

# Fake News Predictor

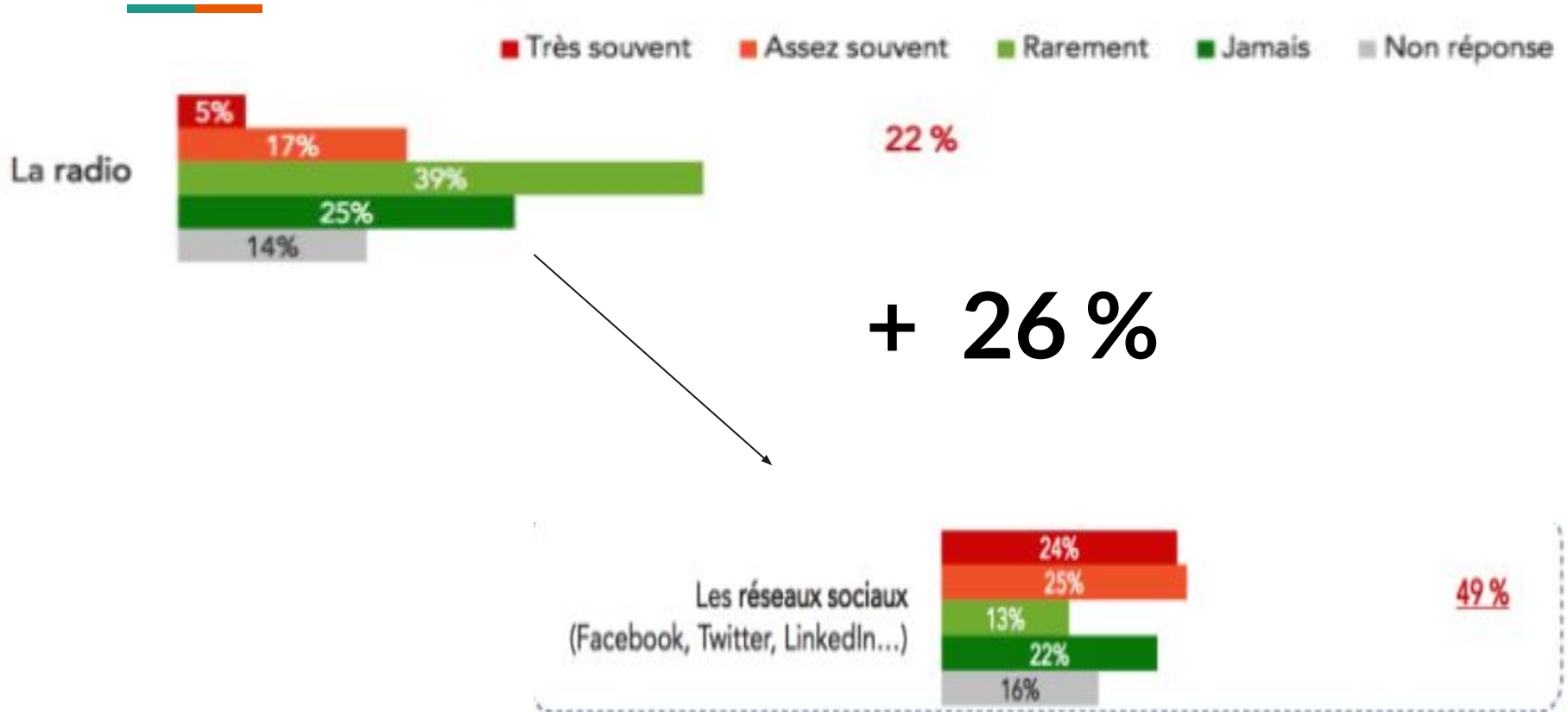


Morgan, Alicia, Alexis.G

# Contexte



# Pourcentage de Fake News publiées





# Streamlit



# Notre équipe et notre organisation

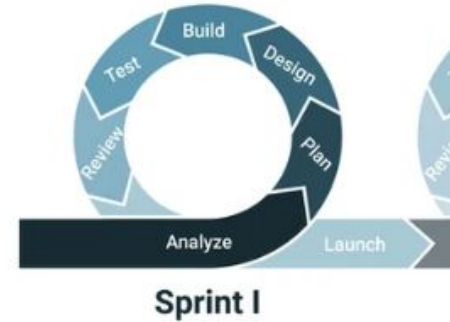
**Alicia**  
chef d'équipe  
data analyste



