Projet Text mining

Réalisé par :

HADJ MAHFOUD Kenza
BOUDJEMAI Ali
AOUCHICHE Salah

Plan

01

Introduction

02

Informations sur l'ensemble de données

03

Processus d’analyse des sentiments

04

Visualisation des données

05

Classification des sentiments

06

Conclusion

Introduction

Contexte du projet

Mise en place du processus Text mining qui consiste à analyser des tweets afin d'en extraire des informations liées aux opinions et aux sentiments et à identifier les informations subjectives d'un tweet.

Objectif du projet

Réaliser une analyse exploratoire et visuelle des tweets présents dans notre jeu de données. Dans un second temps, le but sera de parvenir à classifier à l'aide de différents modèles de Machine Learning les sentiments des tweets.

Informations sur l'ensemble de données

- **ID :** identifiant du tweet
- **Tweet :** Contenu d'un tweet
- **Label :** Attribut de classification

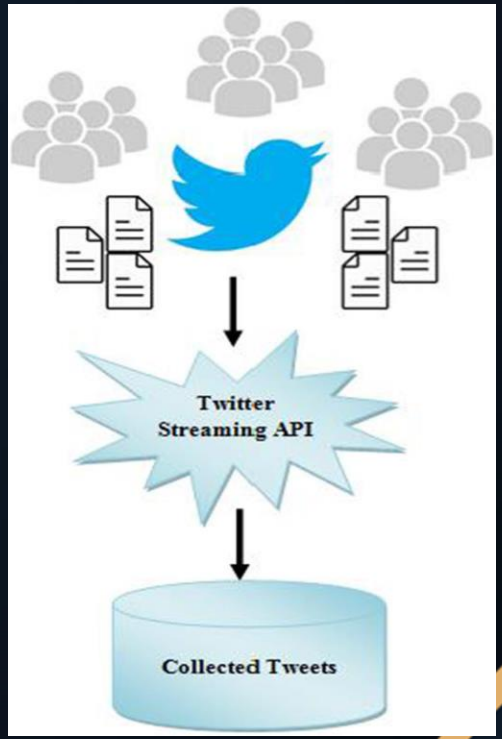
	id	label	tweet
0	1	0	@user when a father is dysfunctional and is s...
1	2	0	@user @user thanks for #lyft credit i can't us...
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in ...
4	5	0	factsguide: society now #motivation

1. **label** : output class **1: Raciste/sexiste** , **0 : No Raciste/sexiste**

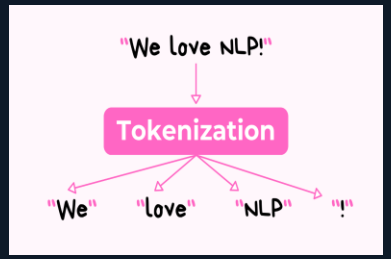
Processus d’analyse des sentiments

Extraction des données

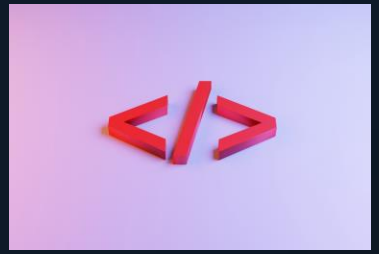
- Collecte des données à partir d'une API.
- Stockage des données sur un fichier csv.
- Exploitation des données pour l'étape du prétraitement



