

Projet Big Data

Sujet

Conception d'un détecteur de fake news avec création d'un ensemble de données associées

Oumar SALL ,N00AD4A20181 BDA

Salimata FAYE ,ET014922, BDA

Fatou DIOUF108 ,ET014722 ,BDA

Table de matières

Partie1 Introduction

1. Introduction
2. Définition du problème
3. Objectif
4. But
5. Besoin du système

Partie2 Les étapes du projet

1. Étape 1 : Collecte et Annotation des données.
2. Étape 2 : Analyse exploratoire des données
3. Étape 3 : Prétraitement des données
4. Étape 4 : Conception, entraînement et évaluation d'un model de detection de fake news

Partie3 Conception de l'application

Définition de Flask

Création d'une interface utilisateur avec Flask.

Conclusion

Sources bibliographiques

Partie# 1

Introduction

Contenus:

- Introduction
- Définition du problème
- Objectif
- But
- Besoin du système

Introduction

Le système de détection de fausses informations est devenu indispensable dans le paysage actuel de l'information numérique, où la propagation de fausses nouvelles peut avoir des conséquences néfastes sur la société. Notre projet vise à concevoir un système robuste et efficace pour détecter les fausses informations, en utilisant des techniques avancées de traitement du langage naturel et d'apprentissage automatique.

Ce système repose sur la création d'un ensemble de données exhaustif et diversifié, qui servira de base à notre modèle de détection de fausses informations. L'ensemble de données comprendra un large éventail d'articles de presse provenant de sources variées, tant réputées que moins fiables, afin de garantir une représentation réaliste des différents types de contenus médiatiques disponibles en ligne.

En utilisant cet ensemble de données, nous mettrons en œuvre une variété de techniques d'apprentissage automatique, notamment des algorithmes de classification tels que la Régression Logistique, pour entraîner notre modèle de détection de fausses informations.

Définition du problème

Le problème abordé par le projet est la prolifération de fausses informations sur Internet, qui peut entraîner des conséquences négatives telles que la panique publique, la méfiance envers les sources crédibles et l'instabilité sociale. Le projet vise à développer un système automatisé capable de distinguer les vraies informations des fausses informations en utilisant des techniques de traitement du langage naturel et d'apprentissage automatique. Cela aidera à prévenir la propagation de la désinformation et à protéger le public contre la manipulation.

Objectif

L'objectif principal de ce projet de conception d'un détecteur de fausses informations avec création d'un ensemble de données associé est de développer un système automatisé capable de détecter et de classer de manière précise les fausses informations présentes dans les articles de presse en ligne.

But

Plus spécifiquement, les buts du projet sont les suivants :

- ✓ Construire un ensemble de données complet et diversifié comprenant une variété d'articles de presse provenant de différentes sources et couvrant différents sujets.
- ✓ Annoter l'ensemble de données pour indiquer la véracité de chaque article, fournissant ainsi une référence pour l'entraînement du modèle de détection de fausses informations.
- ✓ Développer et mettre en œuvre un algorithme de détection de fausses informations basé sur des techniques avancées de traitement du langage naturel et d'apprentissage automatique.
- ✓ Optimiser le modèle de détection de fausses informations pour garantir des performances élevées en termes de précision et de rappel.

- ✓ Créer une interface utilisateur conviviale permettant aux utilisateurs de soumettre des articles pour analyse et de visualiser les résultats de la détection de fausses informations.
- ✓ Sensibiliser les utilisateurs à la propagation de la désinformation en ligne et les aider à développer des compétences critiques pour évaluer la fiabilité des sources d'information.

Besoin du projet

Le besoin du projet de détection de fausses informations découle de la propagation croissante de fausses nouvelles et de désinformation à travers les médias sociaux et d'autres plateformes en ligne. Le système vise à résoudre le problème des fausses informations en détectant automatiquement et en filtrant les informations fausses des vraies. Avec l'aide d'algorithmes d'apprentissage automatique le système peut identifier efficacement les fausses nouvelles, ce qui peut empêcher les gens de prendre des décisions basées sur des informations trompeuses. Ce système est crucial pour maintenir l'authenticité des informations et peut être utilisé par diverses industries telles que les agences de presse, les plateformes de médias sociaux et les institutions éducatives.

Partie# 2

Les étapes du projet

Contenus:

- Étape 1 : Collecte et annotation des données.
- Étape 2 : Prétraitement des données
- Étape 3 : Analyse exploratoire des données
- Etape 4: Conception, entraînement et évaluation d'un model de detection de fake news

