TRAITEMENT AUTOMATIQUE DE LANGAGE

RAPPORT: MINI-PROJET "FAKE NEWS"

Mohamed DHLIMA , Mohamed TLILI

Etudiants en MPDS 2

Faculté des Sciences de Bizerte

Université de Carthage

Décembre 2021

Contents

ABSTRACT

Les fausses nouvelles sur les réseaux sociaux et divers autres médias se répandent largement et constituent un sujet de grave préoccupation en raison de sa capacité à causer beaucoup de dommages sociaux et nationaux avec effets destructeurs. De nombreuses recherches se concentrent déjà sur sa détection. Ce papier fait un analyse des recherches liées à la détection des fake news et explore la machine traditionnelle modèles d'apprentissage pour choisir le meilleur, afin de créer un modèle de produit avec supervision algorithme d'apprentissage automatique, qui peut classer les fausses nouvelles comme vraies ou fausses, en utilisant des outils comme python scikit-learn, PNL pour l'analyse textuelle. Ce processus entraînera l'extraction de caractéristiques et vectorisation ; nous proposons d'utiliser la bibliothèque Python scikit-learn pour effectuer la tokenisation et la fonctionnalité extraction de données texte, car cette bibliothèque contient des outils utiles comme Count Vectorizer et Tiff Vectoriseur. Ensuite, nous effectuerons des méthodes de sélection de fonctionnalités, pour expérimenter et choisir le meilleur ajuster les caractéristiques pour obtenir la plus haute précision, selon les résultats de la matrice de confusion.

1 Introduction

Les fausses nouvelles contiennent des informations trompeuses qui pourraient être vérifiées. Cela maintient le mensonge sur un certain statistique dans un pays ou le coût exagéré de certains services pour un pays, ce qui peut entraîner des troubles pour certains pays comme au printemps arabe. Il y a des organisations, comme la Chambre des communes et le Projet de recoupement, en essayant de résoudre les problèmes en confirmant que les auteurs sont responsables. Cependant, leur la portée est si limitée car ils dépendent de la détection manuelle humaine, dans un globe avec des millions d'articles qu'ils soient supprimés ou publiés toutes les minutes, cela ne peut pas être responsable ou faisable manuellement. UNE la solution pourrait être, par le développement d'un système pour fournir une notation d'indice automatisée crédible, ou évaluation de la crédibilité des différents éditeurs et contexte de l'actualité. Cet article propose une méthodologie pour créer un modèle qui détectera si un article est authentique ou faux sur la base de ses mots, phrases, sources et titres, en appliquant des algorithmes d'apprentissage automatique supervisés sur un ensemble de données annotées (étiquetées), classées et garanties manuellement. Ensuite, sélection de fonctionnalités des méthodes sont appliquées pour expérimenter et choisir les meilleures caractéristiques d'ajustement pour obtenir la plus haute précision, selon les résultats de la matrice de confusion. Nous proposons de créer le modèle en utilisant différentes classifications algorithmes. Le modèle de produit testera les données invisibles, les résultats seront tracés et, par conséquent, le produit sera un modèle qui détecte et classe les articles contrefaits et peut être utilisé et intégré à n'importe quel système pour une utilisation future.

2 Chapitre 1 : Contexte du Projet

2.1 Introduction

Le but de ce chapitre est de présenter d'une part quelques rappels indispensables et nécessaires à la compréhension de ce mémoire et d'autre part faire une synthèse bibliographique de quelques problèmes liés au domaine des Fake News, ceux dédiés plus précisément à l'étude de la gestion des réseaux sociaux. Nous présentons dans un premier temps tout ce qui concerne les fausses nouvelles et ensuite, la façon de les détecter. Nous allons mettre l'accent sur le Web Mining en se basant sur l'Opinion Mining

2.2 Problématique

L'expression « fake news » fleurit depuis quelques mois pour désigner une information délibérément fausse circulant dans les médias ou dans les réseaux sociaux. Les « fake news » font partie d'un phénomène mondial et leur impact est planétaire. Les hommes politiques, les stars du show business, les entreprises, les institutions en sont souvent victimes, l'agriculture aussi. C'est pourquoi nous avons pensé utile de donner quelques éclairages sur ce phénomène médiatique qui interpelle aussi le fonctionnement de notre société.

L'expression, en provenance directe des Etats-Unis, n'a pas vraiment d'équivalent en français. Les « fake news » sont des informations délibérément fausses ou truquées (en anglais « fake » veut dire faux, truqué) émanant des médias, d'un groupe organisé ou d'un individu. Elles peuvent être de simples canulars mais aussi participer à des tentatives de désinformation avec l'intention d'induire en erreur le récepteur dans le but d'obtenir de sa part un avantage financier ou politique. Claire Wardle de First Draft a établit une typologie de « fake news », qui va du mauvais journalisme à la propagande en passant par la parodie ou le contenu politique orienté.

Les « fake news » font aussi partie des instruments de la guerre moderne. Les Russes en sont les champions.

Il existe aussi des gens qui diffusent des « fake news » dans une optique purement mercantile, pour faire de l'argent. Une « fake news » peut ainsi être conçue comme « appeau à clics » pour attirer les consultations des internautes et accroître les revenus publicitaires d'une page web.

Elles sont parfois utilisées dans l'hameçonnage par courriel, en présentant du contenu très attractif ou sensationnaliste pour inciter les utilisateurs à cliquer sur un lien, ce qui permet ensuite à l'envoyeur d'infecter leur ordinateur.

Elles peuvent être le fait d'un site humoristique qui lance un canular. Exemple : le « projet » de la dirigeante de l'extrême droite Marine Le Pen « d'entourer la France d'un mur payé par l'Algérie » inventé par le site parodique Le Gorafi, repris par erreur dans un journal algérien.

2.3 Réseaux sociaux et fausses nouvelles

Les médias sociaux comprennent les sites Web et les programmes consacrés aux forums, aux sites Web sociaux, au microblogging, aux signets sociaux et aux wikis [1][2]. D'un autre côté, certains chercheurs considèrent les fausses nouvelles comme le résultat de problèmes accidentels tels qu'un choc éducatif ou d'actions involontaires comme ce qui s'est passé dans le cas du tremblement de terre au Népal [3][4]. En 2020, il y a eu de fausses nouvelles répandues concernant la santé qui ont exposé la santé mondiale à un risque. L'OMS a publié un avertissement début février 2020 selon lequel l'épidémie de COVID-19 a provoqué une « infodémie » massive, ou une vague de vraies et de fausses nouvelles, qui incluaient beaucoup de désinformation.

2.4 Les composants des fausses nouvelles

Le terme Fake News se base sur quatre composants principaux: créateur/épandeur, victime cible, contenu de l'actualité et contexte social.

- Créateur / Diffuseur: les créateurs de fausses nouvelles en ligne peuvent être des humains ou non.
- Victimes cibles: les victimes des fausses nouvelles peuvent être des utilisateurs de médias sociaux ou d'autres récentes plates-formes. Selon les objectifs de la nouvelle, les cibles peuvent être des étudiants, des électeurs, des parents, des personnes âgées...etc.
- Contenu de l'actualité: le contenu de l'actualité fait référence au corps de la nouvelle. Il contient à la fois un contenu physique (par exemple, titre, corps de texte, multimédia) et contenu non physique (par exemple, but, sentiment, sujets).
- Contexte social: le contexte social indique comment les nouvelles sont diffusées sur Internet. L'analyse du contexte social inclut l'utilisateur, l'analyse du réseau (comment les utilisateurs en ligne sont impliqués dans les actualités) et l'analyse du modèle de diffusion.

2.4.1 Créateur et Diffuseur de fausses nouvelles

Il est important de montrer qui est derrière la fausse nouvelle et pourquoi elle est écrite et partagée tout au long de la vie sociale. Le créateur/diffuseur des fausses nouvelles peut être soit un humain, soit non.