**Alexis**  
data scientist

**Morgan**  
Développeur

Méthode agile





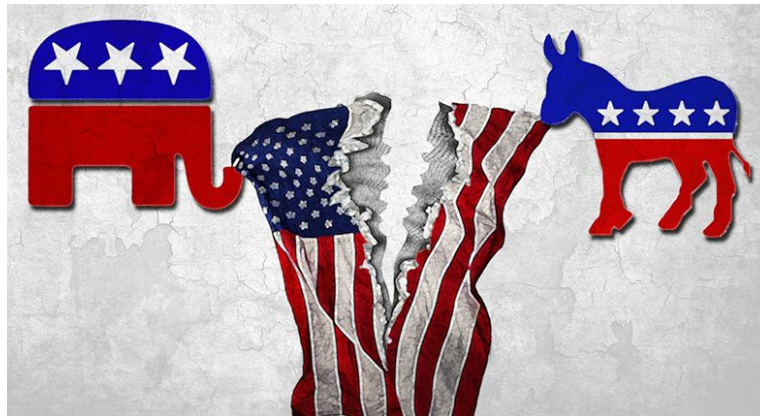
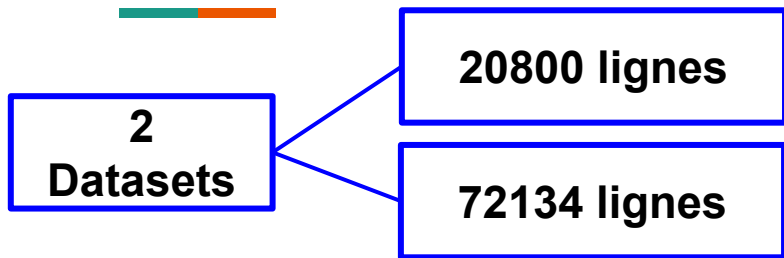
# Planning previsionnel

The image displays a project planning application with a grid of tasks and progress bars. The tasks are organized into four columns, each representing a different phase of the project. The progress bars are color-coded: red for 'livrable' (deliverable), yellow for 'Organisation/ DL' (organization/delivery), blue for 'model', and green for 'Dev Morgan' (development/Morgan). The tasks are as follows:

- Column 1 (Critère de perf):**
  - compréhension du jeux de données
  - performance des modèles de prédiction
  - démo qui fonctionne
  - capacité à apporter une solution dans le temps imparti
  - rédaction du rapport technique
  - qualité de la synthèse du travail
  - + Ajouter une carte
- Column 2 (livrable):**
  - document technique qui explique l'outil
  - analyse de données (notebook\_EDA)
  - procédure suivie pour trouver un modèle adapté (notebook\_model)
  - un modèle d'IA entraîné et adapté au problème (format pickle ou h5)
  - déployé l'API avec le modèle entraîné sur Azure
  - présentation qui explique votre démarche et les résultats obtenus
  - interface utilisateur
  - Fait
- Column 3 (Organisation/ DL):**
  - Jour 1: Adaptation
    - choix dataset
    - repo git
    - Squelette API
  - Nouvelle environnement de travail
  - requirements.txt pour le git
  - Jour 2: Base line
    - EDA
    - model
    - streamlit
  - Preparation des données
  - Jour 3: Amélioration
    - Faire le petit chef
    - pipeline ?
  - base de donnée
  - API
  - pause
  - Jour 4: Halp
    - model
    - base de donnée
    - API
- Column 4 (model):**
  - reconnait une FKCN du dataset
  - reconnait une FKCN d'un article
  - reconnait une FKCN quand on tape une phrase
  - bot discord
  - + Ajouter une carte
- Column 5 (Dev Morgan):**
  - repos git API
  - Squelette API
  - set-up la bdd
  - + Ajouter une carte
- Column 6 (Alexis data):**
  - choix dataset
  - Création du notebook EDA
  - EDA
  - + Ajouter une carte
- Column 7 (Alicia chef):**
  - repo git
  - création du notebook
  - écriture rapport
  - Préparation des données
  - + Ajouter une carte

The background of the application features a repeating pattern of the text "YOU ARE FAKE NEWS" overlaid on a grid of images of Donald Trump. The bottom of the application shows a "Fait" (Done) button and a small image of a group of people walking in a field.

