l'utilisateur[81, 4].

Le choix des techniques d'extraction de caractéristiques est crucial et doit être soigneusement étudié [42, 4]. Différentes méthodes d'extraction de caractéristiques offrent des avantages distincts et capturent divers aspects de l'expression des sentiments dans les tweets [75, 102].

4.6.1 Longueur du texte

La caractéristique *"longueur du texte"* dans l'analyse du sentiment fait référence à la mesure de la longueur ou de la taille d'un texte donné, généralement en termes de nombre de caractères, de mots, ou de phrases qu'il contient. Dans l'analyse du sentiment, cette caractéristique est utilisée pour saisir la notion de corrélation entre la longueur du texte et

l'expression du sentiment [83, 63]. Elle peut fournir des indications sur la quantité d'informations véhiculées dans le texte et peut potentiellement avoir un impact sur le sentiment exprimé. En incorporant la caractéristique de longueur du texte, les modèles d'analyse des sentiments peuvent prendre en compte l'influence de la longueur du texte sur les prédictions de sentiment et améliorer leur précision globale [21, 67].

4.6.2 Caractéristiques spécifiques au domaine

Les caractéristiques spécifiques à un domaine dans l'analyse des sentiments dans les tweets font référence aux caractéristiques ou aux attributs du texte qui sont spécifiques à un domaine ou à un sujet particulier [48, 107]. Lors de l'analyse des sentiments dans les tweets, les caractéristiques spécifiques au domaine capturent le langage, le vocabulaire, ou les expressions uniques associés à un sujet ou à un secteur spécifique [48, 107]. Ces caractéristiques peuvent inclure des mots-clés spécifiques au domaine, des hashtags, des mentions d'entités pertinentes, du jargon industriel, ou des modèles linguistiques spécifiques couramment utilisés dans le domaine [13, 40]. En incorporant des caractéristiques spécifiques à un domaine dans les modèles d'analyse des sentiments, ceux-ci peuvent mieux capturer les nuances et le contexte des sentiments exprimés dans les tweets liés à un domaine spécifique, ce qui permet d'obtenir des prédictions de sentiments plus précises et des informations adaptées à ce domaine particulier [56].

4.6.3 POS Taging features

Les caractéristiques d'étiquetage POS dans l'analyse du sentiment des tweets font référence à l'utilisation de l'étiquetage Part-of-Speech (POS) (Partie Du Discours) pour extraire et analyser les catégories grammaticales ou les classes de mots du texte dans les tweets [81, 4, 42]. L'étiquetage POS attribue une étiquette spécifique à chaque mot d'un tweet pour identifier son rôle grammatical, comme les noms, les verbes, les adjectifs, les adverbes, etc. En incorporant des caractéristiques d'étiquetage POS, les modèles d'analyse des sentiments peuvent prendre en compte la structure syntaxique du texte et la relation entre les différentes classes de mots, ce qui peut fournir des informations contextuelles supplémentaires pour l'analyse des sentiments [81, 4, 42]. Les caractéristiques d'étiquetage POS aident à saisir les nuances de l'expression du sentiment en identifiant comment les différentes parties du discours contribuent au sentiment général dans un tweet, ce qui permet une analyse du sentiment et une interprétation du texte plus précises [28, 110].

4.7 Description de l'approche proposée

Dans ce code, nous avons développé une approche hybride puissante qui fusionne de manière transparente les capacités des réseaux neuronaux récurrents (RNN) et des réseaux neuronaux convolutifs (CNN) pour la tâche d'analyse des sentiments, en se concentrant particulièrement sur l'analyse des sentiments dans les données des tweets(Voir La figure 4.1)

Cette approche hybride exploite les forces combinées des RNN et des CNN, ce qui nous permet de capturer et d'interpréter efficacement des modèles complexes dans les textes des tweets. Alors que les RNN sont compétents pour comprendre les données séquentielles, les CNN excellent dans la reconnaissance des modèles spatiaux. En combinant ces deux architectures d'apprentissage profond, nous obtenons la capacité d'analyser simultanément les aspects temporels et spatiaux du langage.

Dans notre code, nous avons incorporé une série d'étapes de prétraitement afin de préparer les données de tweet pour ce modèle hybride. Nous calculons notamment la longueur des textes et extrayons les caractéristiques spécifiques au domaine, telles que les hashtags, les mentions et les URL. En outre, nous appliquons des techniques essentielles de prétraitement du texte afin de nous assurer que les données sont propres et prêtes pour l'analyse.

Notre approche se distingue par l'utilisation parallèle des couches LSTM (Long Short-Term Memory) et GRU (Gated Recurrent Unit) dans le cadre du RNN. Ces couches fonctionnent harmonieusement pour traiter et comprendre les nuances du texte du tweet. En outre, les résultats de ces unités LSTM et GRU sont judicieusement concaténés pour créer une représentation holistique des informations contextuelles du texte.

En plus de cette approche parallèle, nous avons introduit diverses techniques telles que l'abandon, la normalisation des lots et la régularisation du noyau afin d'améliorer la robustesse du modèle et d'éviter l'ajustement excessif.