- Humain: les robots sociaux ou les Cyborgs ne sont que les porteurs de fausses nouvelles sur les médias sociaux. Ces comptes automatiques sont programmés pour diffuser de faux messages par des humains, qui cherchent à perturber la crédibilité de la communauté sociale en ligne.
- Non-humain: les robots sociaux ou les Cyborgs sont les créateurs non-humains les plus courants, de fausses nouvelles. Ce sont des programmes informatiques conçus pour acher des comportements similaires à ceux des humains, et pour produire automatiquement du contenu et interagir avec les humains via les médias sociaux, diffuser des rumeurs, des spams, des logiciels malveillants, des informations erronées.

2.4.2 Contenu d'actualité

Chaque nouvelle est constituée d'un contenu d'actualité physique et d'un contenu d'actualité non-physique.

- Contenu d'actualité physique: le contenu physique des fausses nouvelles contient le titre de la nouvelle, le corps de la nouvelle, des images, des vidéos, des hashtags, des signaux de mention, des emojis,...etc. En raison de certaines significations et fonctionnalités, ces composants sont des fonctionnalités importantes pour la détection de fausses nouvelles.
- Contenu d'actualité non-physique: le contenu non physique concerne les opinions, les émotions, attitudes et sentiments que les créateurs de nouvelles veulent exprimer.
  Exemple, chaque jour, des millions de commentaires sont publiés sur des plateformes de vente en ligne, ces avis biaisés sont de gros problèmes pour les marques et les clients en ligne, non seulement ils vont aecter la décision du processus de la fabrication, mais aussi ils peuvent facilement détruire la réputation d'une marque.

2.4.3 Contexte social

Le contexte social fait référence à l'ensemble du système d'activité et de l'environnement social dans lequel la diusion de la nouvelle a lieu. Aujourd'hui, les modes de partage de la nouvelle sont de plus en plus dominés par les technologies interactives sur les médias sociaux. Les utilisateurs en ligne peuvent non seulement apprendre d'avantage sur les tendances, mais aussi partager leurs histoires et défendre leurs intérêts. S'ils partagent ces expériences et interactions au sein de certains groupes sociaux, et les membres de ces groupes partageant les mêmes idées, l'influence de ces dernières peut être amplifiée. Ceci facilite la diusion de fausses nouvelles.

2.5 Solution et Contexte du Projet

2.5.1 Le Natural Language Processing (NLP)

Pour obtenir la collection d'informations classées avec précision comme réelles ou fausses, nous devons créer un modèle d'apprentissage automatique à l'aide de domaine du traitement automatique de langue.

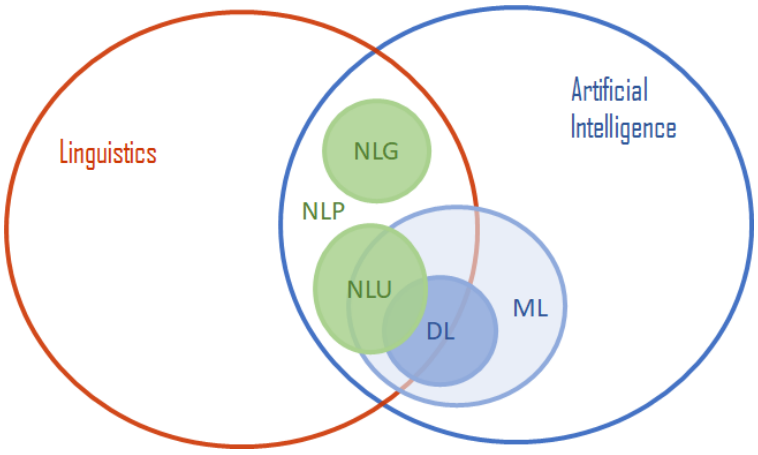
Pour gérer la détection de fausses ou de vraies nouvelles, nous développerons le projet en python à l'aide de 'sklearn', nous utiliserons 'TfidfVectorizer' dans nos données d'actualités que nous collecterons à partir des médias en ligne.

Une fois la première étape terminée, nous allons initialiser le classificateurs, transformer et ajuster le modèle. En fin de compte, nous calculerons les performances du modèle en utilisant les matrices de performances appropriées. Une fois que nous calculerons les matrices de performance, nous pourrons voir les performances de notre modèle.

La mise en œuvre pratique de ces outils est très simple et sera expliquée étape par étape dans cet article.

2.5.2 Natural Language Processing et Intelligence Artificielle

Le natural language processing (NLP) est une branche du machine learning qui vise à doter des programmes informatiques de la capacité de comprendre le langage humain naturel. Plusieurs techniques et modèles existent pour parvenir à cet objectif ambitieux.



NLP et IA

Le Natural Language Processing (NLP), c'est quoi ?
Le natural language processing (NLP), ou traitement du langage naturel, est une branche de l'intelligence artificielle qui s'attache à comprendre le langage humain tel qu'il est écrit et/ou parlé. Pour ce faire, des programmes informatiques spécifiques sont développés. En effet, un ordinateur typique réclame qu'on lui parle dans un langage de programmation bien précis, balisé, structuré, sans ambiguïté. Le langage naturel humain est, lui, imprécis, équivoque, confus. Pour permettre à un programme de comprendre le sens

des mots, il faut employer des algorithmes capables d'analyser le sens et la structure pour "désambiguïser" les mots, de reconnaître certaines références, puis de générer du langage sur cette base.

Quelles sont les différentes techniques de natural language processing (NLP) ?
Les algorithmes de NLP pratiquent différentes analyses syntaxiques et sémantiques, pour évaluer le sens d'une phrase en fonction de règles grammaticales fournies au préalable, en opérant une segmentation des mots et des groupes de mots ou en étudiant la grammaire d'une phrase complète. Pour déterminer le sens et le contexte, ils comparent en temps réel le texte avec toutes les bases de données dont ils disposent. Ayant besoin de quantités importantes de data (étiquetées) pour identifier les corrélations pertinentes, ils ont recours aux techniques modernes d'apprentissage du machine learning ou du deep learning. Diverses techniques sont employées par ces algorithmes telles que la reconnaissance des entités nommées (noms de personnes, lieux...), l'analyse des sentiments (positif, négatif, neutre), la synthèse de texte, l'extraction d'aspects (ciblage de l'intention du texte), et la modélisation de sujets.

Quels sont les principaux modèles de NLP ?
Si le traitement du langage naturel existe depuis longtemps, les progrès réalisés récemment sont considérables avec une multiplication des programmes de NLP, surtout chez les géants du numérique. Parmi les modèles les plus en pointe, on peut citer dans la section suivante:

2.5.3 Les modèles de détection des fausses nouvelles

Nous allons maintenant discuter des modèles de détection des fausses nouvelles .

- Modèles du contenu d'actualités: les modèles de contenu des actualités peuvent être classifiés en ce qui suit:
  1. Modèle basé sur les connaissances: l'approche fondée sur les connaissances a pour objectif d'utiliser des sources pour vérifier les faits dans le contenu des actualités.

  2. Modèle basé sur le style: les éditeurs des fausses nouvelles utilisent certains styles d'écriture spécifique nécessaire pour faire appel à un véritable article de presse, tel que les mots neutres, décrivant les événements avec des faits. Une meilleure qualité d'écriture (en tenant compte des mots épinglés, ponctuation et longueur des phrases).

- Modèles du contexte social: les réseaux sociaux fournissent des ressources supplémentaires aux chercheurs pour compléter et améliorer les modèles de contexte
  d'actualité. Les modèles du contexte social sont aussi utilisés pour la détection des rumeurs et l'identification des faux contenus sur Facebook. Ces modèles sont en fonction de la position et de la propagation.

  1. Basé sur la position: c'est un processus qui peut déterminer les sentiments dégagés par l'utilisateur, en faveur, contre ou neutre.
     Il existe deux façons pour représenter la position de l'utilisateur explicite ou implicite. Les positions explicites sont celles où les lecteurs ont donné des expressions directes, comme le pouce vers le haut ou le pouce vers le bas. Les positions implicites sont les positions où les sentiments extraits des publications sur les réseaux sociaux.