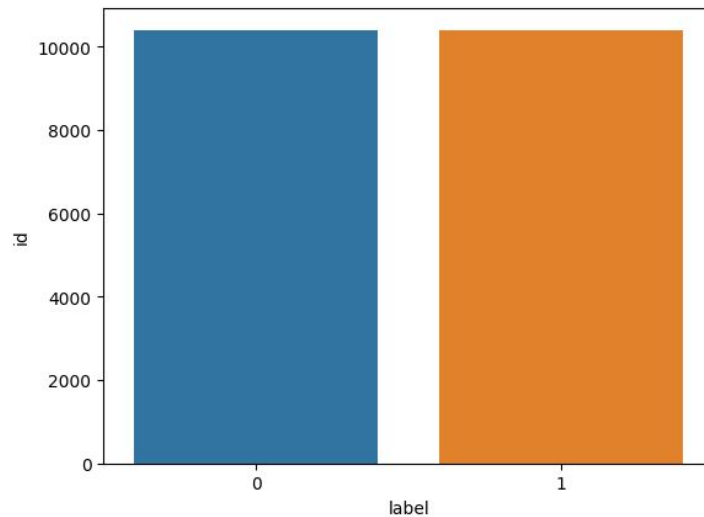
# Dataset



Fake = 1  
real = 0

no duplicated

NaN

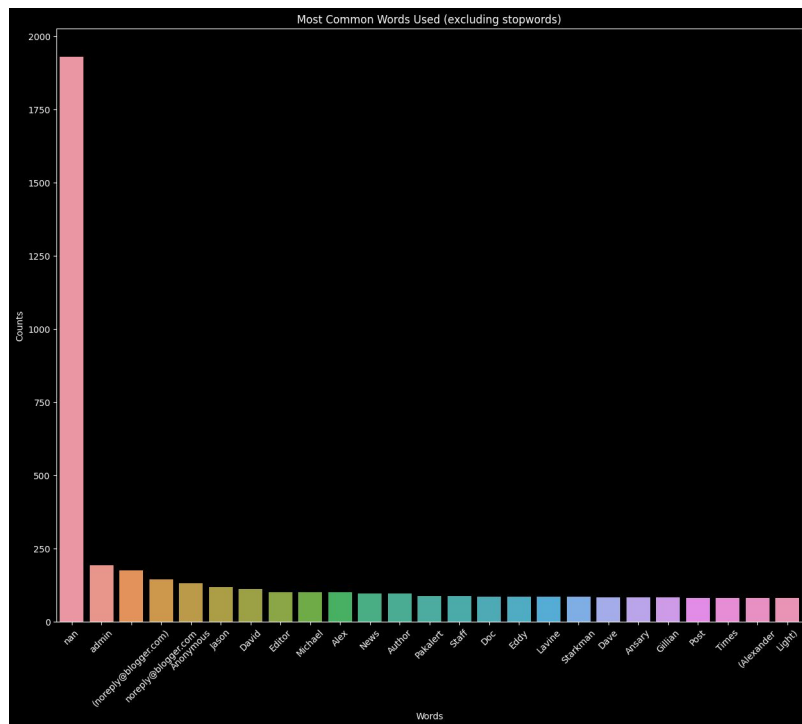




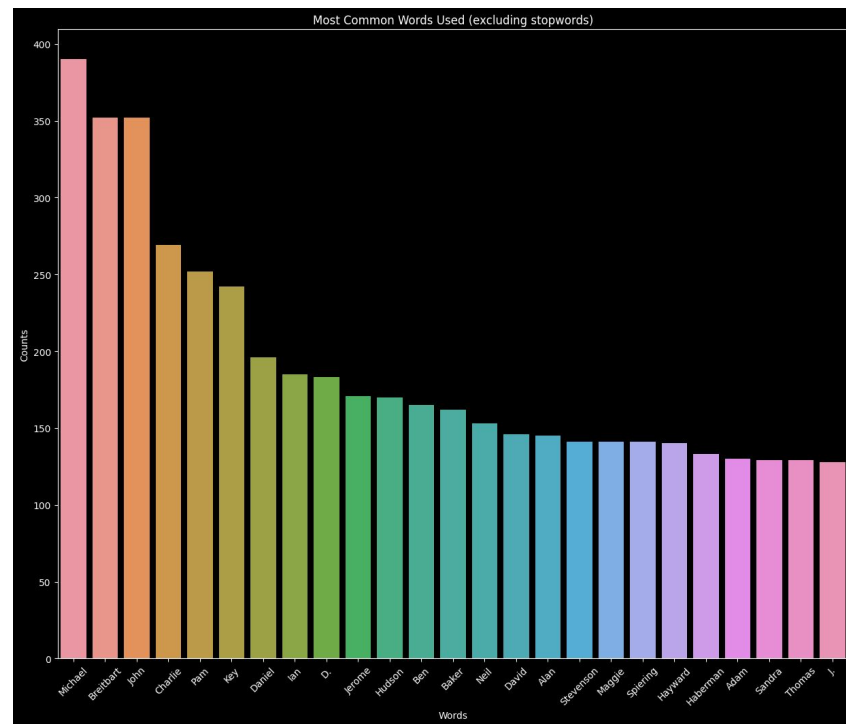


# Des auteurs bien différents

Fake



Real





## Etude de la fréquence des mots

**This is Big Data AI Book**

*Uni-Gram*

This	Is	Big	Data	AI	Book
