République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université A. Mira de Béjaïa

Faculté des Sciences Exactes

Département d'Informatique

Mémoire de Fin d'Etude

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

Thème

Analyse des sentiments envers la vaccination contre le coronavirus dans les réseaux sociaux

---

Réalisé par :

M^{elle} BELMEHDI Sara

M^{elle} BENZAID Mounia

Examinatrice :	*Dr AISSANOU* Karima Epse ADEL	M.C.A Université de Béjaïa.
Examinatrice :	*Dr TIGHIDET* Soraya Epse ALOUI	M.C.A Université de Béjaïa.
Encadrante :	*Dr EL BOUHISSI* Houda Epse BRAHAMI	M.C.A Université de Béjaïa.

Promotion 2020 - 2021

Remerciements

*N*ous souhaitons vivement remercier notre encadreur Madame EL BOUHISSI Houda pour l'assistance qu'elle nous a témoignée, pour sa disponibilité, son orientation et conseils sans lesquels ce travail n'aurait pas vu le jour, qu'elle trouve ici l'expression de notre gratitude.

Nos vifs remerciements vont également aux membres du jury pour l'intérêt qu'ils ont porté à notre recherche en acceptant d'examiner notre travail et de l'enrichir par leurs propositions.

Enfin, Nous remercions aussi toutes nos familles respectives et tous nos amis et collègues qui nous ont soutenus, ainsi que tous ceux qui ont contribué de près ou de loin à la réalisation de ce travail.

Dédicace

Je dédie ce modeste travail,

A la mémoire de mon père,

A ma tendre et chère mère,

Qui n'a jamais cessé de formuler des prières à mon égard, de me soutenir et de m'épauler pour que je puisse atteindre mon objectif.

A mes valeureuses sœurs Katia et Dania,

A mon cher grand frère Elias, à son épouse Lydia

A mes chers petits neveux Ali et Adam

A toute la famille,

A mes amies Dahia et Melissa,

A mon binôme et amie Mounia et à toute sa famille,

BELMEHDI Sara

Dédicace

Je dédie ce modeste travail,

A toutes les personnes qui ont contribué de prés ou de loin à la réalisation de ce travail

Mes pensées vont avant tout à mes parents, qui ont été une source de motivation pour moi tout au long de mon cursus

A mon frère Samy et à ma valeureuse sœur Lisa,

A toute la famille,

Mes deux grands-méres que j'aime énormément

Mes très chéres cousines Manel et Yasmine qui ont toujours été là, dans les bons et les mauvais moments,

Mes très chers amis

Ines et Mélissa,qui m'ont apporté un soutien inconditionnel

A mes amies d'enfance Lina et Celyana, et à mon cher ami Yacine, qui n'ont jamais cessé de me soutenir et de m'encourager,

A mon amie Dahia, je tiens à la remercier pour l'aide qu'elle m'a apportée durant tout mon cursus universitaire,

A ma chére binôme et amie Sara et à toute sa famille,

BENZAID Mounia

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

AS	**A**nalyse de **S**entiment
OM	**O**pinion **M**ining
TALN	**T**raitement **A**utomatique du **L**angage **N**aturel
IA	**I**ntelligence **A**rtificielle
OMS	**O**rganisation **M**ondiale de **S**anté
CDC	**C**enters for **D**isease **C**ontrol and **P**revention
BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers
SVM	**S**upport **V**ector **M**achine
NB	**N**aive **B**ayes
GLoVe	**G**lobal **V**ectors
Bi-LSTM	**Bi**directional **L**ong **S**hort-**T**erm **M**emory
NLP	**N**atural **L**anguage **P**rocessing
LR	**L**ogistic **R**egression
LSTM	**L**ong **S**hort-**T**erm **M**emory
API	**A**pplication **P**rogramming **I**nterface
RNN	**R**ecurrent **N**eural **N**etwork
TF-IDF	**T**erm **F**requency-**I**nverse **D**ocument **F**requency
BNB	**B**ernouli **N**aive **B**ayes
CNB	**C**omplement **N**aive **B**ayes
MNB	**M**ultinomial **N**aive **B**ayes
NBC	**N**aive **B**ayes **C**lassifier
JSON	**J**avaScript **O**bjet **N**otation
XML	**E**xtensible **M**arkup **L**anguage
RDF	**R**esource **D**escription **F**ramework
CSV	**C**omma-**S**eparated **V**alues
GMT	**G**reenwich **M**ean **T**ime
HTML	**H**yper**T**ext **M**arkup **L**anguage
W3C	**W**orld Wide Web **C**onsortium
WHATWG	**W**eb **H**ypertext **A**pplication **T**echnology **W**orking **G**roup
NLTK	**N**atural **L**anguage **T**oolKit
BSD	**B**erkeley **S**oftware **D**istribution
TSV	**T**abulation-**S**eparated **V**alues

Résumé

Avec l'émergence du web 2.0, les réseaux sociaux représentent des Plateformes et outils d'échange d'informations incontournables, ils sont utilisés à des fins personnels et professionnels.

Les internautes expriment leurs sentiments et échangent des idées concernant différents sujets d'actualité, y compris celui de la santé, par conséquent, il est primordial pour les professionnels de ce domaine de s'intéresser aux différentes opinions partagées sur les réseaux sociaux et de les analyser particulièrement en pleine pandémie.

Les méthodes d'analyse de sentiment basiques ne sont plus efficaces compte tenu de la quantité de données énorme présente sur le web (2.5 exaoctets par jour), c'est pour cela que le machine learning devient nécessaire, ce dernier consiste à créer des systèmes qui apprennent ou améliorent les performances en fonction des données qu'ils traitent, c'est un outil d'aide à la décision grâce à son pouvoir de prédiction.

Notre projet portera sur l'analyse des sentiments dans les réseaux sociaux envers la vaccination contre le coronavirus, étant donné que la campagne de vaccination contre le virus covid-19 est un sujet sensible qui ne met pas tout le monde d'accord, notre travail consistera à déterminer la tonalité émotionnelle des discours des internautes en les classifiant dans trois catégories : positif, neutre et négatif.

mots clés : analyse de sentiments,classification, coronavirus, opinions, pandémie, réseaux sociaux, vaccination.

Abstract

With the emergence of web 2.0, social networks represent essential information exchange platforms and tools, they are used for personal and profesional purposes.

Internet users express their feeling and exchange ideas on various topical issues, including that of the health, therefore, it is essential for professionals in this field to be interested in the different opinions shared on social networks and analyze particulary in the midst of a pandemic.

The basic sentiment analisis methods are no longer effective given the huge amount of data present on the web (2.5 exabytes per day), this is why machine learning becomes necessary, the latter consists in creating systems that learn or improve performance based on the data it processes, it is a decision support tool thanks to its power to prediction.

Our project will focus on the analysis of feelings in social networks towards the coronavirus vaccination, given that the vaccination compaign against the covid-19 virus is a sensitive subject that does not make everyone agree, our work will consist in determining the emotional tone of the speeches of internet users by classifying them in 3 categories : positive, neutral and negative.

Keywords : classification,opinions, pandemic, sentiment analysis, social networks, vaccination covid-19, .

Chapitre 1

Introduction générale

1.1 Introduction

La COVID-19 est la maladie causée par un nouveau coronavirus, le SARS-CoV-2. L'OMS a appris l'existence de ce nouveau virus le 31 décembre 2019 lorsqu'un foyer épidémique de cas de « pneumonie virale » a été notifié à Wuhan, en République populaire de Chine [3]. Depuis l'apparition du coronavirus, le monde entier fait face à une crise sanitaire majeure. De l'économie mondiale aux simples gestes du quotidien, la pandémie a bouleversé la planète entière.

En Algérie, le virus a commencé à se propager à partir du 25 février 2020 lorsqu'un salarié d'Eni originaire de Lombardie a été testé positif au SARS-CoV-2. A partir de là une famille habitant à Blida a été touchée, et le virus s'est propagé dans tout le pays.

1.2 Problématique

La vaccination contre la maladie covid-19 suscite énormément de débats, ce qui n'est surprenant, étant donné que la contestation vaccinale existe depuis toujours.

Si les vaccins sont apparus dans un premier temps comme l'instrument universel idéal d'une politique de santé globale, ils sont au fil du temps devenus « des produits inno-

vants distribués par des filières commerciales dans un monde dit globalisé où des maladies émergentes surviennent qui elles, ne connaissant pas de vaccins, impliquant de nouvelles recherches ». Ce qui entraîne surveillance des maladies qui peut être assimilée à de l'espionnage ! D'autant que déferlent sur le Net, quantité d'informations non contrôlées qui peuvent affoler les populations. Ce qui se passe également au Nord où les réseaux sociaux s'enflamment à propos des vaccins, autour de leur efficacité, leur sécurité et leur réelle opportunité à être utilisés [4].

Les campagnes de vaccination ont débuté dans plusieurs pays, certaines ont été ralenties par les suspicions d'effets secondaires graves liés à l'injection des premières doses de vaccins, tandis que dans d'autres pays, notamment dans le continent africains, le démarrage des campagnes de vaccination est lent, faute de moyens.

Dans l'ensemble, il existe plus de cent états qui n'ont toujours pas administré les premières doses de vaccin comme le japon qui est présenté comme l'un des pays les plus vaccinosceptiques au monde.

Plusieurs débats concernant la vaccination contre le coronavirus ont lieu sur les réseaux sociaux, ces derniers amplifient la confusion déjà alimentée par la désinformation basée sur les vaccins. Plus de la moitié de la planète utilise les réseaux sociaux et échange des idées dans les différentes plateformes telles que Twitter, ce chiffre a considérablement augmenté depuis le début de la pandémie. C'est pourquoi, nous estimons que l'utilisation des données présentes sur les réseaux sociaux nous permettra de collecter les informations nécessaires à notre étude et ainsi de répondre à nos besoins.

1.3 Objectifs

Dans le but de comprendre l'opinion du grand public et de répondre aux préoccupations des vaccino-sceptiques, une analyse des sentiments envers la vaccination contre le coronavirus est nécessaire.

Les techniques d'intelligence artificielle et d'apprentissage automatique viennent améliorer et automatiser les anciennes méthodes utilisées par les gouvernements comme les

enquêtes et les sondages. L'analyse des sentiments consiste à catégoriser les opinions subjectives à partir de sources textuelles, audio et vidéo [5]pour déterminer les polarités (par exemple, positive, négative et neutre), les émotions (par exemple, la colère, la tristesse et le bonheur) ou les états d'esprit (par exemple, l'intérêt contre le désintérêt) envers des sujets, des thèmes ou des aspects d'intérêt cibles [6].

Enfin, notre objectif est d'analyser les sentiments du grand public envers la vaccination contre le coronavirus en utilisant des données extraites du réseau social 'Twitter', dans le but de classer les sentiments selon leurs polarités et d'estimer le pourcentage de personnes qui portent un avis positif, négatif ou neutre envers la vaccination contre la maladie de la covid-19.

1.4 Contributions

Les contributions principales de ce mémoire sont :

— Proposition d'un outil qui permet de visualiser le dataset à partir d'une page web.
— Classification des sentiments analysés selon leurs polarités.
— Visualisation des résultats de l'analyse des sentiments à partir de l'interface d'une application web.
— Elaboration de tests de précision.

L'analyse est conclue par l'évaluation de l'algorithme de classification utilisé, cette étape permet de déterminer la qualité du processus établi.

1.5 Méthodologie de travail

Dans ces grandes lignes, la démarche adoptée est basée sur l'intelligence artificielle pour analyser et comprendre la cause de l'hésitation des populations face à la vaccination contre le virus du covid-19. en tenant compte du fait que notre application de machine learning est dédiée à une certaine catégorie de personnes, nous avons mis l'accent sur

l'efficacité de l'algorithme en gardant une interface graphique simple qui servira à afficher les résultats de manière claire. Notre démarche de travail repose plus précisément sur les étapes suivantes :

— étape de recherche du dataset adéquat et nécessaire à l'analyse des sentiments.
— Etape d'analyse des sentiments : cette étape comprend plusieurs traitements différents avant d'arriver au résultat final.
— Etape de conception d'une page web : création de l'interface graphique et affichage des résultats de l'analyse.

1.6 Organisation du mémoire

Les prochains chapitres du mémoire sont présentés comme ceci :

Chapitre 2 : Ce chapitre porte sur les généralités et notions relatives à notre thème :analyse des sentiments, vaccination contre le virus du covid-19 et l'importance de l'analyse concernant ce sujet.

Chapitre 3 : Présente une revue de littérature sur l'analyse des sentiments envers la vaccination contre le coronavirus tentant de couvrir la majorité des approches connexes à nos travaux de recherche.

Chapitre 4 : porte sur la méthodologie suivie dans la conception du programme d'analyse des sentiments.

Chapitre 5 : porte sur les principaux outils utilisés tout au long de la conception de la page web et de l'environnement de développement.

Chapitre 6 : le dernier chapitre du mémoire représente la conclusion générale, nous évaluerons ce travail et donnerons ainsi les perspectives liées à la poursuite de ce dernier.

Chapitre 2

Généralités

2.1 Introduction

Ce chapitre portera dans un premier temps sur l'analyse des sentiments, ses différentes disciplines, ses différents niveaux d'analyse ainsi que sa complexité. Nous allons ensuite aborder le sujet de la vaccination contre le virus Covid-19 et de la réticence du public face à ce dernier, et pour finir nous parlerons de l'utilité de l'analyse des sentiments dans les réseaux sociaux envers la vaccination contre le virus Covid-19.

2.2 Analyse des sentiments et opinion mining

2.2.1 Le sentiment et l'opinion

Dans le domaine de l'analyse des sentiments, une opinion et un sentiment sont souvent confondus, nous allons donc définir brièvement chacun des deux termes.

2.2.1.1 Sentiment

Le sentiment est avant tout l'acte et le résultat du sentir, lequel désigne la prise de conscience immédiate, sans intermédiaire, sans distance, des choses et de nous-mêmes ; l'objet du sentiment est toujours ce qui nous «touche» [7].

Le sentiment est la composante de l'émotion qui implique les fonctions cognitives de l'organisme, la manière d'apprécier. Le sentiment est à l'origine d'une connaissance immédiate ou d'une simple impression. Il renvoie à la perception de l'état physiologique du moment. Le sens psychologique de sentiment qui comprend un état affectif est à distinguer du sens propre de la sensibilité.

2.2.1.2 Catégorisation et polarité des sentiments

La polarité peut être définie par des catégories telles que « positif », « neutre », et « négatif » La polarité d'une opinion exprime la positivité, la négativité ou une information de cette dernière. On dit d'une opinion positive qu'elle possède une polarité positive, et inversement, on dit d'une opinion négative qu'elle possède une polarité négative[1].

La figure 2.1 est un organigramme qui représente l'exemple d'un processus de catégorisation de la polarité des sentiments [1].

