Fake News Predictor

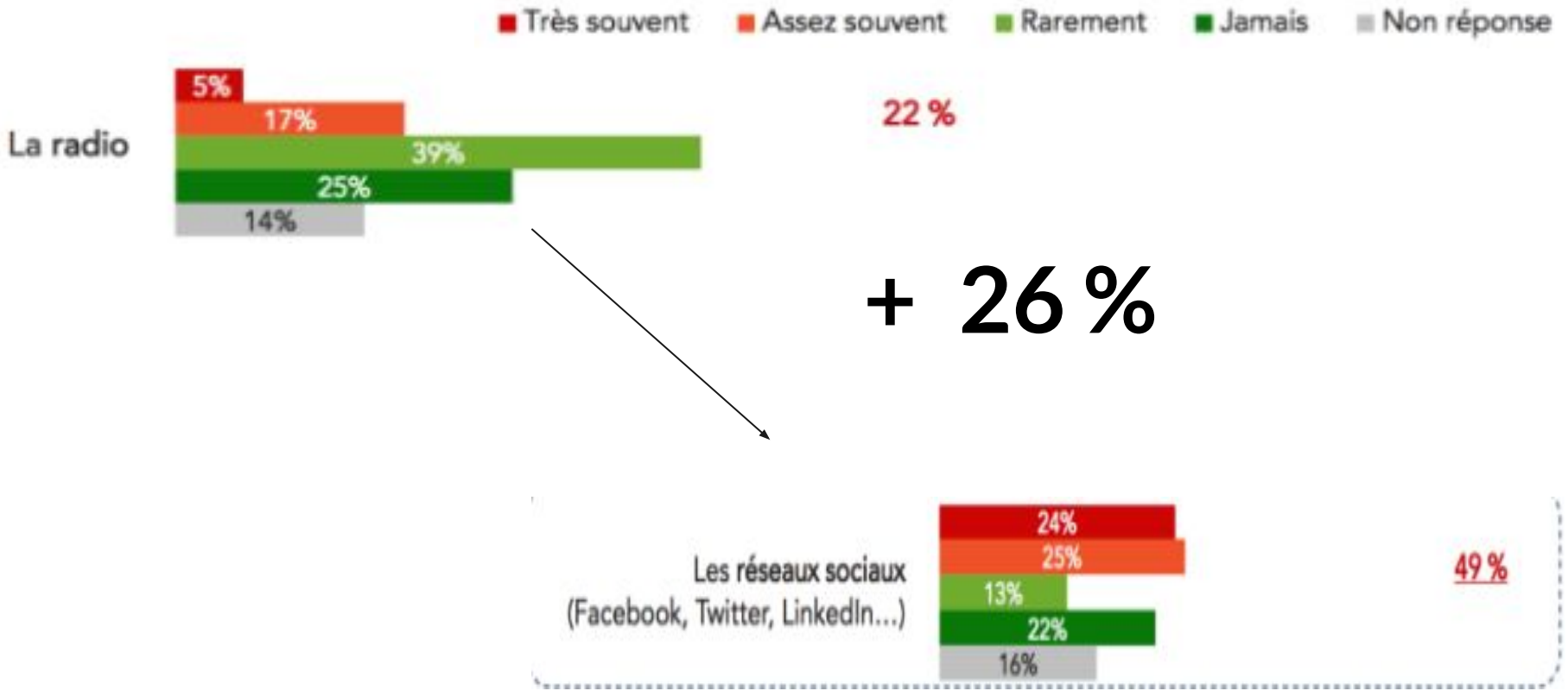
Morgan, Alicia, Alexis.G

