ENSIAS

Master de Recherche BIO-MSCS

Data mining / text mining project

Analyse du sentiment des tweets du vaccin COVID-19

Table of Contents

Liste des figures

Introduction au sujet

1. Compréhention du problème

1.1. Problématique

Avec la pandémie de coronavirus, qui a vraiment changé notre vie, le monde s'est battu pour gérer et contrôler la capacité de sa propagation rapide. Les scientifiques essaient tout, des médicaments comme des nouveaux traitements, mais le moyen le plus probable de mettre fin à cette pandémie est un vaccin. Heureusement, plus d'une centaine de vaccins sont en cours d'élaboration par des scientifiques du monde entier, à l'aide de techniques et d'approches éprouvées qui n'ont jamais été testées auparavant.

Alors que les scientifiques font beaucoup d'efforts et sacrifient leur vie pour trouver un vaccin efficace afin de mettre fin à cette pandémie et de ramener la vie normale, il existe de l'autre côté un mouvement anti-vaccins qui travaille agressivement à promouvoir la désinformation sur les vaccins COVID-19, jusqu'à promouvoir de fausses allégations de décès dus aux vaccins. Et dès le début, les militants anti-vaccins se sont engagés à défendre l'idée que les vaccins COVID-19 ne fonctionneraient pas, qu'ils seraient dangereux et qu'ils seraient promus par une conspiration mondiale malfaisante. Ils continuent de répandre ces allégations, par exemple en utilisant le fait que les vaccins COVID-19 bénéficient de protections en matière de responsabilité pour laisser entendre que les vaccins sont dangereux. Les protections en matière de responsabilité des fabricants de vaccins COVID-19 sont réelles, mais elles ne constituent pas une preuve que les vaccins sont dangereux.

La nouvelle d'un nouveau vaccin, efficace, a fait réfléchir, et c'est ce qui nous a fait penser pourquoi ne pas essayer et chercher à savoir ce que les gens en pensent, s'ils y croient à son efficacité et s'ils l'essaieraient ou non.

Nous choisissons d'extraire et d'utiliser les données Twitter, car il s'agit d'une mine d'or de données, et contrairement à d'autres plates-formes sociales, presque tous les tweets des utilisateurs sont entièrement publics et extractibles. C'est un énorme avantage si nous essayons

d'obtenir une grande quantité de données sur lesquelles exécuter des analyses. Les données de Twitter sont également assez spécifiques.

1.2. Objectif du sujet

L'objectif est mener une analyse des sentiments (sentiment analysis), également appelée opinion Mining qui consiste à recueillir l'opinion des gens sur tout événement survenant dans la vie réelle et à analyser les sensations, les attitudes et les émotions des individus vis-à-vis des entités, et pour notre cas elle va nous aider à suivre le sentiment et les émotions du public concernant les mesures actuelles et potentielles pour contenir et traiter le COVID-19, et à étudier les attitudes à l'égard de la vaccination contre le Covid-19, et détecter s'il existe un mouvement anti-vaccination et une diffusion de fausses informations sur le vaccin.
Nous utilisons des données des médias sociaux pour examiner la perception et le sentiment des mesures de distanciation sociale de COVID-19 et des outils médicaux potentiels et découvrir et décrire le sentiment du public concernant la vaccination contre le COVID-19.
Un autre objectif c'est d'insister sur la nécessité d'employer des "mécanismes de contrôle" pour empêcher la diffusion de la psychologie négative dans l'esprit des utilisateurs des médias sociaux et des fausses idées sur la vaccination contre la pandémie sans aucune références ou épreuves scientifiques.

1.3. Les Réseaux Sociaux

Les réseaux sociaux sont des services web permet aux utilisateurs de construire un profil public ou semi-public dans un système, d'articuler une liste d'autres utilisateurs avec lesquels ils partagent une connexion, voir et parcourir la liste de connexions et contenus effectués par d'autres au sein du même système et s'interconnecter avec eux.

Ils permettent aux gens d'échanger d'information à grand échelle tant au niveau personnel que professionnel, de s'exprimer et partager et leur idées, expériences ou leur opinion et points de vue sur des différents sujets, avec une caractéristique d'interactivité en donnant la possibilité de commenter ou repartager.

Les réseaux sociaux mettent en œuvre trois éléments le caractère communautaire, le modèle participatif et la personnalisation de l'information.

1.4. Twitter

Twitter est un micro-blogue qui permet à l'utilisateur d'écrire et lire des messages sur internet, il est devenu l'un des médias sociaux les plus utilisés. Twitter a été créé le 21/03/2006 par Jack Dorsey , Evan Williams ,Biz Stone et Noah Glass , et lancé en 07/2006. Twitter compte 326 millions d'utilisateurs acifs par moi avec 504 millions de tweets par jour .

Les messages publiés sont appelés des tweets ou gazaouillis et limités à 280 caractères .Twitter offre beaucoup de fonctionnalités comme par exemple poster un message sous forme "tweet", retweet les messages , s'abonner à d'autres utilisateurs.

1.5. Analyse des sentiments

Le mot sentiment signifie une opinion, un avis que l'on porte sur quelque chose.