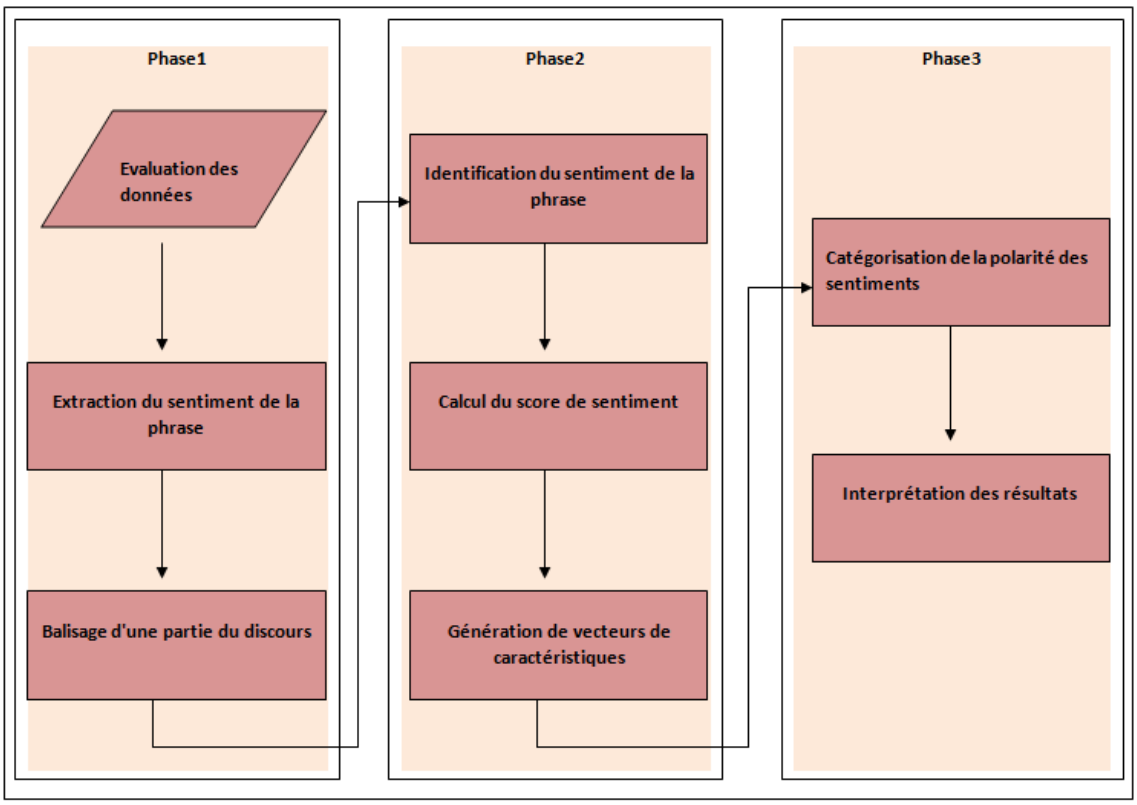


FIGURE 2.1 – Le processus de catégorisation de la polarité des sentiments [1].

2.2.1.3 Opinion

Une opinion est un quadruple (g ; s ; h ; t) : où g est la cible du sentiment, s est le sentiment de l'opinion sur la cible g, h est le détenteur de l'opinion (la personne ou l'organisation qui détient l'opinion), et t est le moment où l'opinion est exprimée. Les quatre éléments ici sont essentiels. Il est généralement problématique si l'un des leur manque. Par exemple, la composante temps est importante en pratique car une opinion d'il y a deux ans n'est pas la même chose qu'une opinion d'aujourd'hui. Ne pas avoir d'opinion titulaire est également problématique. Par exemple, une opinion d'une personne très importante (par exemple, le président américain) est probablement plus importante que celui du Joe moyen dans la rue [8].

2.2.1.4 Types Opinion

On peut distinguer généralement deux catégories d'opinions : usuelle/comparative et explicite/implicite [9] :

1. **Opinion usuelle ou comparative**

opinion usuelle : Une opinion usuelle est une expression simple d'un avis, pouvant viser de manière directe ou indirecte le sujet principal. Les opinions directes sont actuellement les plus exploitées dans la plupart des études de recherches en opinion mining pour leur simplicité et facilité à déterminer les différentes parties de l'opinion elle-même.

Opinion comparative L'opinion comparative représente l'expression d'un avis sur un sujet en le comparant à un autre. Cette relation de comparaison aide à déterminer la valeur de l'opinion émise à l'égard du premier sujet.

2. **Opinion explicite ou implicite**

Opinion explicite Une opinion explicite est souvent un avis subjectif, exprimé de manière simple ou à travers une comparaison.

Opinion implicite Une opinion implicite est un avis généralement objectif qui sous-entend l'expression d'une opinion usuelle ou comparative. Cette catégorie d'opinion étant plus difficile à déterminer, et clairement moins explorée de par le nombre d'études à son compte.

2.2.2 Définition de l'analyse des sentiments

L'analyse des sentiments consiste essentiellement à juger le sentiment et l'émotion qui se cache derrière un écrit. Le processus consiste à agir sur un texte, une phrase ou un article complet, et à analyser l'émotion que l'auteur exprime. Les sentiments sont généralement classés en trois types : négatifs, neutres et positifs.

Les deux expressions analyse de sentiments(AS) ou opinion mining(OM) sont interchangeables. Ils expriment une signification mutuelle. Cependant, certains chercheurs ont déclaré que l'OM et l'AS ont des notions légèrement différentes : Opinion Mining extrait et analyse l'opinion des gens sur une entité tandis que l'analyse des sentiments identifie le sentiment exprimée dans un texte puis l'analyse. Par conséquent, l'objectif de SA est de trouver des opinions, d'identifier les sentiments qu'ils expriment, puis classez leur polarité.

Nous avons schématisé un exemple d'un processus d'analyse de sentiment dans la

figure 2.2.

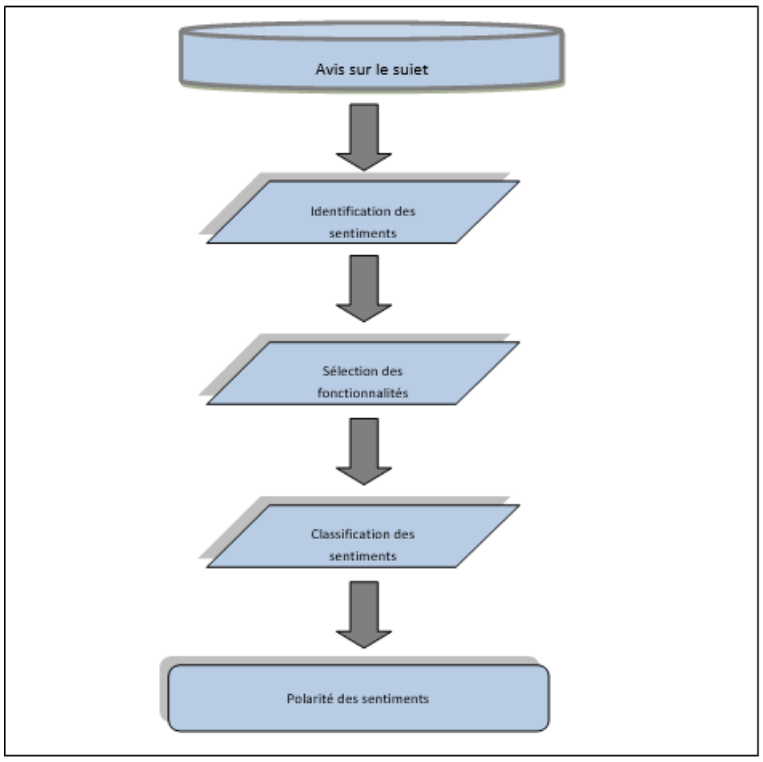


FIGURE 2.2 – Processus d'analyse des sentiments sur un sujet quelconque.

2.2.3 Niveau d'analyse des sentiments

En général, l'analyse des sentiments a été étudiée principalement à trois niveaux :

2.2.3.1 Niveau du document :

Connue sous le nom de classification des sentiments au niveau du document. Ce niveau d'analyse suppose que chaque document exprime des opinions sur une seule entité. Par exemple, étant donné un produit revu, le système détermine si la revue exprime un avis positif ou négatif sur le produit. Cette tâche est généralement appelée analyse des sentiments et exploration d'opinion [10].

2.2.3.2 Niveau de la phrase :

Ce niveau d'analyse est étroitement lié à la classification de la subjectivité, qui distingue les phrases objectives (qui expriment des informations provenant de phrases subjectives) et les phrases subjectives (qui expriment points de vue et opinions subjectifs).

Cependant, il faut noter que la subjectivité n'est pas équivalente au sentiment comme de nombreuses phrases objectives peuvent impliquer opinions. Par exemple, nous avons acheté la voiture le mois dernier et l'essuie-glace est tombé. Les chercheurs ont également analysé les clauses mais le niveau de la clause n'est toujours pas suffisant [11].

2.2.3.3 Niveau des aspects :

Contrairement au niveau du document et au niveau de la phrase, le niveau d'aspect effectue une analyse plus fine. Le niveau d'aspect était plus tôt appelé niveau de fonctionnalité (extraction et synthèse d'opinions basées sur les fonctionnalités).

Au lieu de regarder les constructions de langage (documents, paragraphes, phrases), niveau d'aspect regarde directement l'opinion elle-même. Il est basé sur l'idée qu'une opinion se compose d'un sentiment (positif ou négatif) et d'une cible (d'opinion).

Un avis sans que son objectif soit identifié est d'une utilité limitée, cependant réaliser l'importance des cibles d'opinion nous aide également à mieux comprendre le sentiment problème .

Par exemple, la phrase « L'iPhone est très bon, mais il faut encore travailler sur la durée de vie de la batterie et les problèmes de sécurité » évalue trois aspects : iPhone (positif), la durée de vie de la batterie (négatif) et la sécurité (négative) [11].

2.2.4 Complexité de l'analyse des sentiments

Toute personne qui a étudié la linguistique ne vous dirait que les langues sont complexes, Il serait trop naïf de simplifier à l'excès le langage en pensant que son sentiment

sous-jacent peut toujours être examiné avec précision par une machine ou un algorithme.

Il y a cinq facteurs principaux qui nous empêchent actuellement de compter aveuglément sur des outils pour l'analyse des sentiments [12].

2.2.4.1 Contexte

Un mot positif ou négatif peut avoir un sens inverse en fonction du contexte. "J'ai fait un excellent travail" peut être interprété comme une affirmation positive.

Cependant, dans "mon fournisseur d'Internet fait un excellent travail quand il s'agit de me voler de l'argent", faire un bon travail n'est plus une chose positive, basée sur le contexte ("me voler de l'argent") [13].

2.2.4.2 Sarcasme et l'ironie

Parfois, les opinions implicites peuvent s'exprimer ironiquement, ce qui complique davantage l'analyse de polarité. Dans le tweet « Valls a appris la mise sur écoute de Sarkozy en lisant le journal.

Heureusement qu'il n'est pas ministre de l'intérieur ironie » extrait du corpus FrIC(Un corpus et un schéma d'annotation multi-niveaux pour l'ironie dans les tweets), l'utilisateur emploie une fausse assertion (texte souligné) qui rend de ce fait le message très négatif envers Valls. On remarquera ici le recours au hashtag ironie qui permet d'aider le lecteur à comprendre que le message est ironique.

Il est important de noter que bien que l'extraction des opinions dans ces exemples est d'une simplicité presque enfantine pour un humain, son extraction automatique est extrêmement complexe pour un programme informatique.

Chaque type de langage figuratif a ses propres mécanismes linguistiques qui permettent de comprendre le sens figuré. L'inversion de la réalité/vérité pour exprimer l'ironie, la présence des effets amusants pour exprimer l'humour, etc. Dans la plupart des cas, l'ensemble des phénomènes figuratifs nécessite le recours au contexte de l'énonciation afin

que le lecteur ou l'interlocuteur réussisse à interpréter le sens figuré d'un énoncé donné.

Par conséquent, il est important de pouvoir inférer des informations au-delà des aspects lexicaux, syntaxiques voir même sémantiques d'un texte. Ces inférences peuvent varier selon le profil du locuteur (comme le genre) ou encore son contexte culturel [14].

2.2.5 Problèmes de l'analyse des sentiments

L'analyse des sentiments connait plusieurs limites comme toutes les disciplines de l'informatique, dans ce qui suit nous allons aborder les différents problémes liés à l'analyse des sentiments.

2.2.5.1 Problèmes de précision

La marge d'erreur est souvent trop élevée et souvent peu explicitée, ce qui indique alors qu'il n'existe même pas de tests de robustesse ou que l'on ne fournit pas l'information [15].

2.2.5.2 Problèmes terminologiques et d'écriture

L'ambiguïté de certains mots positifs ou négatifs selon les contextes ne peut pas toujours être levée, le vocabulaire évaluatif n'est pas le même d'une langue à l'autre, l'écriture web 2.0 entraîne des traitements de chaînes de caractères inédits [15].

2.2.5.3 Problème de scalabilité et de transposabilité

L'extension d'un dictionnaire d'un domaine à un autre n'est en rien évidente, ce qui entraîne un travail de reprise et d'enrichissement à chaque nouveau domaine [15].

2.2.6 Différentes disciplines de l'analyse de sentiments

2.2.6.1 Fouille de texte :

Le texte mining, également appelé traitement automatique du langage, peut être défini comme étant un ensemble de techniques issues de l'intelligence artificielle, alliant plusieurs domaines : la linguistique, la sémantique, le langage, les statistiques et l'informatique, combinées ensemble, ces techniques permettent d'extraire des données pour recréer de l'information à partir de corpus de textes en les classifiant et les analysant de manière à établir des tendances.
Le texte mining est notamment beaucoup employé dans le secteur du marketing, mais également dans de nombreux autres domaines tels que la communication, les sciences politiques et la recherche.

Le texte mining, ou fouille de textes, respecte deux étapes principales :
La première est l'étape d'analyse, qui consiste à analyser les corpus de textes de manière à en reconnaître les mots, les phrases, les rôles grammaticaux ainsi que les relations et les sens de ces derniers entre eux.
Cette première étape, commune à tous les traitements, ne trouve sa pertinence que lorsqu'elle est couplée à la seconde étape : l'interprétation de l'analyse. Cette étape permet de sélectionner des textes en particuliers parmi d'autres.

Un exemple d'application concret de cette seconde étape étant la classification de courriers mails en spam, c'est-à-dire dans la catégorie des mails non sollicités, ou bien en non spam, c'est-à-dire en mails devant être lus par le destinataire.

Les outils de texte mining ont donc pour vocation d'automatiser la structuration de documents faiblement structurés, afin de générer de l'information sur le contenu d'un document texte, cette information n'étant alors pas présentée de manière explicite dans la forme initiale du document [16].

2.2.6.2 Traitement automatique du langage naturel (TALN) :

Le traitement automatique du langage naturel, abrégé en TALN, est une discipline s'appliquant au domaine de l'informatique et du langage. Il est utilisé par exemple pour les traductions, la reconnaissance vocale ou encore les réponses automatiques aux questions.

Ces domaines représentent des défis majeurs, car les mots du langage sont souvent traités un à un par l'ordinateur. Or, de nombreux mots sont polysémiques et recouvrent différentes réalités, en s'inscrivant notamment dans des contextes ou expressions qui peuvent changer complètement leur sens d'origine.

Grâce au traitement du langage naturel, une cohérence tente d'être apportée aux textes en s'attachant au sens des phrases et formules.
Ces avancées ne sont pas uniquement utiles pour les traducteurs mais aussi lorsque les ordinateurs exécutent des ordres oraux ou communiquent de manière vocale afin de faciliter par exemple la communication pour les personnes aveugles.

Pour pouvoir résumer des textes longs, ou extraire des informations précises, les ordinateurs ont besoin également de comprendre la cohérence linguistique des textes.
Qu'il s'agisse de traduction automatique ou d'une discussion les méthodes de TALN prêtent attention aux hiérarchies afin de mettre en cohérence les mots entre eux.

Le traitement automatique du langage naturel représente donc un défi colossal dans le domaine de l'informatique. Le langage peut en effet être à double sens et pour le comprendre, il est nécessaire de bien connaître le contexte dans lequel il s'insère. De nombreux utilisateurs ont d'ailleurs déjà fait l'expérience de conversations quelque peu chaotiques avec les chatbots qui sont fréquemment utilisés pour les chats des services clients. Toutefois, les ordinateurs comprennent de mieux en mieux le langage humain[17] .

2.2.6.3 Apprentissage automatique :

L'apprentissage automatique, également appelé apprentissage machine ou apprentissage artificiel et en anglais machine learning, est une forme d'intelligence artificielle

(IA) qui permet à un système d'apprendre à partir des données et non à l'aide d'une programmation explicite.
Cependant, l'apprentissage automatique n'est pas un processus simple.

Au fur et à mesure que les algorithmes ingèrent les données de formation, il devient possible de créer des modèles plus précis basés sur ces données.

Un modèle de machine learning est le résultat généré lorsque vous entraînez votre algorithme d'apprentissage automatique avec des données. Après la formation, lorsque vous fournissez des données en entrée à un modèle, vous recevez un résultat en sortie, par exemple, un algorithme prédictif crée un modèle prédictif, ensuite, lorsque vous fournissez des données au modèle prédictif, vous recevez une prévision qui est déterminée par les données qui ont servi à former le modèle[18].