  2. Basé sur la propagation: c'est un processus qui peut déterminer la relation entre les événements pertinents sur les publications. Il existe deux catégories, propagation homogène qui contient une entité unique, comme un message ou un événement, ou propagation hétérogène qui contient plusieurs entités en même temps.

  3. Détection des rumeurs: la détection des rumeurs a pour objectif de classer une nouvelle comme rumeur ou non. C'est un modèle en quatre étapes: détection, suivi, position et véracité.

2.6 Impact global sur différents domaines

Les nouvelles visent des situation réels actuelles et des histoires complètes, couvrant diérentes questions, (ex; la criminologie, la santé, le sport, la politique...etc). Voici quelques impactes des fausses nouvelles sur des diérents domaines .

1. Dans les médias en ligne, les fausses nouvelles permettent de donner de l'ampleur aux faux comptes, (ex; récupération de maximum de clicks et d'adhérents).

2. Influence les propagandes politiques, en aectant les décisions de vote lors des élections;

3. Influencer les marchés financiers, où des millions pourraient être perdus.

2.7 Data Mining

Les techniques d'exploration de données sont classées en deux méthodes principales, à savoir : supervisé et non supervisé. La méthode supervisée utilise les informations d'entraînement afin de prévoir les activités cachées. L'exploration de données non supervisée tente de reconnaître les modèles de données cachés fournis sans fournir de données d'apprentissage, par exemple des paires d'étiquettes et de catégories d'entrée. Un exemple de modèle pour l'exploration de données non supervisée sont les mines agrégées et une base de syndicat [12].

- **Opinion Mining** Aujourd'hui, les internautes ont pris l'habitude de consulter les commentaires déposés par les autres dès qu'ils doivent prendre une décision d'achat pour un produit technique,ou encore pour une réservation d'hôtel. Et donc un commentaire peux détruire facilement la réputation d'une marque ou un produit. D'où les outils de Web Mining s'intéresse a analyser les opinions des internautes qui s'y expriment spontanément et en temps réel.
  - Définition L'Opinion Mining c'est l'étude qui analyse les opinions, les sentiments, les évaluations, les attitudes et les émotions des gens à partir de leurs écrits. L'importance croissante de l'analyse des sentiments coïncide avec la croissance des réseaux sociaux tels que Facebok et Twitter.
    Pour la première fois dans l'histoire de l'humanité, nous avons maintenant un énorme volume de données appelé big data d'opinion enregistrées sous forme numérique pour les classer en trois catégories: positive, négative et neutre.
  - Méthodes d'analyse Le but de l'analyse d'opinion est de déterminer si le sentiment dégagé par une phrase est positif ou négatif, qui dépend du contexte dans lequel elle est utilisée, du type de langage, ainsi que de la personne qui l'a écrite. Il existe plusieurs façons d'analyser automatiquement les sentiments.
    Parmi les méthodes les plus populaires, on cite :
    • Apprentissage automatique supervisé: l'objectif ici est de faire entraîner le système sur une base de telle façon qu'il puisse classer par la suite de façon automatique les nouvelles informations.
    Par exemple, un expert détermine pour chaque commentaire si l'utilisateur est « très énervé », « moyennement énervé » et « pas du tout énervé ». Pour que l'apprentissage automatique fonctionne, il faut annoter un grand jeu de données. Ensuite, un algorithme analyse cette annotation pour en faire ressortir des règles qui permettront de classer d'autres commentaires selon le degré d'énervement.
    • Règles linguistiques: dans ce cas, un linguiste analyse un échantillon de commentaires pour déterminer quels sont les mots ou les expressions qui indiquent l'énervement et ses différents degrés. Une fois cette analyse linguistique effectuée, il crée des règles de grammaire qui permettront au système de classer chaque commentaire dans l'une ou l'autre des catégories.

2.8 Conclusion

Dans ce chapitre, les différents concepts des fausses nouvelles ont été présentés. Nous avons vu leur signification, caractéristiques ainsi que les différents modèles utilisés pour leur détection.
L'application du Data Mining dans le Web et particulièrement l'Opinion Mining a permis d'apporter des solutions intéressantes a ce problème. Dans le chapitre suivant, nous allons nous concentrer précisément sur les méthodes d'apprentissage automatique et choisir la méthode qui convient le mieux au problème de détection des fausses nouvelles.

3 Chapitre 2 : Conception et Modélisation

3.1 Introduction

Dans ce chapitre nous aborderons une description générale de notre système, en mettant en évidence son coté conceptuel du pré-traitement qui constitue une étape fondamentale avant la mise en oeuvre de notre système, ensuite nous détaillerons chaque phase en citant les principaux algorithmes et techniques utilisées dans chacune des phases.

3.2 Logiciels utilisés

Dans notre travail nous avons utilisé généralement l'environnement Anaconda et google colab.

- Anaconda : Anaconda est une distribution scientifique de Python : c'est-à-dire qu'en installant Anaconda, vous installerez Python, Jupyter Notebook (que nous présenterons plus en détail au prochain chapitre) et des dizaines de packages scientifiques, dont certains indispensables à l'analyse de données !

- Google Colab ou aussi appelé Colaboratory est un service dans le cloud, proposé par Google gratuitement. Il est basé sur l'environnement Jupyter Notebook et est destiné à la formation et à la recherche en apprentissage automatique.Cette plateforme permet de former des modèles de machine learning directement dans le cloud . Il n'est pas nécessaire de l'installer sur l'ordinateur, les ressources informatiques peuvent donc être utilisées pour d'autres tâches

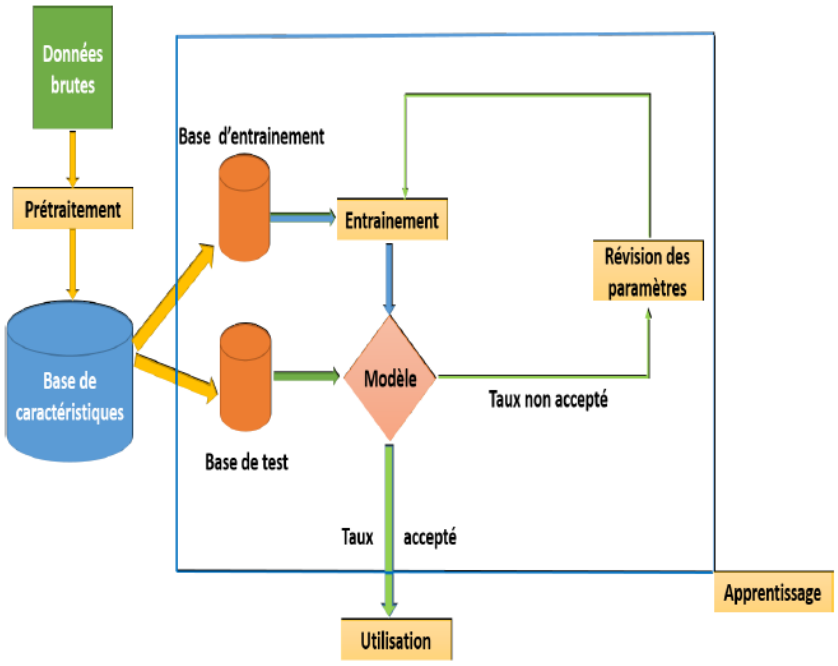
3.3 Langages utilisés

Dans notre travail nous avons le langage Python

- Python : est le langage de programmation open source le plus employé par les informaticiens. Ce langage s'est propulsé en tête de la gestion d'infrastructure, d'analyse de données ou dans le domaine du développement de logiciels. En effet, parmi ses qualités, Python permet notamment aux développeurs de se concentrer sur ce qu'ils font plutôt que sur la manière dont ils le font. Il a libéré les développeurs des contraintes de formes qui occupaient leur temps avec les langages plus anciens. Ainsi, développer du code avec Python est plus rapide qu'avec d'autres langages.