La fusion des RNN et des CNN, en particulier l'utilisation simultanée des unités LSTM et GRU avec concaténation, présentée dans ce code, représente une approche avancée et sophistiquée pour traiter les tâches d'analyse des sentiments. Elle souligne le potentiel des architectures de réseaux neuronaux hybrides.

Nous allons nous lancer dans une exploration détaillée du code, en exposant les intrications de ce modèle d'analyse des sentiments, y compris l'incorporation des unités LSTM et GRU.

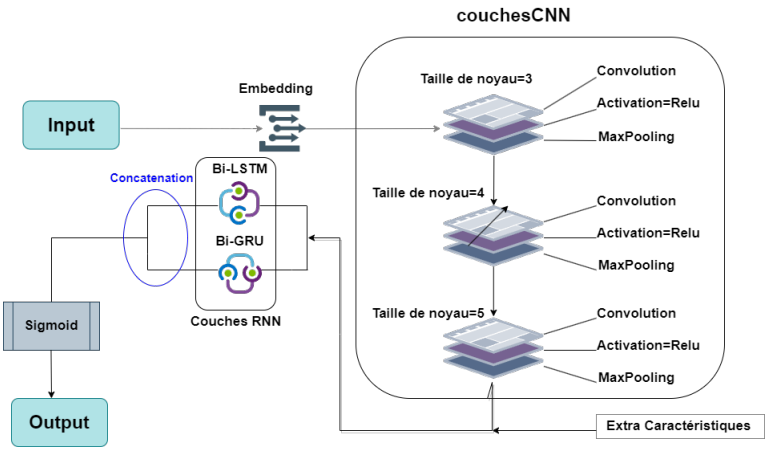

FIGURE 4.1 – Modéle Proposé.

4.7.1 Importations et Configuration

Nous avons importé les bibliothèques et modules nécessaires. Plus précisément, nous avons utilisé `pandas` et `numpy` pour la manipulation des données, `sklearn.metrics` pour le calcul des métriques d'évaluation, et `tensorflow.keras` pour la construction et l'entraînement de notre modèle d'apprentissage profond. Nous avons également configuré nos hyperparamètres avec les valeurs suivantes :

— `MAX_SEQ_LENGTH = 100`[23, 15]
— `MAX_NB_WORDS = 20000`[15]
— `EMBEDDING_DIM = 100`[16]
— `FILTER_SIZES = [3, 5, 7]`[60]
— `NUM_FILTERS = 90`[15]
— `LSTM_UNITS = 128`[15]
— `GRU_UNITS = 128`[15]
— `UNITÉS_DENSES = 1`[89]
— `DROPOUT_RATE = 0,5`[89]

4.7.2 Mélange de l'ensemble de données

Nous avons chargé notre ensemble de données CSV, qui contenait des données sur les tweets, et nous avons mélangé les lignes pour garantir le caractère aléatoire de notre formation.

4.7.3 Prétraitement des données

Pour préparer nos données de tweets à la modélisation, nous avons procédé en plusieurs étapes :

Tout d'abord, nous avons calculé la longueur de chaque texte de notre ensemble de données et l'avons stockée dans un tableau appelé `text_lengths`. Ce tableau représente la longueur (nombre de caractères) de chaque texte tweeté. Ensuite, nous avons extrait des textes tweetés des caractéristiques spécifiques au domaine, notamment le nombre de hashtags par tweet, le nombre de mentions (noms d'utilisateur Twitter) par tweet, et le nombre d'URL (liens web) dans chaque tweet.

Nous avons ensuite défini une fonction appelée `preprocess_text` pour prétraiter le texte du tweet, notamment en convertissant le texte en minuscules, en supprimant les caractères spéciaux, en symbolisant le texte, et en réunissant les symboles en une seule chaîne.

Après avoir défini la fonction de prétraitement, nous l'avons appliquée à chaque texte de tweet dans la liste des textes à l'aide d'une compréhension de liste, ce qui a permis de prétraiter toutes les données textuelles.

Nous avons ensuite procédé à la tokenisation des données textuelles prétraitées à l'aide d'un tokenizer dont la taille maximale du vocabulaire est `MAX_NB_WORDS`. Ensuite, nous avons reconverti ces séquences tokenisées en séquences numériques et les avons complétées pour atteindre une longueur de séquence maximale de `MaX_SEQ_LENGTH`. Nous avons spécifié le

remplissage à droite ("post") et la troncature à droite ("post") pour nous assurer que toutes les séquences ont la même longueur.

Enfin, nous avons combiné les données textuelles tokénisées et capitonnées avec les caractéristiques spécifiques au domaine extraites. Pour ce faire, nous avons empilé horizontalement les données textuelles avec les tableaux de caractéristiques (text lengths, hashtags, mentions, urls) à l'aide de `np.hstack`. Cette étape a permis d'intégrer les données textuelles avec des caractéristiques supplémentaires pour la modélisation.