2.3 Le coronavirus

La pandémie de COVID-19 qui a frappé le monde bat aujourd'hui son plein. Parallèlement aux actions menées par l'Organisation mondiale de santé (OMS) et ses partenaires pour riposter à cette pandémie une course aux vaccins a été engagée.

Les vaccins sauvent des millions de vies chaque année, le 18 février 2021, au moins sept vaccins différents avaient été mis à disposition dans les pays par l'intermédiaire de trois plateformes. La vaccination doit viser en priorité les populations vulnérables dans tous les pays.

2.3.1 Qu'est ce que le coronavirus ?

L'épidémie de maladie à coronavirus (COVID-19) a bouleversé la vie des gens dans le monde entier. Le COVID-19 est causé par le coronavirus 2 du syndrome respiratoire aigu sévère (SRAS-CoV-2), un nouvel agent pathogène humain qui, selon les virologues, a émergé des chauves-souris et a finalement sauté aux humains via un hôte intermédiaire

[19].
Les manifestations cliniques vont de symptômes légers ou inexistants à une maladie plus grave pouvant entraîner une insuffisance pulmonaire et même la mort [20].

Le 11 mars 2020, l'OMS a déclaré le COVID-19 pandémique [21]. Le 23 juin, l'OMS a signalé 8 993 659 cas confirmés de COVID-19 dans le monde et 469 587 décès [22], et les « Centers for Disease Control and Prevention » (CDC) ont signalé plus de 2 millions de cas confirmés aux États-Unis et plus de 120 000 décès[23].

2.3.2 Symptômes et évolution

Les manifestations des coronavirus font leur apparition moins de 24 heures après l'infection. Le plus généralement, le virus entraîne des maladies respiratoires légères à modérées comme le rhume avec des symptômes tels que :

- Maux de tête.
- Toux.
- Gorge irritée.
- Fièvre.
- Un sentiment général de malaise.
- Perte de goût et d'odorat.

Plus gravement, le nouveau coronavirus peut provoquer des maladies respiratoires comme la pneumonie ou la bronchite, particulièrement chez les personnes atteintes d'une maladie cardio-pulmonaire, chez celles dont le système immunitaire est affaibli et chez les personnes âgées.

2.3.3 Statistiques et évolution quotidienne du Covid-19

L'une des caractéristiques importantes des maladies infectieuses, en particulier celles causées par un nouveau pathogène, est leur gravité, dont la mesure ultime est leur capacité à entraîner la mort[22].

Voici quelques statistiques qui montrent le nombre de décès, nombre de personne infecté ainsi que le nombre de personne guéries du COVID-19 dans le monde.

Les statistiques représentées dans la figure 2.3 montrent le nombre de personnes infectées par le coronavirus COVID-19 dans le monde le 17 mai 2021. On observe que depuis la fin du mois de mars, ce sont les États-Unis qui comptabilisent le plus grand nombre de personnes contaminées au COVID-19 [24].

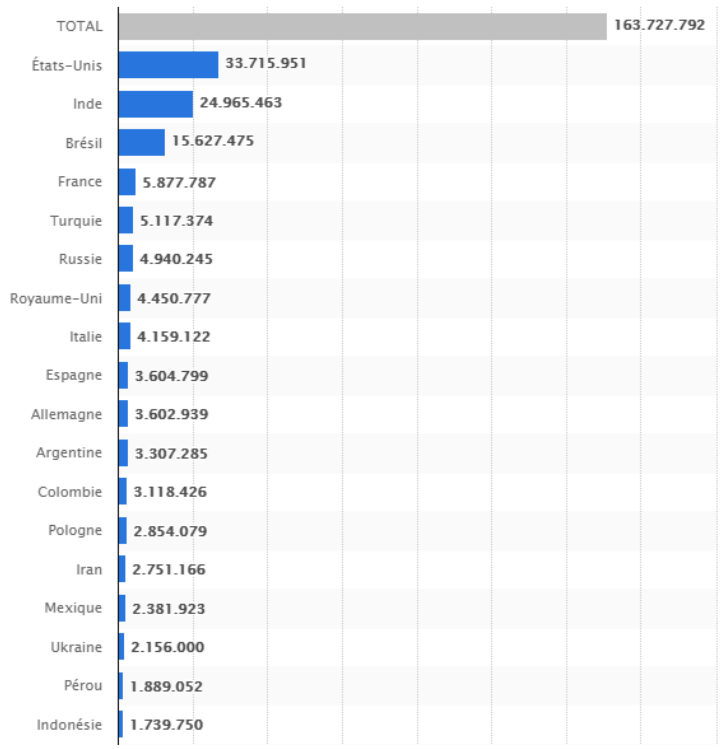


FIGURE 2.3 – nombre de personnes infectées par le COVID-19 [24].

Néanmoins, ce virus n'est pas systématiquement fatal pour les personnes contaminées : plusieurs cas de guérisons ont aussi été répertoriés. Les statistiques représentées dans la figure 2.4 présentent le nombre de personnes guéries du coronavirus (COVID-19) dans le monde le 17 mai 2021. On observe que sur un total de 5,9 millions personnes infectées en France, 5,1 millions ont été guéries [24].

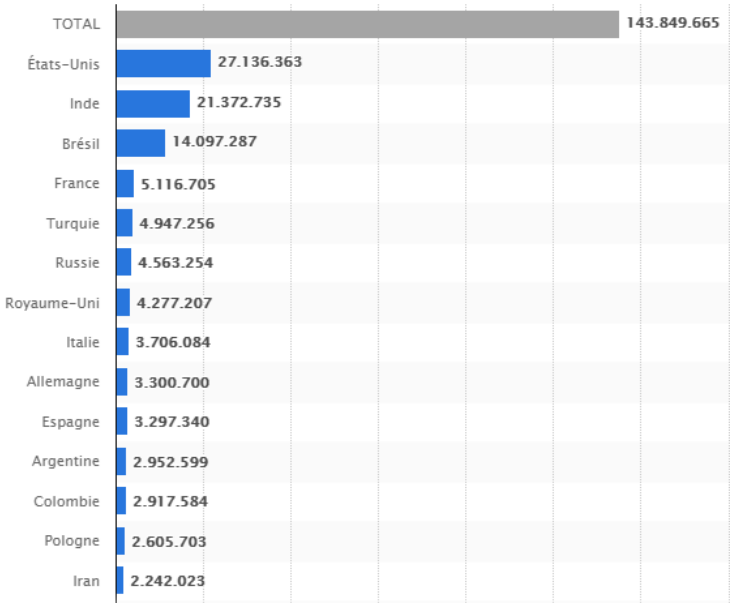


FIGURE 2.4 – nombre de personnes guéries du COVID-19 dans le monde [24] .

Sur un total de 163,7 millions d'infections liées au virus dans le monde, 3,4 millions sont à ce jour décédées. Avec plus d'un demi-million mort, ce sont les États-Unis qui dénombrent le plus de victimes. Voici donc les statistiques des 15 pays les plus touchés ainsi que le nombre de décès pour chacun d'eux.
Les statistiques représentées dans la figure 2.5 montrent le nombre de personnes décédées à cause du coronavirus (COVID-19) [24].

2.3.4 Les vaccins contre le coronavirus

Les vaccins sauvent des millions de vies chaque année. Leur mode d'action consiste à entraîner et à préparer le système immunitaire (défenses naturelles de l'organisme) à reconnaître et à combattre les virus et les bactéries qu'ils ciblent. Ainsi, si l'organisme se trouve par la suite exposé à ces mêmes agents pathogènes, il est immédiatement prêt à les détruire, ce qui permet de prévenir la maladie[3].

Le 18 février 2021, au moins sept vaccins différents avaient été mis à disposition

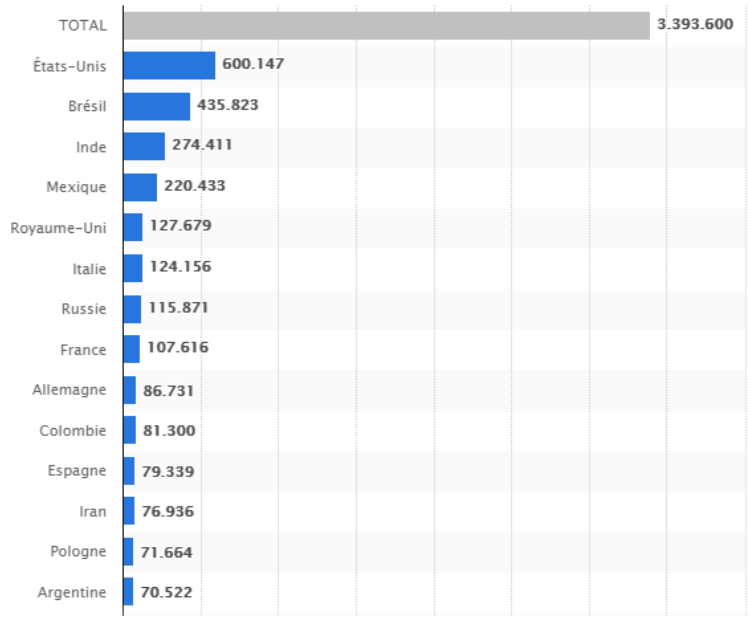


FIGURE 2.5 – nombre de décès due au COVID-19 dans le monde.
[24]

dans les pays par l'intermédiaire de trois plateformes. La vaccination doit viser en priorité les populations vulnérables dans tous les pays.

2.4 Analyse des sentiments dans les réseaux sociaux envers la vaccination contre le covid-2019

La disponibilité imminente des vaccins COVID-19 pose un besoin pressant de surveiller en permanence et de mieux comprendre les sentiments du public afin de développer des niveaux de base de confiance en eux parmi le grand public et de permettre l'identification des signaux d'alerte précoce de perte de confiance [25].
Cela aidera à répondre aux préoccupations des sceptiques en matière de vaccins[26, 27] et de développer la confiance requise du public dans la vaccination [28, 29]. pour atteindre l'objectif de l'immunité collective [30].

En effet, depuis la sortie des premiers vaccins contre le covid-19, les réseaux sociaux sont le premier lieu de débat quant à la l'efficacité et à la confiance que le public accorde à ce dernier. La préoccupation du public face à cela n'est généralement pas prouvé même

s'il y a eu des cas où les débats sur les vaccins ont été délibérément polarisés, exploitant ainsi les faiblesses du public et du système douteux à des fins politiques, tandis que la diminution de la confiance en matière de vaccins peut être influencée par une méfiance générale à l'égard du gouvernement et des élites scientifiques.

Pour comprendre l'attitude du public, les gouvernements utilisent des sondages, or, ceux-ci sont limités en termes de tailles d'échantillon et de granularité spatio-temporelle, de plus, les questions posées sont des questions fermées. Afin de surmonter ces limites, nous soutenons que les données issues des réseaux sociaux peuvent être utilisées pour obtenir davantage d'informations en temps réel sur les sentiments et les attitudes du public avec une granularité spatio-temporelle considérable.
C'est pour cette raison, qu'une analyse des sentiments envers la vaccination contre le covid-19 avec les données des réseaux sociaux est nécessaire pour comprendre les raison pour lesquelles la population est perplexe et hésitante par rapport aux vaccins conte le covid-19.

2.5 Conclusion

Dans ce chapitre, nous avons abordé le thème de l'analyse des sentiments et du Covid-19 . Nous avons également parlé de l'intérêt d'effectuer une analyse des sentiments sur les réseaux sociaux envers la vaccination contre le covid-19. Ainsi, nous avons défini les grands titres et le contenu général de notre mémoire.

Dans le prochain chapitre, nous allons élaborer un état de l'art des principales contributions relative au domaine de l'analyse des sentiments sur les réseaux sociaux envers la vaccination contre le coronavirus.

Chapitre 3

État de l'art

3.1 Introduction

Le coronavirus est un virus soupçonné de provenir des chauves-souris, avec la transmission du virus, il s'est donc répondu dans 188 pays et 25 territoires à travers le monde depuis novembre 2019 suscitant des tentatives d'élaboration d'anti-virus avec au moins 166 vaccins candidats depuis le début de l'épidémie.

Depuis l'apparition des vaccins, les premiers commentaires anti-vaccination ont vu le jour. L'utilisation des médias comme source d'informations (ex. Tweeter) ont permis de faire circuler les informations sur la santé, engendrant de la réticence à la vaccination en raison du contenu anti-vaccination largement disponible sur les réseaux sociaux. L'identification des commentaires anti-vaccination pourrait être utile dans le développement des stratégies pour réduire ces sentiments anti-vaccination.

L'utilisation croissante des réseaux sociaux comme source d'information sur la santé peut contribuer à l'hésitation de la population par rapport à la vaccination en raison du contenu anti-vaccination largement disponible sur ces derniers [31].

Un rapport a révélé qu'environ 31 millions de personnes suivaient les comptes Facebook des « antivaxxers » en 2019, et environ 17 millions de personnes étaient abonnés à des comptes similaires sur YouTube [32]. Depuis, le nombre de personnes suivant des

comptes "anti-vaxxer" sur les réseaux sociaux a augmenté d'au moins 7,8 millions de personnes.
Le rapport soulignait également que ceux qui ont reçu des informations sur la pandémie due au covid-19 à partir des réseaux sociaux étaient plus susceptibles d'être hésitants vis-à-vis du vaccin [32]. Une autre étude a révélé que la consommation du vaccin contre la grippe était inversement associé à l'utilisation de Twitter et de Facebook [33].

Dans le présent chapitre, nous allons présenter un état de l'art des principales contributions relatives à l'analyse des sentiments envers la vaccination contre le coronavirus. Cette dernière portera sur la capture des avis des internautes sur les réseaux sociaux. Cet état de l'art sera accompagné d'une étude comparative afin d'avoir une vue globale sur les différents travaux réalisés. Cependant les publications choisies sont différentes quant à l'objectif final de l'étude.

3.2 Travaux connexes

Un nouveau paradigme de développement de vaccin pandémique a été suggéré en Indonésie pour répondre au besoin immédiat d'un vaccin. Une équipe nationale a été créée par le Président de la République d'Indonésie pour accélérer la production de vaccins contre le COVID-19 et lance une campagne de vaccination pour lutter contre la pandémie. Les statistiques montrent que L'Indonésie compte 27 203 décès avec plus de 951 651 cas signalés.

[31] ont utilisé l'analyse des sentiments en explorant les données Twitter avec le mot-clé « COVID-19 », les Tweets indonésiens ont également été filtrés suivant les mots-clés : « vaccin » et «sinovac».

L'analyse des sentiments a été réalisée par les auteurs en prenant en compte les opinions exprimées sur Twitter concernant les événements ou les problèmes liés à la vaccination contre le coronavirus afin que la polarité des sentiments puisse être conclue à l'aide de l'algorithme Naïve Bayes.

La méthode Naïve Bayes est une méthode de classification pour l'exploration de

texte utilisée dans l'analyse des sentiments. Les auteurs affirment que cette approche est efficace en termes de cohérence des données et de classification des calculs, en particulier sur Twitter, en utilisant une variété de méthodes telles que Unigram Naïve Bayes, Multinomial Naïve Bayes et Maximum Entropy Classification. La principale caractéristique est d'obtenir une hypothèse forte de toute condition ou événement.

Après la collecte des données, les auteurs sont passés à la suppression des mots vides et à la méthode de tokenisation. Le processus d'étiquetage est effectué pour décider si un tweet appartient à une classe positive ou à une classe négative.
Une étude plus approfondie devrait utiliser divers algorithmes pour obtenir des résultats plus précis dans l'analyse des opinions.

[34] ont présenté une étude qui vise à évaluer les performances de différents modèles de traitement du langage naturel afin d'identifier les tweet anti-vaccination qui ont été publiés durant la pandémie COVID-19.
Cela consiste à comparer entre les performances des représentations d'encodeurs bidirectionnels à partir de transformateurs (BERT) et des réseaux de mémoire bidirectionnels à long terme avec des plongements Global Vectors(GLoVe) pré-entraînés Bi-LSTM(bidirectional long short-term memory) avec des méthodes d'apprentissage automatique classiques, y compris la machine à vecteurs de support (SVM) et le Naïf Bayes (NB).