------	----	-----	------	----	------

*Bi-Gram*

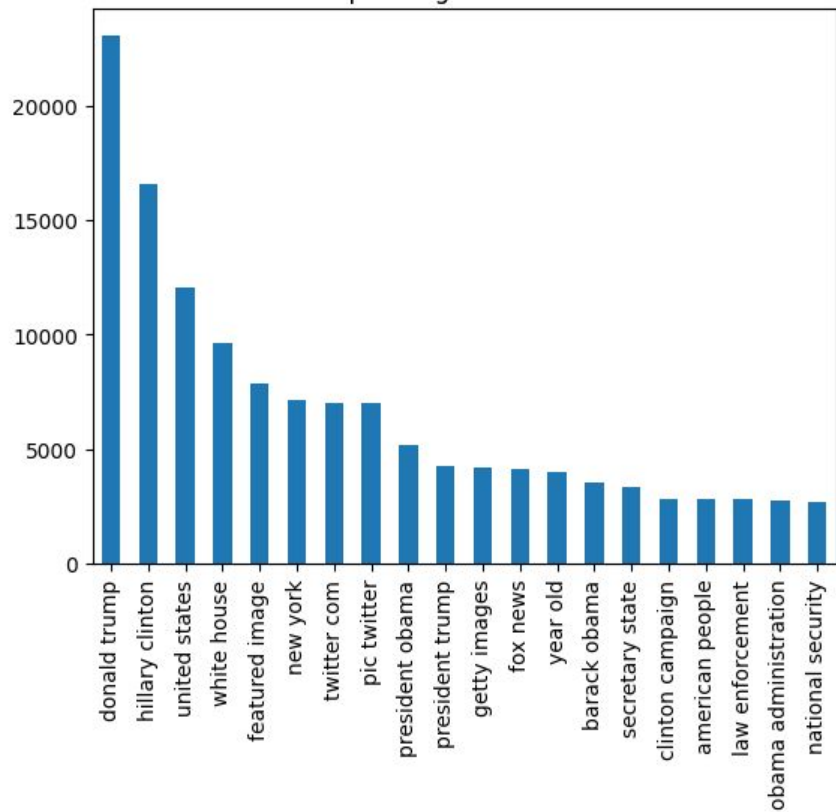
This Is	Is Big	Big Data	Data AI	AI Book
---------	--------	----------	---------	---------

*Tri-Gram*

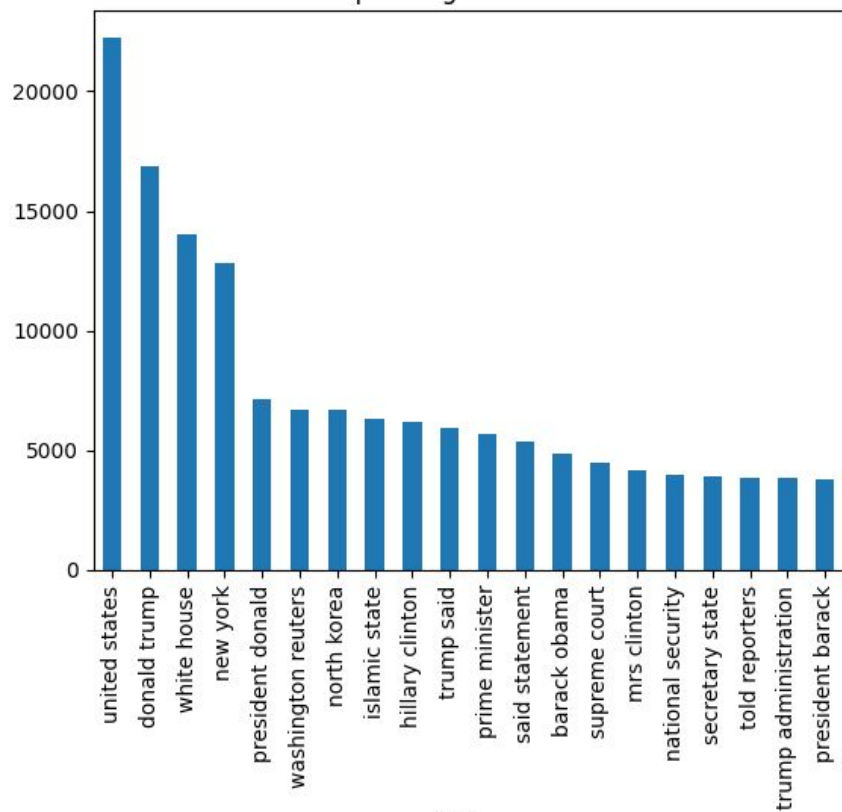
This is Big	Is Big Data	Big Data AI	Data AI Book
-------------	-------------	-------------	--------------

