Introduction :

La vaccination est une pierre angulaire de la prévention des maladies infectieuses transmissibles.

Cependant, les vaccins, ses effets et son efficacité ont traditionnellement rencontré la peur et l'hésitation du public, et les vaccins COVID-19 ne font pas exception. Il a été démontré que l'utilisation des médias sociaux joue un rôle dans la faible acceptation des vaccins.

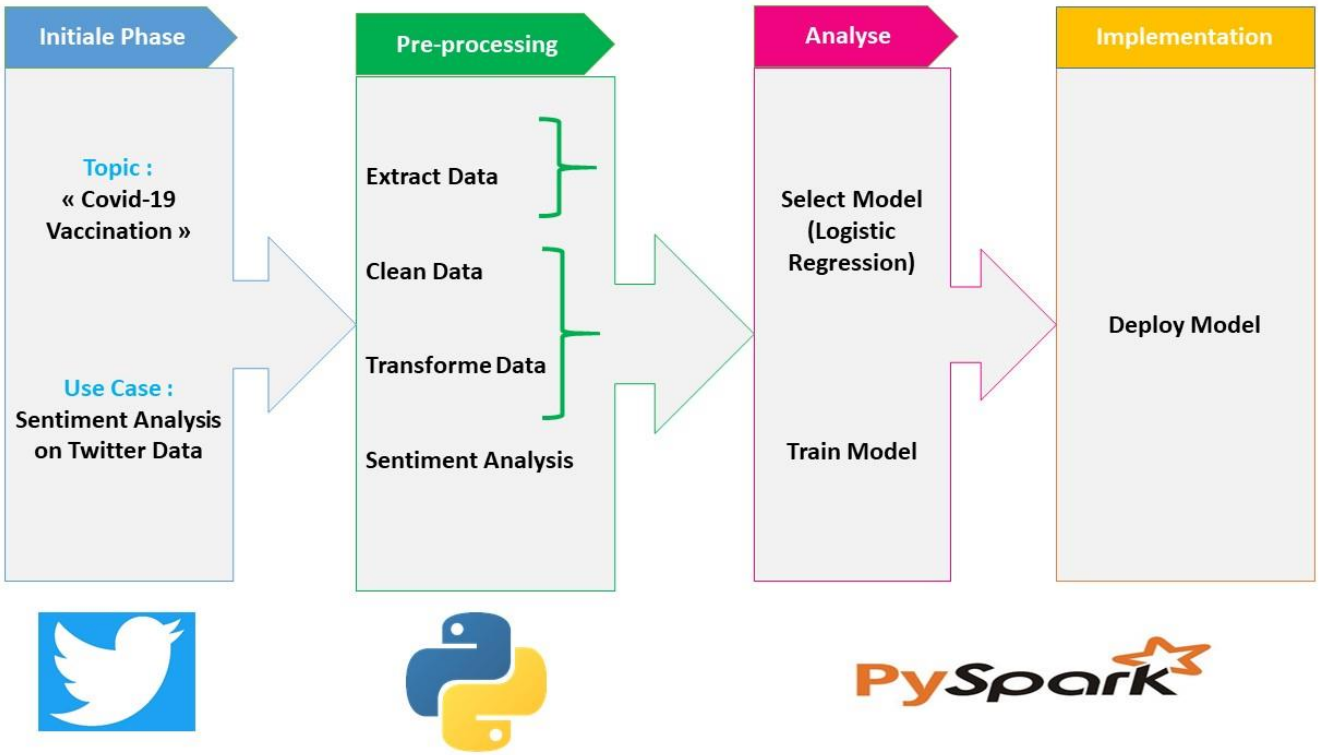
Objectif :

- Extraire des informations de tweets liés au vaccin contre COVID 19 où les opinions sont très peu structurées, hétérogènes et sont soit positives, négatives ou neutres.

- Analyser les sentiments dans la discussion publique afin de mieux comprendre les perceptions, les préoccupations et les émotions du public qui peuvent influencer l'atteinte des objectifs d'immunité collective.

- Création d’un classificateur de sentiments en utilisant la régression logistique.

Architecture :



Les méthodes :

Dans le cadre de la création d'un classificateur de sentiments à l'aide de la régression logistique, nous entraînons le modèle sur un exemple de jeu de données Twitter.

L'ensemble de données disponible est dans son format humain naturel de tweets, ce qui n'est pas facile à comprendre pour un modèle.

Ainsi, nous devrons effectuer un prétraitement et un nettoyage des données pour décomposer le texte donné en un format facile à comprendre pour le modèle.

La tâche	technique	package
Collection des données	Extraction de Tweets de Twitter	snscrape
Pré-traitement des données	Suppression de la ponctuation, des mots vides,	re, nltk,CountVectorizer, pandas, numpy

	des URL, des emojis, de la lemmatisation	
Analyse de texte	Analyse des sentiments	textblob
Classification des sentiments	La régression logistique	pyspark.ml.classification

Extraction des données :

snscrape est une bibliothèque qui permet à quiconque de récupérer des tweets sans avoir besoin de clés API personnelles.

L'ensemble de données utilisé pour l'analyse constitué de 16169 tweets contenant les mots-clés vaccination, vaccinations, vaccine, vaccines.

Pré-traitement des données :

Le nettoyage des données a compris :

a. La suppression des valeurs manquantes.
b. La suppression des hashtags.
c. La représentation de texte par des expressions régulières (RegEx).
d. Le nettoyage des contractions.
e. La suppression de la ponctuation, les liens, les crochets et convertir les mots du texte en minuscule en utilisant la bibliothèque « re ».
f. La réduction de mots en suppriment les suffixes, les préfixes, les terminaisons … en utilisant « nltk.WordNetLemmatizer() » et «ntlk.LancasterStemmer() »
g. La suppression des stopwords en utilisant « nltk ».