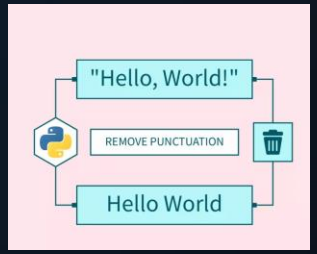
Text pré-processing



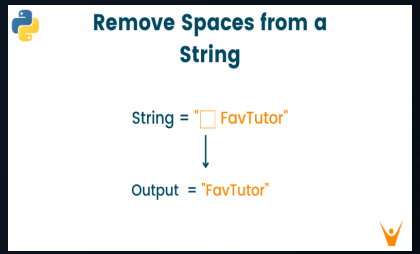
Tokenisation



Removal of HTML tags



Delete special characters

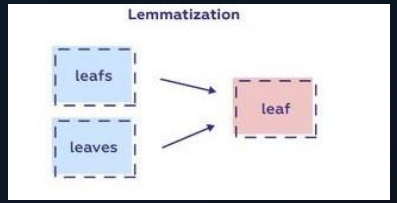


Removal of strings

Text pré-processing

StopWords

Spell checker

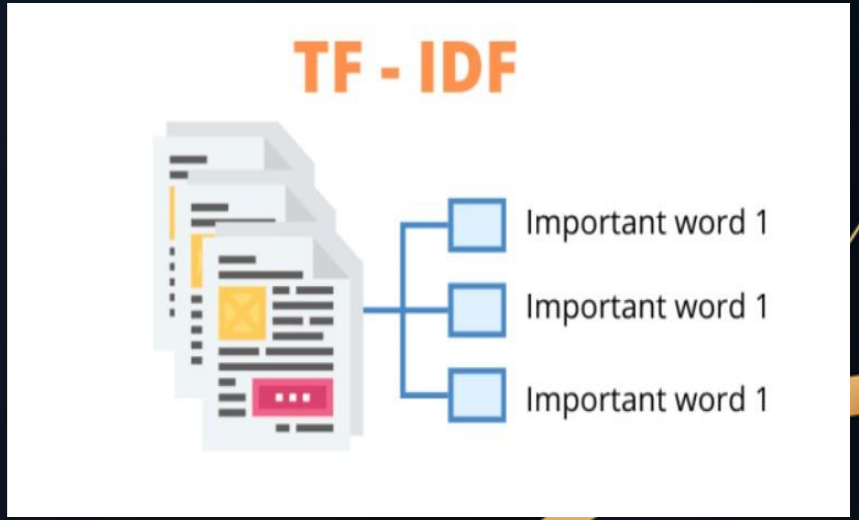


Lemmatization

Abréviation

Embedding Technique(TF-IDF)

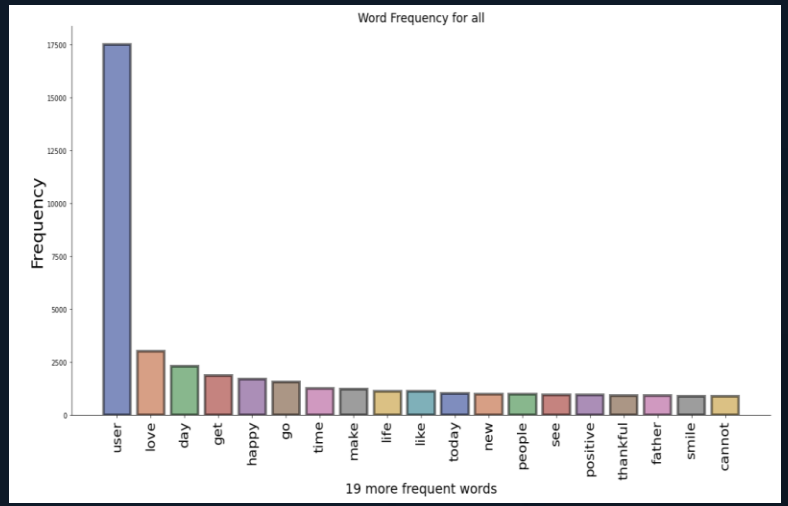
- Convertir les mots ou les phrases en un format numérique
- Représenter les données textuelles sous forme de vecteurs
- Refléter l'importance d'un mot