# Les bigrams

Top 20 bigrams in fake



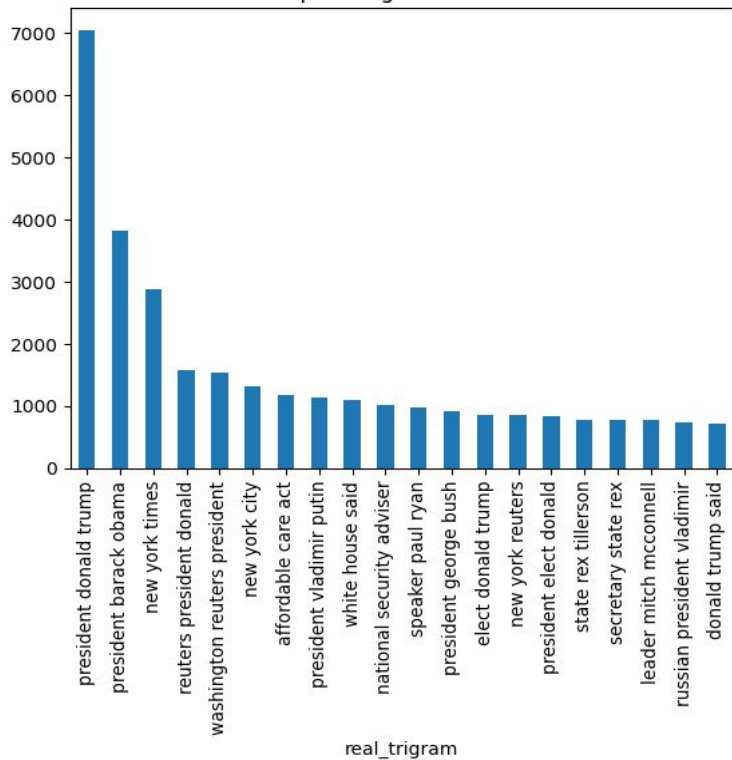
Top 20 bigrams in real



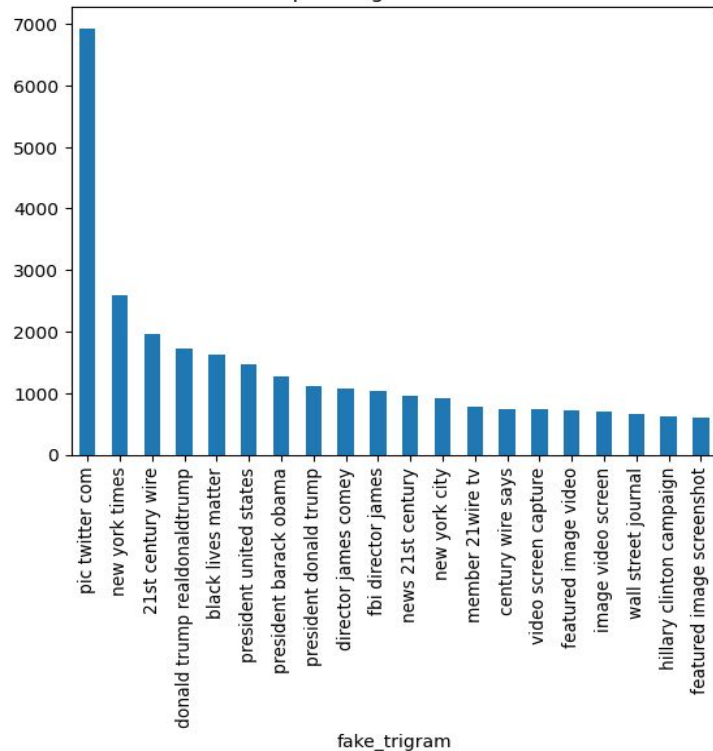
# Les trigrams



Top 20 trigrams in real



Top 20 trigrams in fake



# Feature Transformation :



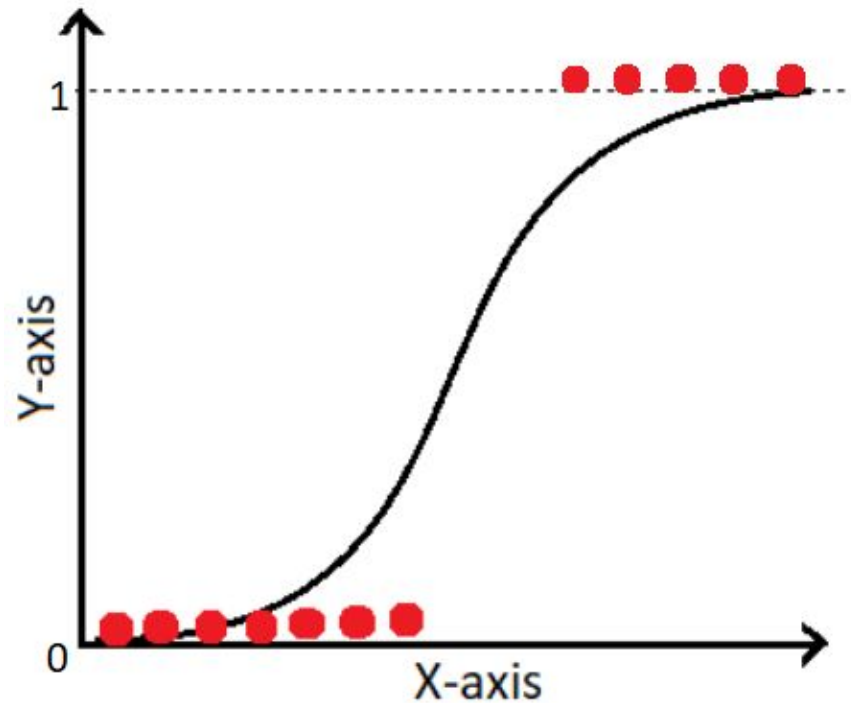
- Elimination des stop\_words
- Stemmatisation (PorterStemmer)
- Vectorization ( TfidfVectorizer )