Selon les auteurs, les résultats obtenus montrent que la performance du modèle BERT a surpassé les modèles Bi-LSTM, SVM et NB dans cette tâche. Ce qui laisse à conclure que BERT est le candidat idéal pour identifier les tweets anti-vaccination. Dans cette étude, une approche a été appliquée afin d'analyser les sentiments de la population indienne à l'égard de la pandémie COVID-19. Afin d'analyser les différents avis des internautes, une collecte de tweets comprise entre le 23 mars 2020 et le 15 juillet 2020 a été mise en place et les divers textes collectés ont été qualifiés de peur, de tristesse, de colère et de joie.

Une technique de traitement naturel (NLP) est appliquée et celle-ci implique plusieurs approches d'exploration de texte. Afin d'améliorer cette étude, l'analyse de ces données a été menée par le modèle BERT, qui est un nouveau modèle d'apprentissage en profondeur pour l'analyse des performances de texte.

Le modèle BERT a été comparé à trois autres modèles d'analyse telle que la régression logistique (LR), les machines vectorielles de support (SVM) et la mémoire à long-court terme (LSTM) Pour l'évaluation des performances du modèle, la précision a été considérée comme le paramètre principal pour cela la précision de chaque sentiment a été calculée séparément. Le modèle BERT a produit 89% de précision et les trois autres modèles ont produit respectivement 75%, 74,75% et 65%, d'après [34] ces résultats encouragent les gouvernements locaux à imposer des vérificateurs de faits sur les réseaux sociaux pour vaincre la fausse propagande.

Ce travail clarifie tout de même l'opinion publique sur les pandémies et oriente les autorités médicales, les travailleurs publics et privés pour surmonter l'anxiété inutile pendant les pandémies.

[35] ont proposé une approche afin d'analyser les sentiments des gens au Royaume-Uni et aux États-Unis à l'égard des vaccins covid-19.

Les gouvernements ont l'habitude d'utiliser des sondages afin de comprendre l'attitude du public mais étant donné que ces derniers sont limités, les auteurs ont utilisé les données non structurées présentes dans les réseaux sociaux et ils y ont appliqué des techniques établies d'intelligence artificielle (IA) telles que l'apprentissage automatique, et le traitement du langage naturel (NLP).

Mise à part l'analyse des sentiments, l'approche « détection de positions » a été utilisée, c'est une approche complémentaire, elle sert à attribuer une étiquette de position (favorable, contre et aucun) à un message sur une cible prédéterminée spécifique. Les informations proviennent de tweet et de publications facebook, l'analyse a été faite entre le 1er et le 22 novembre 2020, pour facebook les données ont été extraites via la plateforme CrowdTangle [1] et pour twitter via une api twitter, un processus de filtrage 1 thématique en 2 étapes a été appliqué avant le traitement et l'analyse.

L'ensemble de données filtrées a été initialement prétraité et un nouveau modèle d'IA basé sur un ensemble hybride hiérarchique a été développé pour l'analyse thématique des

1. CrowdTangle est un outil pour effectuer une veille sur les contenus des médias sociaux, surveiller son e-réputation et détecter facilement les sujets tendances qui peuvent être pertinents pour sa marque. Cette plateforme de social listening est gérée par Facebook.

sentiments.

Dans un premier temps, les résultats obtenus concernaient les tendances du sentiment temporel : les différentes polarités concernant le début de la campagne de vaccinations, l'annonce du vaccin Pfizer, le vaccin russe. . . , il a été noté que les sentiments positifs et négatifs étaient plus prononcés sur facebook que sur twitter, ainsi tous les pics dans les tendances positives et négatives dans les deux pays ont été mentionnés et enregistrés et une analyse statistique a été faite.
Ensuite, [35] ont proposé une cartographie géospatiale des sentiments moyens du public sur les réseaux sociaux aux États-Unis et au Royaume-Uni à l'égard des vaccins COVID-19.

Enfin l'obtention des sentiments moyens globaux dans une figure selon leurs polarités, les deux réseaux sociaux et les 2 pays étudiés.
Cependant, les auteurs affirment que l'approche proposée est contrainte aux sources de données : les données utilisées ne sont pas forcément représentatives, mais les limites techniques ont également été citées : la détermination de l'emplacement géographique des utilisateurs et les problèmes liés à l'exactitude des techniques d'IA.

L'analyse rétrospective de deux plateformes Facebook et Twitter démontre le potentiel de la surveillance des réseaux sociaux en temps réel activée par l'IA des sentiments et des attitudes du public pour aider à détecter et prévenir ces craintes et aussi pour permettre aux décideurs politiques de comprendre les raisons pour lesquelles certaines personnes peuvent hésiter à se faire vacciner contre le COVID-19, cette analyse peut servir à élaborer des politiques plus efficaces et à promouvoir un dialogue participatif sur des questions complexes de déploiement de vaccins.

Les auteurs [36]ont proposé une étude qui vise à évaluer les performances de différents modèles de traitement du langage naturel pour identifier les tweets anti-vaccination qui ont été publiés pendant la pandémie COVID-19 et à comparer les performances du modèle d'apprentissage non supervisé GLoVE (global vector for word representation) ainsi que le modèle de langage BERT(Bidirectional Encoder Representations from Transformers) aux méthodes d'apprentissage automatique classiques, notamment SVM (machines à vecteurs de support) et NB(Classification naïve bayésienne).

Les données entrées sont des tweets ont été collectés entre le 1er janvier et le 23 août 2020 à l'aide d'une API Twitter Stream qui permet au public d'accéder à un échantillon de 1% du flux quotidien de Twitter.
Les données ont subi plusieurs traitements puis ont été sélectionnées avec une méthode d'échantillonnage aléatoire systématique (20 854 tweets sélectionnés parmi 1 474 276 tweets à étiqueter). Les tweets ont été étiquetés comme « anti-vaccination » ou « autres ». Enfin les données ont ensuite été divisées en trois parties : ensemble de formation (70%), ensemble de développement (15%) et ensemble de tests (15%). Les ensembles de formation et de développement ont été utilisés pour construire le modèle, dont les performances ont été évaluées sur l'ensemble de tests.

Les auteurs ont utilisé les réseaux de neurones récurrents (RNN) dans de nombreuses tâches de traitement du langage naturel en raison de leur capacité à gérer des données séquentielles de différentes longueurs.

Les méthodes d'incorporation dynamiques telles que BERT qui produisent des représentations vectorielles pour les mots conditionnels au contexte de la phrase ont pallié aux problèmes de GloVe et word2vec.

Dans l'étude, le terme « frequency-inverse document frequency » a été utilisé pour vectoriser les données textuelles.
Les données représentent :

— Les performances des modèles Bi-LSTM-128 sur l'ensemble de développement .

— Les performances des modèles BERT sur l'ensemble de développement.

— Les performances des modèles SVM et NB sur l'ensemble de développement.

— Les performances entre Bi-LSTM, BERT, SVM et NB sur l'ensemble de test.

Pour conclure, l'étude effectuée par[36] a démontré que les modèles BERT ont surpassé les modèles Bi-LSTM, SVM et NB pour cette tâche. De plus, le modèle BERT a obtenu d'excellentes performances et peut être utilisé pour identifier les tweets anti-vaccination dans de futures études.

3.3 Analyse comparative

Dans cette section, nous allons présenter les approches et les méthodes citées par les auteurs des différentes contributions présentes dans l'étude bibliographique, premièrement en faisant une analyse comparative,ensuite en dressant un tableau comparatif.

3.3.1 Comparaison entre les différentes approches

[31] ont effectué une étude dont le but est d'analyser les sentiments de la population indonésienne à l'égard du vaccin contre le covid-19, il a été noté que les données sont cohérentes mais que l'analyse manque de précision, comparée à l'analyse effectuée par [35].

Cette dernière est également limitée à une zone géographique précise : Royaume Uni et USA, mais est meilleure en termes de précision et de représentation de résultats, cela grâce à l'utilisation de plusieurs techniques en plus des algorithmes de classification.

Néanmoins, le point négatif de l'analyse des sentiments de[35] concerne les sources de données qui, d'après les auteurs, ne sont pas assez représentatives en raison des lois de confidentialité de Twitter.

[34] , ont effectué une étude différente des deux autres citées précédemment, car elle a pour objectif d'analyser les tweets indiens concernant la vaccination contre le coronavirus en utilisant le modèle Bert (Bidirectionnel Encoder Representations from Transformers), mais elle vise également à évaluer les performances du nouveau modèle d'apprentissage en profondeur utilisé, et à le comparer à trois autres modèles différents.

Enfin le principal avantage de l'étude menée par[36] est l'utilisation des réseaux de neurones récurrents (RNN) dans le traitement du langage naturel. Une fois de plus, les auteurs affirment que les modèles BERT surpassent les autres modèles, ce qui a été également prouvé dans les contributions citées précédemment.

Notre approche se basera sur l'algorithme d'apprentissage automatique Naive Bayes car il est plus performant que d'autres modèles tels que : Random Forest, Support Vectors

Machine et XGBoost.
De plus, les données que nous allons utilisées ne sont pas restreintes à une zone géographique précise car les tweets concernant la vaccination contre le coronavirus proviennent du monde entier. la première étape de notre analyse est représentée par le tableau ci-dessous, ce dernier regroupe tous les travaux dont on s'est servis dans l'étude bibliographique et leurs principales caractéristiques.

3.3.2 Tableau comparatif

Nous avons procédé à une étude comparative des différentes contributions relatives à l'analyse des sentiments envers la vaccination contre le coronavirus.

Dans cette dernière, nous avons comparé entre les différents algorithmes, approches et modèles de classification utilisés dans l'analyse des sentiments, ainsi que la précision de chacun des modèles de classification. Cette comparaison discute également les résultats de l'analyse des sentiments : La polarité des sentiments exprimés sur les réseaux sociaux concernant la vaccination contre le virus du covid-19.
Le tableau contient des colonnes qui indiquent les critères de comparaison qui sont :

— Approche : désigne l'approche de chaque papier synthétisé.
— Données en entrée : désigne la source des données en entrée dans l'analyse des sentiments.
— Données en sortie : désigne le résultat de l'analyse des sentiments.
— Techniques utilisées : indique les méthodes et algorithmes utilisés.
— Avantages : présente les principaux avantages de l'étude.
— Inconvénient : indique les principaux inconvénients de l'étude.
— Outil logiciel : désigne si l'analyse des sentiments a été implémentée avec un outil logiciel.

Approche	input	output	Techniques utilisées	Avantages	Inconvénients	outils logiciels
[31]	Tweets indonésiens.	Polarité des sentiments des indonésiens concernant les vaccins contre le coronavirus : 39% de sentiments positifs, 56% de sentiments négatifs et 1% neutre.	- Algorithme naïf Bayes ; - Tokenisation ; - Etiquetage.	-Cohérence des données. - Classification des calculs.	-Manque de précision.	Oui
[34]	Tweets des internautes Indiens	Précision de chaque modèle dans l'analyse des tweets	-TLN ; -BERT ; - Régression logistique) ; -SVM) -LSTM.	-Evaluation et performance des modèles ; - Précision de chaque modèle.	-Analyse des sentiments limitée à une zone géographique restreinte.	oui
[35]	-Tweeter ; -Facebook.	-L'évolution des polarités de sentiments -cartographie géospatiale des sentiments moyens du public -figure représentant les sentiments moyens globaux	-Bert ; -Vader ; -TextBlob ; -NLP.	-Conforme aux lois sur la confidentialité et aux politiques des réseaux sociaux ;	-Sources de données ne sont pas forcément représentatives ; -problèmes liés à l'exactitude des techniques d'IA	oui
[36]	Ensemble de données Twitter collectées entre le 1er janvier et le 23 août 2020	Performances des modèles (sur l'ensemble de développement) : Bi-LSTM-128, BERT, SVM et NB. -Performances des modèles (sur l'ensemble de Test) : Bi-LSTM,BERT, SVM et NB.	- Méthode d'échantillonn age aléatoire systématique ; -Modèles : BERT,Bi-LST M, SVM et NB.	Évaluation des performances des modèles d'apprentissage automatique sur l'identification des tweets anti-vaccination.	Limite de réglages de certains paramètres des modèles.	oui

TABLE 3.1 – Tableau comparatif des différentes approches

3.4 Conclusion

Plusieurs études ont montré que les techniques les plus utilisées dans l'apprentissage supervisé pour les microblogs sont : la Classification Bayésienne et la machine à vecteur de support.

L'étude bibliographique faite dans de chapitre concerne des zones géographiques précises et les méthodes et algorithmes utilisés sont différents d'un document à un autre.

Dans le chapitre suivant, nous allons décrire, étape par étape, le processus de conception de notre application web grâce aux données de la plate-forme de microblogging Twitter.

Chapitre 4

Méthodologie

4.1 Introduction

L'analyse des sentiments à partir de flux de données vise à détecter l'attitude, les émotions et les opinions des auteurs à partir de textes en temps réel. Dans notre cas, elle permet aux médecins, aux chercheurs et aux autorités d'identifier l'opinion de la population vis-à-vis de la vaccination contre le coronavirus.

Au fil du temps, divers points de vue liés à la vaccination ont fait l'objet de discussions sur des plateformes de réseaux sociaux telles que Twitter et Facebook.

Dans ce chapitre, nous allons parler de la méthodologie utilisée dans l'extraction et le traitement des tweets afin que l'on puisse les classer selon la polarité des sentiments exprimés.

4.2 Méthodologie

La pandémie du COVID-19 affecte des millions de vies dans le monde entier et en tant que problème majeur qui touche à la santé des habitants du monde entier, nous devons trouver des solutions efficaces certes mais également des solutions fiables qui rassureront les populations.

Les scientifiques et les chercheurs ont redoublé d'efforts depuis le début de la pandémie dans le but de créer un vaccin efficace qui puisse être distribué de manière égale

et large, en effet, comme aujourd'hui, il existe plusieurs vaccins dont : Pfizer, Moderna, AstraZeneca, Janssen, depuis le début des tests et de la commercialisation de ces derniers, les avis des gens les concernant divergent sur les réseaux sociaux, certains sont pro-vaccination et d'autres sont anti-vaccination.

L'objectif de notre travail est de développer et d'appliquer une approche basée sur un algorithme de classification pour analyser les sentiments du public sur les réseaux sociaux dans le monde entier envers les vaccins COVID-19 afin de mieux comprendre l'attitude et les préoccupations du public concernant les vaccins COVID-19. Notre étude concerne un ensemble de données (dataset) qui regroupe des tweets réalisés avec le hashtag covidVaccine dont la collecte a commencé le 14/02/2021 et a été mise à jour le 19-06-2021.

Afin de mener à bien notre étude, nous avons suivi un processus d'analyse des données pour déterminer la population pro-vaccination et la population anti-vaccination.