4.7.4 Architecture du modèle

L'architecture de notre réseau neuronal commence par une couche d'intégration, qui met en correspondance les mots avec des vecteurs à 100 dimensions. Ensuite, nous avons utilisé des couches de réseau neuronal convolutif (CNN) avec des tailles de filtre de 3, 5 et 7, chacune avec 90 filtres.

Pour les couches récurrentes, nous avons utilisé des couches LSTM bidirectionnelles et GRU bidirectionnelles, toutes deux avec 256 unités. Nous avons également procédé à une concaténation, appliqué un taux d'abandon de 0.2, et inclus une normalisation par lots. La couche de sortie finale comportait une unité et nous avons utilisé un terme de régularisation à noyau d'une valeur de 0.1(voir la figure 4.1.

4.7.5 l'entraînement du modèle

Nous avons appliqué l'arrêt premature pendant l'entraînement afin d'éviter le surajustement. L'arrêt premature surveillait la précision de la validation et arrêtait la formation lorsqu'elle ne s'améliorait pas. Le processus de formation a consisté en 10 époques avec une taille de lot de 32.

4.8 Algorithmes compares

Naive Bayes : Classificateur probabiliste basé sur le théorème de Bayes. Il suppose que les caractéristiques sont indépendantes, et fonctionne souvent bien pour la classification de texte.

Machine à vecteurs de support (SVM) Algorithme de classification puissant qui trouve l'hyperplan optimal pour séparer les données en classes. Convient aux données linéaires et non linéaires.

Régression logistique (Logistic Regression) Modèle linéaire couramment utilisé pour la classification binaire, modélise la probabilité d'appartenance à une classe.

Adaboost (Adaptive Boosting) Technique d'apprentissage d'ensemble combinant des apprenants faibles en un apprenant fort en ajustant séquentiellement les poids des exemples en fonction des erreurs.

Random Forest (Forêt aléatoire) Méthode d'apprentissage d'ensemble construisant plusieurs arbres de décision et combinant leurs prédictions. Efficace pour la classification et la régression, réduit le surajustement.

Mémoire à long terme (LSTM) Type de réseau neuronal récurrent (RNN) capturant les dépendances à long terme dans les données séquentielles. Utilisé pour le traitement du langage naturel et l'analyse des séries temporelles.

Unité récurrente à portes (GRU) Autre type de RNN similaire à LSTM mais avec une architecture plus simple. Plus efficace en calcul et adapté à diverses tâches de modélisation de séquences.

4.9 Résultats expérimentaux

Dans le domaine de l'analyse des sentiments appliquée aux tweets, la recherche d'une meilleure compréhension et de prédictions précises a conduit à l'évaluation méticuleuse de divers modèles d'apprentissage profond. Le tableau comparatif 4.1 fournit un aperçu concis mais complet des résultats obtenus lors des tests rigoureux de ces modèles dans le contexte de l'analyse du sentiment des tweets. En se concentrant sur les indicateurs de performance et les résultats clés, ce tableau permet une comparaison simplifiée qui dévoile les attributs distinctifs et les limites inhérentes à chaque modèle.

La juxtaposition de ces résultats ne met pas seulement en lumière les aspects quantitatifs de la performance, mais offre également un aperçu de la nature nuancée de l'analyse des sentiments lorsqu'elle est appliquée au monde dynamique et concis des tweets. Cette ressource est conçue pour les chercheurs, les praticiens de l'analyse des sentiments et les décideurs qui souhaitent exploiter la puissance de l'apprentissage profond pour décoder les sentiments dans les tweets.

modèles	Accuracy	Precision	Recall	F1-score
Naïve bayes	72,59	72.13	73,02	72,81
Random Forest	74,10	74,68	75,01	74,80
Logistic Regression	77	75,96	76,15	75,02
SVM	70,99	71,62	71,89	71,23
AdaBoost	62,69	56,34	57	56,93
LSTM	71,24	71,73	71,56	71,4
GRU	73,89	73.32	73,23	73,54
Bi-LSTM	75,98	76,08	76,97	76,22
Bi-GRU	77,40	77,27	77,01	77,5
CNN Bi-LSTM	80,57	80,39	80,42	80,45
CNN Bi-GRU	81,23	81,42	81,03	81,1
modèle proposé(CNN-BiLSTM-BiGRU)	**83,57**	**84**	**83,58**	**83,50**

TABLE 4.1 – Résultats de l'étude comparative

4.10 Discussion des résultats

Dans notre analyse comparative complète de divers modèles d'apprentissage automatique et d'apprentissage profond, y compris Naïve Bayes, Random Forest, Logistic Regression, SVM, AdaBoost, LSTM, GRU, Bi-LSTM, Bi-GRU, CNN Bi-LSTM, et CNN Bi-GRU, le modèle que nous proposons émerge comme un modèle performant avec des avantages convaincants. Il atteint une précision impressionnante de 83,57%, ce qui démontre sa capacité à faire des prédictions globales correctes. En outre, notre modèle fait preuve d'une précision élevée de 84,00%, ce qui indique qu'il classifie correctement les instances positives. Il présente également un rappel louable de 83,58%, ce qui montre son efficacité à capturer les vrais cas positifs. Le score F1, qui harmonise la précision et le rappel, atteint 83,50%, ce qui renforce les performances équilibrées de notre modèle.