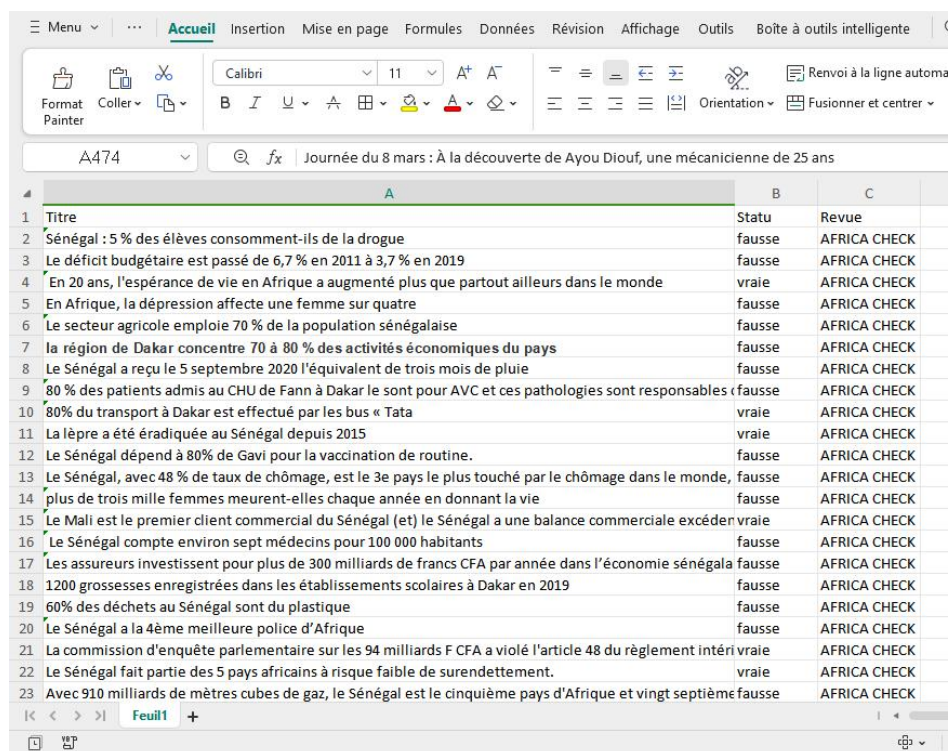
Étape 1 : Collecte et étiquetages des données

Les données utilisées pour la réalisation du projet sont collectées à partir de sites d'information bien connus tels que Senego, Seneweb, Chech Afrika, et PressAfrik. Le processus de collecte et de préparation des données a été effectué comme suite :

- ✓ Récupération des données à partir de sites d'information en ligne, notamment Senego, Seneweb, Chech Afrika, et PressAfrik. Cela a été fait manuellement



- ✓ Stockage des données dans un fichier Excel. Chaque article est enregistré avec des informations telles que le titre, la revue, id, etc.
- ✓ Étiquetage des données : Les articles stockés sont ensuite examinés et étiquetés en fonction de leur véracité ou de leur statut de vérité. Cela a été fait manuellement par nous.



	A	B	C
1	Titre	Statu	Revue
2	Sénégal : 5 % des élèves consomment-ils de la drogue	fausse	AFRICA CHECK
3	Le déficit budgétaire est passé de 6,7 % en 2011 à 3,7 % en 2019	fausse	AFRICA CHECK
4	En 20 ans, l'espérance de vie en Afrique a augmenté plus que partout ailleurs dans le monde	vraie	AFRICA CHECK
5	En Afrique, la dépression affecte une femme sur quatre	fausse	AFRICA CHECK
6	Le secteur agricole emploie 70 % de la population sénégalaise	fausse	AFRICA CHECK
7	la région de Dakar concentre 70 à 80 % des activités économiques du pays	fausse	AFRICA CHECK
8	Le Sénégal a reçu le 5 septembre 2020 l'équivalent de trois mois de pluie	fausse	AFRICA CHECK
9	80 % des patients admis au CHU de Fann à Dakar le sont pour AVC et ces pathologies sont responsables	fausse	AFRICA CHECK
10	80% du transport à Dakar est effectué par les bus « Tata	vraie	AFRICA CHECK
11	La lèpre a été éradiquée au Sénégal depuis 2015	vraie	AFRICA CHECK
12	Le Sénégal dépend à 80% de Gavi pour la vaccination de routine.	fausse	AFRICA CHECK
13	Le Sénégal, avec 48 % de taux de chômage, est le 3e pays le plus touché par le chômage dans le monde,	fausse	AFRICA CHECK
14	plus de trois mille femmes meurent-elles chaque année en donnant la vie	fausse	AFRICA CHECK
15	Le Mali est le premier client commercial du Sénégal (et) le Sénégal a une balance commerciale excédent	vraie	AFRICA CHECK
16	Le Sénégal compte environ sept médecins pour 100 000 habitants	fausse	AFRICA CHECK
17	Les assureurs investissent pour plus de 300 milliards de francs CFA par année dans l'économie sénégalaise	fausse	AFRICA CHECK
18	1200 grossesses enregistrées dans les établissements scolaires à Dakar en 2019	fausse	AFRICA CHECK
19	60% des déchets au Sénégal sont du plastique	fausse	AFRICA CHECK
20	Le Sénégal a la 4ème meilleure police d'Afrique	fausse	AFRICA CHECK
21	La commission d'enquête parlementaire sur les 94 milliards F CFA a violé l'article 48 du règlement intérieur	vraie	AFRICA CHECK
22	Le Sénégal fait partie des 5 pays africains à risque faible de surendettement.	vraie	AFRICA CHECK
23	Avec 910 milliards de mètres cubes de gaz, le Sénégal est le cinquième pays d'Afrique et vingt septième	fausse	AFRICA CHECK

Après avoir collecté et stocké les données à partir des différents sites d'information, nous avons procédé à la lecture de la base de données dans Jupyter Notebook pour commencer les analyses.