La figure 4.1 représente la méthodologie à suivre pour notre étude :

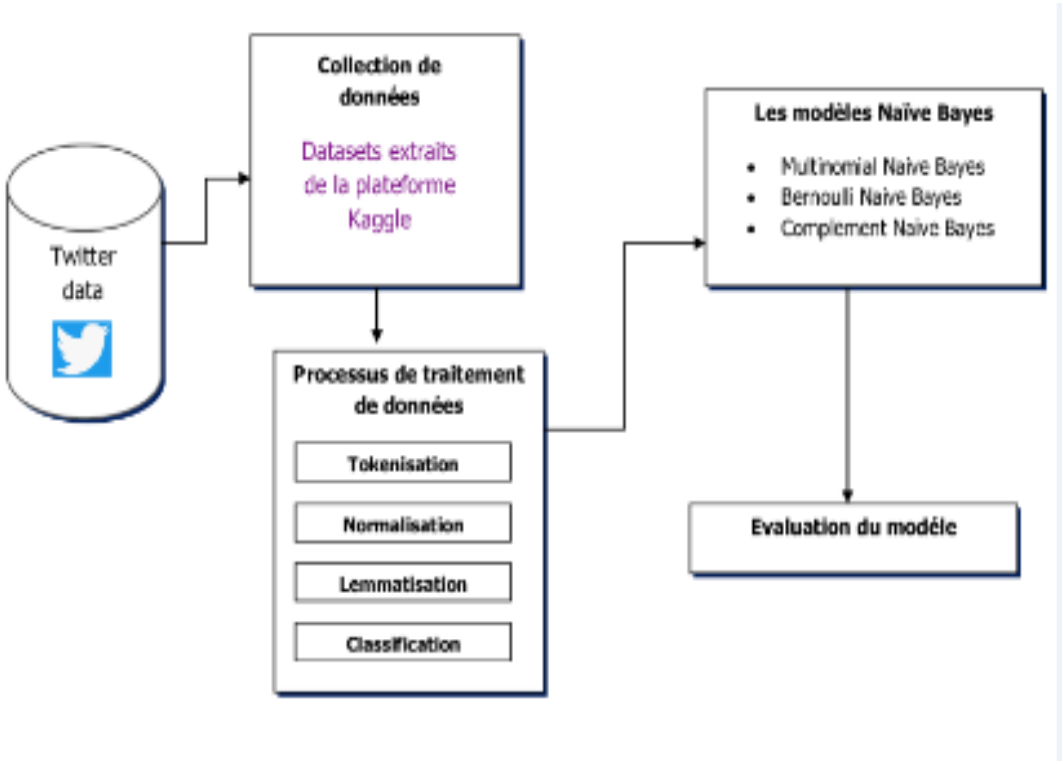


FIGURE 4.1 – Méthodologie appliquée à la conception du système

Dans ce qui suit,nous exposerons l'approche que nous avons adopté pour la mise en oeuvre de nos algorithmes d'analyse de sentiments,nous allons ainsi détailler les étapes de notre système. Cette approche est illustrée dans la figure 4.1 et comprend principalement quatre étapes. La première étape sert à rassembler les données nécessaires à l'analyse, ensuite les quatre étapes du processus de traitement des données(Tokenisation,normalisation, lemmatisation et classification) servent à traiter les données collectées, enfin la dernière étape "Evaluation du modéle" sert à évaluer la précision des algorithmes utilisés.

4.2.1 Collecte des données

La première étape consiste à collecter les données avec lesquelles nous allons faire notre étude d'analyse des sentiments.
Les données(dataset)[37] sont des tweets récents sur les vaccins COVID-19 utilisés dans le monde entier à grande échelle. Dans le tableau 4.2.1, nous avons listé tous les vaccins mentionnés dans le dataset, le pays et l'année ou les vaccins ont été crées.

Vaccin	***Pays***	***Date de création***
Pfizer/BioNtech	Allemagne/Etats-Unis	17 mars 2020
Sinopharm	Chine	07 mai 2021
Sinovac	Chine	Avril 2021
Moderna	Etat-unis	1er trimestre 2021
Oxford/AstraZeneca	Angleterre	mi-janvier 2021
Covaxine	Inde	3 Janvier 2021
Spoutnik V	Russie	Décembre 2020

TABLE 4.1 – Les différents vaccins mentionnés dans le dataset

La base de l'analyse des sentiments qu'effectue notre système est la source de donnée que ce dernier utilise.

Les données sont collectées à l'aide du package Python tweepy pour accéder à l'API Twitter. Pour chacun des vaccins, j'utilise le terme de recherche pertinent (le plus souvent utilisé sur Twitter pour désigner le vaccin respectif). Le dataset a été crée le 14-02-2021 [37].

Les données initiales ont été fusionnées à partir de tweets sur le vaccin Pfizer/BioNTech.ensuite

des tweets ont été ajoutés : concernant Sinopharm, Sinovac (tous deux vaccins fabriqués en Chine), Moderna, Oxford/Astra-Zeneca, Covaxin et Sputnik V. La collecte était dans les premiers jours deux fois par jour, jusqu'à ce que le créateur identifie approximativement le nouveau quota de tweets puis la collecte (pour tous les vaccins) s'est stabilisée à une fois par jour, pendant les heures du matin (GMT)[37].
Une fois la collecte de données faite, nous avons supprimé toutes les colonnes du dataset afin de garder uniquement le texte des tweets.

La figure 4.2 est un aperçu (les 5 premières lignes) du dataset que nous avons utilisé.

id	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text	hashtags	source	retweets
1340539111971516416	Rachel Roh	La Crescenta-Montrose, CA	Aggregator of Asian American news; scanning di...	2009-04-08 17:52:46	405	1692	3247	False	2020-12-20 06:06:44	Same folks said daikon paste could treat a cyt...	['PfizerBioNTech']	Twitter for Android	0
1338158543359250433	Albert Fong	San Francisco, CA	Marketing dude, tech geek, heavy metal & '80s ...	2009-09-21 15:27:30	834	666	178	False	2020-12-13 16:27:13	While the world has been on the wrong side of ...	NaN	Twitter Web App	1
1337858199140118533	eli.treu	Your Bed	heil, hydra	2020-06-25 23:30:28	10	88	155	False	2020-12-12 20:33:45	#coronavirus #SputnikV #AstraZeneca #PfizerBio...	['coronavirus', 'SputnikV', 'AstraZeneca', 'Pf...	Twitter for Android	0
1337855739918835717	Charles Adler	Vancouver, BC - Canada	Hosting "CharlesAdlerTonight" Global News Radi...	2008-09-10 11:28:53	49165	3933	21853	True	2020-12-12 20:23:59	Facts are immutable, Senator, even when you're...	NaN	Twitter Web App	446
1337854064604966912	Citizen News Channel	NaN	Citizen News Channel bringing you an alternati...	2020-04-23 17:58:42	152	580	1473	False	2020-12-12 20:17:19	Explain to me again why we need a vaccine @Bor...	['whereareallthesickpeople', 'PfizerBioNTech']	Twitter for iPhone	0

FIGURE 4.2 – Aperçu du dataset utilisé

Dans ce qui suit, nous expliquerons chaque étape du processus de traitement dt texte que subissent les données collectées. Dans le but d'illustrer ces dernières, nous prenons le tweet publié le 13/07/2021 par le compte officiel @Reuters_Health, présenté dans la figure 4.3.

Pour cet exemple, le texte du tweet subira premièrement le processus de tokenisation illustré dans le tableau 4.2, ensuite le processus de normalisation, illustré dans le tableau 4.3 et enfin le processus de Lemmatisation illustré dans le tableau 4.4.

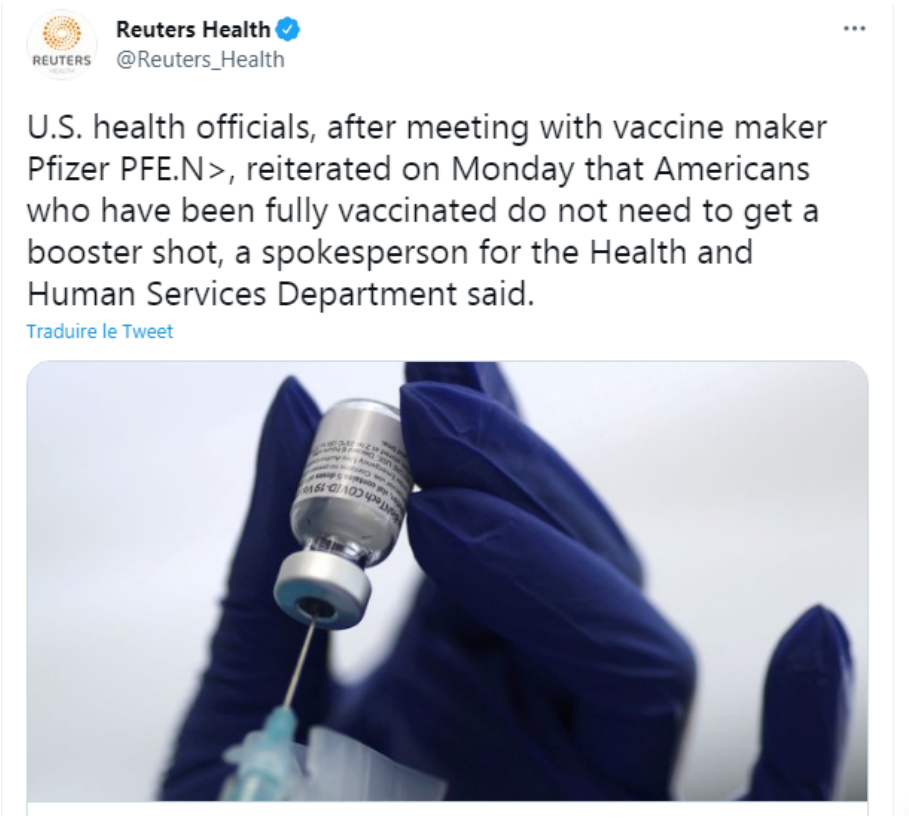


FIGURE 4.3 – Aperçu d'un tweet

4.2.2 Tokenisation

La tokenisation est utilisée pour séparer la séquence de chaînes des Tweets en morceaux tels que des mots, des phrases et des symboles. Cette étape est importante pour séparer le Gestionnaire Twitter des comptes et des liens du Tweet. Après la séparation, le gestionnaire Twitter sera supprimé[38].

C'est la première étape pour transformer des données non structurées en données structurées, plus faciles à analyser[3].

Cette première phase de lissage, de correction, de nettoyage, des données textuelles laisse la place à une seconde série d'opérations qui vise à définir l'objet de l'analyse : c'est la tokenisation des documents qui consiste à identifier les unités de textes élémentaires qui peuvent être des mots, mais aussi des lettres, des syllabes, des phrases, ou des séquences de ces éléments. Chaque document devient alors une liste ordonnée (ou non) de termes élémentaires : les tokens. Nous passons d'un plan de données qui liste des documents (et les associe éventuellement à des auteurs), à un plan qui associe un document à une série

d'attributs qui sont ses éléments unitaires (lettres, mots, syllabes, phrases)[39]. Il existe deux types de tokenisation :

Tokenization par mot : Les mots sont comme les atomes du langage naturel. Ils sont la plus petite unité de sens. La tokenisation du texte par mot permet d'identifier les mots qui reviennent particulièrement souvent.

Tokenisation par phrase : lorsque la phrase est segmentée, on analyse la relation entre les mots qu'elle contient afin d'avoir le contexte.

Le tableau 4.2 représente le texte du tweet illustré dans la figure 4.3 avant et après l'étape de tokenisation.

Tweet avant tokenization	tweet après tokenization
U.S. health officials, after meeting with vaccine maker Pfizer PFE.N>, reiterated on Monday that Americans who have been fully vaccinated do not need to get a booster shot, a spokesperson for the Health and Human Services Department said.	'U','.','S','.','health', 'officials',',', 'after', 'meeting', 'with', 'vaccine', 'maker', 'Pfizer', 'PFE','.','N','>', 'reiterated', 'on', 'Monday', 'that', 'Americans', 'who', 'have','been' 'fully', 'vaccinated', 'do', 'not', 'need', 'to', 'get', 'a' ,'booster', 'shot', 'a', 'spokesperson', 'for', 'the', 'Health', 'and', 'Human', 'Services', 'Department', 'said'.

TABLE 4.2 – Tweet avant et aprés la tokenisation

4.2.3 Normalisation

Les mots vides, les ponctuations, les lettres majuscules sont supprimés lors de la normalisation. Les mots vides sont des mots couramment utilisés("le", "un", "un", "dans"). En regardant les tweets, la possibilité de la gestion des hashtags est grande, ce qui explique donc pourquoi la normalisation est en effet importante dans cette procédure. La plupart des ponctuations et des mots vides sont supprimés, ce qui laisse quelques mots pour une lemmatisation ultérieure[38].

Dans le tableau 4.3, un exemple du tweet précédent après la normalisation :

Tweet avant normalisation	*Tweet aprés normalisation*
U’,’.’,’S’,’.’,’health’, ‘officials’, ’,’,’, ‘after’,‘meeting’, ‘with’, ‘vaccine’, ‘maker’, ‘Pfizer’,‘PFE’,’.’,’N’,’>’, ‘reiterated’, ‘on’, ‘Monday, ‘that’, ‘Americans’, ‘who’, ‘have’,’been’ ‘fully’,‘vaccinated’, ‘do’, ‘not’, ‘need’, ‘to’, ‘get’,‘a’ ,’booster’, ‘shot’, ‘a’, ‘spokesperson’, ‘for’, ‘the’, ‘Health’, ‘and’, ‘Human’, ‘Services’, ‘Department’, ‘said’.	’health’, ‘officials’,‘meeting’, ‘vaccine’,‘maker’,‘Pfizer’, ‘PFE’, ‘reiterated’, ‘Monday, ‘Americans’, ‘vaccinated’, ‘need’, ‘get’,’booster’, ‘shot’, ‘spokesperson’, ‘Health’, ‘Human’, ‘Services’, ‘Department’,‘said’.

TABLE 4.3 – Tweet avant et aprés la normalisation

Le **stemming** et la **lemmatisation** sont deux techniques pour normaliser les jetons. En anglais et dans la plupart des langues, un même mot peut avoir plusieurs formes selon le contexte dans lequel il est utilisé. Lorsque l’on symbolise un texte, chacune de ces formes est considérée comme différente, mais en tant qu’utilisateur de la langue, nous savons que toutes les formes se réfèrent à un seul concept ou à une seule idée.

4.2.3.1 Stemming

Le stemming est une tâche de traitement de texte dans laquelle on réduit les mots à leur racine, qui est la partie centrale d’un mot. Il met le focus sur le sens de base d’un mot plutôt que sur tous les détails de son utilisation par exemple les mots « automatiser », « automatique », « automatisme », « automatismes » seront réduits à « automatisme » de manière à ce que toutes ces formes fassent référence à un seul jeton : « automatisme »[40].

4.2.3.2 Lemmatisation

La lemmatisation, c’est où les mots sont générés jusqu’à leur racine . Contrairement au stemming, la lemmatisation préserve la partie du discours sans diviser les suffixes[41]. Par exemple, le mot « caring» en anglais se transforme en « care » avec la lemmatisation contrairement au stemming qui le transforme en ‘car’.

Enfin le stemming peut être utilisé lorsque le radical est constant dans toutes les formes possibles, mais lorsque le radical n'est pas constant dans toutes les formes du mot, la lemmatisation est la meilleure option. La lemmatisation vise à obtenir une "tige" de base similaire pour un mot, mais vise à dériver le véritable mot racine du dictionnaire, pas seulement une version tronquée du mot. Par exemple, « était », « étaient », « est », « sont » sera lemmatisé en « être ».

Pour le traitement de nos données, nous avons utilisé la lemmatisation.
Dans le tableau 4.4, le tweet cité précédemment après avoir subi le processus de Lemmatization

Tweet avant Lemmatisation	Tweet aprés Lemmatisation
'health', 'officials','meeting', 'vaccine','maker','Pfizer', 'PFE', 'reiterated', 'Monday, 'Americans', 'vaccinated','need', 'get','booster', 'shot', 'spokesperson', 'Health', 'Human', 'Services', 'Department','said'.	'health', 'official','meet', 'vaccine','make','Pfizer', 'PFE', 'reiterate', 'Monday, 'American', 'vaccinate','need', 'get','booster', 'shot', 'spokesperson', 'Health', 'Human', 'Service', 'Department','say'.

TABLE 4.4 – Tweet avant et aprés la lemmatisation

4.2.4 Vectorisation

La vectorisation c'est créer une matrice clairsemée de tous les mots et du nombre de fois où ils sont présents dans un document avec soit un vectoriseur de comptage ou un vectoriseur TF-IDF [42].

Comme indiqué ci-dessus, la vectorisation est le processus de conversion de texte en entrées numériques sous forme matricielle.
Dans la technique de vectorisation, une matrice de termes de document est générée où chaque cellule indique le nombre de fois qu'un mot apparaît dans un document, également connu sous le nom de fréquence de terme.