Au-delà des mesures de précision, l'architecture basée sur les réseaux neuronaux convolutifs (CNN) que nous proposons excelle dans l'extraction de caractéristiques, car elle exploite la puissance des réseaux neuronaux convolutifs (CNN) pour l'apprentissage automatique de caractéristiques à partir de données brutes. Cette capacité d'extraction de caractéristiques est particulièrement avantageuse, car elle réduit le besoin d'ingénierie manuelle des caractéristiques, qui est souvent nécessaire dans les modèles d'apprentissage automatique traditionnels tels que la régression logistique et Naïve Bayes.

En outre, la capacité de notre modèle à capturer les dépendances à long terme, grâce à l'inclusion des mémoires à long terme bidirectionnelles (bi-LSTM) et des unités récurrentes gérées (bi-GRU), en fait un excellent choix pour les tâches impliquant des données séquentielles. Cela contraste avec les modèles LSTM et GRU simples, qui peuvent avoir du mal à capturer efficacement les dépendances à très longue portée.

Par comparaison, les modèles traditionnels d'apprentissage automatique tels que Naïve Bayes et SVM sont limités dans le traitement des données séquentielles et l'extraction de caractéristiques significatives directement à partir des données brutes. En outre, si les réseaux neuronaux récurrents tels que LSTM et GRU sont efficaces pour modéliser les dépendances séquentielles, ils peuvent ne pas exploiter pleinement les points de données passés et futurs, comme le fait notre modèle bi-LSTM-GRU.

En résumé, le modèle CNN-bi-LSTM-GRU que nous proposons présente des performances supérieures en termes d'exactitude, de précision, de rappel et de capacités d'extraction de caractéristiques. Il surmonte les limites associées aux modèles traditionnels d'apprentissage automatique et présente des avantages par rapport aux architectures RNN plus simples, ce qui en fait un choix convaincant pour un large éventail d'applications.

4.11 Conclusion

Dans ce chapitre, nous avons réalisé une analyse comparative de divers modèles d'apprentissage automatique et d'apprentissage profond pour relever les défis posés par les données séquentielles complexes et l'extraction de caractéristiques. Notre évaluation a porté sur des modèles allant des algorithmes traditionnels d'apprentissage automatique tels que Naïve Bayes, Random Forest, Logistic Regression, et SVM aux réseaux neuronaux

récurrents tels que LSTM et GRU, ainsi qu'à des architectures plus avancées telles que Bi-LSTM et Bi-GRU. En outre, nous avons exploré l'intégration des réseaux neuronaux convolutifs (CNN) avec les réseaux récurrents sous la forme de modèles CNN Bi-LSTM et CNN Bi-GRU.

Au-delà de ces mesures de précision, notre modèle a excellé dans l'extraction de caractéristiques, réduisant la nécessité d'une ingénierie manuelle des caractéristiques. Cet attribut est inestimable pour faire face à la complexité des ensembles de données modernes. En outre, son aptitude à capturer les dépendances à longue portée dans les données séquentielles le distingue des modèles LSTM et GRU plus simples, qui ont souvent du mal à gérer les dépendances étendues.

En comparaison, les modèles traditionnels d'apprentissage automatique tels que Naïve Bayes et SVM ont montré des limites dans le traitement des données séquentielles et l'extraction de caractéristiques significatives directement à partir des données brutes. De même, les modèles LSTM et GRU simples ont montré des contraintes dans la capture des dépendances à très longue portée et dans l'exploitation du contexte séquentiel complet. En outre, ils s'appuient sur l'ingénierie manuelle des caractéristiques, un processus à forte intensité de main-d'œuvre.

En conclusion, le modèle CNN-bi-LSTM-GRU que nous proposons est non seulement plus performant que d'autres modèles, mais il s'attaque également aux limites associées aux modèles d'apprentissage automatique traditionnels et aux architectures de réseaux neuronaux récurrents plus simples. Ses capacités d'extraction de caractéristiques, son efficacité à capturer les dépendances à long terme, et son aptitude à traiter des ensembles de données complexes et de haute dimension en font un choix intéressant pour un large éventail d'applications.

Conclusion générale

En conclusion, ce mémoire nous a permis de dresser un panorama plus ou moins complet, précis et détaillé de l'analyse de sentiments sur les réseaux sociaux, en examinant les différentes étapes clés de cette discipline ainsi que les principaux défis et applications.

Nous avons également étudié les différentes approches utilisées pour la catégorisation des sentiments, en particulier les méthodes d'apprentissage automatique. Grâce à notre analyse comparative, nous avons pu mettre en évidence les avantages et les limites de chaque méthode, et dresser un tableau des approches les plus performantes pour l'analyse De sentiments dans des contextes spécifiques.