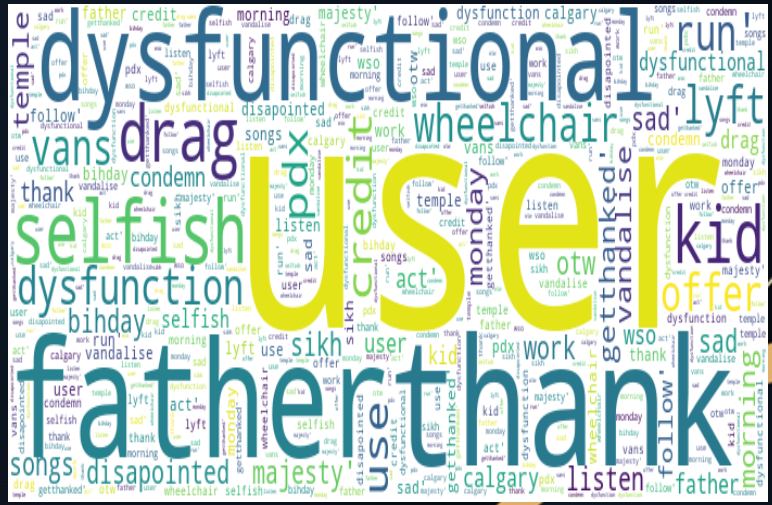
Visualisation des données

Les mots les plus fréquents

Histogramme des 19 mots les plus fréquents

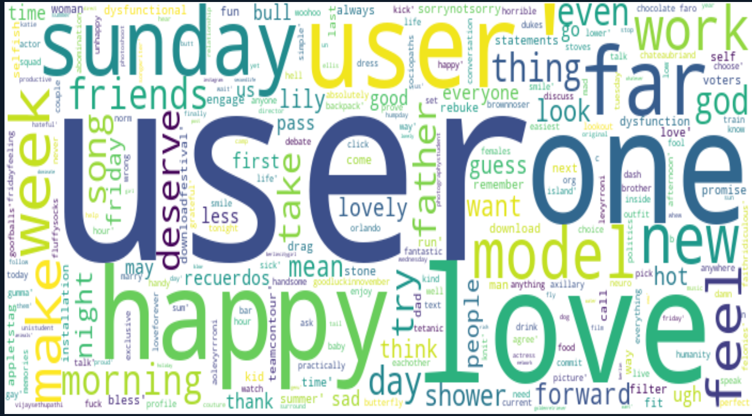


WordCloud des mots les plus fréquents

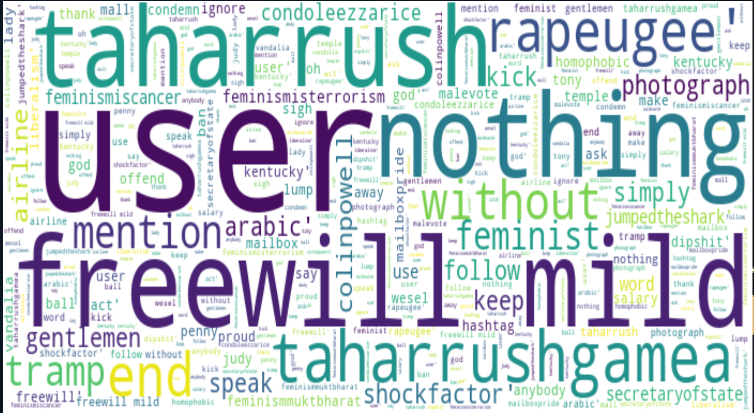


Les mots les plus fréquents sur chaque classe.

WordCloud des mots No raciste/sexiste les plus fréquents

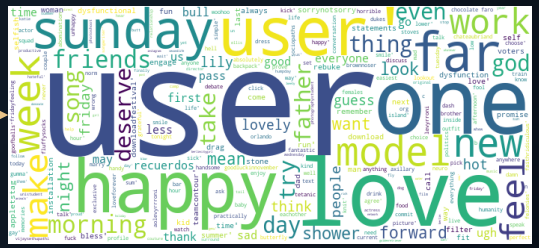


WordCloud des mots raciste/sexiste les plus fréquents



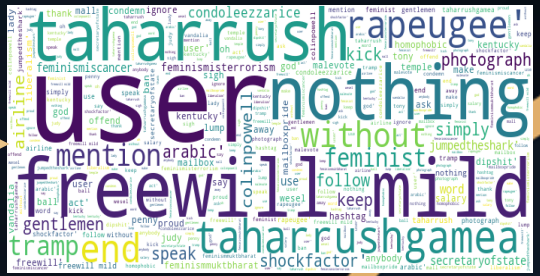
Suppression des mots les plus fréquents inutiles

NON Raciste/Sexiste



- Le mot User est le plus fréquent dans les deux classes.
- Le mot User n'est pas significatif.