Natural Language Processing  
with Python!  
Stop Words

[**“This”**, **“is”**, **“a”**, **“test”**]  
✓ X X ✓



# La régression logistique



# Model : Logistic Regression ( Author )



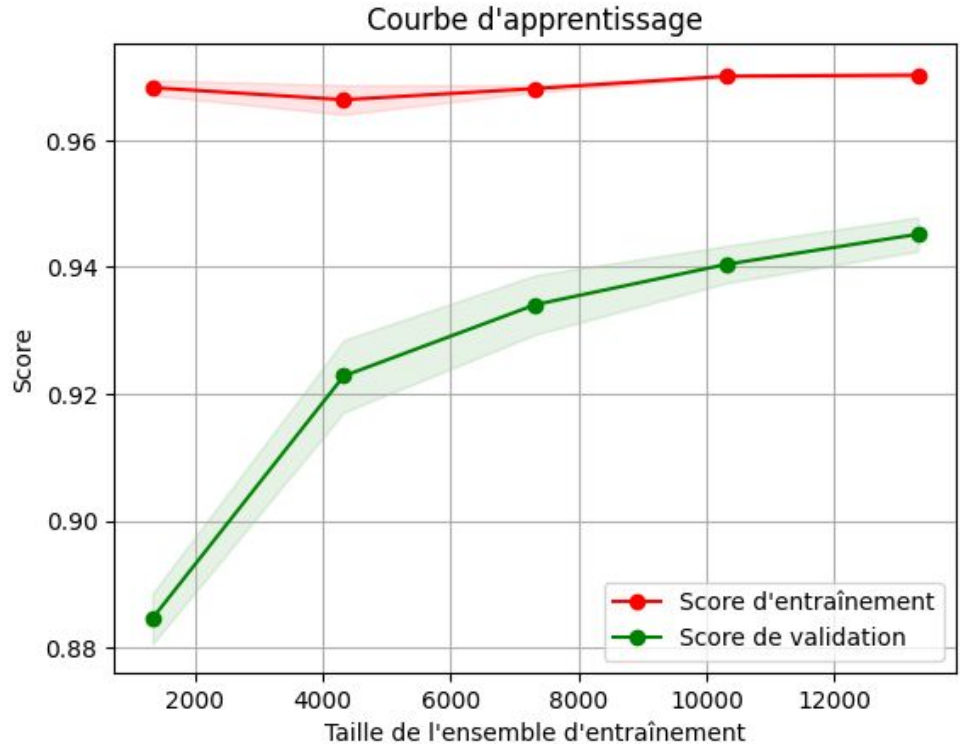
Accuracy score: 0.9865985576923076

	precision	recall	f1-score	support
Real	0.99	0.98	0.99	8310
Fake	0.98	0.99	0.99	8330
accuracy			0.99	16640
macro avg	0.99	0.99	0.99	16640
weighted avg	0.99	0.99	0.99	16640

# Model : Logistic Regression ( Text )

Accuracy score: 0.9401442307692308

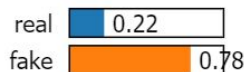
	precision	recall	f1-score	support
Real	0.94	0.94	0.94	2077
Fake	0.94	0.94	0.94	2083
accuracy			0.94	4160
macro avg	0.94	0.94	0.94	4160
weighted avg	0.94	0.94	0.94	4160



# Explicabilité du modèle: lime

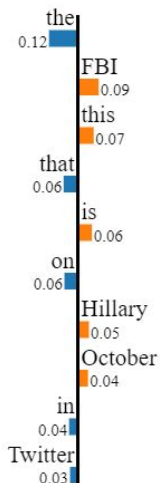
```
exp = explainer.explain_instance(text, pipeline.predict_proba, num_features=10)
exp.show_in_notebook(text=True)
```