L'analyse des sentiments est le traitement informatique d'un texte qu'on explore afin d'identifier le sentiment ou l'opinion générale exprimée. C'est une classification de polarité (positive, négative ou neutre) d'un texte en se basant sur plusieurs caractéristiques telles que la fréquence d'un terme, l'importance d'un terme dans le texte, les négations dans les phrases.

L'exploration se fait concernant des individus, des produits, des évènements ou encore des sujets variés qui sont susceptibles d'être couverts par des commentaires, critiques ou appréciations.

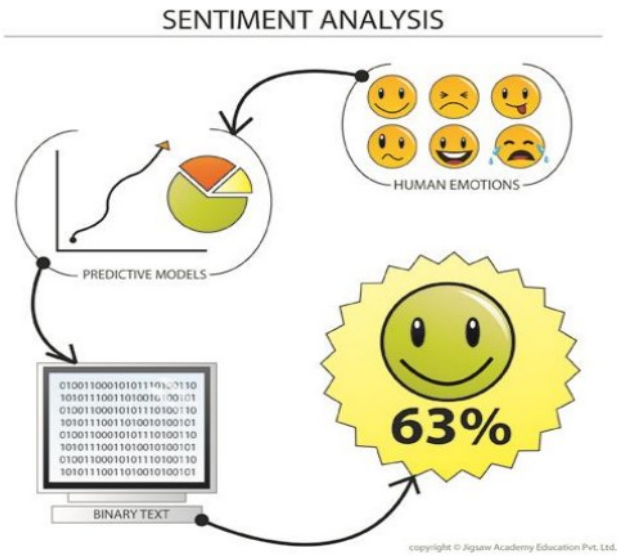

Figure 1: Analyse des sentiments

2. Compréhension des données

2.1. Extraction des données

D'abord , Nous avons demandé un compte de développeur Twitter pour avoir accès à l'utilisation de l'api Twitter , une fois approuvé, nous avons créé une application de développement qui nous a fourni un ensemble d'informations d'identification , des clés (keys) et des jetons (tokens), que nous utiliserons pour authentifier toutes les demandes adressées à l'API et nous aideront à récupérer les données de Twitter .

2.2. Twitter API

L'API Twitter peut être utilisée pour récupérer et analyser des données de manière programmatique, ainsi que pour participer à la conversation sur Twitter.

Cette API permet d'accéder à une variété de ressources différentes, dont les tweets, utilisateurs, messages directs , listes , tendances , médias , lieux .

L'API de Twitter nous permet d'effectuer des requêtes complexes, comme extraire chaque tweet sur un certain sujet au cours des vingt dernières minutes, ou extraire les tweets non retweetés d'un certain utilisateur.

2.3. Description des données

Nous avons extrait nos données de tweets qui parlent principalement du vaccin COVID-19, ces données sont des données prédictives, elles contiennent huit des fonctionnalités qui sont :

- **User.screen_name** : c'est une variable d'objet utilisateur qui montre le nom d'écran qui a publié ce Tweet
- **Text** : c'est une variable de chaîne qui affiche le texte UTF-8 réel de la mise à jour de l'état.
- **Created_at** : c'est une variable d'objet utilisateur qui montre la date et l'heure le tweet a été publié par l'utilisateur .
- **Retweet_count** : c'est une variable de type Entier qui représente le nombre de fois que ce Tweet a été retweeté.
- **Source** : c'est une variable de chaîne utilisé pour afficher le Tweet sous forme de chaîne au format HTML.

- **Favorite_count** : c'est une variable de type Entier et il indique approximativement combien de fois ce Tweet a été apprécié par les utilisateurs de Twitter.
- **User.location** : Représente la localisation géographique de l'utilisateur.
- **User.followers_count** : c'est une variable de type Entier, qui représente le nombre des utilisateurs qui suivent le propriétaire de tweet .

2.4. Rapport de qualité des données

Total de 10612 entries , 0 to 10611 , en 9 colonnes :

Colonne	Compte valeurs non nul	Comp valeurs manquantes	Type de donnée
User	10612 non-nul	0	Objet
Text	10612 non-nul	0	Objet
Created_at	10612 non-nul	0	Objet
Location	10612 non-nul	0	Entier
User_followers	10590 non-nul	22	Objet
Retweet_count	10612 non-nul	0	Entier
Source	8982 non-nul	2530	Objet
Favorite_count	10612 non-nul	0	Entier

Type des données: (4) entier, (5) objet

Utilisation de la mémoire: 746 .3+ KB

Prétraitement:

Le prétraitement de texte est une approche permettant de nettoyer et de préparer des données de texte à utiliser dans un contexte spécifique. Par l'identification et l'extraction des caractéristiques qui sont utilisées pour transformer les données textuelles non structurées en un format intermédiaire structuré qui est stocké dans une base de données. Les techniques de prétraitement représentent la phase la plus longue de tous les processus de découverte des connaissances. La complexité du prétraitement des données dépend des sources de données utilisées.