Out[26]:

	Titre	Statu	Revue
0	Sénégal : 5 % des élèves consomment-ils de la ...	fausse	AFRICA CHECK
1	Le déficit budgétaire est passé de 6,7 % en 20...	fausse	AFRICA CHECK
2	En 20 ans, l'espérance de vie en Afrique a au...	vraie	AFRICA CHECK
3	En Afrique, la dépression affecte une femme su...	fausse	AFRICA CHECK
4	Le secteur agricole emploie 70 % de la populat...	fausse	AFRICA CHECK

Etape 2 : Prétraitement des données

◆ Prétraitement et Normalisation du texte

C'est un processus important qui comprend la suppression des ponctuations, la réduction de tous les caractères en majuscules, la suppression de tous les mots vides et la radicalisation. La plupart du temps, nous appelons ce processus la tokenisation.

Exemple:

Si on a un texte il nous faut d'abord effectuer un prétraitement. Les étapes de la préparation du texte sont :

Texte exemple:

"Les bienfaits de l'exercice physique sur la santé» «L'exercice physique régulier est essentiel pour maintenir une bonne santé physique et mentale»..."

Conversion en minuscules:

"les bienfaits de l'exercice physique sur la santé» «l'exercice physique régulier est essentiel pour maintenir une bonne santé physique et mentale»..."

Suppression des balise html et bracket:

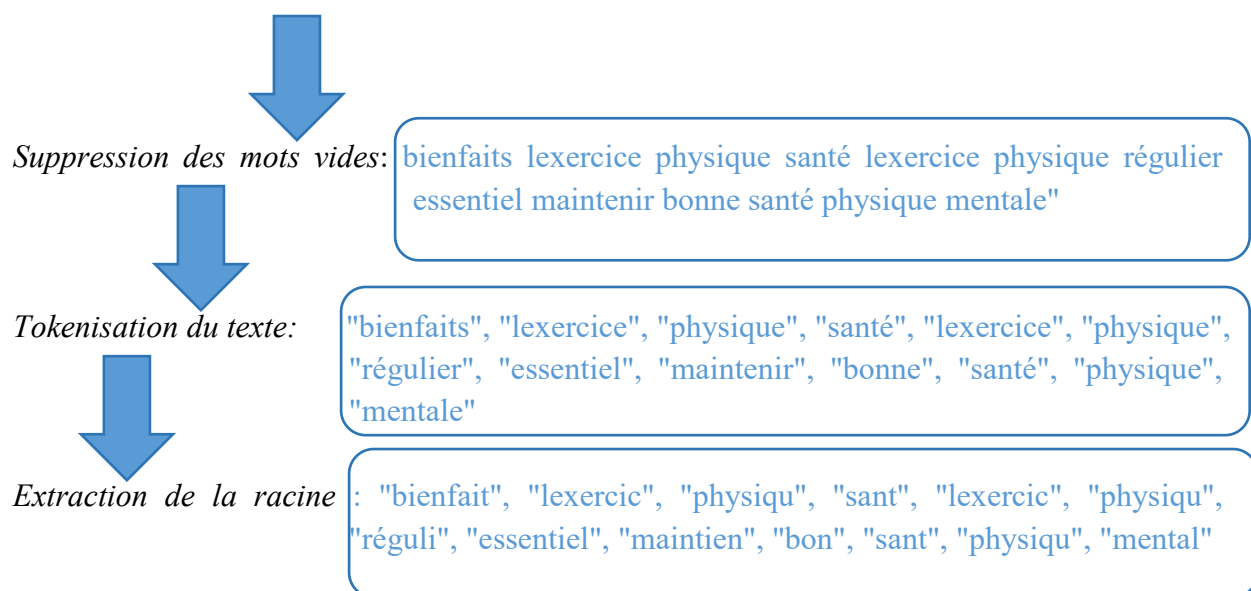
"les bienfaits de l'exercice physique sur la santé
l'exercice physique régulier est essentiel pour
maintenir une bonne santé physique et mentale..."

Suppression de la ponctuation:

"les bienfaits de lexercice physique sur la santé
lexercice physique régulier est essentiel pour
maintenir une bonne santé physique et mentale"

Suppression des caractères spéciaux:

les bienfaits de lexercice physique sur la santé
lexercice physique régulier est essentiel pour
maintenir une bonne santé physique et mentale



Donc au sortie du résultat on aurons ceci:

Out[35]:

	Titre	Statu	Revue	titre_nettoyée
0	Sénégal : 5 % des élèves consomment-ils de la ...	fausse	AFRICA CHECK	sngal 5 lves consommentils drogue
1	Le déficit budgétaire est passé de 6,7 % en 20...	fausse	AFRICA CHECK	dficit budgtaire pass 67 2011 37 2019
2	En 20 ans, l'espérance de vie en Afrique a au...	vraie	AFRICA CHECK	20 ans lesprance vie afrique a augment plus pa...
3	En Afrique, la dépression affecte une femme su...	fausse	AFRICA CHECK	afrique dpression affecte femme quatre
4	Le secteur agricole emploie 70 % de la populat...	fausse	AFRICA CHECK	secteur agricole emploie 70 population sngalaïse