Après avoir exploré différentes méthodes d'analyse de sentiment, nous avons pu proposer notre propre méthode fondée sur l'apprentissage automatique. Les résultats obtenus ont montré que notre modèle surpassait de loin les approches précédentes en termes de précision et de performances. Notre approche a également permis de mieux comprendre les dimensions des sentiments exprimés sur Twitter concernant le sujet abordé.

Il est important de souligner que l'analyse de sentiments est devenue un outil incontournable pour comprendre l'opinion publique et aider les entreprises à prendre des décisions éclairées, et cela grâce aux résultats encourageants obtenus à l'aide des différentes méthodes d'analyse de sentiments. Néanmoins, il convient également de signaler des limites qui peuvent se dresser aux différentes méthodes d'analyse de sentiments notamment en termes de qualité et de quantité des données collectées. Des efforts devront ainsi être faits pour améliorer l'efficacité de l'analyse des sentiments sur les réseaux sociaux, et pour développer de nouveaux outils et approches capables de traiter les données à grande échelle en temps réel.

En somme, ce mémoire met en lumière l'importance stratégique de l'analyse de sentiments pour les entreprises et organisations cherchant à comprendre et à répondre aux besoins de leurs clients et communautés utilisateurs présents sur les réseaux sociaux. En fournissant une vue d'ensemble claire et détaillée des pratiques actuelles dans ce domaine, cette étude pourrait servir de guide pratique et de référence pour les professionnels et chercheurs intéressés par cette problématique complexe et passionnante.

Bibliographie

[1] An exploration of airline sentimental tweets with different classification models. doi : 10.18231/2454-9150.2018.0124.

[2] Comment effectuer une analyse des sentiments basée sur les aspects, Année de publication. URL https://www.voxco.com/fr/blog/comment-effectuer-une-analyse-des-sentiments-basee-sur-les-aspects/.

[3] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38, 2011.

[4] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*, pages 30–38, 2011.

[5] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*, pages 30–38, 2011.

[6] B. Agarwal and N. Mittal. Sentiment analysis of movie reviews : A study using machine learning techniques. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(7) :228–234, 2013.

[7] A. Agresti. *Categorical data analysis*, volume 482. John Wiley & Sons, 2013.

[8] R. AHUJA and S. SHARMA. B^2grua : Bertweet bi-directional gated recurrent unit with attention model for sarcasm detection. *Journal of Information Science & Engineering*, 39(4), 2023.

[9] T. M. S. Akshi Kumar, A. Kumar, and T. M. Sebastian. Sentiment analysis on twitter. *IJCSI Int. J. Comput. Sci. Issues*, 9(4) :372–378, 2012.

[10] P. Allison. *Logistic regression using SAS : Theory and application*. SAS Institute, 2012.

[11] E. Alpaydin. *Introduction to Machine Learning*. MIT Press, 2nd edition, 2010.

[12] R. N. Arku, A. Buttazzoni, K. Agyapon-Ntra, and E. Bandauko. Highlighting smart city mirages in public perceptions : A twitter sentiment analysis of four african smart city projects. *Cities*, 130 :103857, 2022.

[13] A. R. Balamurali and V. Sinha. Sentiment analysis on twitter data using domain-specific feature selection. *International Journal of Engineering & Technology*, 7(3.21) : 346–349, 2018.

[14] S. Banerjee, A. Jayapal, and S. Thavareesan. Nuig-shubhanker@ dravidian-codemix-fire2020 : Sentiment analysis of code-mixed dravidian text using xlnet. *arXiv preprint arXiv :2010.07773*, 2020.

[15] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

[16] J. Bergstra, D. Yamins, D. D. Cox, et al. Hyperopt : A python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in science conference*, volume 13, page 20. Citeseer, 2013.

[17] M. Bibi, W. A. Abbasi, W. Aziz, S. Khalil, M. Uddin, C. Iwendi, and T. R. Gadekallu. A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis. *Pattern Recognition Letters*, 158 : 80–86, 2022.

[18] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python : analyzing text with the natural language toolkit*. O'Reilly Media, 2009.

[19] A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7) :1145–1159, 1997.

[20] L. Breiman. Bagging predictors. *Machine learning*, 24(2) :123–140, 1996.

[21] E. Cambria and A. Hussain. *Sentic Computing : Techniques, Tools, and Applications for Sentiment Analysis*. Springer, 2012.

[22] E. Cambria and B. White. Jumping nlp curves : A review of natural language processing research. *IEEE Computational Intelligence Magazine*, 9(2) :48–57, 2014.

[23] A. Candelieri. A gentle introduction to bayesian optimization. In *2021 Winter Simulation Conference (WSC)*, pages 1–16. IEEE, 2021.

[24] L. Chen and H. Wang. A shuffled sequential learning algorithm for microblog sentiment analysis. pages 372–376, 2012.

[25] B. communication. Les réseaux sociaux qui comptent en 2022, 2022. URL https://www.bridge-communication.com/2022/01/10/les-reseaux-sociaux-qui-comptent-en-2022/.