Dans notre projet nous avons utilisées ces techniques pour nettoyer et extraire le maximum d'information à partir des opinions des gens sur le vaccin anti-COVID-19 en anglais, L'un des plus grands défis de l'exécution de tâches d'apprentissage automatique sur les données des réseaux sociaux et sur les messages texte est l'anglais qu'est très différent de l'anglais standard. Il y a des argots, des acronymes, des abréviations, des initialismes, des hashtags, des URL et, entre autres, des fautes d'orthographe dans les messages tweets et d'autre plateformes. Qui n'apparaissent pas en anglais standard. Alors nous couvrirons certaines étapes de prétraitement des données spécifiques aux données de tweets ou nécessitant une attention particulière.

1. Removing Twitter Handles

Twitter Handles est le nom d'utilisateur qui apparaît à la fin d'URL Twitter unique. Ce n'est pas la même chose qu'un nom Twitter. Twitter handles apparaissent après le signe @. Il n'ajoute aucune valeur à l'ensemble de données.

Exemple des données :

- RT @USAID_NISHTHA: Batting misinformation surrounding #COVID19Vaccine! Leveraging VHND sessions to disseminate accurate information on the…

Après le traitement:

- RT Batting misinformation surrounding #COVID19Vaccine! Leveraging VHND sessions to disseminate accurate information on the…

2. Normalisation du texte

Sur notre data nous avons supprimée les ponctuations et caractères spéciaux, Ces caractères se trouvent le plus souvent dans les commentaires, les références, les numéros de devises, etc. Ces caractères n'ajoutent aucune valeur à la compréhension du texte et induisent du bruit dans les algorithmes. Et remplacé les caractères majuscules par des caractères minuscule

Exemple des données :

- RT Batting misinformation surrounding #COVID19Vaccine! Leveraging VHND sessions to disseminate accurate information on the…

Après le traitement:

- RT Batting misinformation surrounding COVID Vaccine Leveraging VHND sessions to disseminate accurate information on the…..

3. Stopwords

Stopwords sont des mots très courants. Des mots comme «we» et «are» n'aident probablement pas du tout dans l'analyse des sentiments ou les classifications de texte. Par conséquent, nous avons supprimer les Stopwords qui ont le nombre de caractères inferieure a 3 pour gagner du temps de calcul et des efforts dans le traitement de gros volumes de texte.

Exemple des données :

- RT Batting misinformation surrounding COVID Vaccine Leveraging VHND sessions to disseminate accurate information on the

Après le traitement:

- Batting misinformation surrounding COVID Vaccine Leveraging VHND sessions disseminate accurate information.

4. Tokenisation

La tokenisation est le processus de division du texte en un ensemble d'éléments significatifs. Ces pièces sont appelées **jetons**. Par exemple, nous pouvons diviser un morceau de texte en mots, ou nous pouvons le diviser en phrases. En fonction de la tâche à accomplir, nous pouvons définir nos propres conditions pour diviser le texte d'entrée en jetons significatifs.

Exemple des données :

- Batting misinformation surrounding COVID Vaccine Leveraging VHND sessions disseminate accurate information

Après le traitement:

- [Batting, misinformation, surrounding, COVID, Vaccine, Leveraging, VHND, sessions, disseminate, accurate, information]

5. Stemming

Stemming est un processus où les mots sont réduits à une racine en supprimant l'inflexion en supprimant des caractères inutiles, généralement un suffixe. L'algorithme de Porter est un des plus connus pour la langue anglaise. Il applique une succession de règles (mécaniques) pour réduire la longueur des mots c.-à-d. supprimer la fin des mots.

Nous avons utilisé cette méthode principalement pour réduire la dimension des données.

- Batting misinformation surrounding COVID Vaccine Leveraging VHND sessions disseminate accurate information

Après le traitement:

- bat misinform surround covid vaccin leverag vhnd session dissemin accur inform

6. Indexation

Nous avons utilisé la méthode Bag-of-Words qui nous donne le meilleur résultat . Bag-of-Words est une représentation qui transforme un texte arbitraire en **vecteurs de longueur fixe** en comptant le nombre de fois où chaque mot apparaît. Ce processus est souvent appelé **vectorisation** .permet de représenter des données textuelles lors de la modélisation de

texte avec des algorithmes d'apprentissage automatique. Elle divise un texte ou une phrase a des mots et représenté comme le sac (multiset) de ses mots. L'idée principale étant de transformer cette masse de texte non structurée en données digestes pour les algorithmes et les capacités de calculs.

7. Pondération

Pour la représentation numérique nous avons utilisés la méthode TF-IDF est une méthode d'analyse qui peut être utilisée dans une stratégie de référencement pour déterminer les mots-clés et les termes qui augmentent la pertinence des textes publiés et donc du projet Web dans son ensemble. C'est une formule dans laquelle les deux valeurs **TF** (Term Frequency) et **IDF** (Inverse Document Frequency) sont multipliées entre elles. Le résultat est la **fréquence relative des termes** (ou « pondération des termes ») d'un document par rapport à tous les autres documents Web qui **contiennent également le mot-clé** en question lors de l'analyse. Avant de pouvoir effectuer l'analyse TF-IDF, les deux facteurs mentionnés doivent d'abord être déterminés.