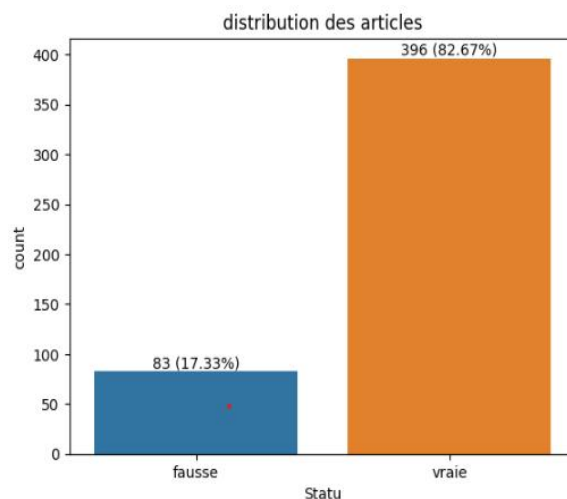
Lien Github

<https://github.com/oumar090199/projet-fake-news/blob/main/fake%20news%20.ipynb>

Etape3 : Analyse exploratoire des données

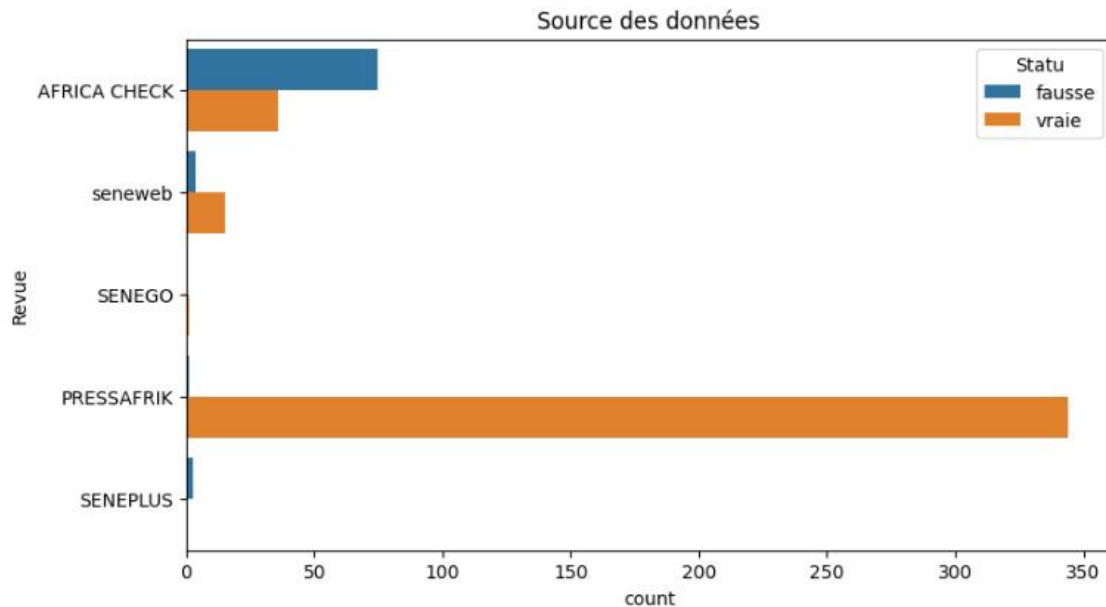
♦ Visualisation des données

Maintenant pour visualiser la distribution des données fausses aux données vraies on va crée un graphique à barres . Cette analyse descriptive permet de mieux comprendre la répartition des différentes valeurs de la colonne "Status" dans les données.



Après visualisation on peut constater que les données avec le statut "fausse" représentent 17,33% (soit 83) tandis que les données avec le statut "vraie" représentent 82.67%(soit 396).

La répartition des revues montre que PRESSAFRIK et AFRIKA CHECK sont le plus représentés.



Cela est dû au fait que la plupart des articles sont issus de ces revues.

◆ Visualisation des nuages de mots des données

Séparation des données (vraies)

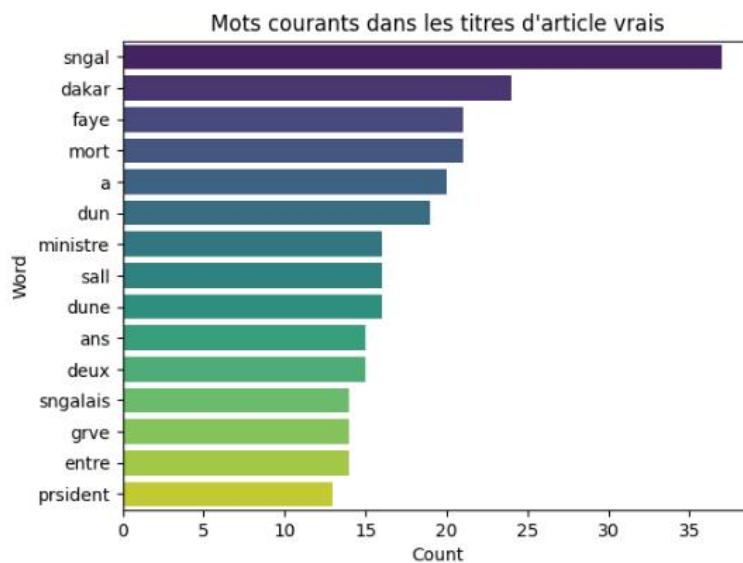
```
Entrée [41]: vrais_articles=data[data['Statu']=='vraie']
vrais_articles.head()
```