3.4 Architecture générale

Notre système se base sur l'utilisation du Machine Learning pour détecter les Fake News. Le système prend en entrée une base brute de commentaires et leurs caractéristiques et la transforme en une base de caractéristiques utilisable par la phase d'apprentissage. Cette transformation est appelée pré-traitement, elle effectue une série d'opérations telles que le nettoyage, le filtrage et l'encodage. La base pré-traitée est subdivisée en deux partie; une pour l'entraînement et l'autre pour le test. Le module d'entraînement utilise la base d'entraînement et un algorithme d'apprentissage pour fournir un modèle de décision qui est appliqué sur la base de test. Si le modèle est accepté, c-à-d a pu atteindre un taux de reconnaissance acceptable, il sera conservé et utilisé par le module d'utilisation et l'entraînement se termine. Dans le cas contraire, les paramètres de l'algorithme d'apprentissage sont révisé dans le but d'améliorer le taux de reconnaissance.



Architecture générale

3.5 Architecture détaillée

Dans ce qui suit, nous détaillons chacune des phases de notre système.

3.5.1 Dataset

Les ensembles de données utilisés pour ce projet ont été tirés de Kaggle . L'ensemble de données d'entraînement a environ 44898 des lignes de données provenant de divers articles sur Internet. Nous avons dû faire pas mal de pré-traitement des comme le montre notre code source, afin d'entraîner nos modèles.
Un ensemble de données d'entraînement complet a les attributs suivants :

- id : identifiant unique pour un article d'actualité
- titre : le titre d'un article d'actualité
- Date : date de création de l'article
- texte : le texte de l'article ; incomplet dans certains cas
- étiquette : une étiquette qui marque l'article comme potentiellement non fiable
- 1 : peu fiable
- 0 : fiable

3.5.2 Analyse des données

Ici, on va expliquer l'ensemble de données.

Dans ce projet python, nous avons utilisé le jeu de données CSV. L'ensemble de données contient 44898 lignes et 2 colonnes.

Cet ensemble de données a quatre colonnes,

titre : il s'agit du titre de l'actualité. auteur : il s'agit du nom de l'auteur qui a écrit la nouvelle. text : cette colonne a les nouvelles elle-même. label : il s'agit d'une colonne binaire représentant si la nouvelle est fausse (1) ou réelle (0). L'ensemble de données est open source et peut être trouvé ici .

3.5.3 Bibliothèques utilisées

Les bibliothèques très basiques de science des données sont sklearn, pandas, NumPy, etc. et certaines bibliothèques spécifiques telles que les transformateurs. Les plongements utilisés pour la majorité de nos modélisations sont générés comme la figure suivante.

```
[ ]  import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import scikitplot.plotters as skplt
     import re
     import string
     import nltk
     from nltk.corpus import stopwords
     from sklearn.model_selection import train_test_split
     from sklearn.feature_extraction.text import TfidfVectorizer
     from sklearn.metrics import classification_report, confusion_matrix
     from sklearn.tree import DecisionTreeClassifier
     from sklearn.ensemble import GradientBoostingClassifier
     from sklearn.ensemble import RandomForestClassifier
     from sklearn.ensemble import AdaBoostClassifier
     from sklearn.neighbors import KNeighborsClassifier
     from sklearn.svm import SVC
     import time
```

- **Scikit-learn** (Sklearn) est la bibliothèque la plus utile et la plus robuste pour l'apprentissage automatique en Python
- **Pandas**Pandas est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données.
- **NumPy** est une bibliothèque pour langage de programmation Python, destinée à manipuler des matrices ou tableaux multidimensionnels ainsi que des fonctions mathématiques opérant sur ces tableaux.
- **Matplotlib** est une bibliothèque du langage de programmation Python destinée à tracer et visualiser des données sous formes de graphiques
- **NLTK** est une plate-forme leader pour la création de programmes Python pour travailler avec des données de langage humain.

Le but est de produire une représentation vectorielle de chaque article. Avant d'appliquer notre modèle nous effectuons un prétraitement de base des données. Cela inclut la suppression des mots vides, la suppression des caractères spéciaux et la ponctuation, et convertir tout le texte en minuscules. Cela produit une liste séparée par des virgules

Avant de continuer, nous devons vérifier si une valeur nulle est présente ou non dans notre ensemble de données.

$\mathrm{df}_fake = df_fake.isnull()$
$df_true = df_true.isnull()$

Il n'y a pas de valeur nulle dans cet ensemble de données. Mais si vous avez des valeurs nulles présentes dans votre ensemble de données, vous pouvez le remplir. Dans le code donné ci-dessous, je vais vous dire comment vous pouvez remplacer les valeurs nulles.

$\mathrm{df}_fake = df_fake.fillna('')$
$df_true = df_true.fillna('')$

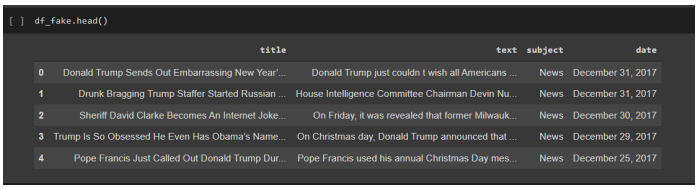
3.5.4 Pré-traitement

Notre objectif est d'extraire les meilleures caractéristiques permettant de détecter une Fake News. On commence par le pré-traitement des données du dataset brute qui sont subdivisées en trois catégories: les données textuelles, les données catégorielles et les données numériques qui représentent respectivement:

le texte de la nouvelle, la source de la nouvelle avec son auteur enfin la date et le sentiment dégagé par la nouvelle. Le pré-traitement de chaque catégorie est effectué à travers un ensemble d'opérations:

Brève description

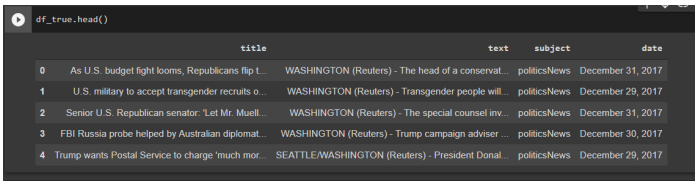
1. L'ensemble de données est prétraité 111M (59.9M pour les fausses nouvelles et 51.1M pour les vraies)



```
[ ] df_fake.head()
```

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017

False news dataframe



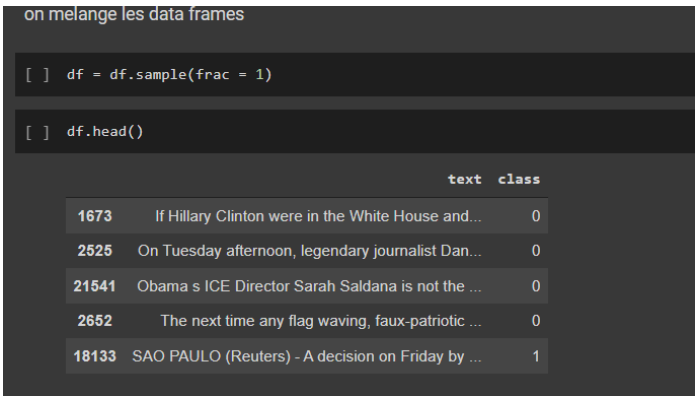
```
df_true.head()
```

	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017

True news dataframe

2. Les textes dans plusieurs contextes proviennent de kaggle.com en format CSV à l'aide de Python.

3. Melanger les deux dataframe



on melange les data frames

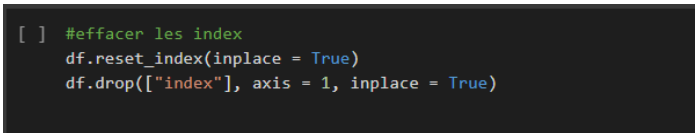
```
[ ] df = df.sample(frac = 1)
```

```
[ ] df.head()
```

	text	class
1673	If Hillary Clinton were in the White House and...	0
2525	On Tuesday afternoon, legendary journalist Dan...	0
21541	Obama s ICE Director Sarah Saldana is not the ...	0
2652	The next time any flag waving, faux-patriotic ...	0
18133	SAO PAULO (Reuters) - A decision on Friday by ...	1