Prediction probabilities



real

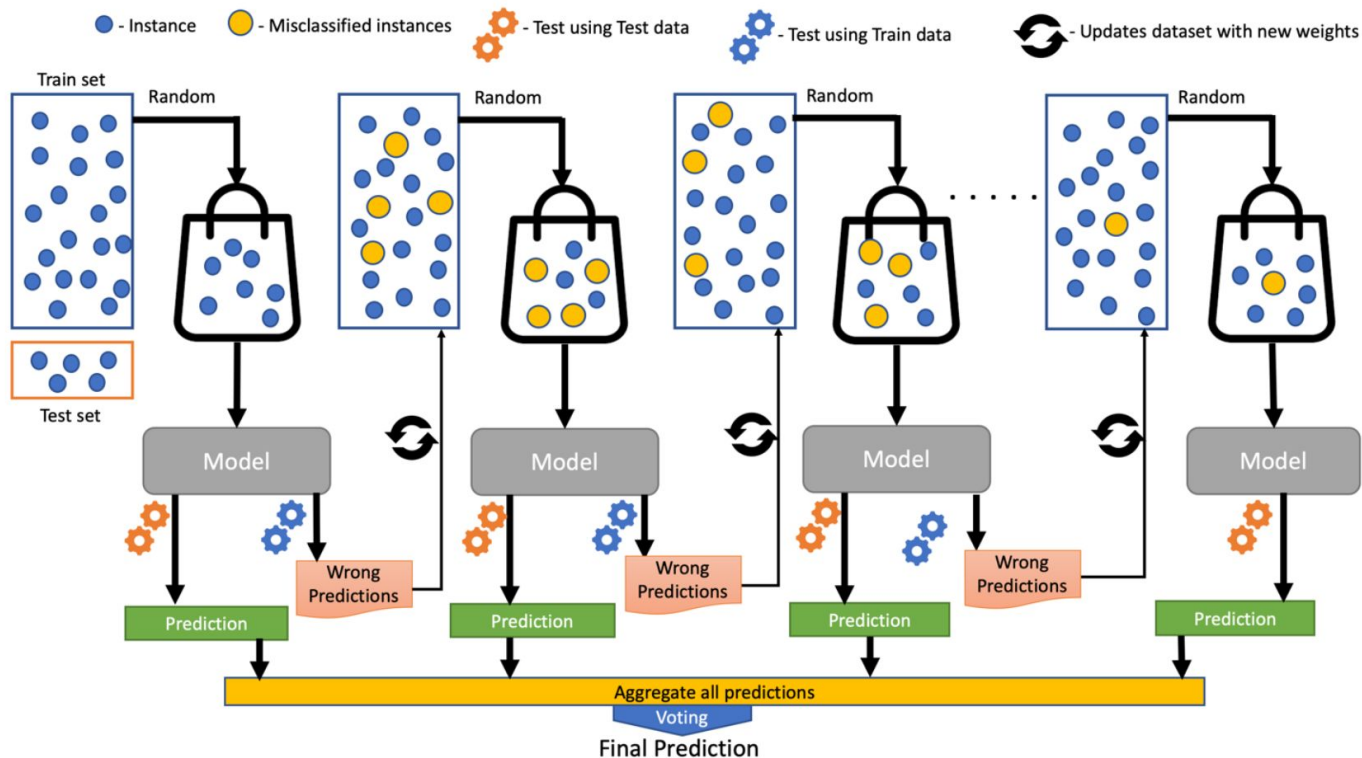
fake



## Text with highlighted words

House Dem Aide: We Didn't Even See Comey's Letter Until Jason Chaffetz Tweeted It By Darrell Lucas on October 30, 2016 Subscribe Jason Chaffetz on the stump in American Fork, Utah ( image courtesy Michael Jolley, available under a Creative Commons-BY license) With apologies to Keith Olbermann, there is no doubt who the Worst Person in The World is this week-FBI Director James Comey. But according to a House Democratic aide, it looks like we also know who the second-worst person is as well. It turns out that when Comey sent his now-infamous letter announcing that the FBI was looking into emails that may be related to Hillary Clinton's email server, the ranking Democrats on the relevant committees didn't hear about it from Comey.