Raciste/Sexiste

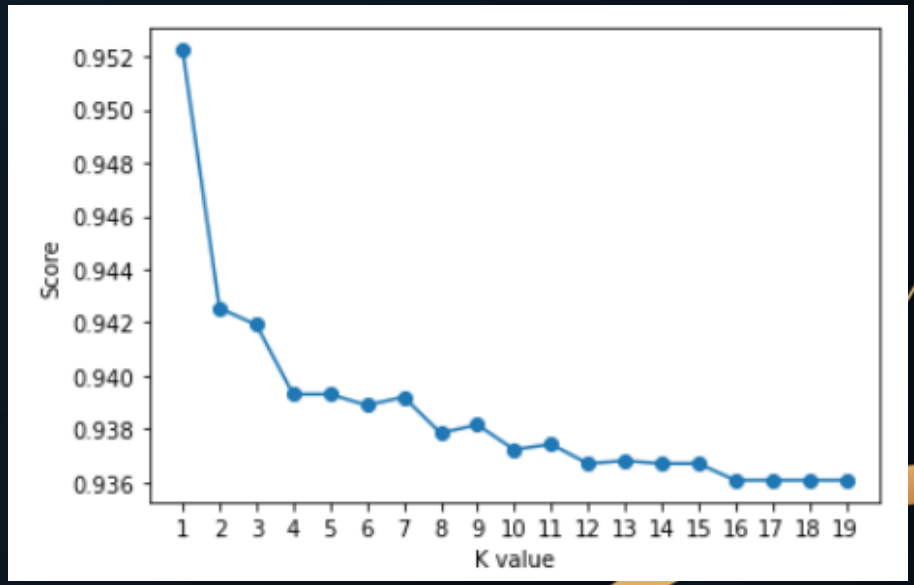


Classification des sentiments

Algorithmes Machine Learning

KNN

On remarque que le GreadSearch nous a donné la meilleure configuration de knn qui est la suivante : K = 1 avec un score de prédiction de 95%



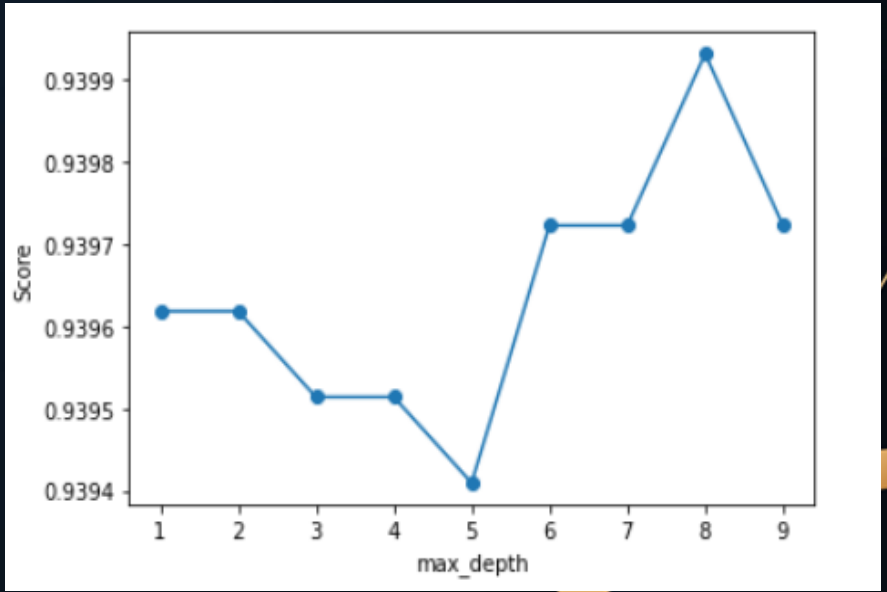
Algorithmes Machine Learning

DecisionTree

On remarque que le GreadSearch nous a donné la meilleure configuration de DecisionTree qui est la suivante :