Melanger les deux dataframes

4. L'étape suivante consiste à nettoyer le bruit à l'aide des bibliothèques NLP NLTK et de la bibliothèque SAFAR v2. Le bruit implique des identifiants, des points, des virgules, des citations,Majuscules et en éliminant les termes, supprimer le suffixe. La prochaine étape consiste à utilisez POS (Part of Speech) qui transformera l'ensemble de données en jetons et en valeurs statistiques.



```
[ ] #effacer les index
    df.reset_index(inplace = True)
    df.drop(["index"], axis = 1, inplace = True)
```

Effacer les index

supprimer les cracteres specieaux, ponctuation, les liens

```
[ ]  # re.sub  pour supprimer les cracteres specieaux

[ ]  # help(re)

[ ]  def supp_char_spe(text):
         text = re.sub('\[.*?\]', '', text)
         text = re.sub("\\W"," ",text)
         text = re.sub('https?://\S+|www\.\S+', '', text)
         text = re.sub('<.*?>+', '', text)
         text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
         text = re.sub('\n', '', text)
         text = re.sub('\w*\d\w*', '', text)
         return text

     df["text"] = df["text"].apply(supp_char_spe)
```

Enlever les caractères spéciaux

enléve des mots vides :Stop words

```
nltk.download('stopwords')
stops = set(stopwords.words('english'))
# print(stops)
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

```
[ ]  df.head()
```

	text	class
0	if hillary clinton were in the white house and...	0
1	on tuesday afternoon legendary journalist dan...	0
2	obama s ice director sarah saldana is not the ...	0
3	the next time any flag waving faux patriotic ...	0
4	sao paulo reuters a decision on friday by ...	1

Enlever les mots vides

5. Effectuer l'extraction de caractéristiques en choisissant des caractéristiques lexicales, telles que le nombre de mots, la longueur moyenne des mots, longueur de l'article, nombre, nombre de sections du discours (adjectif).

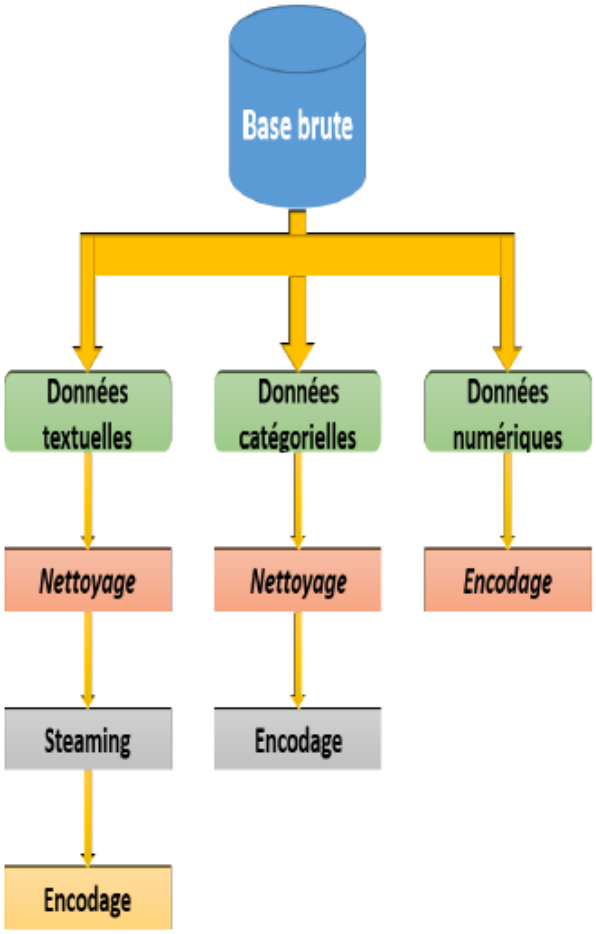
6. Extraire les fonctionnalités unigram et bigram en utilisant la fonction Tfidf Vectorizer de python sklearn. Caractéristique bibliothèque d'extraction pour générer des fonctionnalités n-gram TF-IDF.

texts to vecteurs

```
[ ]  # tf-idf
     vectorization = TfidfVectorizer(stop_words = stops)
     xv_train = vectorization.fit_transform(x_train)
     xv_test = vectorization.transform(x_test)
     xv_train
```