# Le XGboost





## Model : XGBoost ( Text )



Accuracy score: 0.9024038461538462

	precision	recall	f1-score	support
Real	0.90	0.90	0.90	2077
Fake	0.90	0.90	0.90	2083
accuracy			0.90	4160
macro avg	0.90	0.90	0.90	4160
weighted avg	0.90	0.90	0.90	4160

# L'application



Client / Application

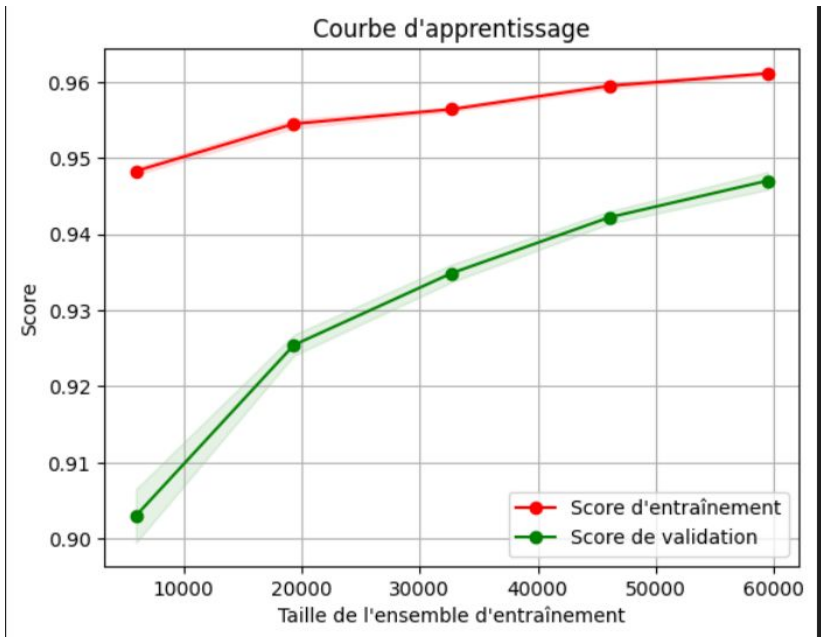


Base de données



# Réentraînement : Logistic Regression

Real	0.96	0.94	0.95	9083
Fake	0.94	0.96	0.95	9504
accuracy			0.95	18587
macro avg	0.95	0.95	0.95	18587
weighted avg	0.95	0.95	0.95	18587





## Perspective

- Implémentation de données pour généralisé
- Ajout des bigrams et trigrams comme features
- Amélioration des modèles ( Deep learning?)

Merci