- max_depth: 8

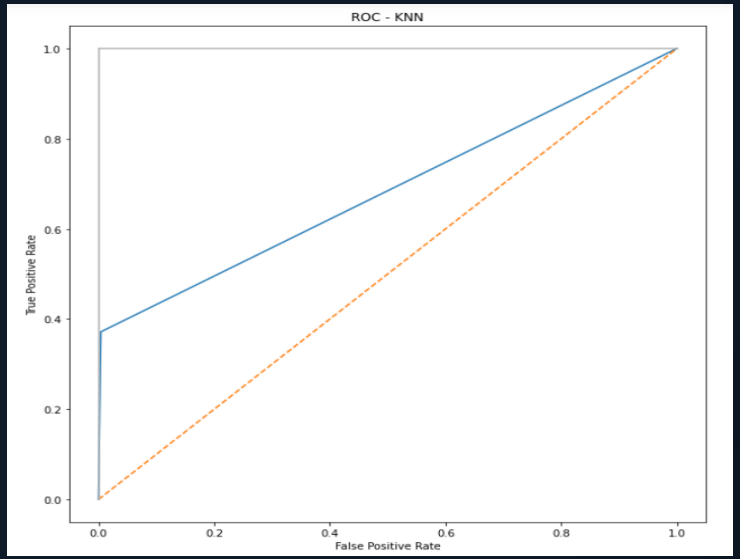
Nous avons un score de prédiction de 94%.