[26] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3) :273–297, 1995.

[27] A. Danday and T. S. Murthy. Twitter data analysis using distill bert and graph based convolution neural network during disaster. 2022.

[28] S. Das and S. Bandyopadhyay. Sentiment analysis on twitter data using pos tagging and lexicon-based features. *International Journal of Computer Applications*, 181(34) : 22–26, 2018.

[29] A. Dhir, Y. Yossatorn, P. Kaur, and S. Chen. Online social media fatigue and psychological wellbeing—a study of compulsive use, fear of missing out, fatigue, anxiety and depression. *International Journal of Information Management*, 40 :141–152, 2018.

[30] T. Dholpuria, Y. Rana, and C. Agrawal. A sentiment analysis approach through deep learning for a movie review. In *2018 8th International Conference on Communication Systems and Network Technologies (CSNT)*, pages 173–181. IEEE, 2018.

[31] T. Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[32] M. Duggan and A. Smith. Social media update 2013, 2013.

[33] N. B. Ellison, C. Steinfield, and C. Lampe. The benefits of facebook "friends" : Social capital and college students' use of online social networking sites. *Journal of Computer-Mediated Communication*, 12(4) :1143–1168, 2007.

[34] H. Fang, G. Xu, Y. Long, and W. Tang. An effective electra-based pipeline for sentiment analysis of tourist attraction reviews. *Applied Sciences*, 12(21) :10881, 2022.

[35] Y. Freund and R. Schapire. Experiments with a new boosting algorithm. In *Thirteenth International Conference on Machine Learning*, pages 148–156, 1996.

[36] K. Gajbhiye and N. Gupta. Real time twitter sentiment analysis for product reviews using naive bayes classifier. In *Proceeding of the International Conference on Computer Networks, Big Data and IoT (ICCBI-2018)*, pages 342–350. Springer, 2020.

[37] J. Gao, Z. Jin, and H. Huang. Cloud-based computational research with google colaboratory : current status and future directions. *Environmental Health Perspectives*, 128(7) :074502, 2020.

[38] A. Garg and R. K. Kaliyar. Psent20 : An effective political sentiment analysis with deep learning using real-time social media tweets. In *2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, pages 1–5. IEEE, 2020.

[39] M. Ghiassi, J. Skinner, and D. Zimbra. Twitter brand sentiment analysis : A hybrid system using n-gram analysis and dynamic artificial neural networks. *Expert Systems with Applications*, 40(16) :6266–6282, 2013.

[40] A. Ghosal and R. Das. Sentiment analysis in social media text using domain-specific sentiment lexicon. In *2020 4th International Conference on Intelligent Sustainable Systems (ICISS)*, pages 733–738, 2020.

[41] D. Ghosh, S. Ghosh, and S. Muresan. Time matters ! exploiting temporal information for automated sarcasm detection in social media. In *Proceedings of the 24th International Conference on World Wide Web*, pages 643–653, 2015.

[42] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12), 2009.

[43] A. Goldbloom. Kaggle : Your machine learning and data science community. *Journal of Data Science*, 18(3) :491–494, 2020.

[44] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[45] E. Haddi, X. Liu, and Y. Shi. The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, 17 :26–32, 2013. ISSN 1877-0509. doi : https://doi.org/10.1016/j.procs.2013.05.005. URL https://www.sciencedirect.com/science/article/pii/S1877050913001385. First International Conference on Information Technology and Quantitative Management.

[46] M. Haffar and O. Azzi. Twitter sentiment analysis using machine learning techniques. pages 131–136, 2019.

[47] D. J. Hand and R. J. Till. A simple generalization of the area under the roc curve for multiple class classification problems. *Machine Learning*, 45(2) :171–186, 2001.

[48] A. Hassan, J. E. Guerrero, and P. Nakov. Semeval-2013 task 2 : Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2 : Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, 2013.

[49] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2017.

[50] A. Hogenboom, D. Bal, F. Frasincar, F. de Jong, and M. Bal. Coreference resolution in sentiment analysis. In *International Conference on Advanced Information Systems Engineering*, pages 341–355. Springer, 2015.

[51] D. Hosmer Jr, S. Lemeshow, and R. Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[52] M. S. Hossen, A. H. Jony, T. Tabassum, M. T. Islam, M. M. Rahman, and T. Khatun. Hotel review analysis for the prediction of business using deep learning approach. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 1489–1494. IEEE, 2021.

[53] C. J. Hutto and E. Gilbert. Vader : A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International Conference on Weblogs and Social Media*, pages 216–225, 2014.

[54] I. Irawanto, C. Widodo, A. Hasanah, P. A. D. Kusumah, K. Kusrini, and K. Kusnawi. Sentiment analysis and classification of forest fires in indonesia. *ILKOM Jurnal Ilmiah*, 15(1) :175–185, 2023.

[55] S. M. Jayasanka, T. Marcus, E. Aberathne, and S. Premaratne. Sentiment analysis for social media. 2013.