Out[41]:

	Titre	Statu	Revue	titre_nettoyée	nombre de mots	nombre de mots nettoyé
2	En 20 ans, l'espérance de vie en Afrique a au...	vraie	AFRICA CHECK	20 ans lesprance vie afrique a augment plus pa...	17	11
8	80% du transport à Dakar est effectué par les ...	vraie	AFRICA CHECK	80 transport dakar effectu bus tata	12	6
9	La lèpre a été éradiquée au Sénégal depuis 2015	vraie	AFRICA CHECK	lpre a radique sngal depuis 2015	9	6
13	Le Mali est le premier client commercial du Sé...	vraie	AFRICA CHECK	mali premier client commercial sngal sngal a b...	24	14
19	La commission d'enquête parlementaire sur les ...	vraie	AFRICA CHECK	commission denquete parlementaire 94 milliards ...	20	15

Entrée [56]: `generer_wordcloud(vrai_article['titre_nettoyée'], "Nuage de mot des articles vrais")`

Nuage de
mots
WordCloud



Les mots
les plus
fréquents

Séparation des données (fausses)

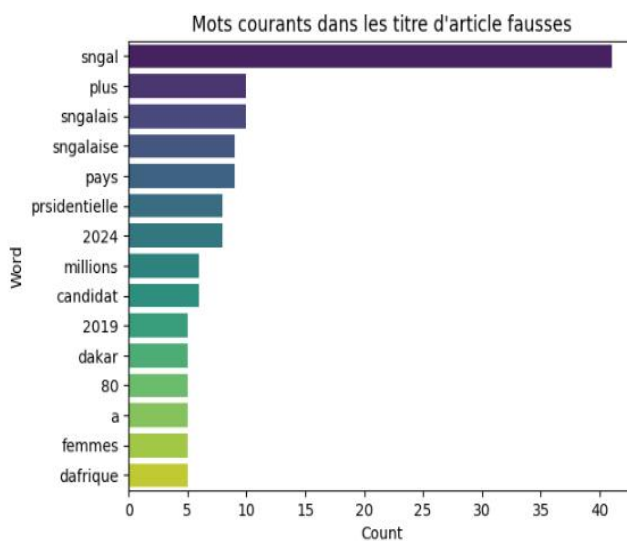
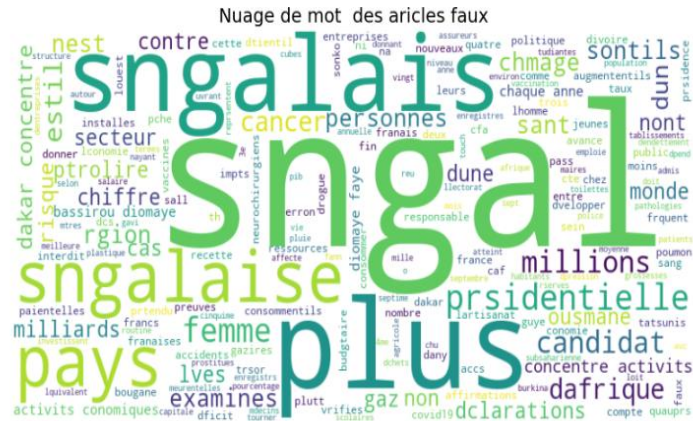
Entrée [42]: `faux_articles=data[data['Statu']=='fausse']
faux_articles.head()`

Out[42]:

	Titre	Statu	Revue	titre_nettoyée	nombre de mots	nombre de mots nettoyé
0	Sénégal : 5 % des élèves consomment-ils de la ...	fausse	AFRICA CHECK	sngal 5 lves consommentils drogue	10	5
1	Le déficit budgétaire est passé de 6,7 % en 20...	fausse	AFRICA CHECK	dficit budgtaire pass 67 2011 37 2019	15	7
3	En Afrique, la dépression affecte une femme su...	fausse	AFRICA CHECK	afrique dpression affecte femme quatre	9	5
4	Le secteur agricole emploie 70 % de la populat...	fausse	AFRICA CHECK	secteur agricole emploie 70 population sngalaise	10	6
5	la région de Dakar concentre 70 à 80 % des act...	fausse	AFRICA CHECK	rgion dakar concentre 70 80 activits onomique...	14	8

```
entrée [62]: generer_wordcloud(faux_articles['titre_nettoyée'], "Nuage de mot des aricles faux")
```

Nuage de
mots
WordCloud



Les mots
les plus
fréquents

Lien Github

<https://github.com/oumar090199/projet-fake-news/blob/main/fake%20news%20.ipynb>

Etape4 : Conception, entraînement et évaluation d'un model de detection de fake news

Tout d'abord avant d'entamer cette étape on va d'abord procéder à un pipeline.