○ TF

TF est l'abréviation de l'anglais term frenquency (fréquence du terme). Il détermine la fréquence relative d'un mot ou d'une combinaison de mots dans un document. Cette fréquence du terme sera comparée à la survenance de tous les autres mots restants du texte, du document ou du site web analysé. Cette formule utilise un logarithme qui se lit comme suit :

$$TF(i) = \frac{log_2(Freq(i,j) + 1}{log_2(L)}$$

○ IDF

L'IDF calcule l'inverse document frequency (la fréquence inverse du document) et complète l'analyse de l'évaluation du mot. Il agit en tant que correctif du TF. L'IDF inclut dans le calcul la fréquence des documents pour un mot précis, autrement dit l'IDF compare le chiffre correspondant à tous les documents connus avec le nombre de textes contenant le mot en question. Le logarithme suivant permet ainsi de "condenser" les résultats :

$$IDF_t = \log\left(1 + \frac{N_D}{f_t}\right)$$

Modélisation

1. Les données d'apprentissage

Pour l'étape de modélisation, et car notre data est non labellisés , nous avons décidé d'utiliser une autre donnée, qui est assez similaire, de kaggle nommée «**Twitter Sentiment Analysis »** .

→ Description de data

Ainsi, les données contiennent un tas de tweets différents. Formellement, étant donné un échantillon de formation de tweets et d'étiquettes, où l'étiquette '1' indique que le tweet exprime un sentiment négatif et l'étiquette '0' indique que le tweet exprime un qu'est positif, et un échantillon de validation .

2. Sélection des techniques de modélisation

Pour la première étape nous avons fait le même prétraitement que pour nos données COVID .

Ensuite ,pour le modèle que nous allons utiliser, nous avons choisi le SVM, sur la base d'une étude que nous avons réalisée dans le cadre de notre projet Machine Learning, dans lequel nous avons comparé plusieurs modèles de classification et le SVM était l'un des meilleurs, Il est également prouvé qu'elle est optimale pour les cas linéairement séparables. De plus, sa stratégie est le meilleure pour réduire l'erreur de prédiction.

2.1. SUPPORT VECTOR MACHINE

Figure 2: Logo de SVM

Est un modèle d'apprentissage automatique supervisé qui utilise des algorithmes de classification pour les problèmes de classification à deux groupes. Après avoir donné à un modèle SVM des ensembles de données d'entraînement étiquetées pour chaque catégorie, il est capable de classer un nouveau texte.

Le SVM classe les données en trouvant le meilleur hyperplan qui sépare tous les points de données d'une classe de ceux de l'autre classe. Le véritable avantage du SVM réside dans sa précision et dans le fait qu'il a tendance à ne pas surajuster les données.

→ **SVM avec la classification du langage naturel**?

Pour appliquer l'algorithme SVM à la classification de texte, la première chose dont nous avons besoin est un moyen de transformer un morceau de texte en un vecteur de nombres afin que nous puissions exécuter le SVM avec eux.

La méthode la plus courante est celle des fréquences de mots. Cela signifie que nous traitons un texte comme un bag of words, et pour chaque mot qui apparaît dans ce sac, nous avons une caractéristique. La valeur de cette caractéristique correspond à la fréquence de ce mot dans le texte.

Cette méthode se résume à compter le nombre de fois que chaque mot apparaît dans un texte et à le diviser par le nombre total de mots.

3. Apprentissage

Et car nous avons déjà fait bag of words, nous avons divisés les données en ensembles de training et de validation, nous transmettons les données d'apprentissage à l'algorithme, qui utilisera la classe SVC, avec un kernel "linéaire" que nous avons choisi, afin de produire un modèle.

4. Evaluation

Pour l'évaluation, nous avons fait l'évaluation de modèle qui a nous donné, un pourcentage d'accuracy de 95% , qu'est bien, Ensuite nous avons fait la classification au data test, notre data de COVID.

Après cela, , nous avons fait les mêmes étapes, avec les données qui ont été exposées à la méthode TF-IDF.

Et on a obtenu presque les mêmes résultats qu'avec la méthode de bag of words.

On a apporté les donnes test résultante avec cette méthode dans un autre fichier csv pour extraire quelques informations, qui sont présentés ci-dessous.

→ Polarity Distribution of negative and positive tweets

Cette figure représente une distribution de polarité des tweets négatives et positives.

Nous pouvons voir que les tweets positifs l'emportent sur les négatifs, avec 92% pour les tweets positifs et 8% pour les tweets négatifs.

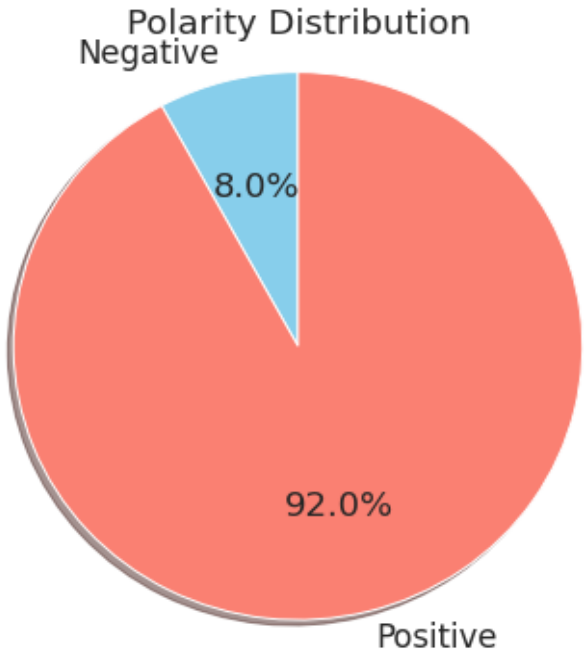

Figure 3: Polarity Distribution

→ **Most Negative Hashtags:**

Cette figure représente les Hashtags les plus négatifs dans les tweets, comme « COVID19Vaccine », « COVID19vaccine », « Pakistan », « Biden ».