```
<35918x97213 sparse matrix of type '<class 'numpy.float64'>'
        with 5774775 stored elements in Compressed Sparse Row format>
```

vectorisation



Architecture Détaillée : Pré-traitement

Données textuelles: représentent le texte brute rédigé par l'auteur et pré-traité par les opérations suivantes:

Description des importantes étapes de pré-traitement

- Nettoyage: consiste à éliminer les mots vides tel que a, about, am, you, are... et les caractères spéciaux tel que !, ?, :, ;, , ... et toute information non utile, dans notre cas il s'agit des chiffres dans le texte.
- Steaming: consiste à transformer les mots utiles en des racines par exemple, les mots actor, acting , reenact sont tous transformés en la même racine act.
- Encodage: transformer l'ensemble des mots du commentaire en un vecteur

numérique en passant par deux étapes:
la combinaison des deux techniques:

Sac à mots: dans ce modèle, le texte est représenté sous forme de vecteur contenant ses mots, sans tenir compte de leur ordre, mais en gardant la multiplicité. Cette technique est principalement utilisée pour

calculer différentes mesures qui caractérisent le texte par exemple, le mot le plus répété dans l'ensemble des documents c-à-d le dataset appelé corpus. Mais le problème ici est que, un mot qui se répète dans tout le corpus ne veut pas dire qu'il est vraiment important ou il caractérise un document précis.

Pour résoudre ce problème, la fréquence du terme est pondérée par l'importance du document dans le corpus.
Dans notre système nous avons utilisé la méthode TF-IDF .

TF-IDF (Term Frequency-Inverse Document Frequency) est une mesure statistique permet d'évaluer l'importance d'un terme contenu au sein d'un document, dans un corpus. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document. Cette méthode est utilisée dans des moteurs de recherche pour apprécier la pertinence d'un document à une requête.

Plusieurs formules de calcul ont été proposé pour cette méthode. Dans notre travail nous avons utilisé la formule suivante:

TF(t) = Nombre d'apparition du terme t dans le document (n) / Nombre total de termes dans le document, en gardant la multiplicité de chaque terme (k).

$$T\cdot Ft = n/k$$

IDF(t)= Nombre totale des documents (D)/ Nombre des documents citant ce terme ($\cdot$Dt).

$$ID\cdot Ft = \log(D\cdot/Dt)$$

- N-gram: dans cette technique, le texte est représenté sous forme de vecteur contenant des blocs de mots, en tenant compte de leur ordre, et en gardant la multiplicité.

- Données catégorielles: représentent dans notre étude la source et l'auteur. Le pré-traitement de ces données passe par deux étapes:
  – Nettoyage: consiste à éliminer les caractères spéciaux et la transformation des lettres en minuscules.

3.5.5 Apprentissage:

Brève description

1. Diviser l'ensemble de données en 80% pour le train et 20% pour le test à l'aide de python sklearn.

2. Produire le fichier ipynb du modèle de classification après avoir appliqué tous les algorithmes.
3. Testez la précision du modèle sur la partie test de l'ensemble de données et produisez une matrice de confusion.

4. Évaluez l'exactitude, la précision, le rappel et le score f1 pour les fausses et les vraies nouvelles.

5. Concevez l'interface à utiliser pour tester les nouvelles invisibles par l'utilisateur.

Description en détails (exemple algorithme SVM)
On regroupe deux modules, l'entraînement et la validation utilisant chacun une partie de la base des caractéristiques subdivisée en deux parties, base d'entraînement et base de test.
Le module d'entraînement utilise la base d'entraînement pour fournir un modèle de décision tandis que le

module de validation utilise la base de test pour mesurer la performance du modèle fourni.
Entraînement: pour entraîner notre modèle, nous avons choisis plusieurs algorithmes que nous allons présenter dans le chapitre suivant. La selection du meilleur algorithme est basée sur deux raisons:

1. Parce qu'il donne les meilleures résultats au niveau du Texte Mining.

2. Pour utiliser la valeur de la fonction de décision comme un degré de confiance pour la classification des nouvelles.

Une valeur positive de la fonction de décision désigne en même temps une nouvelle vraie ainsi que son degré de vérité et vise versa, une valeur négative de la fonction de décision désigne une fausse nouvelle ainsi que son degré de faux.
On calcule donc, lors de l'entraînement le maximum et le minimum de la fonction de décision. Lors de l'utilisation le degré de vérité ou de faux est calculé par la fonction suivante:(algorithme SVM comme exemple

$$\begin{cases} Dec > 0 & p = \frac{Dec}{Max_{dec}} * 100 \\ Sinon & p = \frac{Dec}{Min_{dec}} * 100 \end{cases}$$

Où:

- Dec: est la valeur de la fonction de décision.
- Maxdec et Mindec: représentent les valeurs maximales et minimales de la fonction de décision.
- p: est le pourcentage de vérité ou de faux.

La figure suivante illustre cette idée:

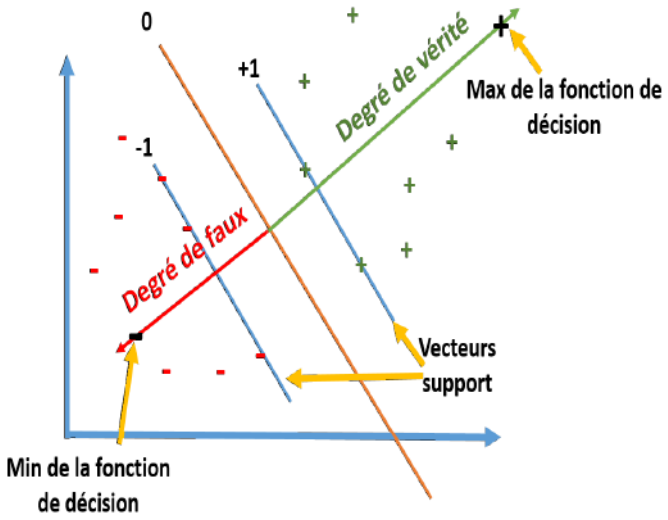


Figure 1: Degré de confiance de la classification des nouvelles

Le résultat de l'entraînement est un modèle ou pattern, qui représente l'analyse des données et leur transformation en informations utiles, en établissant des relations entre elles. Plusieurs métriques sont utilisées pour estimer la qualité du modèle qui se basent sur les valeurs suivantes:

1. VP: les exemples positifs classés correctement.
2. FP: les exemples positifs mal classés .
3. VN: les exemples négatifs classés correctement.
4. FN: les exemples négatifs mal classés.

- Précision: proportion des exemples positifs correctement classés dans l'ensemble des exemples positifs.

  $P = \frac{VP}{VP+FP}$
- Rappel: proportion des exemples positifs correctement classés VP par rapport aux exemples classés positifs (VP +FN).

  $R = \frac{VP}{VP+FN}$
- Fmesure: moyenne harmonique de la précision et du rappel. Elle mesure la capacité du système à donner toutes les solutions pertinentes et à refuser les autres.

  $R = \frac{2*P*R}{P+R}$