C'est quoi le pipeline ?

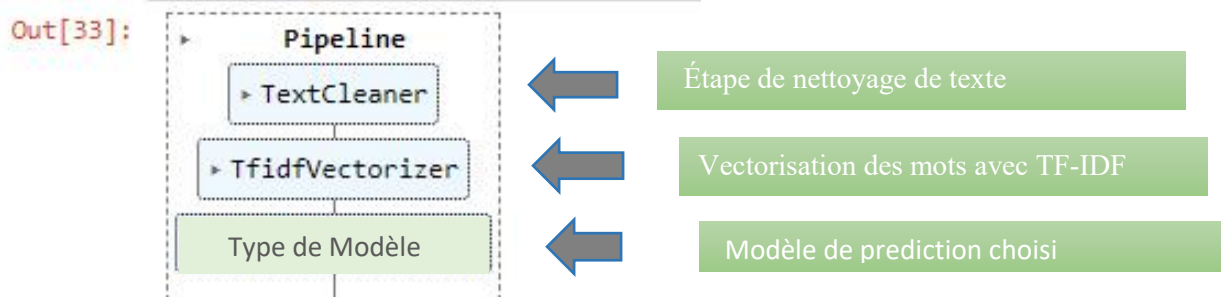
Un pipeline en machine learning est une séquence d'étapes qui permet de transformer et de modéliser les données de manière automatique et cohérente. Voici les composants principaux d'un pipeline typique :

- ✓ **Prétraitement des données** : Cette étape comprend le nettoyage des données (par exemple, la suppression des valeurs manquantes, le traitement des caractères spéciaux etc...), la transformation des variables (par exemple, la normalisation, la discrétisation, l'encodage), et la sélection des caractéristiques pertinentes.
- ✓ **Modélisation** : Une fois que les données sont préparées, le pipeline inclut l'étape où le modèle de machine learning est sélectionné et entraîné sur les données traitées. Cela peut

inclure l'ajustement des hyperparamètres du modèle pour obtenir de meilleures performances.

- ✓ Évaluation : Après avoir entraîné le modèle, le pipeline évalue sa performance en utilisant des métriques appropriées pour la tâche (par exemple, précision, rappel, F1-score pour la classification).
- ✓ Déploiement : Dans certains cas, le pipeline peut également inclure des étapes pour déployer le modèle entraîné dans un environnement de production, prêt à être utilisé pour des prédictions en temps réel.

Dans cadre de notre pipeline on procédera ainsi:



Une fois que nous avons effectué le nettoyage des données, il est essentiel de diviser l'ensemble de données en ensembles d'entraînement et de test. Cette étape est cruciale pour évaluer la performance du modèle de manière impartiale.

```
from sklearn.model_selection import train_test_split

# Diviser les données en ensembles d'entraînement et de test de façon stratifiée
X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.3, random_state=42)
```

Cela nous permettra de faire ensuite une conversion du texte en données numériques exploitables par les algorithmes de machine learning. Cette conversion est nécessaire car la plupart des modèles de machine learning ne peuvent pas traiter directement le texte brut. du texte en numérique.

Cette vectorisation du texte consiste à représenter chaque document textuel par un vecteur numérique. Il existe plusieurs techniques de vectorisation, mais pour nous on utilisera le **TF-IDF (Term Frequency - Inverse Document Frequency)**.

Qu'est ce que sait??

C'est une méthode qui permet de convertir chaque jeton en vecteur calculant l'importance d'un mot dans un document par rapport à une collection de documents. Elle combine la fréquence du terme (TF) et l'inverse de la fréquence dans les documents (IDF).

TF mesure le nombre de fois qu'un terme apparaît dans un document, tandis que IDF évalue l'importance d'un terme dans la collection globale de documents en réduisant le poids des termes fréquents.

En utilisant TF-IDF, on peut attribuer une valeur numérique à chaque terme, reflétant à la fois sa fréquence dans le document et son importance relative dans la collection de documents.

$$w_{i,j} = tf_{i,j} \times \log \left(\frac{N}{df_i} \right)$$

$tf_{i,j}$ = number of occurrences of i in j
 df_i = number of documents containing i
 N = total number of documents

Supposons que nous ayons trois documents :

	Document 1 :	Document 2	Document 3
Phrases	"Le chat est mignon."	"Le chien est fidèle."	"Le chien et le chat jouent ensemble."