La **matrice de termes du document** est un ensemble de variables fictives qui indiquent si un mot particulier apparaît dans le document. Une colonne est dédiée à chaque mot du corpus. Le décompte est directement proportionnel à la corrélation de la

catégorie du titre de l'actualité. Cela signifie que si un mot particulier apparaît plusieurs fois dans de faux titres d'actualités ou de vrais titres d'actualités, alors le mot particulier a un pouvoir prédictif élevé pour déterminer si le titre d'actualités est faux ou réel [2]

4.3 Modéle de classification

La catégorisation de texte sert à déterminer si un document appartient à une série de documents de classe pré-spécifiés. Le système de classification automatique peut grandement favoriser le processus de classification.

Parallèlement à la croissance rapide de l'information sur Internet, la classification des textes est une tendance générale et importante dans le domaine de la recherche d'informations. La plupart des approches traitant des problèmes de classification de texte ont été proposées pour améliorer la précision des classificateurs de texte. Pour gérer les tâches de classification de texte, les documents sont caractérisés par des mots apparaissant dans le texte.

Une technique consiste à utiliser l'apprentissage automatique pour classer les documents et traiter l'absence de chaque mot comme un attribut logique comme dans l'un des modèles statistiques initiaux du langage : le modèle multivarié de Bernoulli Naive Bayes (BNB) . BNB est bien pensé mais ne se soucie que de l'apparence des mots qui en font une référence pour la classification des textes. En BNB, quand le mot apparaît dans le document, la valeur de l'attribut équivalente à ce dernier est soit 1 sinon 0. Comme méthode améliorée de BNB, c'est la méthode multinomial Naive Bayes (MNB) qui a été proposée. MNB prend le document comme un sac de mots et prend en compte la fréquence des mots et des informations. Afin de surmonter les problèmes du système auxquels MNB est confronté, c'est le Complément Naive Bayes (CNB) qui a été proposé [43].

De nos jours, les modèles multinomiaux sont considérés comme l'approche de modélisation dominante et ils sont plus efficaces que le modèle de Bernoulli multivarié qui introduit la modélisation du langage dans la recherche d'informations. On constate que les modèles multivariés de Bernoulli sont significativement meilleurs que le modèle multinomial dans les tâches de recherche de phrases.

Les classificateurs de Naive Bayes (NB) sont une famille de classificateurs basés sur le théorème de probabilité populaire de Bayes. Connu pour créer des modèles simples et puissants, en particulier dans les domaines de la classification des documents et de la prédiction des maladies. La classification textuelle de NB est le plus souvent utilisée pour catégoriser le texte car elle est rapide et facile à mettre en œuvre.

Fondamentalement, l'algorithme NB est un algorithme d'apprentissage automatique. Il est principalement utilisé pour catégoriser le texte, y compris les ensembles de données d'entraînement multidimensionnel. Certains exemples sont connus pour la classification de documents, le filtrage de portée et surtout, l'analyse des sentiments. L'algorithme NB est appelé « naïf » car il suppose que l'apparence d'une caractéristique n'a aucun rapport avec l'apparence d'autres caractéristiques.

Sélectionner le classificateur Naive Bayes (NBC) est plus souhaitable en raison de sa grande vitesse. Il est même suggéré d'utiliser NB, plutôt que d'autres algorithmes, pour cette taille de problème, car ils ont une implémentation parallèle de map-reduce.

Le Naive Bayes Classifier(NBV) a d'excellents résultats pour l'analyse de données textuelles. c'est-à-dire le traitement du langage naturel. Il nécessite des connaissances mathématiques en probabilité conditionnelle et théorème de Bayes, mais c'est un concept extrêmement simple et « naïf » [2].

Supposons que nous ayons des données d'entrée sur les caractéristiques du temps (perspectives, température, humidité, vent) et si vous pouvez jouer au golf ou non :

perspectives	**Temperature**	**Humidité**	**venteux**	**jouer**
Soleil	Chaleur	Élevée	Faux	Non
Soleil	Chaleur	Élevée	Vrai	Non
Nuageux	Chaleur	Élevée	Faux	Oui
pluie	Douce	élevée	Faux	Oui
pluie	Fraîche	Normale	Faux	Oui
pluie	fraîche	Normale	Vrai	Non
Nuageux	Fraîche	Normale	Vrai	Oui
Soleil	Douce	Élevée	Faux	Non
Soleil	Fraîche	Normale	Faux	Non
Pluie	Douce	Normale	Faux	Non
Soleil	Douce	Normale	Vrai	Oui
Nuageux	Douce	Élevée	vrai	Oui
Nuageux	Chaleur	Normale	Faux	Oui
Pluie	Douce	Élevée	Vrai	Non

TABLE 4.5 – Exemple de caractéristique du temps[2].

Ce que fait essentiellement **Naive Bayes** , c'est comparer la proportion entre chaque variable d'entrée et les catégories dans la variable de sortie. Cela peut être montré dans le tableau 4.6.

Perspectives			**Température**			**Humidité**			**Venteux**			**Jouer**	
	Oui	Non		Oui	Non		Oui	Non		Oui	Non	Oui	Non
soleil	2/9	3/5	Chaleur	2/9	2/5	Elevé	3/9	3/5	Faux	6/9	2/5	9/14	7/14
Nuage	4/9	0/5	Douce	4/9	2.5	Normale	6/9	1/5	Vrai	3/9	3/5		
Pluie	3/9	2/5	Fraiche	3/9	1/5								

TABLE 4.6 – Comparaison faite par l'algorithme de Naive Bayes[2].

Pour donner un exemple qui aide à lire ceci, dans la section température, il faisait chaud pendant deux jours sur les neuf jours où vous avez joué au golf (c'est-à-dire oui). La probabilité est essentielle pour comprendre le reste. Une fois que vous avez cela, vous pouvez prédire si vous jouez au golf ou non pour toute combinaison de caractéristiques météorologiques.

Imaginez que nous ayons un nouveau jour avec les caractéristiques suivantes : Perspectives : ensoleillées,Température : douce,Humidité : normale, Venteux : faux [2] .

Tout d'abord, nous allons calculer la probabilité que vous jouiez au golf étant donné X, P(oui|X) suivi de la probabilité que vous ne jouez pas au golf étant donné X, P(no|X).

En utilisant le tableau 4.6, nous pouvons obtenir les informations suivantes :

P(oui)= 9/14

p(perspectives=ensoleillée | *oui*) = 2/9

$p(temprature = douce \mid oui) = 4/9$

$p(humidit = normale \mid oui) = 6/9$

$p(venteux = faux \mid oui) = 6/9$

On peut écrire ces informations sous cette forme :

$p(oui \mid X)(X \mid y) \times p(y)$

$p(oui \mid X) \times p(x1 \mid y) \times p(x2 \mid y) \times p(x3 \mid y) \times p(x4 \mid y) \times p(x \mid y)$

$p(oui \mid X) \propto p(ensoleill|oui) \times p(douce \mid oui) \times p(normale \mid oui) \times p(faux \mid oui) \times p(oui)$

$p(oui \mid x) \propto 2/9 \times 4/9 \times 6/9 \times 9/14$

De même, vous effectueriez la même séquence d'étapes pour

$P(non \mid X)$.

$P(non \mid X) \propto 0.0069$

$PuisqueP(oui \mid x) > P(non \mid x)$,

alors vous pouvez prédire que cette personne jouerait au golf vu les perspectives.

Nous avons choisi le classificateur de naïve bayes pour classer les données que nous avons traitées et à la fin de notre travail nous avons évalué chacun de ces modèles cités précédemment, entre autres : MNB, BNB, CNB [2].

4.4 Conclusion

Dans ce chapitre, nous avons présenté la méthodologie que nous avons suivie pour l'analyse des sentiments. Notre système passe par plusieurs étapes, à savoir le traitement de données et la classification.

Nous avons défini ce qu'est l'algorithme de Naive Bayes et pourquoi nous l'avons choisi pour l'analyse des sentiments des tweets concernant les vaccins contre le COVID-19.

Le prochain chapitre portera sur les outils utilisés pour l'analyse des sentiments, et sur l'environnement de développement utilisé pour l'application web que nous avons mise en œuvre dans le but d'afficher le résultat de l'analyse.

Chapitre 5

Expérimentation

5.1 Introduction

Après avoir effectué une analyse détaillée de nos données, nous allons à présent entamer l'implémentation de l'ontologie de notre application en mettant en place un corpus de connaissances utilisable et lisible par les utilisateurs.

Nous présenterons alors, au cours de ce chapitre :

— Présentation du dataset,

— Environnement et outils de développement,

de données,

— Évaluation.

5.2 Présentation du dataset

Dataset : est un ensemble cohérent de données produites dans le cadre d'un même projet, sur un même objet d'étude et/ou recueillies sur un même lieu. Toutes les données d'un dataset peuvent donc être décrites avec une majorité de métadonnées communes et peuvent être composées d'une centaine de fichiers, sa taille pouvant atteindre quelques dizaines de giga-octets.

Elles sont représentées avec différentes structures, on peut avoir une structure tabulaire, par exemple un fichier CSV, une structure d'arbre, comme dans un fichier JSON

ou XML, ou encore une structure de graphe, comme dans le RDF. Lorsque les données sont tabulaires, en principe, chaque ligne correspond à une observation et chaque colonne à une variable.

Le dataset utilisé est extrait du site kaggle représenté sous format CSV, plus adapté pour le traitement et l'analyse de sentiment sous python. Il est composé de 88978 avis concernant la vaccination au COVID-19 réalisé avec l'hashtag covidVaccine dont la collecte à commencé le 14/02/2021 et a été mise à jour le 19/06/2021.
La collecte des données initiales a été fusionnées à partir de tweets sur le vaccin Pfizer/BioNTech, des tweets de Sinopharm, Sinovac (tous deux vaccins fabriqués en Chine), Moderna, Oxford/Astra-Zeneca, Covaxin et Sputnik V ont été ajoutés. La collecte de données se faisait deux fois par jour au début, jusqu'à l'identification du nouveau quota de tweets, puis la collecte des tweets concernant tous les vaccins s'est stabilisée à une fréquence d'une fois par jour, durant la matinée (GMT) [37].

5.3 Outils et environnement de développement

Avant d'entamer l'implémentation de l'architecture de notre application, on a choisi un ensemble d'outils qui peuvent répondre aux exigences de développement de notre système en vue des possibilités et des avantages qu'ils offrent.

5.3.1 Environnements de développements

Dans cette section nous allons définir tous les logiciels que nous avons utilisé pour le développement de notre système.

5.3.1.1 Anaconda

Anaconda est une plate-forme open source qui rassemble des meilleurs d'outils pour les professionnels de la science des données avec plus de 100 packages populaires prenant en charge les langages Python, Scala et R [44], visant à simplifier la gestion des paquets et de déploiement.

La figure 5.1 ci-dessous est une captures de l'interface graphique d'Anaconda :

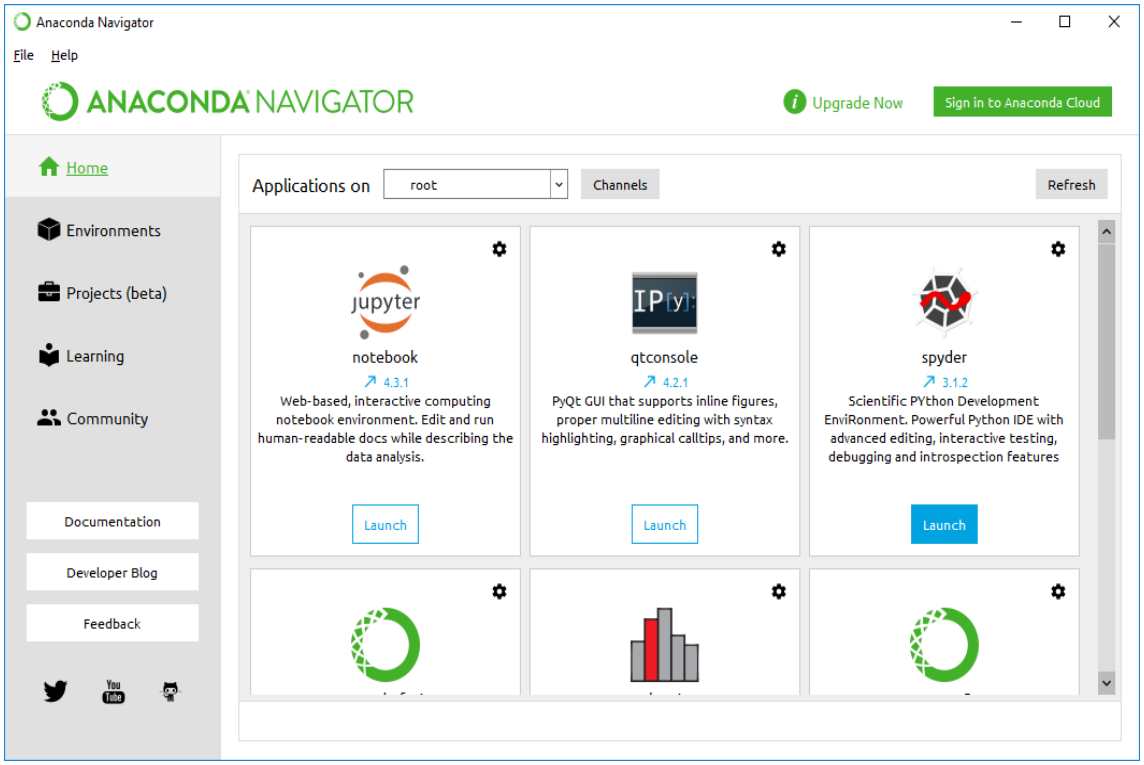


FIGURE 5.1 – capture d'écran de l'environnement Anaconda.

5.3.1.2 Jupyter Notebook

est une application Web open source qui vous permet de créer et de partager des documents contenant du code en direct, des équations, des visualisations et du texte narratif. Les utilisations incluent : le nettoyage et la transformation des données, la simulation numérique, la modélisation statistique, la visualisation des données, l'apprentissage automatique et bien plus encore [45].

La figure ci-dessous représente l'environnement de Jupyter notebook .

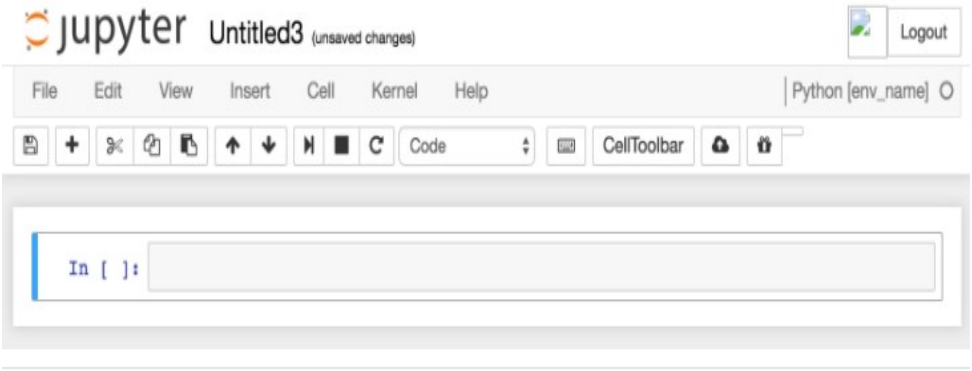


FIGURE 5.2 – capture d'écran de l'environnement Jupyter.

5.3.2 Outils de développements

Python est un langage de programmation open source créé par le programmeur Guido Van Rossum en 1991, il s'agit d'un langage puissant et facile à apprendre efficace pour la programmation orientée objet. Il possède des structures de données de haut niveau efficaces, syntaxe élégante, le typage dynamique, ainsi qu'une nature interprétée. En effet, c'est un langage idéal pour les scripts et le développement rapide d'applications dans de nombreux domaines sur la plupart des plates-formes notamment dans le domaine du machine learning. L'interpréteur Python est facilement étendu avec de nouvelles fonctions et types de données implémentés en C ou C++, python convient également comme langage d'extension pour les applications personnalisables [46].