[56] S. Jindal and B. Liu. Opinion spam and analysis. In *Proceedings of the International Conference on Web Search and Data Mining (WSDM)*, pages 219–230, 2008.

[57] T. Joachims. Text categorization with support vector machines : Learning with many relevant features. In *European conference on machine learning*, pages 137–142, 1998.

[58] Y. G. Jung, K. T. Kim, B. Lee, and H. Y. Youn. Enhanced naive bayes classifier for real-time sentiment analysis with sparkr. In *2016 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 141–146. IEEE, 2016.

[59] D. Jurafsky and J. Martin. *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, volume 2. 02 2008.

[60] R. Khalid and N. Javaid. A survey on hyperparameters optimization algorithms of forecasting models in smart grid. *Sustainable Cities and Society*, 61 :102275, 2020.

[61] S.-M. Kim and E. Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 1367–1373, 2004.

[62] Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1746–1751, 2014.

[63] S. Kiritchenko and S. Mohammad. Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 811–817, 2016.

[64] D. J. Kuss and M. D. Griffiths. Online social networking and addiction—a review of the psychological literature. *International Journal of Environmental Research and Public Health*, 8(9) :3528–3552, 2011.

[65] B. Liu. *Sentiment analysis and opinion mining*, volume 5 of *Synthesis lectures on human language technologies*. 2012.

[66] B. Liu. *Sentiment analysis and opinion mining*, volume 5. Synthesis Lectures on Human Language Technologies, 2012.

[67] B. Liu. *Sentiment analysis : Mining opinions, sentiments, and emotions*. Cambridge University Press, 2015.

[68] B. Liu, Z. Cao, and L. Zhang. Fuzzy random forests for uncertain data. *IEEE Transactions on Fuzzy Systems*, 23(4) :996–1008, 2015.

[69] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics : Human language technologies*, volume 1, pages 142–150, 2011.

[70] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.

[71] M. Mansouri and R. Khelifi. A new hybrid approach for text classification based on a combination of support vector machine (svm) and random forest (rf). *Intelligent Automation & Soft Computing*, 25(2) :375–384, 2019.

[72] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[73] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, ..., and S. Petersen. Human-level control through deep reinforcement learning. *Nature*, 518 (7540) :529–533, 2015.

[74] S. Modak and A. C. Mondal. Sentiment analysis of twitter data using clustering and classification. In P. K. Singh, S. T. Wierzchoń, S. Tanwar, J. J. P. C. Rodrigues, and M. Ganzha, editors, *Proceedings of Third International Conference on Computing, Communications, and Cyber-Security*, pages 651–664, Singapore, 2023. Springer Nature Singapore. ISBN 978-981-19-1142-2.

[75] S. M. Mohammad and P. D. Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3) :436–465, 2013.

[76] MédiaNet. Étude : Réseaux sociaux en algérie, Année de publication. URL https://www.medianet.tn/fr/actualites/detail/etude-reseaux-sociaux-en-algerie/all/1.

[77] B. Naderi, N. GhasemAghaee, and M. Dehghani. Deep shuffled cnn for twitter sentiment analysis. pages 23–27, 2017.

[78] T. E. Oliphant. Python for scientific computing. *Computing in Science & Engineering*, 9(3) :10–20, 2007.

[79] D. Orzech. Réseaux sociaux : quelles perspectives pour 2023 ?, 2023. URL https://www.agorapulse.com/fr/blog/reseaux-sociaux-quelles-perspectives-pour-2023/?fbclid=IwAR0w6bZ.

[80] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf.

[81] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC 2010)*, 2010.

[82] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends® in information retrieval*, 2(1-2) :1–135, 2008.

[83] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2) :1–135, 2008.

[84] I. Pantic, A. Damjanovic, J. Todorovic, D. Topalovic, D. Bojovic-Jovic, S. Ristic, and S. Pantic. Association between online social networking and depression in high school students : Behavioral physiology viewpoint. *Psychiatria Danubina*, 24(1) :90–93, 2012.

[85] D. Paul, A. Dey, and P. Nandy. A comparative analysis of feature selection techniques for sentiment analysis of twitter data. pages 88–93, 2015.

[86] J. M. Perkel. Why jupyter is data scientists' computational notebook of choice. *Nature*, 563(7729) :145–146, 2018.

[87] S. Poria, E. Cambria, and A. Hussain. A review of affective computing : From unimodal analysis to multimodal fusion. *Information Fusion*, 37 :98–125, 2017.

[88] M. A. Rai, R. Jain, G. Das, and P. Bharadwaj. Sentimental analysis and detection of rumours for social media data using logistic regression. *International Journal of Innovative Technology and Exploring Engineering*, 9(1) :2123–2126, 2019. doi : 10.35940/ijitee.a4670.119119.

[89] M. Rhanoui, M. Mikram, S. Yousfi, and S. Barzali. A cnn-bilstm model for document-level sentiment analysis. *Machine Learning and Knowledge Extraction*, 1(3) :832–847, 2019.