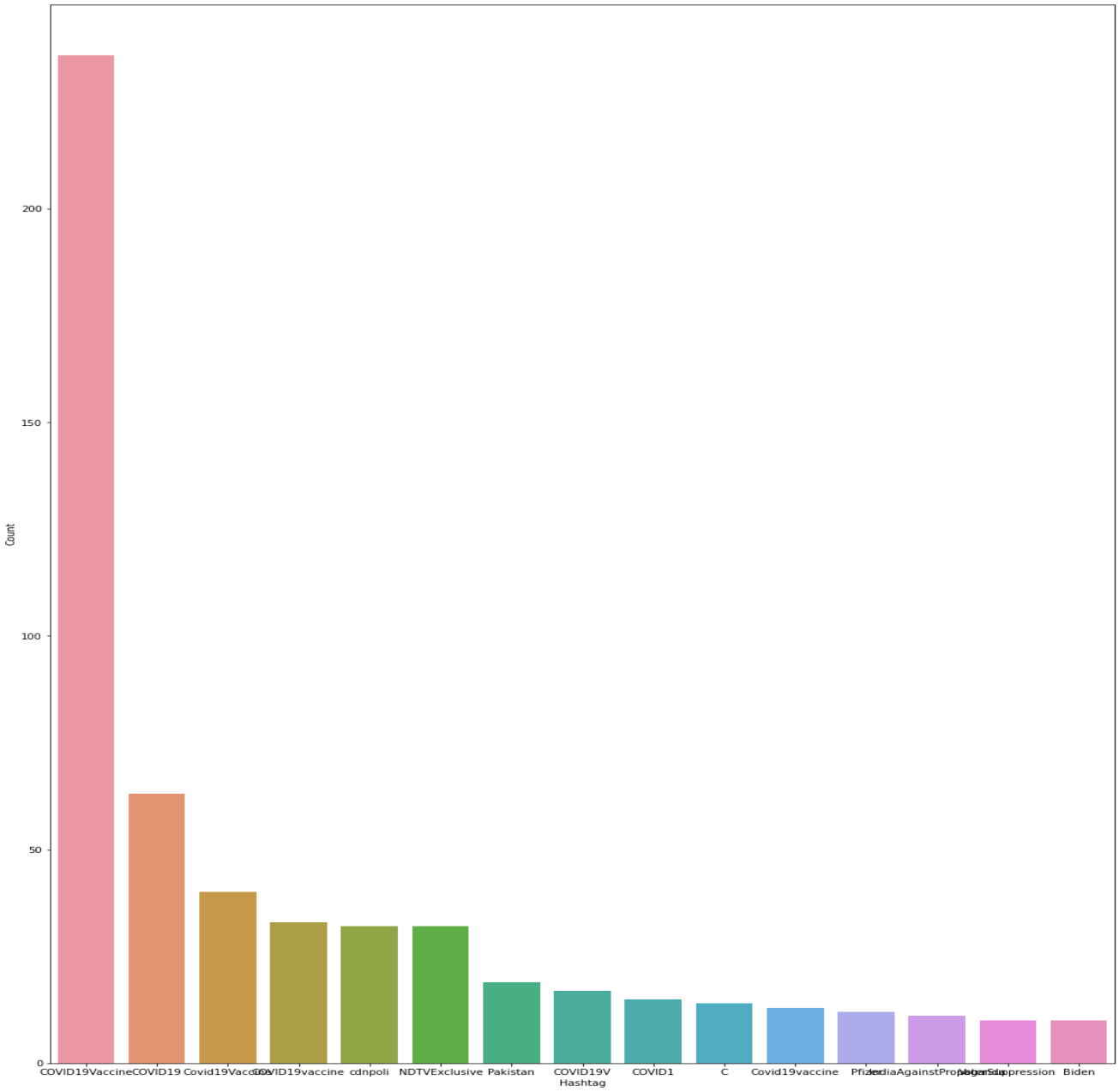

Figure 4:Plot de Hashtags les plus Négatives

→ Word cloud of Negative Tweets:

Cette figure représente un nuage des tweets négative.

Le nuage de mots de la figure présente les mots les plus fréquemment rencontrés dans le corpus de tweets, comme « COVID Vaccine », « Vaccine Manufacturing », « German Chancellor ». Plus la taille du mot est grande dans le nuage, plus il apparaît fréquemment dans le corpus.

