REPUBLIQUE DU CAMEROUN

MINISTERE DE L'ENSEIGNEMENT SUPERIEURE

ECOLE NATIONALE SUPERIEURE POLYTECHNIQUE DE YAOUNDE

REPUBLIC OF CAMEROUN

MINISTRY OF HIGHER EDUCATION

NATIONAL ADVANCE POLYTECHNIC SCHOOL OF YAOUNDE

ANI-IA 467 Technologies et défies de l'IA de demain

PROJET : ANALYSE ET PREDICTIONS DES SENTIMENTS

Ecrit par :

NGUAZONG TSAFACK AUREL 20P001 AIA4

Superviseur :

M. MBITHA

Année académique :

2023/2024

Table des matières

INTRODUCTION

I. Présentation des données

Cette première étape comporte l'importation des librairies utilisées dans tout le projet, l'importation des datasets.

1. Importation des bibliothèques

Dans ce projet sont concernés les bibliothèques suivantes :

- Pandas : Pour manipuler le jeu de données
- Numpy : Pour effectuer des opérations dans le jeu de données
- PathLib : Pour la gestion du chemin d'accès dans les fichiers
- Seaborn :
- Mathplotlib : Pour la visualisation graphique
- Plotly
- Nltk :
- String
- Sklearn
- Spicy
- Warning
- Re
- Tqdm

2. Importation du jeu de donnée

Le jeu de donnée fourni tel quel n'est pas exploitable donc il faut faire certaines opérations préalables comme :

- Importation des fichiers .tsv
- Concaténation de tous ces fichiers sous une dataset qui servira de jeu de donnée principal sous .csv
- Importation de la dataset et début de la manipulation

II. Investigation et prétraitement de la dataset

Cette partie se concentre sur l'exploration et le traitement de la dataset pour la fin avoir un dataset propre et exploitable.

1. Investigation

Cette investigation nous donne des résultats comme quoi le jeu de donnée ne présente aucune valeur manquante mais possède des valeurs dupliquées que nous allons nous atteler à supprimer.

2. Prétraitement des données

Avec le jeu de donnée privé des valeurs manquantes, c'est le moment de traiter ces données pour mieux les exploités, et ce par :

- Suppression des ponctuations : En utilisant la fonction string, l'on annote toutes les ponctuations possibles et l'on écrit une fonction qui parcoure le jeu de donnée tout en supprimant ces ponctuations
- Tokenisation du jeu de donnée : Important pour la suite car sans transformer le jeu de donnée en liste de chaîne de caractère, il y'aura une erreur dans le processus de stemming.
- Méthode de Stemming : Cette méthode sert à raccourcir les mots du jeu de donnée à leur racine pour mieux analyser notre dataset. Contrairement à la méthode de lemmatisation qui ne marche pas pour une raison incompréhensible.

III. Analyse et entraînement du modèle

1. L'analyse du modèle

L'analyse utilisé dans ce cas est l'analyse univariée qui nous donne le résultat : Les tweets neutre sont les dominant avec une occurrence de +20000, ensuite vient les tweets positifs avec un occurrence compris entre 15000 et 20000 et enfin, les tweets négatifs compris entre 5000 et 10000 occurrences.

Par la suite nous utilisons la méthode word cloud visualisation qui sous forme d'un tableau affiche l'occurrence de chaque mot selon la taille qu'il a sur le tableau en question.

Ensuite, l'on compte ces mots grâce à une fonction et l'on affiche leurs occurrences en forme d'histogramme.

2. Entraînement du modèle

- La première étape serait de compter les mots qui sont considérés comme positifs ou négatifs et ce, en nous aidant d'une dataset déjà préconçu téléchargeable sur le site (http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html). Ce fichier contient les mots qui peuvent déterminer si un tweet est positif ou négatif. L'on compte par la suite le nombre de mots total, le nombre de mots positifs puis le nombre de mots négatifs. L'on remarque que les nombres de positives words et de negatives words peuvent être un indicateur pour prédire la classe d'appartenance d'un tweet. L'on constate aussi que les tweets classés négatifs n'ont pas en général beaucoup de mots comparés aux tweets des classes positive et neutral.
- La deuxième étape vient celui de la séparation des datasets d'entraînement puis celui de de test
- La troisième étape est celui de la vectorisation des données au travers de la méthode TF-IDF qui est la plus adapté pour le traitement de texte.
- La quatrième et dernière étape est l'utilisation de différente méthode d'entraînement et déterminé quel sera celle qui produit les meilleurs résultats. Méthode utilisée : Linear SCV, Random Forest et Regression logistic.

IV. Résultat d'entraînement

Le résultat se montre sous ce tableau :

Modèle	Training Accuracy	Test Accuracy
Logistic regression	65%	63%
Random Forest	60%	57,9%
Linear SCV	89%	65%

*

Nous remarquons forcement que le meilleur algorithme d'entraînement est celui de Linear SCV avec les meilleurs rendus