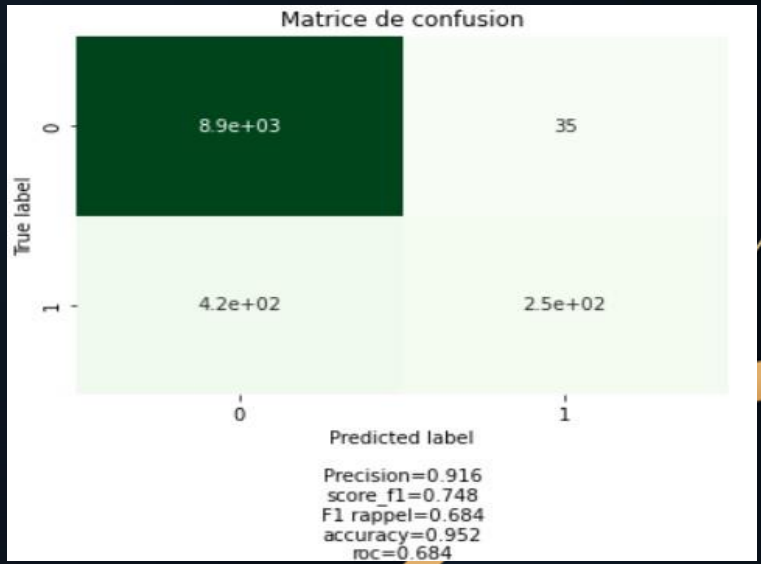
Évaluation des Algorithmes

KNN

La courbe ROC

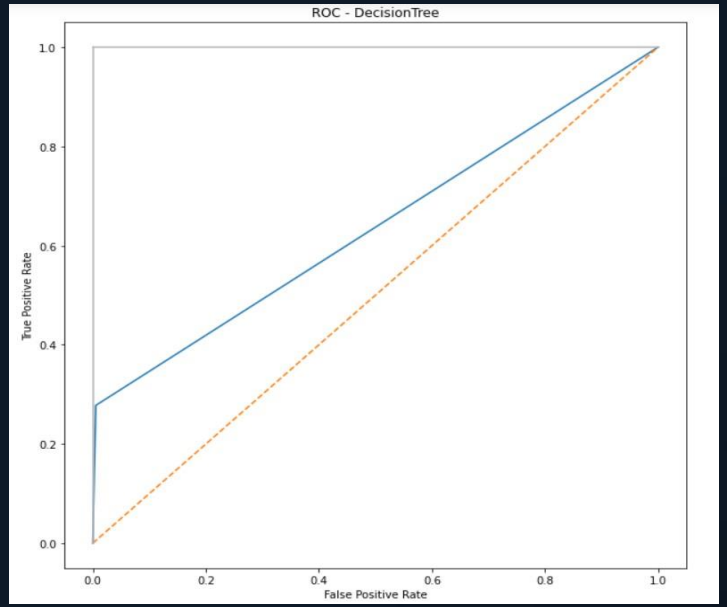


Matrice de confusion et les métriques

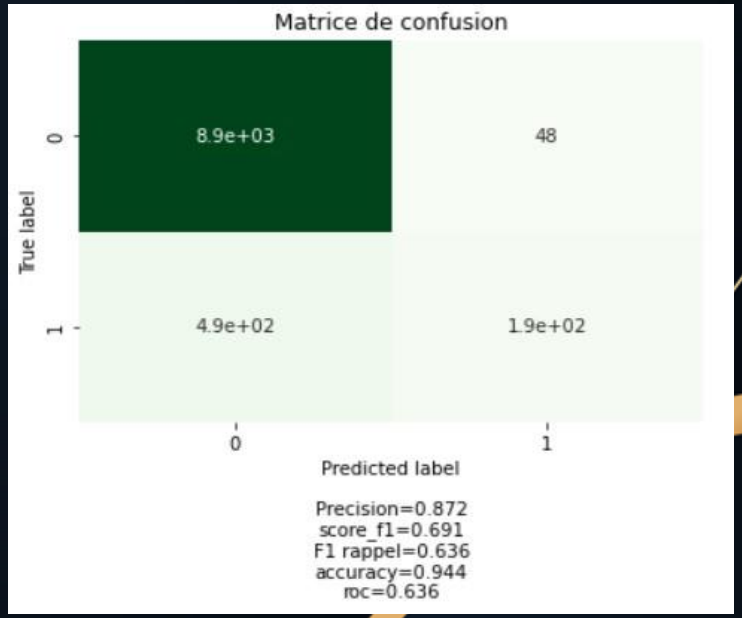


DecisionTree

La courbe ROC

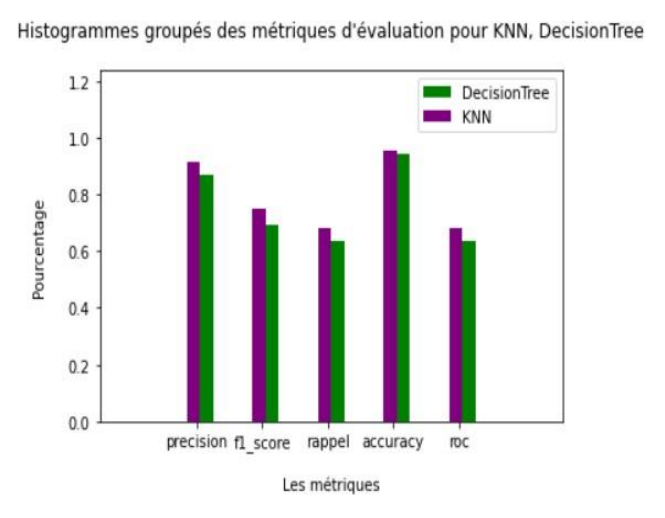


Matrice de confusion et les métriques

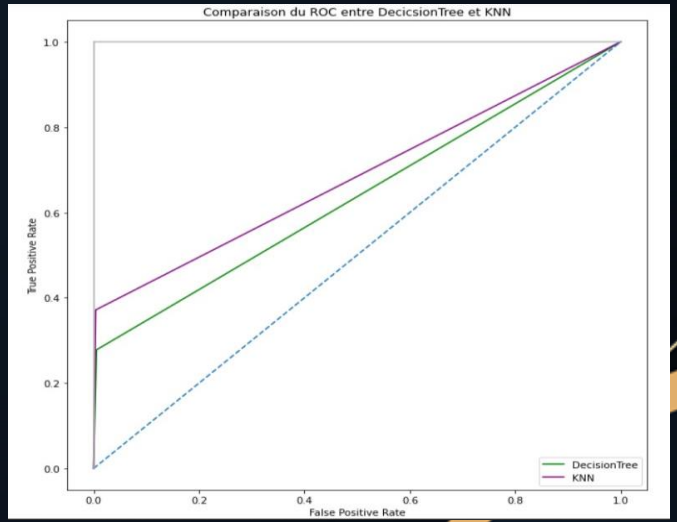


Comparaison des modèles

Métriques



La courbe ROC



Conclusion

L'analyse des sentiments peut aider à comprendre les opinions et les émotions des gens à propos d'un sujet donné, ce qui peut être utile pour prendre des décisions informées en matière de marketing, de relations publiques, de politique, etc. Cependant, il est important de prendre en compte les limitations de l'analyse des sentiments, telles que la subjectivité de la polarité et la complexité de la compréhension du langage humain.

Merci !