Figure 5: Nuage des tweets Négative

→ Most Positive Hashtags:

Cette figure représente les Hashtags les plus positives dans les tweets, comme « COVID19Vaccine », « COVID19vaccine », « Pzifer ».

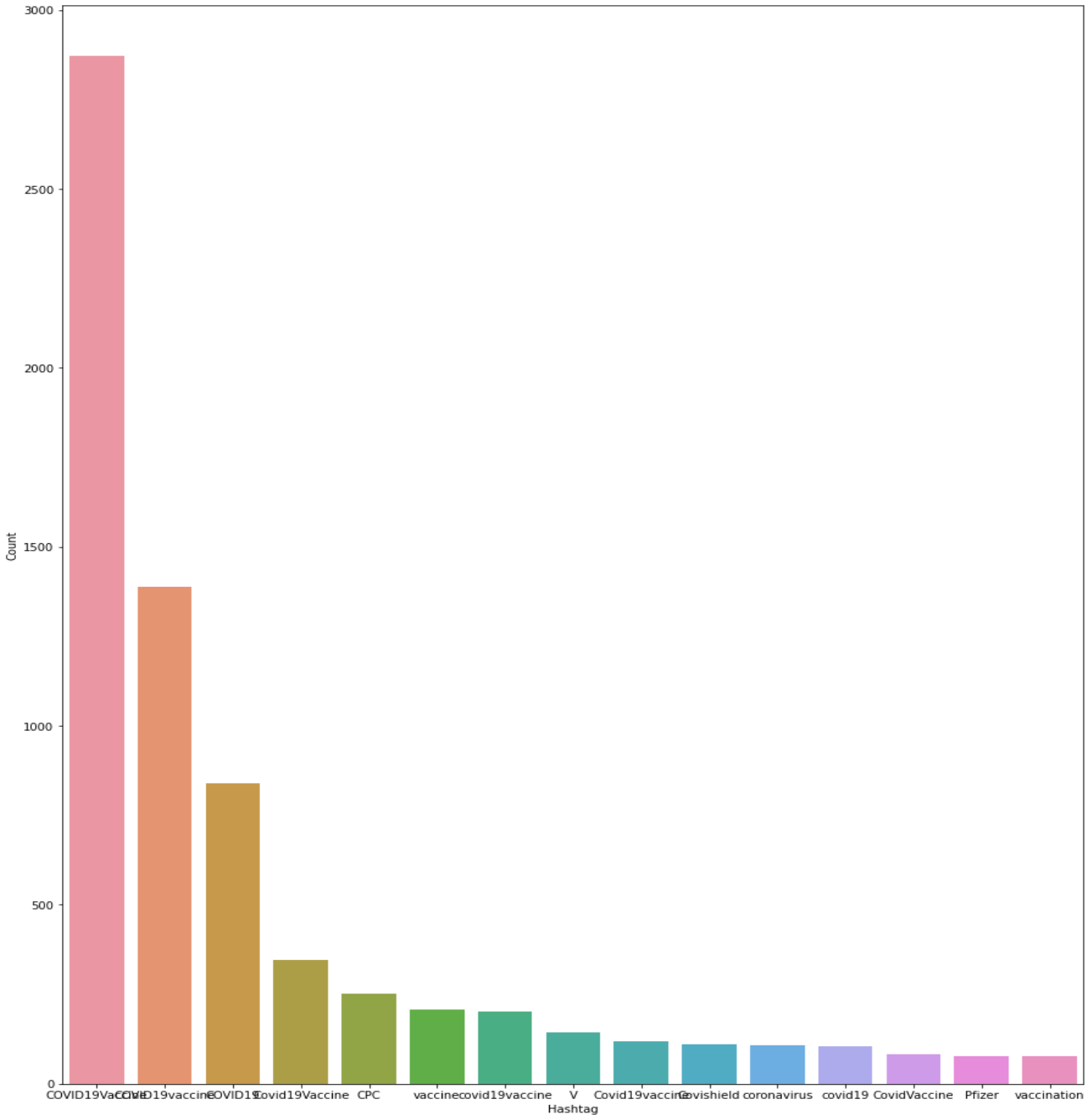

Figure 6: Plot des Hashtags les plus Positives

→ Word cloud of Positive Words:

Cette figure représente un nuage des tweet positives.

Le nuage de mots de la figure présente les mots les plus fréquemment rencontrés dans le corpus de tweets, comme les mots « Vaccinations »,« COVID », « CPC Leader ».

Plus la taille du mot est grande dans le nuage, plus il apparaît fréquemment dans le corpus.