Analyse des sentiments:

L'analyse des sentiments peut nous aider à déchiffrer l'humeur et les émotions du grand public et à recueillir des informations pertinentes sur le contexte.

Nous utilisons **textblob** pour la classification du texte des tweets.

TextBlob est un module NLP sur Python utilisé pour l'analyse de sentiment. La fonction de **TextBlob** qui nous intéresse permet pour un texte donné de déterminer le ton du texte et le sentiment de la personne qui l'a écrit.

TextBlob renvoie la **polarité** et la **subjectivité** d'une phrase. La polarité se situe entre [-1,1], -1 définit un sentiment négatif et 1 définit un sentiment positif. Les mots de négation inversent la **polarité**.

TextBlob a des étiquettes sémantiques qui aident à une analyse fine. Par exemple, les émoticônes, point d'exclamation, emojis, etc.

La subjectivité se situe entre [0,1]. La subjectivité quantifie la quantité d'opinions personnelles et d'informations factuelles contenues dans le texte.

La plus grande subjectivité signifie que le texte contient une opinion personnelle plutôt que des informations factuelles. **TextBlob** a un autre paramètre : « l'intensité ».

Il calcule la subjectivité en regardant "l'intensité".

L'intensité détermine si un mot modifie le mot suivant. Pour l'anglais, les adverbes sont utilisés comme modificateurs ("très bien").

Le cas de notre ensemble de données :

- Nombre de tweets avec un sentiment positif 3597
- Nombre de tweets avec un sentiment négatif 4917
- Nombre de tweets avec un sentiment neutre 1487

La régression logistique :

Désormais, nous allons commencer notre travail de classification de sentiments des tweets à l'aide d'une régression logistique.

Pour ce faire, l'idée est comme d'habitude de diviser notre jeu de données en un échantillon d'entraînement (70%) dans lequel nous allons apprendre les paramètres du modèle et un échantillon test (30%) dans lequel nous allons les tester.

Néanmoins, les variables explicatives étant des données textuelles (les tweets), nous allons au préalable créer un nouveau variable explicative numérique en lien avec les tweets (sentiment) afin de pouvoir ensuite prédire les labels.

Pour évaluer notre modèle, on a utilisé :

Precision score : représente la capacité du modèle à prédire correctement les positifs parmi toutes les prédictions positives qu'il a faites.

Recall score: représente la capacité du modèle à prédire correctement les positifs parmi les positifs réels. En d'autres termes, il mesure la capacité de notre modèle d'apprentissage automatique à identifier tous les positifs réels parmi tous les positifs qui existent dans un ensemble de données. Plus le score de rappel est élevé, plus le modèle d'apprentissage automatique est efficace pour identifier les exemples positifs et négatifs.

Accuracy score : est utilisé pour mesurer les performances du modèle en termes de mesure du rapport de la somme des vrais positifs et des vrais négatifs sur toutes les prédictions faites.

F1 score : représente le score du modèle en fonction de la précision et Recall score, il permet de traduire l'équilibre entre la précision et le rappel.

Nous affichons ces valeurs avec ‘’MulticlassClassificationEvaluator ‘’:

LR classifier Accuracy (test) = 0.806935

LR classifier F1 (test) = 0.791702

LR classifier weightedPrecision (test) = 0.80467

LR classifier weightedRecall (test) = 0.806935

Cela signifie que cette modèle a une bonne mesure de séparabilité des classes.

Conclusion :

Cette étude s'est concentrée sur la démonstration des conversations autour des vaccins COVID-19 sur Twitter à l'aide d'un ensemble de données créé avec des tweets d'individus tirant parti de l'approche de l'apprentissage automatique et de l'analyse de texte.

L'ensemble de données a été analysé plus avant pour les sentiments positifs et négatifs. Nous avons également effectué une classification des sentiments pour une compréhension plus approfondie.

Références :

1. Twitter sentiment analysis using Logistic Regression | by Kolamanvitha | Nerd For Tech | Medium
2. Analyze The Sentiment of Tweets From Twitter Data and Tweepy in Python | Earth Data Science - Earth Lab
3. Introduction au NLP avec Python pour l'analyse de sentiments (larevueia.fr)
4. NLP Twitter - Analyse de Sentiment - DataScientest
5. Text Data Cleaning - tweets analysis | Kaggle
6. Final Project Report.pdf
7. Sentiment Analysis using TextBlob · Twitter Sentiment Analysis Visualization Tutorial (gitbooks.io)
8. Lizd9o textblob sentiment
9. Sentiment Analysis of tweets, Wordclouds, Textblob | Kaggle
10. Sentiment-Analysis-using-Pyspark-on-Multi-Social-Media-Data/pyspark_logistic_reg_twitter.ipynb at master · chaithanya21/Sentiment-Analysis-using-Pyspark-on-Multi-Social-Media-Data (github.com)
11. pyspark-twitter-sentimental-analysis/Sentimental_Analysis.ipynb at master · apurva-modi/pyspark-twitter-sentimental-analysis (github.com)
12. Twitter Sentiment PySpark | Kaggle
13. Sentiment Analysis using TextBlob | by Parthvi Shah | Towards Data Science
14. Accuracy, Precision, Recall & F1-Score - Python Examples - Data Analytics (vitalflux.com)