- Validation: consiste à mesurer la capacité du modèle a reconnaître des nouveaux exemples. Pour cela, on écarte dès le départ une partie des exemples pour les utiliser pour le test du modèle. La base des caractéristiques est alors subdivisée en deux parties, une partie d'entraînement et une partie de test. Son utilité consiste a éviter le sur-apprentissage, c-à-d tester le modèle sur la même base d'entraînement.
  La subdivision n'est pas faite au hasard mais selon un échantillonage particulier :

  1. Holdout method: le dataset de taille n est subdivisé en deux partie, la première généralement de 60test.

  2. K-fold cross-validation: le dataset est subdivisé en m parties, m-1 parties pour l'apprentissage et une pour le test. Cette opération est répétée m fois et à chaque fois on obtient un taux de reconnaissance. À la fin on calcul la moyenne et l'écart types de ces taux pour estimer la performance du modèle.

  3. Leave-one-out cross-validation: le dataset est subdivisé en m partie tel que m=k exemples où k représente le nombre d'exemples total de la base. À chaque opération l'apprentissage se fait sur k-1 et le test sur l'exemple qui reste. C'est un cas particulier de la validation croisée.

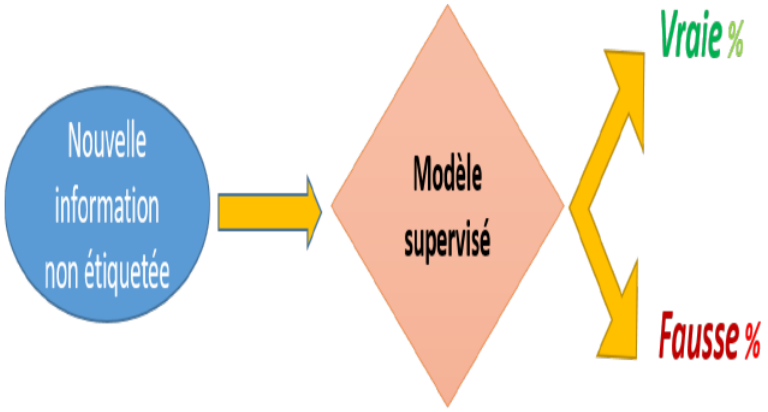
- Révision des paramètres: cette opération a pour objectif d'améliorer le modèle, avec le tuning ou le réglage des paramètres de la machine à vecteurs de support et de changer la variante de validation croisée ou la valeur de k au cas de k-folds cross validation.
  Il existe de nombreux paramètres de l'svm mais, les plus important sont:

  – Cost: ce paramètre désigne l'optimisation de l'svm pour éviter de mal classer les données d'entraînement. Pour des valeurs élevées de C, l'optimisation choisira un hyper-plan à plus petite marge, inversement, une très petite valeur de C amènera l'optimisation à rechercher un hyper-plan de séparation à plus grande marge. – Gamma: ce paramètre définit jusqu'où l'influence d'une seule donnée d'entraînement atteint, avec des valeurs faibles une signification «loin» et des valeurs élevées

une signification «proche».
– degré: ce paramètre présente le dégrée du noyau.
– Epsilon: ce paramètre détermine la tolérance du critère de terminaison. C'est le taux d'erreur permis.

3.5.6 Utilisation

C'est la dernière phase et la plus importante dans notre système. Après être arrivé au meilleur taux de reconnaissance, ou après avoir construit le meilleur modèle dans la phase précédente, nous devons l'utiliser sur des nouvelles informations non étiquetées, et le modèle nous permet de prédire la classe de la nouvelle si elle est fausse ou vraie avec un degré de confiance comme suit:



Utilisation

3.6 Conclusion

Ce chapitre a décrit la conception de notre système et il a présenté la démarche suivie dans ses différentes phases. Dans le chapitre suivant, nous allons décrire le fonctionnement de notre modèle mettant en oeuvre l'environnement et les algorithmes proposés avec une comparaison.

4 Chapitre 3: Implémentions et résultats

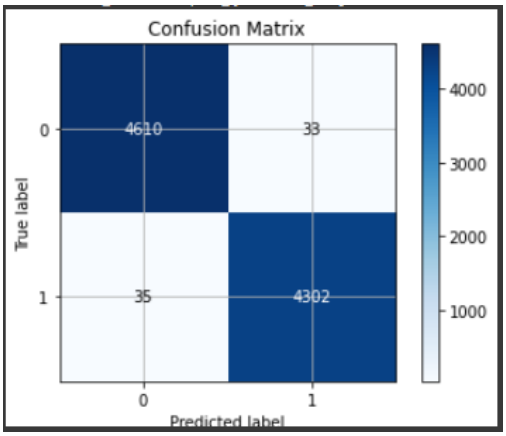
4.1 Introduction

implémentation d'un système informatique, désigne sa réalisation et sa mise en oeuvre, pour passer à l'interaction avec les utilisateurs. L'objectif de ce chapitre est de présenter les outils (les logiciels, les langages, les bibliothèques et les données utilisés dans notre système, ensuite nous aborderons l'application, puis nous discuterons les résultats.

4.2 Algorithmes utilisés

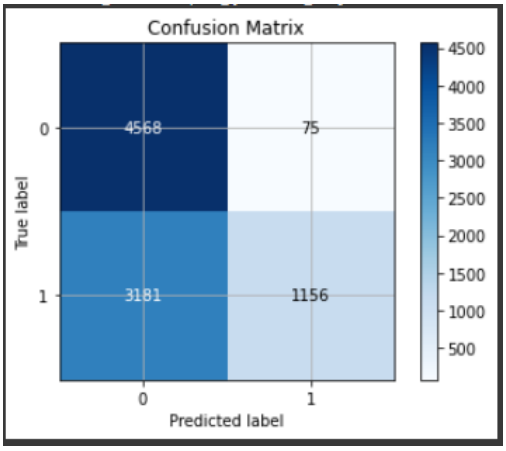
Support Vector Machine "SVM"
La machine à vecteurs de support (SVM) originale a été proposée par Vladimir N. Vapnik et Alexey Ya. Chervonenkis en 1963. Mais ce modèle ne peut faire qu'une classification linéaire, il ne convient donc pas à la plupart des des problèmes pratiques. Plus tard en 1992, Bernhard E. Boser, Isabelle M. Guyon et Vladimir N. Vapnik a introduit l'astuce du noyau qui active le SVM pour la classification non linéaire. Qui fait le SVM beaucoup plus puissant. Nous utilisons le noyau de la fonction de base radiale dans notre projet. La raison pour laquelle nous utilisons ce noyau est que deux Les vecteurs de caractéristiques Doc2Vec seront proches les uns des autres si leurs documents correspondants sont similaires, donc la distance calculée par la fonction noyau doit toujours représenter la distance d'origine. Depuis le La fonction de base radiale est

$K(x, x') = exp(-\frac{||x-x'||^2}{2\delta^2})$



matrice de confusion pour SVM

KNN (k- Nearest Neighbors)
KNN classe les nouvelles positions en fonction de la plupart des sons du k voisin par rapport à celles-ci. La position attribuée dans la classe est hautement mutuellement exclusive entre les voisins les plus proches K, comme mesuré par le rôle de la distance [6].
KNN appartient à la catégorie de l'apprentissage supervisé et ses principales applications sont la détection d'intrusion, la reconnaissance de formes. Il n'est pas paramétrique, donc aucune distribution spécifique n'est attribuée aux données ou à tout l'hypothèse est faite à leur sujet. Par exemple GMM, suppose une distribution gaussienne des données données.

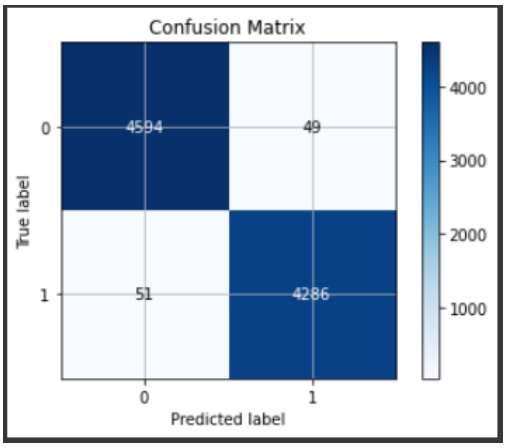


matrice de confusion pour KNN

Random Forest

Random Forest est construit sur le concept de construction de nombreux algorithmes d'arbre de décision, après quoi les arbres de décision obtiennent un résultat séparé. Les résultats, qui sont prédits par un grand nombre d'arbres de décision, sont repris par la forêt aléatoire. Pour assurer une variation des arbres de décision, la forêt aléatoire sélectionne aléatoirement une sous-catégorie de propriétés de chaque groupe [16][17]

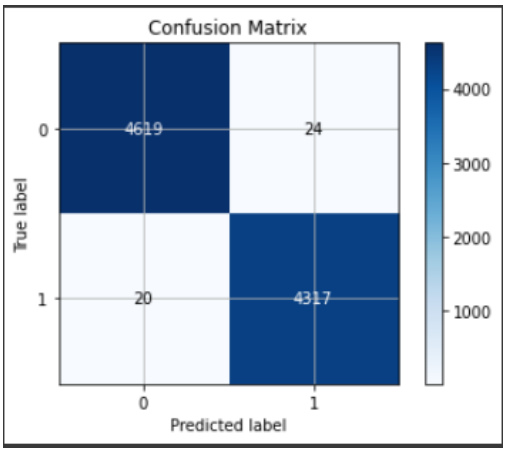
L'applicabilité de la forêt aléatoire est meilleure lorsqu'elle est utilisée sur des arbres de décision non corrélés. S'il est appliqué sur des arbres similaires, le résultat global sera plus ou moins similaire à un seul arbre de décision. Des arbres de décision non corrélés peuvent être obtenus par bootstrap et caractéristiques aléatoires.



matrice de confusion pour Random Forest

Decision Tree