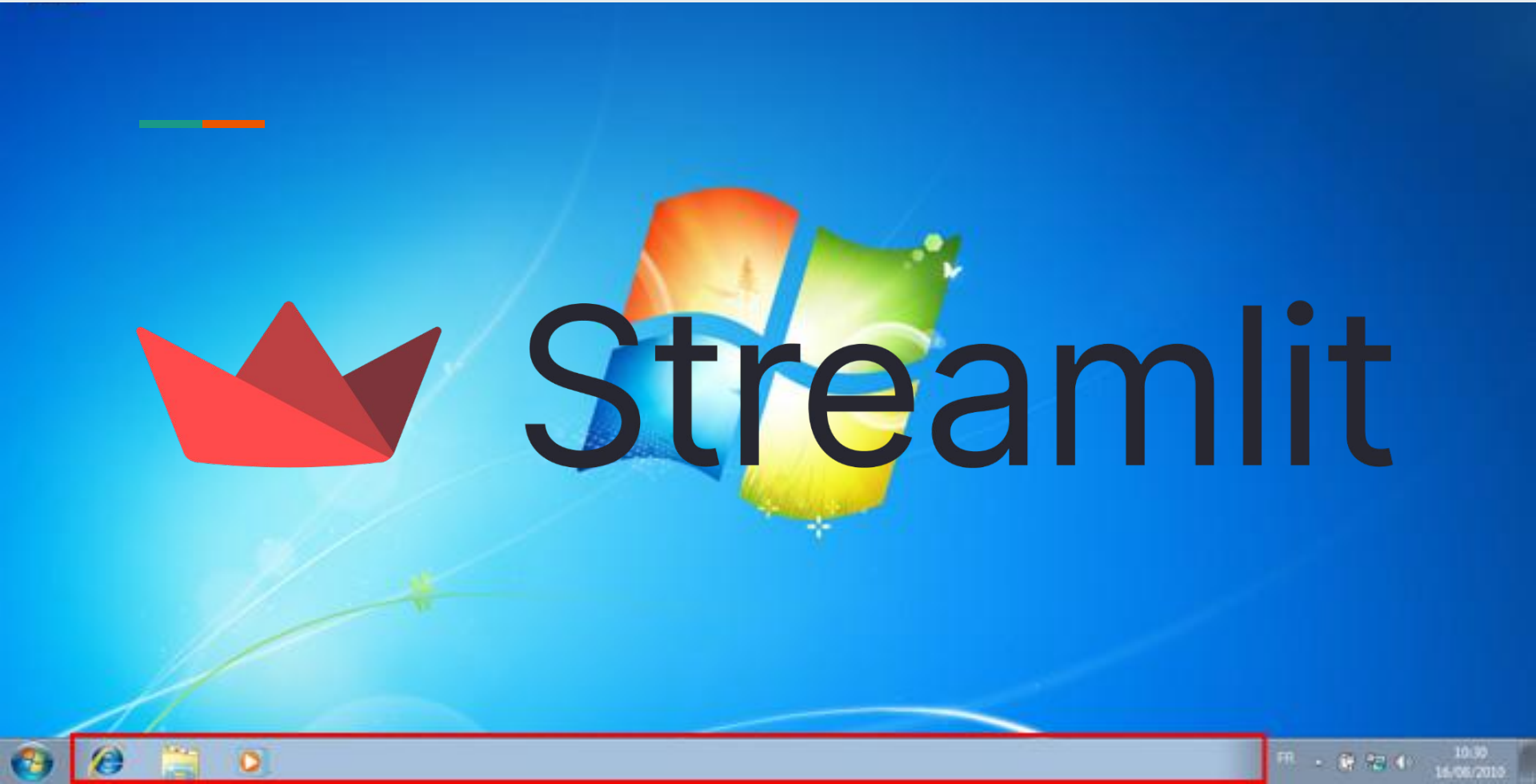
Contexte


FAKE NEWS
?