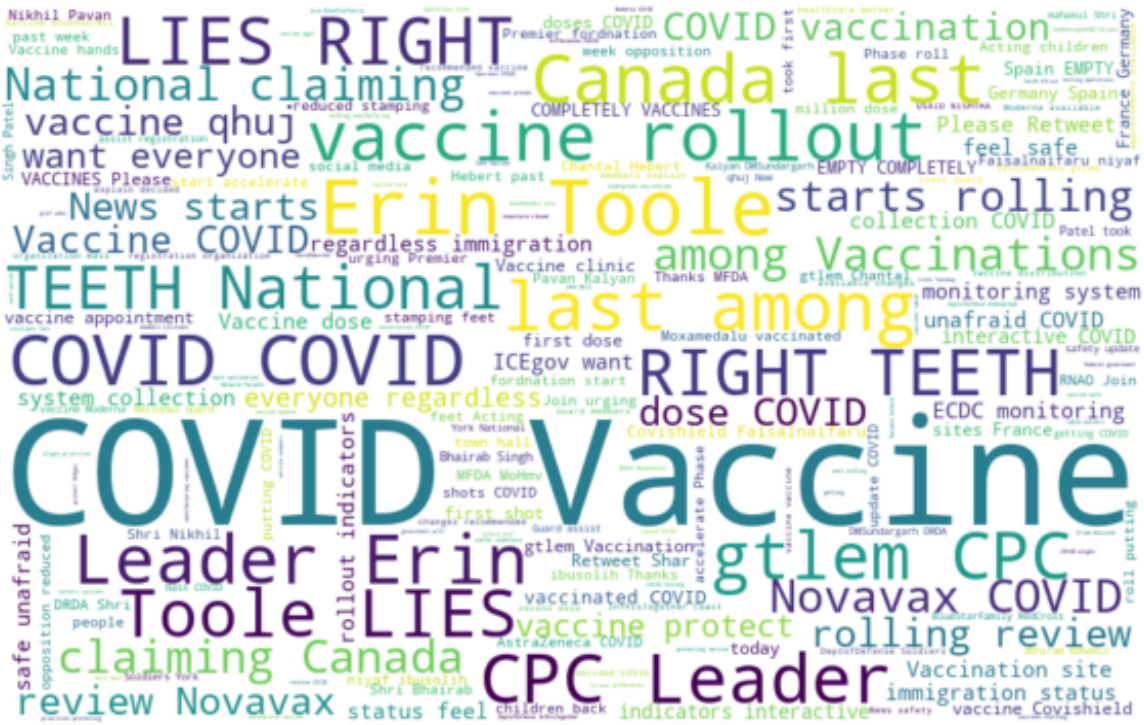

Figure 7: Nuage des mots Positive

D’après les résultats obtenus, Nous pouvons voir que les tweets positifs l'emportent sur les tweets négatifs.

5. Données statistiques sur la vaccination contre covid-19

Cette figure représente le nombre de personnes ayant reçu au moins une dose de vaccin. Il est possible que les personnes vaccinées aient reçu plusieurs doses.

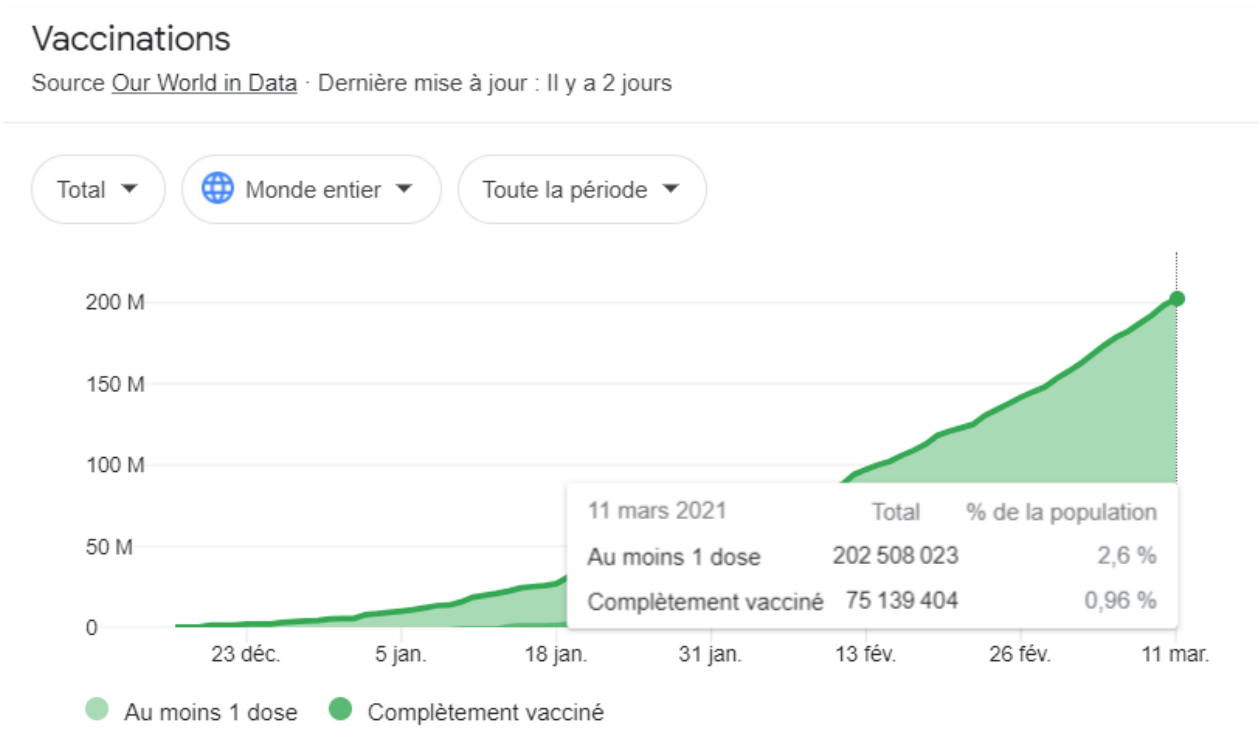

Figure 8: Nombre de personnes ayant reçu le vaccin

Sources : Our world in Data

Pour cette figure ,il représente le nombre total de doses administrées dans chaque zone géographique. Plusieurs doses étant nécessaires pour certains vaccins, le nombre de personnes vaccinées est souvent moins élevé.