L'arbre de décision est un outil important qui fonctionne sur la base d'une structure de type organigramme qui est principalement utilisée pour les problèmes de classification. Chaque nœud interne de l'arbre de décision spécifie une condition ou un « test » sur un attribut et le branchement se fait sur la base des conditions et du résultat du test. Enfin, le nœud feuille porte une étiquette de classe qui est obtenue après avoir calculé tous les attributs. La distance de la racine à la feuille représente la règle de classification. La chose étonnante est qu'il peut fonctionner avec la catégorie et la variable dépendante. Ils sont bons pour identifier les variables les plus importantes et ils décrivent également la relation



matrice de confusion pour Decision Tree

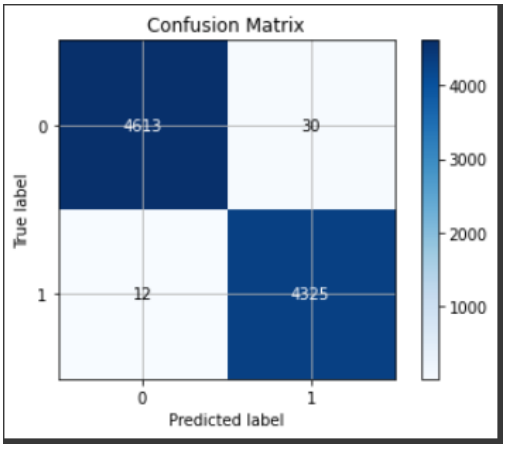
Ada-boost

Ada-boost ou Adaptive Boosting est l'un des classificateurs d'amplification d'ensemble proposés par Yoav Freund et Robert Schapire en 1996. Il combine plusieurs classificateurs pour augmenter la précision des classificateurs. AdaBoost est une méthode d'ensemble itérative.

Le classificateur AdaBoost construit un classificateur fort en combinant plusieurs classificateurs peu performants afin que vous obteniez un classificateur fort de haute précision. Le concept de base d'Adaboost consiste à définir les poids des classificateurs et à entraîner l'échantillon de données à chaque itération de manière à garantir des prédictions précises des observations inhabituelles. Tout algorithme d'apprentissage automatique peut être utilisé comme classificateur de base s'il accepte des poids sur l'ensemble d'apprentissage. Adaboost doit remplir deux conditions :

Le classificateur doit être entraîné de manière interactive sur divers exemples d'entraînement pesés.

À chaque itération, il essaie de fournir un excellent ajustement pour ces exemples en minimisant les erreurs d'apprentissage.



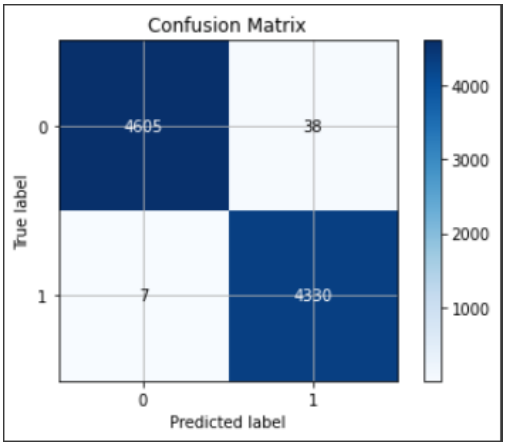
matrice de confusion pour Ada-boost

Gradient Boost

Gradient Boost : Le classificateur d'amplification de gradient dépend d'une fonction de perte . Une fonction de perte personnalisée peut être utilisée et de nombreuses fonctions de perte normalisées sont prises en charge par des classificateurs à gradient d'amplification, mais la fonction de perte doit être différentiable.

Les algorithmes de classification utilisent fréquemment une perte logarithmique, tandis que les al-

gorithmes de régression peuvent utiliser des erreurs quadratiques.
Les systèmes d'amplification à gradient n'ont pas à dériver une nouvelle fonction de perte à chaque fois que l'algorithme d'amplification est ajouté, mais toute fonction de perte différentiable peut être appliquée au système.



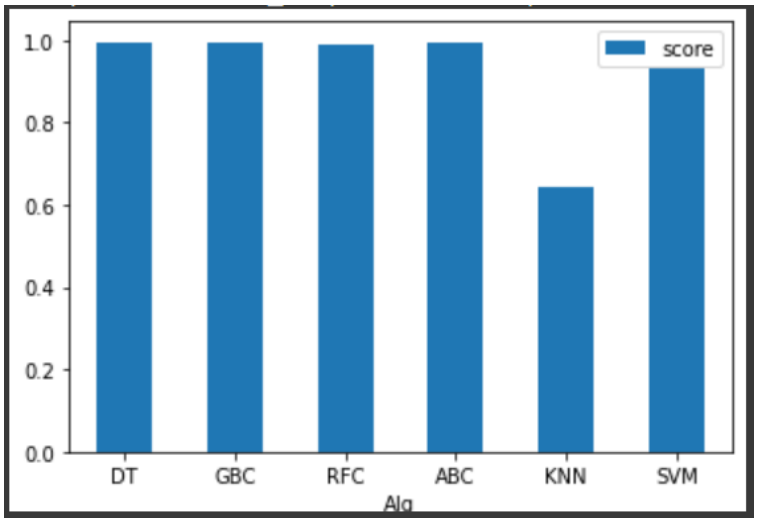
matrice de confusion pour Gradient Boost

4.2.1 Comparaison des Resultats

Nous avons comparé nos modèles en utilisant leurs matrices de confusion pour calculer la précision, le rappel et le Les scores de la F1. Le tableau 1 montre nos résultats

Algorithmes	Précession (accuracy)	Temps d'exécution	erreur de classification
Gradient Boost	0.994	284.8 s	38+7 | 8980
Support Vector Machine "SVM"	0.992	2285.1 s	33+35 | 8980
Random Forest	0.988	71.5 s	49+51 | 8980
Decision Tree	0.995	28.4	24+20 | 8980
Ada-boost	0.995	56.2 s	12+30 | 8980
KNNClassifier	0.637	66.6 s	3181+75 | 8980

Graphe de comparaison selon l'accuracy de chaque algorithme

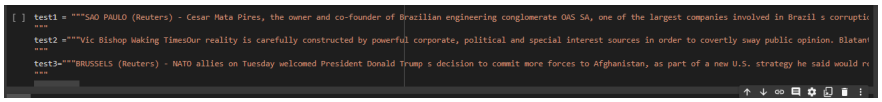


taux d'accuracy de chaque algorithme

=====> on recommande alors L'algorithme Arbre de décision "Decesion Tree" grâce à sa meilleur performance en temps d'exécution minimum et avec taux d'erreur de classification minimum

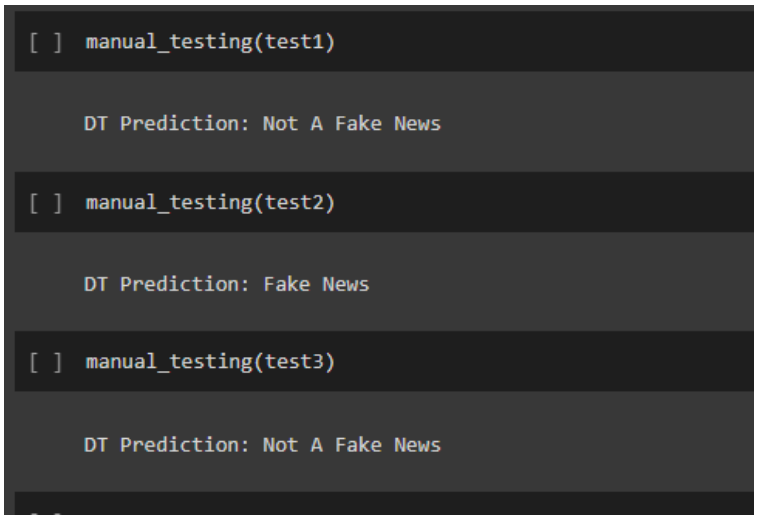
4.2.2 Phase de test

Dans cette partie on propose trois corpus de test pour vérification



3 corpus de test

Résultats:



Résultats du test

4.3 Conclusion

Nous avons détaillé dans ce chapitre les différents outils utilisés pour la mise en oeuvre de notre proposition ainsi que les résultats obtenus et leur discussion, Nous avons montré qu'une représentation du texte des nouvelles par une combinaison de la méthode de L'algorithme Arbre de décision "Decesion Tree" donnent des taux de reconnaissance très élevés.

5 Conclusion générale

Les techniques de détection de fausses informations peuvent être divisées en celles basées sur le style et celles basées sur le contenu, ou la vérification des faits. Trop souvent, on suppose qu'un mauvais style (mauvaise orthographe, mauvaise ponctuation, vocabulaire limité, utilisation de termes abusifs, agrammaticalité, etc.) est un indicateur sûr de fake news.

Il s'agit plus que jamais d'un cas où l'opinion de la machine doit s'appuyer sur des indications claires et parfaitement vérifiables du fondement de sa décision, en fonction des faits vérifiés et de l'autorité par laquelle la véracité de chaque fait a été déterminée.

La collecte des données une seule fois ne va pas suffire compte tenu de la rapidité avec laquelle les informations se propagent dans le monde connecté d'aujourd'hui et du nombre d'articles produits.

J'espère que vous pourriez trouver cela utile. Vous pouvez commenter dans les sections de commentaires pour toute question.

6 Réferences

[1]. Economic and Social Research Council. Using Social Mmedia. Available at: https://esrc.ukri.org/research/impact-toolkit/social-media/using-social-media

. Gil, P. Available at: https://www.lifewire.com/what-exactly-is-twitter-2483331. 2019, April 22.

. E. C. Tandoc Jr et al. "Defining fake news a typology of scholarly definitions". Digital Journalism , 1–17. 2017.

. J. Radianti et al. "An Overview of Public Concerns During the Recovery Period after a Major Earthquake: Nepal Twitter Analysis." HICSS '16 Proceedings of the 2016 49th Hawaii International Conference on System Sciences (HICSS) (pp. 136-145). Washington, DC, USA : IEEE. 2016.

. RAY, S. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/ 2017, September

. Researchgate.net. Available at: https://www.researchgate.net/figure/Pseudocode-for-KNNclassification$_fig7_260397165$, 2014.