Growth of Fake News
2018
HARD TO TELL THE DIFFERENCE BETWEEN REAL AND FAKE NEWS
2017
FAKE NEWS BECOMES AN OFFICIAL WORD THAT DEFINED 2017
FAKE NEWS
2016
USAGE OF THE TERM INCREASED BY 365%

Pourcentage de Fake News publiées




Streamlit

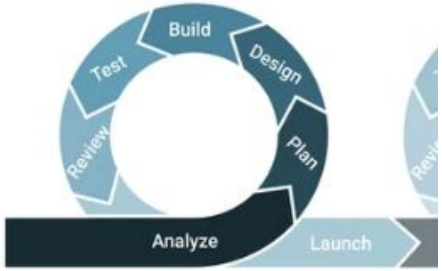
Notre équipe et notre organisation

Alicia
chef d'équipe
data analyste

Alexis
data scientist

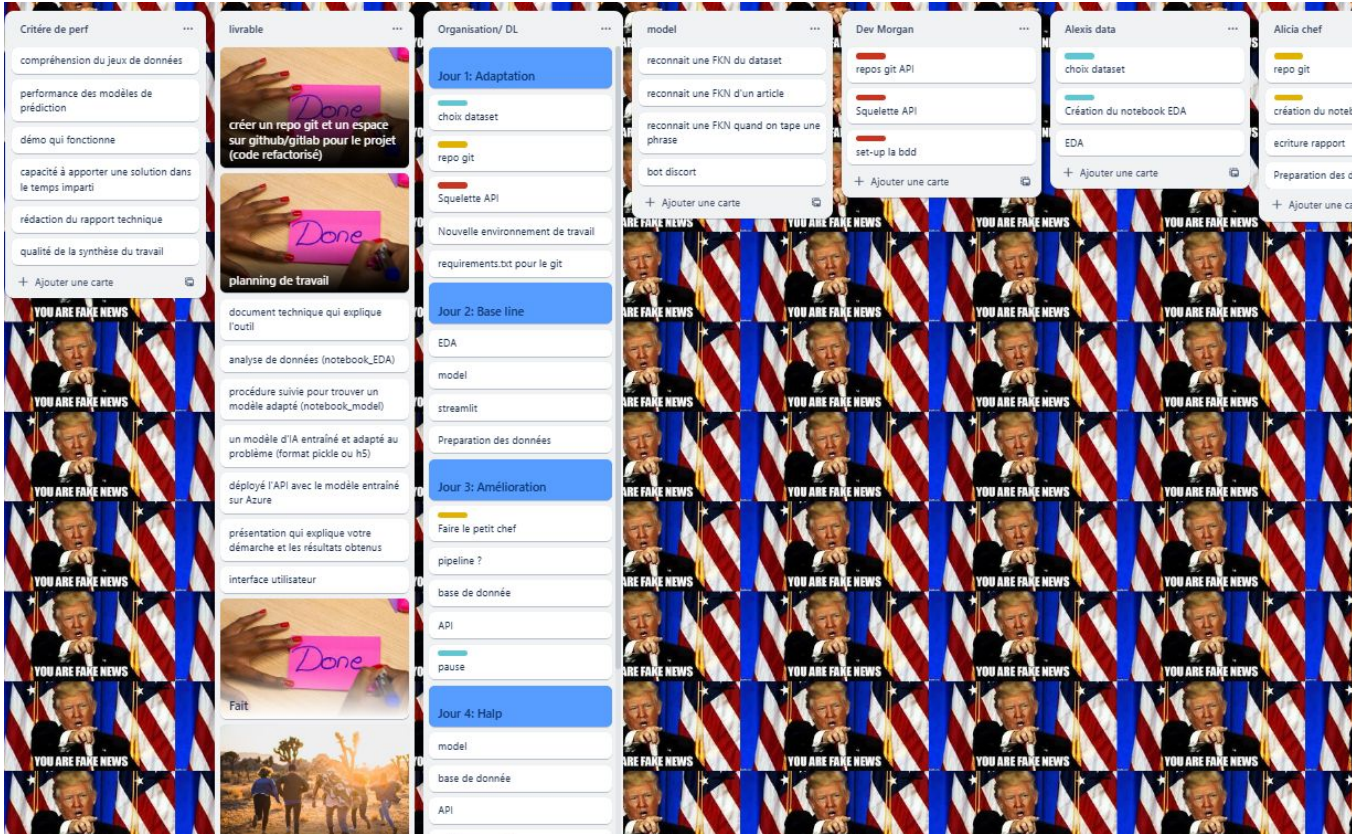
Morgan
Développeur

Méthode agile



Sprint I

Planning previsionnel



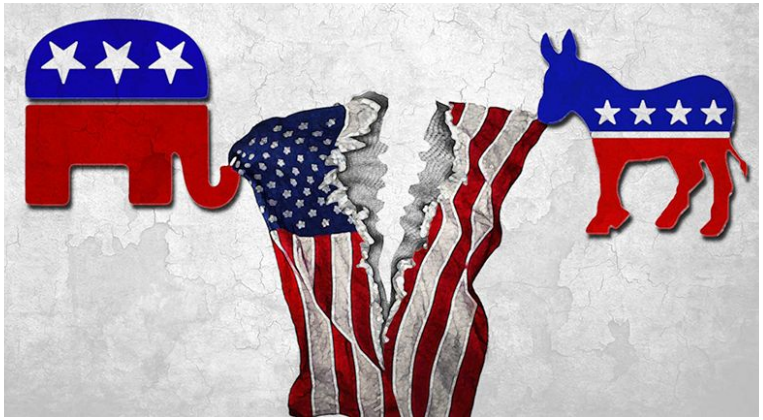
Dataset

2 Datasets