HTML (HyperText Markup Language) désigne un type de langage informatique descriptif. Il s'agit plus précisément d'un format de données utilisé dans l'univers d'Internet pour la mise en forme des pages Web. Il permet, entre autres, d'écrire de l'hypertexte, mais aussi d'introduire des ressources multimédias dans un contenu.
Développé par le W3C (World Wide Web Consortium) et le WHATWG (Web Hypertext Application Technology Working Group), le format ou langage HTML est apparu dans les années 1990. Il a progressivement subi des modifications et propose depuis 2014 une version HTML5 plus aboutie.
HTML est ce qui permet à un créateur de sites Web de gérer la manière dont le contenu de ses pages Web va s'afficher sur un écran, via le navigateur. Il repose sur un système de balises permettant de titrer, sous-titrer, mettre en gras, etc., du texte et d'introduire des éléments interactifs comme des images, des liens, des vidéos... Il est plus facilement

compris des robots de crawl des moteurs de recherche que le langage JavaScript, aussi utilisé pour rendre les pages plus interactives [47].

VOILA : voila est une bibliothèque python open source qui est utilisée pour transformer le notebook jupyter en une application web autonome. Il prend en charge les widgets pour créer des tableaux de bord interactifs, des rapports, etc.
Voila lance un noyau lorsqu'il est connecté à un notebook et exécute toutes les cellules mais il n'arrête pas le noyau là pour que l'utilisateur puisse interagir avec la sortie. Il convertit le bloc-notes jupyter en HTML et le renvoie à l'utilisateur sous forme de tableau de bord ou de rapport avec toutes les entrées exclues et les sorties incluses.
Voila prend en charge toutes les bibliothèques python pour les widgets tels que bqplot, plotly, ipywidgets, etc. Enfin, il est indépendant du framework et du langage, ce qui signifie q'il peut fonctionner avec n'importe quel noyau jupyter, que ce soit C++ ou Python, car il est construit sur les protocoles standard jupyter[48].

5.3.3 Bibliothèques utilisées

NLTK (Natural Language Toolkit) : est une suite de modules de programmes open source, de didacticiels et d'ensembles de problèmes, fournissant des didacticiels de linguistique informatique prêts à l'emploi et facile à utiliser, ainsi couvre le traitement symbolique et statistique du langage naturel et est interfacé à des corpus annotés.

Il représente une suite de modules Python fournissant un ensemble de bibliothèques de traitement de texte pour la classification, la tokenisation, le radicalisme, le balisage, l'analyse, le raisonnement sémantique et de nombreuses données NLP [49].
NLTK a été appelé "un outil merveilleux pour enseigner et travailler dans la linguistique informatique à l'aide de Python" et "une bibliothèque incroyable pour jouer avec le langage naturel".

Numpy : est le package fondamental pour le calcul scientifique en Python. Il s'agit d'une bibliothèque Python qui fournit un objet tableau multidimensionnel, divers objets dérivés (tels que des tableaux et des matrices masqués) et un assortiment de routines pour des opérations rapides sur des tableaux, notamment mathématiques, logiques,

algèbre linéaire de base, opérations statistiques de base, simulation aléatoire et bien plus encore [50].

Pandas : est une bibliothèque open source sous licence BSD fournissant des structures de données et des outils d'analyse de données hautes performances et faciles à utiliser pour le langage de programmation Python [51]. Son installation en ouvrant le Shell de commande et en interpellant la commande : Pip install pandas.
Celle-ci permet de manipuler facilement des tableaux de données avec des étiquettes de variables et d'individus, ces tableaux sont appelés « DataFrames » (stockées dans des fichiers CSV, TSV...), similaires aux dataframes sous R. on peut facilement lire et écrire ces dataframes à par tir ou vers un fichier tabulé ainsi tracer des graphes à partir de ces DataFrames grâce à matplotlib [52].

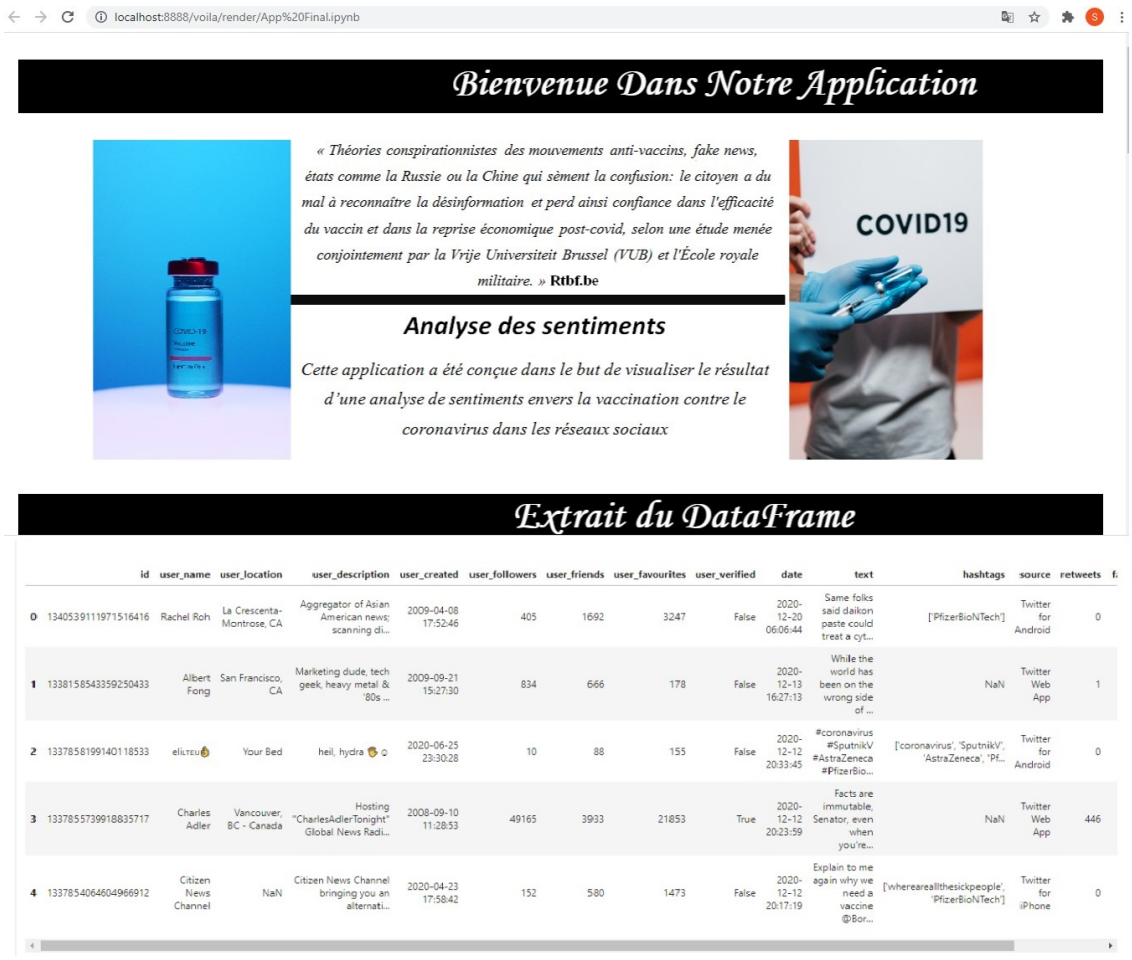
Matplotlib : est une bibliothèque complète permettant de créer des visualisations statiques, animées et interactives en Python [53].
Elle peut être combinée avec les bibliothèques python de calcul scientifique NumPy et SciPy. Matplotlib est distribuée librement et gratuitement sous une licence de style BSD.

Ipywidget : ipywidgets , également connu sous le nom de jupyter-widgets ou simplement widgets, sont des widgets HTML interactifs pour les notebooks Jupyter et le noyau IPython.

Les blocs-notes prennent vie lorsque des widgets interactifs sont utilisés. Les utilisateurs prennent le contrôle de leurs données et peuvent visualiser les changements dans les données[54].

5.4 Jeux de données

Dans cette section, nous allons présenter les interfaces graphiques de notre système sous forme de captures d'écran. La figure 5.3 est la capture d'écran de l'entête de notre application web suivi de l'affichage d'un extrait du dataset.



	id	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text	hashtags	source	retweets	f
0	1340539111971516416	Rachel Roh	La Crescenta-Montrose, CA	Aggregator of Asian American news; scanning di...	2009-04-08 17:52:46	405	1692	3247	False	2020-12-20 06:06:44	Same folks said daikon paste could treat a cyt...	['PfizerBioNTech']	Twitter for Android	0	
1	1338158543359250433	Albert Fong	San Francisco, CA	Marketing dude, tech geek, heavy metal & '80s ...	2009-09-21 15:27:30	834	666	178	False	2020-12-13 16:27:13	While the world has been on the wrong side of ...	NaN	Twitter Web App	1	
2	1337858199140118533	eli.treu	Your Bed	hell, hydra	2020-06-25 23:30:28	10	88	155	False	2020-12-12 20:33:45	#coronavirus #SputnikV #AstraZeneca #PfizerBio...	['coronavirus', 'SputnikV', 'AstraZeneca', 'Pf...	Twitter for Android	0	
3	1337855739918835717	Charles Adler	Vancouver, BC - Canada	Hosting "CharlesAdlerTonight" Global News Radi...	2008-09-10 11:28:53	49165	3933	21853	True	2020-12-12 20:23:59	Facts are immutable, Senator, even when you're...	NaN	Twitter Web App	446	
4	1337854064604966912	Citizen News Channel	NaN	Citizen News Channel bringing you an alternati...	2020-04-23 17:58:42	152	580	1473	False	2020-12-12 20:17:19	Explain to me again why we need a vaccine @Bor...	['whereareallthesickpeople', 'PfizerBioNTech']	Twitter for iPhone	0	

FIGURE 5.3 – Interface de l'application web.

Le dataset sur lequel nous avons construit notre analyse a subi toutes les étapes de prétraitement de texte nécessaires , à savoir : la tokenisation, la normalisation, la lemmatisation et la classification et le résultat de toutes ces étapes a été affiché dans notre application web dans un seul et unique tableau. Le dataset utilisé contient environ 90000 lignes(tweet), il est possible d'afficher uniquement les 5 premiéres lignes dans une application web, nous avons donc permis à l'utilisateur de notre application de visualiser le résultat du prétraitement des données des 5 premiéres lignes du dataset prétraité avant de montrer le résultat, la figure 5.4 est une capture d'écran de l'interface graphique de l'application montrant le résultat du prétraitement du texte.

	text	tokenized	No_stopwords	stemmed_porter	stemmed_snowball	lemmatized
0	Same folks said daikon paste could treat a cyt...	[same, folks, said, daikon, paste, could, trea...	[folks, said, daikon, paste, could, treat, cyt...	[folk, said, daikon, past, could, treat, cytok...	[folk, said, daikon, past, could, treat, cytok...	[folk, said, daikon, paste, could, treat, cyto...
1	While the world has been on the wrong side of ...	[while, the, world, has, been, on, the, wrong,...	[world, wrong, side, history, year, hopefully,...	[world, wrong, side, histori, year, hope, bigg...	[world, wrong, side, histori, year, hope, bigg...	[world, wrong, side, history, year, hopefully,...
2	#coronavirus #SputnikV #AstraZeneca #PfizerBio...	[, coronavirus, sputnikv, astrazeneca, pfizerb...	[, coronavirus, sputnikv, astrazeneca, pfizerb...	[, coronaviru, sputnikv, astrazeneca, pfizerbi...	[, coronavirus, sputnikv, astrazeneca, pfizerb...	[, coronavirus, sputnikv, astrazeneca, pfizerb...
3	Facts are immutable, Senator, even when you're...	[facts, are, immutable, senator, even, when, y...	[facts, immutable, senator, even, ethically, s...	[fact, immut, senat, even, ethic, sturdi, enou...	[fact, immut, senat, even, ethic, sturdi, enou...	[fact, immutable, senator, even, ethically, st...
4	Explain to me again why we need a vaccine @Bor...	[explain, to, me, again, why, we, need, a, vac...	[explain, need, vaccine, borisjohnson, matthan...	[explain, need, vaccin, borisjohnson, matthanc...	[explain, need, vaccin, borisjohnson, matthanc...	[explain, need, vaccine, borisjohnson, matthan...

FIGURE 5.4 – Interface de l'application web affichant le résultat du prétraitement de texte.

L'utilisation du framework voila ainsi que la bibiliothéque ipywidget nous a permis d'augmenter l'intéractivité de notre application web, dans la figure 5.5, on visualise l'état de l'interface avant de cliquer sur les boutons" calculer la polarité" et " voir graphe" qui est relatif au calcul de la polarité des sentiments.

FIGURE 5.5 – Interface de l'application web affichant les boutons calcul de polarité et voir graphe.

Nous avons programmé le bouton **Calculer la polarité**, lorsqu'on clique sur le bouton, un événement est produit et ce dernier est le calcul de la polarité des sentiments des tweet, l'affichage que l'on visualise dans l'interface graphique est alors un mini-tableau ayant deux colonnes, la premiére représente le sentiment exprimé dans le tweet(1=Neutre, 2= Positif, 0= Négatif) et la deuxiéme colonne représente le nombre de mots relatif à chaque sentiment.

Ensuite le bouton **Voir graphe** nous permet de visualiser le résultat du calcul précédent sous forme d'un cercle(pie dans matplotlib) où chaque partie du cercle représente la polarité des sentiments en pourcentage. La figure 5.6 montre le résultat de l'événement que les deux boutons produisent.

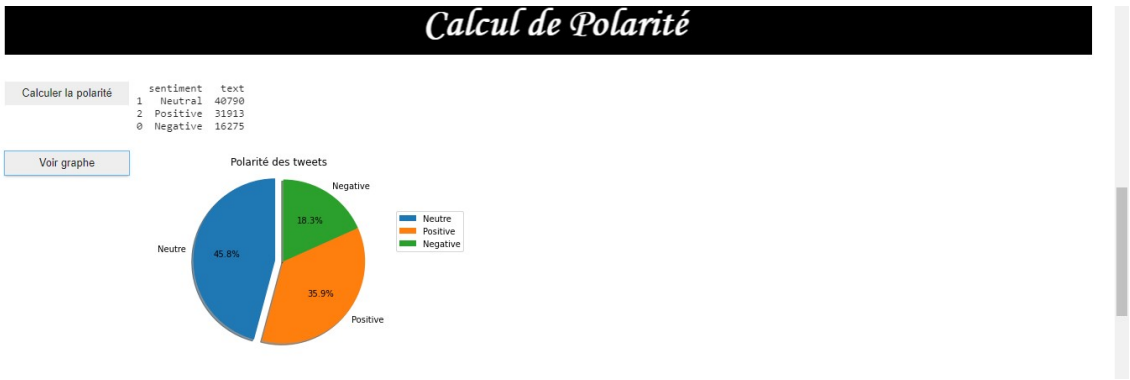


FIGURE 5.6 – Interface de l'application web affichant le résultat de l'évènement crée par deux boutons.

Dans la figure , 5.7 on voit la partie de notre application web affichant les boutons suivant :

— **Prédiction score :** qui affiche le résultat prédit pour l'algorithme multinomial de Naive Bayes que nous avons utilisé pour notre analyse ;

— **Précision score :** qui affiche la précision(pourcentage) de trois algorithmes de Naive Bayes à savoir, le multinomial Naive Bayes, Bernouli Naive Bayes et ComplementNaive Bayes ;

— **Voir graphe :** qui affiche un deuxiéme graphe représentant la différence de précisions entre les trois algorithmes de Naive Bayes : sur l'axe vertical le pourcentage de précision et sur l'axe horizental les différents algorithmes évalués.

FIGURE 5.7 – Interface de l'application web affichant les boutons prédictions, précision et voir graphe.

Enfin la figure 5.8 montre l'événement produit par chacun des boutons : **Prédiction score**, **Précision score**, **Voir graphe**.