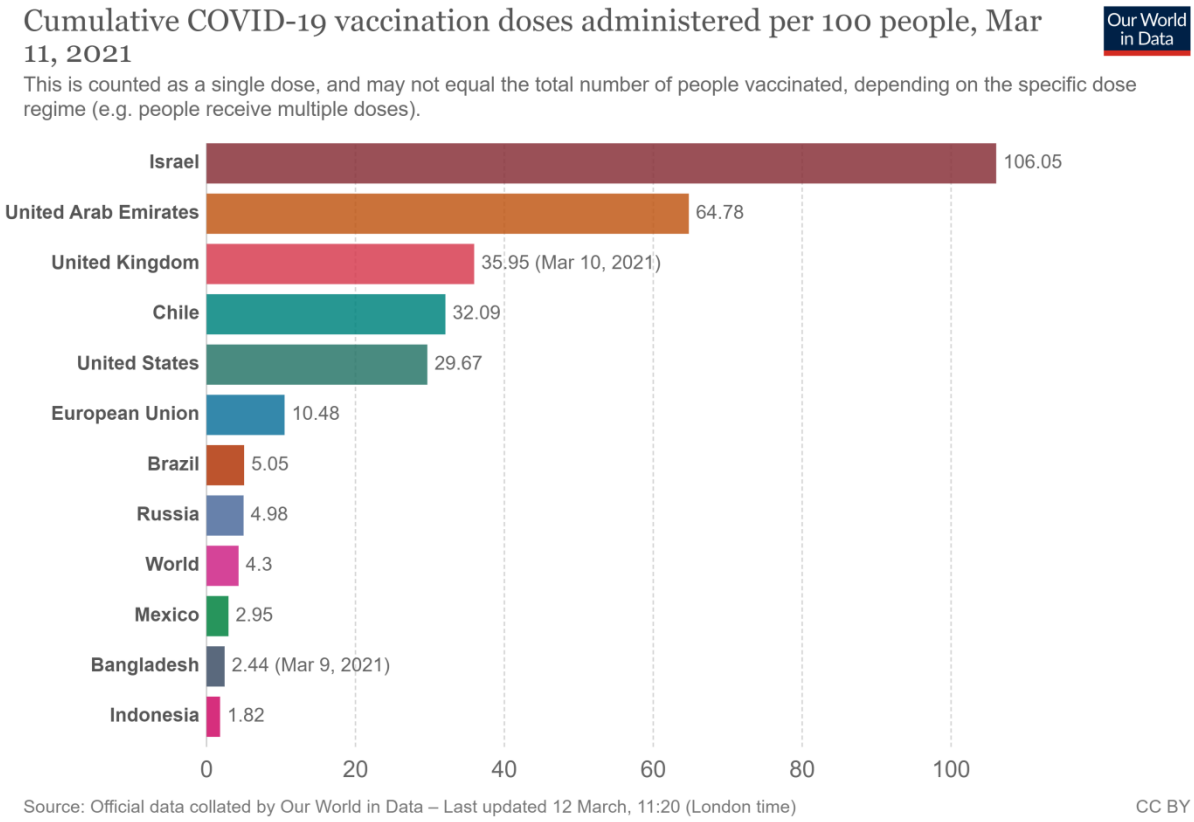

Figure 9: Le nombre total de doses administrées dans chaque zone géographique

Sources: our world in Data, autorisés sanitaires.

Nous pouvons en conclure que, selon nos recherches et notre modélisation utilisant l'analyse des sentiments, avant la mise sur le marché du vaccin, et d'après les résultats de ces chiffres, qui représentent la situation actuelle après la mise sur le marché et l'utilisation du vaccin dans plusieurs pays du monde, notre prédiction est acceptable, c'est-à-dire que la plupart des gens avaient des pensées positives à ce sujet.

Conclusion

Pour conclure, on peut bien dire que on a pu à réaliser notre principal objectif de faire une analyse de sentiment, opinion Mining afin d'analyser les sensations et les attitudes à l'égard de la vaccination contre Covid-19, et détecter s'il existe un mouvement anti-vaccination et une diffusion de fausses informations sur le vaccin.

D'après notre traitement et modélisation, nous pouvons dire que la plupart des personnes ont une bonne vue concernant le vaccin contre la COVID-19.

Pour l'autre objectif, nous avons heureusement trouvé un petit groupe de tweets négatifs, qui n'ont pas affecté les tweets positifs.

Nous pouvons dire qu'en utilisant l'analyse des sentiments, nous avons pu à détecter, même si c'est de façon mineure, la psychologie et les pensées des gens sur la vaccination COVID.