Calculons maintenant le score TF-IDF pour les termes **"chat"** et **"chien"** dans chaque document :

	Document 1 :	Document 2	Document 3
Phrases	"Le chat est mignon."	"Le chien est fidèle."	"Le chien et le chat jouent ensemble."
Fréquence des termes (TF)	TF("chat")= 1	TF("chat") = 0	TF("chat") = 1
	TF("chien")= 0	TF("chien") = 1	TF("chien") = 1
Fréquence inverse dans les documents (IDF)	IDF("chat") = $\log(3/2) = \log(1.5) \approx 0.18$		
	IDF("chien") = $\log(3/2) = \log(1.5) \approx 0.18$		
	TF-IDF("chat") = $1 * 0.18 \approx 0.18$	TF-IDF("chat") $= 0 * 0.18 = 0$	TF-IDF("chat") $= 1 * 0.18 \approx 0.18$

TF-IDF	TF-IDF("chien") = $0 * 0.18 = 0$	TF-IDF("chien") $= 1 * 0.18 \approx 0.18$	TF-IDF("chien") $= 1 * 0.18 \approx 0.18$
--------	---	--	--

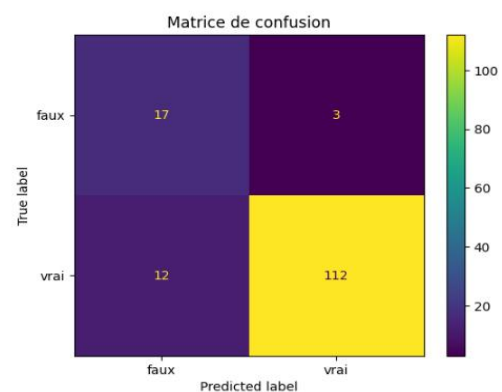
Choix du modèle

Le choix du modèle ici repose sur l'évaluation des performances d'algorithmes d'apprentissages. Donc pour se faire nous allons effectuer une comparaison de ces différents algorithmes d'apprentissages considéré et en prendre le plus performant.

Résultats: accuracy et matrix de confusion

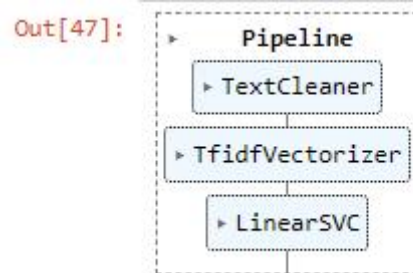
 SVM

	precision	recall	f1-score	support
fausse	0.59	0.85	0.69	20
vraie	0.97	0.90	0.94	124
accuracy			0.90	144
macro avg	0.78	0.88	0.82	144
weighted avg	0.92	0.90	0.90	144



Score de précision moyenne: 0.8958333333333334 soit: 89.58%

Selon l'évaluation établie sur différentes models , le SVM semble le plus performant. Donc notre choix se portera sur ce model.



Maintenant on peut faire des prédictions pour une entrée donnée en vue de voir si cette information est fiable ou non.

Ces prédictions seront faite sur une application que l'on réalisera.

Lien Github

<https://github.com/oumar090199/projet-fake-news/blob/main/fake%20news%20.ipynb>

Partie# 3

Mise en Application

Contenus:

- Définition de Flask
- Création d'une interface utilisateur avec Flask.
 - Installer Flask dans votre environnement Python
 - Création un fichier Python, et importation de Flask
 - Création un dossier de répertoire de travail, puis création d'un fichier HTML
 - Exécution de l'application Flask

Dans cette partie nous allons voir comment sera déployer notre application avec l'utilisation de Flask.

C'est quoi Flask?

Flask est un framework web léger et flexible écrit en Python. Il est conçu pour faciliter la création d'applications web, des plus simples aux plus complexes. Voici quelques points clés à connaître sur Flask :

- ✧ Léger et minimaliste : Flask est conçu pour être simple et minimaliste. Il fournit les fonctionnalités de base nécessaires au développement d'applications web, sans imposer de structure ou de dépendances excessives.
- ✧ Facilité d'utilisation : Flask est réputé pour sa facilité d'utilisation. Son API est simple et intuitive, ce qui permet aux développeurs de se concentrer sur la création d'applications plutôt que sur la configuration et la gestion du framework.
- ✧ Extensibilité : Bien que Flask soit léger, il est également très extensible. Il offre une grande flexibilité pour étendre ses fonctionnalités à l'aide d'extensions (ou "plugins") qui ajoutent des fonctionnalités spécifiques telles que l'authentification utilisateur, la gestion des formulaires, les bases de données, etc.
- ✧ Basé sur Werkzeug et Jinja2 : Flask repose sur deux bibliothèques principales : Werkzeug, qui fournit des outils pour gérer les requêtes HTTP, les sessions, les cookies, etc., et Jinja2, un moteur de templates puissant qui facilite la création de pages HTML dynamiques.
- ✧ Développement rapide : Grâce à sa simplicité et à sa facilité d'utilisation, Flask permet un développement rapide d'applications web. Il est idéal pour les projets de petite à moyenne taille, les prototypes et les applications simples où la rapidité de développement est une priorité.

En résumé, Flask est un framework web léger, flexible et facile à utiliser, idéal pour développer des applications web en Python de manière efficace et rapide.



Création d'une interface utilisateur avec Flask.