[90] A. Samuels and J. Mcgonical. News sentiment analysis. *arXiv preprint arXiv :2007.02238*, 2020.

[91] Sentiment140. Sentiment140 help for students, 2023. URL http://help.sentiment140.com/for-students. May,21st,2023.

[92] J. Sokolova, N. Japkowicz, and S. Szpakowicz. Beyond accuracy, f-score and roc : A family of discriminant measures for performance evaluation. In *Proceedings of the 2006 Conference of the Canadian Society for Computational Studies of Intelligence*, pages 101–108. Springer, 2006.

[93] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4) :427–437, 2009.

[94] D. Sunitha, R. K. Patra, N. Babu, A. Suresh, and S. C. Gupta. Twitter sentiment analysis using ensemble based deep learning model towards covid-19 in india and european countries. *Pattern Recognition Letters*, 158 :164–170, 2022.

[95] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim. Roberta-lstm : a hybrid model for sentiment analysis with transformer and recurrent neural network. *IEEE Access*, 10 :21517–21525, 2022.

[96] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 63(1) :406–418, 2012.

[97] Y. Tian, J. Liu, and X. Zeng. A hybrid method of random forest and convolutional neural network for sentiment analysis. *IEEE Access*, 7 :154369–154376, 2019.

[98] P. D. Turney. Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, 2002.

[99] S. Urologin. Sentiment analysis, visualization and classification of summarized news articles : a novel approach. *International Journal of Advanced Computer Science and Applications*, 9(8), 2018.

[100] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, ..., and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[101] J. Vitak, N. B. Ellison, and C. Steinfield. The ties that bond : Re-examining the relationship between facebook use and bonding social capital. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, pages 571–580, 2011.

[102] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120, 2012.

[103] Y. Wang, M. Huang, L. Zhao, and J.-R. Wen. Group-based context-aware sarcasm detection in social media. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1947–1950, 2012.

[104] Wikipédia, l'encyclopédie libre. Plutchik's wheel of emotions (french). https://fr.wikipedia.org/wiki/Fichier:Plutchik-wheel_fr.svg, Dernière modification le 10 septembre 2023. Page Wikipédia.

[105] D. Wolpert. Stacked generalization. *Neural networks*, 5(2) :241–259, 1992.

[106] Y. Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1-2) :69–90, 1999.

[107] L. Yu, Y. Zhang, and Z. Zhang. Domain-specific sentiment classification with word embeddings. In *2017 International Conference on Service Systems and Service Management (ICSSSM)*, pages 1–6, 2017.

[108] J. Yuan, W. Wang, and J. Luo. Sentiment analysis of tweets with topic models. *Expert Systems with Applications*, 41(14) :6315–6321, 2014.

[109] L. Zhang and R. Ghosh. Sentiment analysis of restaurant reviews using machine learning approaches. *Journal of Big Data*, 5(1) :1–17, 2018.

[110] Y. Zhang and X. Ding. A hybrid system of sentiment classification based on tokenizing words by pos tagging. In *2010 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 757–762, 2010.

[111] Y. Zhang and Y. Yang. A survey on deep learning for sentiment analysis. *Information Fusion*, 42 :146–157, 2018.

Abstract

Social networks, especially Twitter, have transformed global communication, elevating sentiment analysis as vital research. This study delves into Twitter sentiment analysis, covering methods, tools, and challenges. We explore the core foundations, primary obstacles, and applications. The practical dissertation section entails a Twitter sentiment analysis using a tweet sample. We scrutinize every facet, offering an overview of research trends. Furthermore, we introduce our classification model (CNN-bi-LSTM-GRU) and contrast its performance with conventional models. Our comparison highlights the superiority of the proposed hybrid CNN and RNN approach in feature extraction and accuracy, underscoring its value for such tasks.
Caractéristiques

Les réseaux sociaux, en particulier Twitter, ont révolutionné la communication mondiale, faisant de l'analyse des sentiments un domaine de recherche essentiel. Cette étude explore l'analyse des sentiments sur Twitter, en se concentrant sur les méthodes, les outils, et les défis. Nous étudions les fondements conceptuels de l'analyse des sentiments, les principaux défis, et les applications.

La partie pratique de ce mémoire consiste en une étude de l'analyse de sentiments sur Twitter, en utilisant un échantillon représentatif de tweets. Nous avons examiné en détail les divers aspects de cette analyse, et avons présenté un aperçu des tendances de la recherche en ce domaine. Par ailleurs nous avons, nous-mêmes, proposé un modèle (CNN-bi-LSTM-GRU) de classification, et en avons comparé les performances avec des modèles classiques. Notre étude comparative montre la supériorité de l'approche proposée, en terme d'extraction de caractéristiques et de précision, soulignant ainsi l'intérêt des solutions hybrides CNN et RNN pour ce genre de tâches.

keywords

Artificial Intelligence (AI),Sentiment Analysis,Social Networks,Machine Learning,Deep Learning,Natural Language Processing (NLP),Emotion recognition..