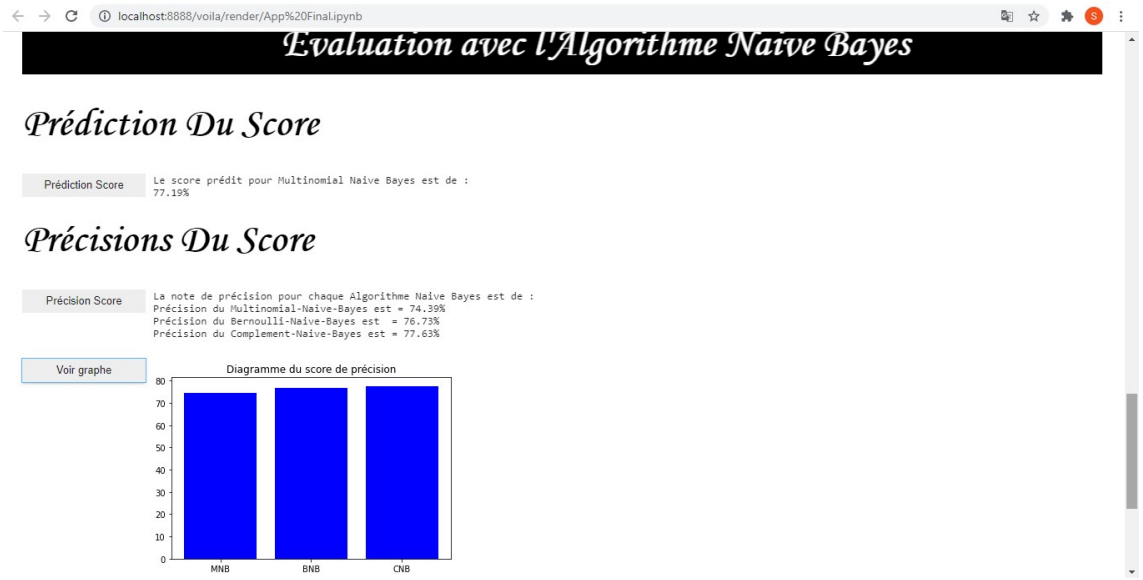


FIGURE 5.8 – Interface de l'application web affichant le résultat des boutons prédictions et précision.

5.5 Évaluation

Après avoir effectué la classification, une étape d'évaluation est nécessaire pour la détermination de la qualité du processus qui a été établi durant notre analyse de sentiment. Les notions de précisions et de rappel correspondent ainsi à une conception et à une mesure de la pertinence, nous avons ainsi élaboré un test de précision sur notre analyse sous les trois différents modèles de l'algorithme Naïve Bayes qui correspondent aux : modèle Multinomial Naïve Bayes, Bernoulli Naive Bayes et Complément Naive Bayes.

Nous avons effectué des tests de précision sur les différents modèles de l'algorithme Naive Bayes :

— Multinomial naive bayes

— Bernoulli Naive Bayes

— Complement Naive bayes

Nous avons remarqué que le modèle Complement Naive Bayes est plus pertinent que les deux autres modèles en terme de précision. Si l'on se réfère au graphe **"Diagramme du score de précision"**, la différence est négligeable mais pour une bonne analyse et pour les travaux futurs, il est important de calculer la précision de chaque algorithme, éventuellement pour les travaux futurs.

5.6 Conclusion

Dans ce chapitre nous avons présenté les divers aspects utilisés pour l'implémentation de notre application à savoir les environnements, les outils de développements ainsi que les différentes bibliothèques évoquées dans notre approche, nous avons également présenté l'interface graphique de notre application web et nous avons expliqué chaque composant que contient cette dernière.

De plus, nous devons noter que l'interface graphique a été construite dans le but de simplifier l'accès aux résultat des notre analyse des sentiments aux utilisateur et de représenter de manière claire et limpide les étapes principales de cette dernière.

Chapitre 6

Conclusion générale

6.1 Rappel du cadre et des objectifs du mémoire

Les campagnes de vaccination ont été ralenties car plusieurs personnes sont dubitatives et hésitent à se faire vacciner. Il existe plusieurs facteurs qui ont semé le doute quant à l'efficacité et la fiabilité des vaccins contre le coronavirus. Le scepticisme du grand public est nourri par la désinformation qui existe sur internet.

Cependant, en pleine pandémie, une meilleure compréhension des doutes et des préoccupations du public nous aidera à avancer plus vite vers une immunité collective. Partant de ce constat, nous avons effectué une analyse des sentiments envers la vaccination contre le coronavirus dans les réseaux sociaux, plus précisément le réseau social "Twitter" carL'utilisation de tweets dans notre étude nous donne une large palette d'avis et de sentiments liés au sujet étudié.

6.2 Principales contributions

Au terme de ce projet, nous rappelons que notre travail porte sur l'étude, l'analyse et le traitement des réactions (tweets) en rapport avec le vaccin contre la Covid-19, ceci afin de déterminer la tonalité émotionnelle des internautes.

Notre application fournit différentes vues à travers les données importées. Elle permet d'intégrer correctement les données dans le but de les analyser.

6.3 Principales limites

Le travail présenté dans ce mémoire est guidé par les nombreuses préoccupations des populations concernant les vaccin contre le virus du covid-19 afin de mieux comprendre la raison de cette grande hésitation à vouloir se faire vacciner en temps de pandémie.

Nous tenons à rappeler que le programme machine-learning que nous avons implémenté a pour résultat la polarité des sentiments exprimés sur twitter à savoir négatif, positif et neutre ainsi que la valeur en pourcentage de chacune de ces polarités. Néanmoins, c'est un sujet d'actualité et les programmes machine-learning sont efficaces et utiles dans d'autres études, ce travail peut être utilisé afin de subir des améliorations, étant donnée que c'est un travail qui demeure incomplet et limité. Il d'agit en particulier à des limites liées à :

— L'analyse des sentiments est faite sur tous les vaccins en général et ne compare pas entre les différents vaccins.
— L'analyse des sentiments n'est pas faite en temps réel car il faut mettre à jour l'application web à chaque fois que le dataset est mis à jour.
— L'analyse nous permet pas de connaître la cause des sentiments négatifs à l'égard des vaccins contre le coronavirus.

6.4 Perspectives et travaux futurs

Conscients des limites que présentent nos travaux et soucieux de la continuité des travaux qui peut être mise en œuvre, nous avons envisagé les perspectives principales suivantes :

— Étudier les réactions des internautes vis-à-vis de chaque vaccin indépendamment des autres et de comparer la popularité de chaque vaccin.
— Utiliser une api Twitter afin d'améliorer le système et de permettre une analyse en

temps réel.

— Ajouter une fonctionnalité qui permet à l'utilisateur d'introduire son propre dataset.

— Analyser les mots et expressions négatives afin de mieux cerner la raison de la méfiance des population face aux différents vaccins.

6.5 Conclusion

Notre étude est basée sur un théme d'actualité qui évolue quotidiennement depuis le début de la pandémie en 2019, c'est pourquoi le résultat de notre analyse risque fortement de changer dans les mois à venir et nécessite par conséquent des mises à jour régulières. Enfin, nous espérons que ce travail sera complété et deviendra l'objet et la réflexion d'autres projets.

Bibliographie

[1] Tim O'reilly. What is web 2.0 : Design patterns and business models for the next generation of software. *Communications & strategies*, (1) :17, 2007.

[2] https ://towardsdatascience.com// consulté le 25/07/2021.

[3] https ://www.who.int/fr/news-room/q-a-detail/coronavirus-disease-covid-19]. consulté le 10/07/2021.

[4] Institut pasteur, paris, 14 novembre 2018 compte rendu : docteur brigitte biardeau, acms.

[5] *Commutation audiovisuelle contextuelle basée sur l'apprentissage profond pour l'amélioration de la parole dans des environnements réels. Fusion d'informations juil ;59 :163-170.].*

[6] *Un cadre d'analyse des sentiments persan hybride : intégrer des règles basées sur la grammaire de dépendance et des réseaux de neurones profonds. Neuroinformatique .*

[7] https ://www.universalis.fr/encyclopedie/sentiment.

[8] Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. *A practical guide to sentiment analysis*. Springer, 2017.

[9] *Diffusion des opinions dans les réseaux.*

[10] Mikalai Tsytsarau and Themis Palpanas. Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3) :478–514, 2012.

[11] Najeh Elouni. *Etude de quelques formes d'expression des émotions et des sentiments dans le contexte des nouvelles formes de communication.* PhD thesis, Université Bourgogne Franche-Comté, 2018.

[12] Despo Georgiou, Andrew MacFarlane, and Tony Russell-Rose. Extracting sentiment from healthcare survey data : An evaluation of sentiment analysis tools. In *2015 Science and Information Conference (SAI)*, pages 352–361. IEEE, 2015.

[13] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications : A survey. *Ain Shams engineering journal*, 5(4) :1093–1113, 2014.

[14] Jihen Karoui, Farah Benamara, and Véronique Moriceau. *Détection automatique de l'ironie : Application à la fouille d'opinion dans les microblogs et les médias sociaux.* ISTE Group, 2019.

[15] Marc Vincent and Grégoire Winterstein. Building and exploiting a french corpus for sentiment analysis (construction et exploitation d'un corpus français pour l'analyse de sentiment)[in french]. In *Proceedings of TALN 2013 (Volume 2 : Short Papers)*, pages 764–771, 2013.

[16] Grzegorz Dziczkowski. Analyse des sentiments : système autonome d'exploration des opinions exprimées dans les critiques cinématographiques. automatique / robotique. École nationale supérieure des mines de paris, 2008. français. ffnnt : 2008enmp1637ff. 2009.

[17] https ://www.universalis.fr/encyclopedie/sentiment consulté le 08/05/2021.

[18] https ://www.ibm.com/ consulté le 10/04/2021.

[19] Michael Rajnik, Marco Cascella, Arturo Cuomo, Scott C Dulebohn, and Raffaela Di Napoli. Features, evaluation, and treatment of coronavirus (covid-19). Technical report, Uniformed Services University Of The Health Sciences, 2021.

[20] Zi Yue Zu, Meng Di Jiang, Peng Peng Xu, Wen Chen, Qian Qian Ni, Guang Ming Lu, and Long Jiang Zhang. Coronavirus disease 2019 (covid-19) : a perspective from china. *Radiology*, 296(2) :E15–E25, 2020.

[21] Man Hung, Evelyn Lauren, Eric S Hon, Wendy C Birmingham, Julie Xu, Sharon Su, Shirley D Hon, Jungweon Park, Peter Dang, and Martin S Lipsky. Social network analysis of covid-19 sentiments : application of artificial intelligence. *Journal of medical Internet research*, 22(8) :e22590, 2020.

[22] https ://www.who.int/emergencies/diseases/novel-coronavirus consulté le 11/07/2021.

[23] Centers for Disease Control, Prevention, et al. Coronavirus disease 2019 (covid-19). centers for disease control and prevention ; 2020, 2020.

[24] https ://fr.statista.com, consulté le 30/05/2021.

[25] Alexandre de Figueiredo, Clarissa Simas, Emilie Karafillakis, Pauline Paterson, and Heidi J Larson. Mapping global trends in vaccine confidence and investigating barriers to vaccine uptake : a large-scale retrospective temporal modelling study. *The Lancet*, 396(10255) :898–908, 2020.

[26] Bruce Gellin. Why vaccine rumours stick—and getting them unstuck. *The Lancet*, 396(10247) :303–304, 2020.

[27] https ://www.stat4decision.com/ consulté le 10/07/2021.

[28] Jeffrey V Lazarus, Scott C Ratzan, Adam Palayew, Lawrence O Gostin, Heidi J Larson, Kenneth Rabin, Spencer Kimball, and Ayman El-Mohandes. A global survey of potential acceptance of a covid-19 vaccine. *Nature medicine*, 27(2) :225–228, 2021.

[29] Barry R Bloom, Glen J Nowak, and Walter Orenstein. "when will we have a vaccine ?"—understanding questions and answers about covid-19 vaccination. *New England Journal of Medicine*, 383(23) :2202–2204, 2020.

[30] Arnaud Fontanet and Simon Cauchemez. Covid-19 herd immunity : where are we ? *Nature Reviews Immunology*, 20(10) :583–584, 2020.

[31] Mulkan Ritonga, Muhammad Ali Al Ihsan, Agus Anjar, Fauziah Hanum Rambe, et al. Sentiment analysis of covid-19 vaccine in indonesia using naïve bayes algorithm. In *IOP Conference Series : Materials Science and Engineering*, volume 1088, page 012045. IOP Publishing, 2021.

[32] Talha Burki. The online anti-vaccine movement in the age of covid-19. *The Lancet Digital Health*, 2(10) :e504–e505, 2020.

[33] Neha Puri, Eric A Coomes, Hourmazd Haghbayan, and Keith Gunaratne. Social media and vaccine hesitancy : new updates for the era of covid-19 and globalized infectious diseases. *Human vaccines & immunotherapeutics*, 16(11) :2586–2593, 2020.

[34] G. ; Amenta F. Chintalapudi, N. ;Battineni. Sentimental analysis of covid-19 tweets using deep learning models. infect.

[35] Amir Hussain, Ahsen Tahir, Zain Hussain, Zakariya Sheikh, Mandar Gogate, Kia Dashtipour, Azhar Ali, and Aziz Sheikh. Artificial intelligence–enabled analysis of public attitudes on facebook and twitter toward covid-19 vaccines in the united kingdom and the united states : Observational study. *Journal of medical Internet research*, 23(4) :e26627, 2021.

[36] Quyen G To, Kien G To, Van-Anh N Huynh, Nhung TQ Nguyen, Diep TN Ngo, Stephanie J Alley, Anh NQ Tran, Anh NP Tran, Ngan TT Pham, Thanh X Bui, et al. Applying machine learning to identify anti-vaccination tweets during the covid-19 pandemic. *International journal of environmental research and public health*, 18(8) :4069, 2021.

[37] https ://www.kaggle.com/gpreda/all-covid19-vaccines-tweets, consulté le 18/07/2021.

[38] NH Abd Rahim and SH Mohd Rafie. Sentiment analysis of social media data in vaccination. *International Journal*, 8(9), 2020.

[39] C. Benavent Sophie Balech. Les techniques du nlp pour la recherche en sciences de gestion. 2019.

[40] M Trupthi, Suresh Pabboju, and G Narasimha. Sentiment analysis on twitter using streaming api. In *2017 IEEE 7th International Advance Computing Conference (IACC)*, pages 915–919. IEEE, 2017.

[41] E. Cambria. *Affective computing and sentiment analysis. IEEE Intelligent Systems*, volume 32. 2016.

[42] H Tran-Ngoc, Samir Khatir, T Le-Xuan, G De Roeck, T Bui-Tien, and M Abdel Wahab. A novel machine-learning based on the global search techniques using vectorized data for damage detection in structures. *International Journal of Engineering Science*, 157 :103376, 2020.

[43] Jiang L. Li C Wang, S. *Adapting naive Bayes tree for text classification*, volume 44. 2015.

[44] James Yan Yuxing Yan. *Science des données pratique avec Anaconda.* 2018.

[45] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian E Granger, Matthias Bus-sonnier, Jonathan Frederic, Kyle Kelley, Jessica B Hamrick, Jason Grout, Sylvain Corlay, et al. *Jupyter Notebooks-a publishing format for reproducible computational workflows.*, volume 2016. 2016.

[46] Guido Van Rossum and Fred L Drake Jr. *Python tutorial*, volume 620. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[47] Steven Pemberton et al. Xhtmltm 1.0 the extensible hypertext markup language. *W3C Recommendations*, 2000.

[48] https ://analyticsindiamag.com/ consulté le 01/09/2021.

[49] Edward Loper and Steven Bird. Nltk : The natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[50] https ://numpy.org/doc/stable/user/whatisnumpy.html le 15/07/2021.

[51] https ://pandas.pydata.org/docs/, consulté le 15/07/2021.

[52] http ://www.python-simple.com/python-pandas/panda-intro.php le 15/07/2021.

[53] https ://matplotlib.org/ le 14/07/2021.

[54] https ://ipywidgets.readthedocs.io/ consulté le 01/09/2021.

[55] *Flask Web Development : Developing Web Applications With Python.*

[56] Bruce Johnson. *outils d'édition et de débogage de bout en bout pour les développeurs Web.* 2019.