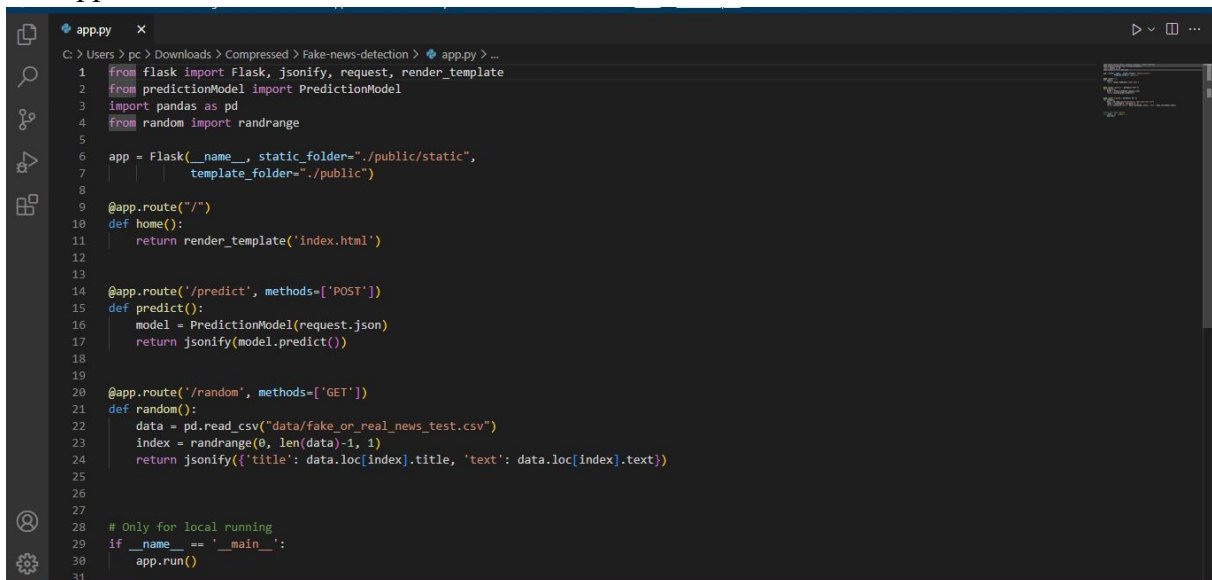
Pour créer une interface utilisateur avec Flask, on a suivi les étapes suivantes:

- ✓ Installer Flask dans votre environnement Python en utilisant pip :



```
MINGW64:/c/Users/pc  
pc@DESKTOP-OGVSLN3 MINGW64 ~  
$ pip install flask
```

- ✓ Créer un fichier Python, par exemple app.py, et importez Flask. Initialisez ensuite votre application Flask :



```
1 from flask import Flask, jsonify, request, render_template  
2 from predictionModel import PredictionModel  
3 import pandas as pd  
4 from random import randrange  
5  
6 app = Flask(__name__, static_folder='./public/static',  
7           template_folder='./public')  
8  
9 @app.route("/")  
10 def home():  
11     return render_template('index.html')  
12  
13  
14 @app.route('/predict', methods=['POST'])  
15 def predict():  
16     model = PredictionModel(request.json)  
17     return jsonify(model.predict())  
18  
19  
20 @app.route('/random', methods=['GET'])  
21 def random():  
22     data = pd.read_csv("data/fake_or_real_news_test.csv")  
23     index = randrange(0, len(data)-1, 1)  
24     return jsonify({'title': data.loc[index].title, 'text': data.loc[index].text})  
25  
26  
27  
28 # Only for local running  
29 if __name__ == '__main__':  
30     app.run()  
31
```

- ✓ Créer un nouveau dossier appelé « templates » dans notre répertoire de travail, puis créer un nouveau fichier HTML appelé « main.html ».



Conclusion

Dans un monde où l'information circule rapidement et largement, la détection des fausses informations devient une nécessité critique. La dynamique changeante de l'environnement numérique nécessite une surveillance continue et des mises à jour régulières pour rester efficace face aux tactiques d'évolution des diffuseurs de fausses informations.

Sources bibliographiques

https://github.com/dots13/fake_news_app/tree/master

https://www.linkedin.com/posts/muhammad-azam-62a875268_liberia-chapter-detecting-fake-news-using-activity-7107571477880672256-OznJ/

<https://omdena.com/chapter-challenges/detecting-fake-news-using-ai/>

<https://liberia-fake-news-detection.streamlit.app/>

<https://towardsdatascience.com/how-to-build-a-fake-news-detection-web-app-using-flask-c0cfd1d9c2d4> *How to Build a Fake News Detection Web App Using Flask*

<https://medium.com/analytics-vidhya/building-a-fake-news-classifier-deploying-it-using-flask-6aac31dfe31d> *Building a Fake News Classifier & Deploying it Using Flask*

<https://dev.to/iyissa/fake-news-detection-with-machine-learning-and-flask-e5l> *Fake News Detection with Machine Learning and Flask*

<https://flask.palletsprojects.com/en/3.0.x/> *Welcome to Flask*

<https://www.youtube.com/watch?v=U6ieiJAhXQ4> *Fake News Detection Using Machine Learning | Machine Learning Projects In Python | Simplilearn*

<https://www.youtube.com/watch?v=RvvYV73JWgg> *Full Stack Fake News Detection Using Machine Learning || Final Year Project*

<https://www.youtube.com/watch?v=cxgC3Zi3Sqw> *Traitement automatique des langues [5.6] Classification de documents - pondération TF-IDF*

<https://www.youtube.com/watch?v=8JcLENGoXL0&t=538s> *NLP Techniques | TF-IDF and bag of words Hands on | Natural Language Processing*