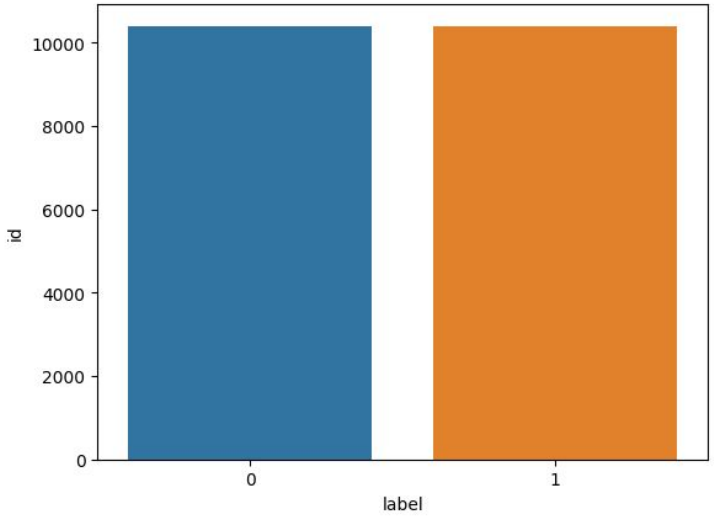
- 20800 lignes
- 72134 lignes

no duplicated

NaN

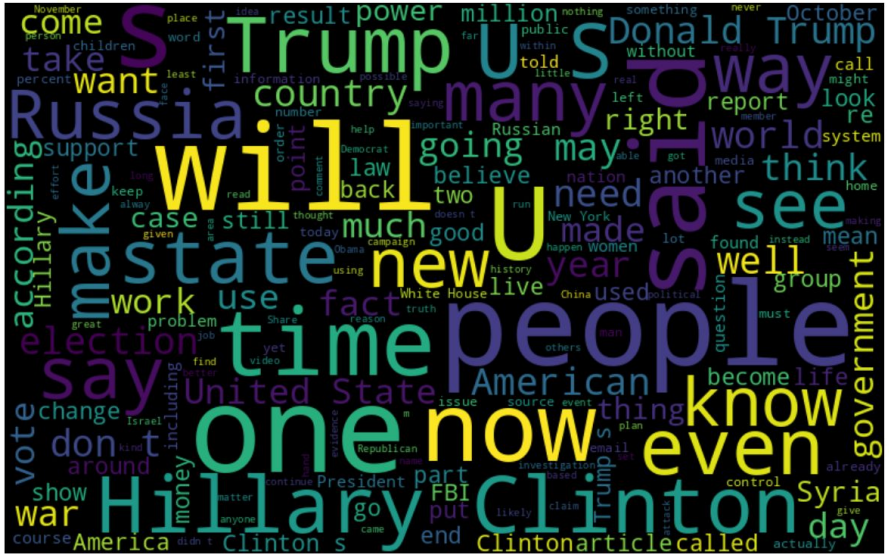


Fake = 1
real = 0

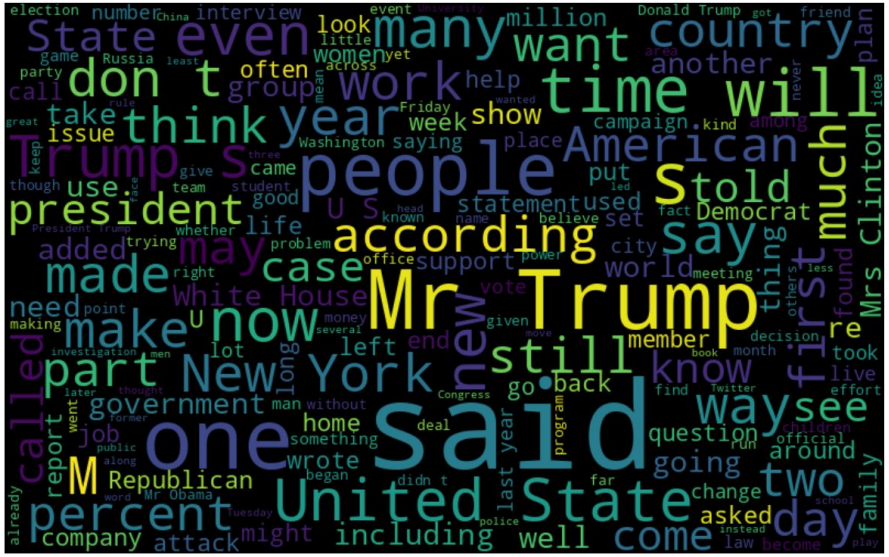


Analyse du dataset

Fake



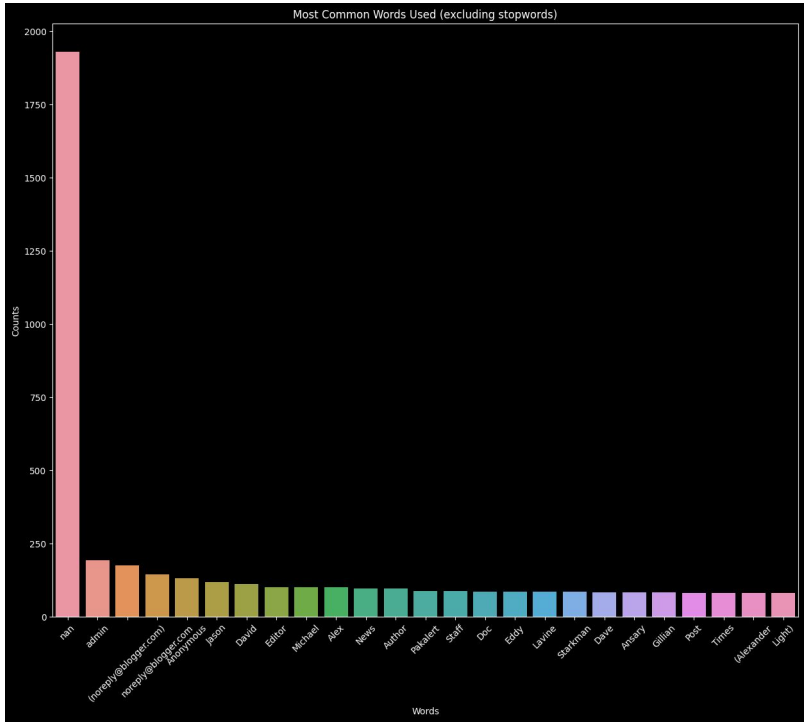
Real



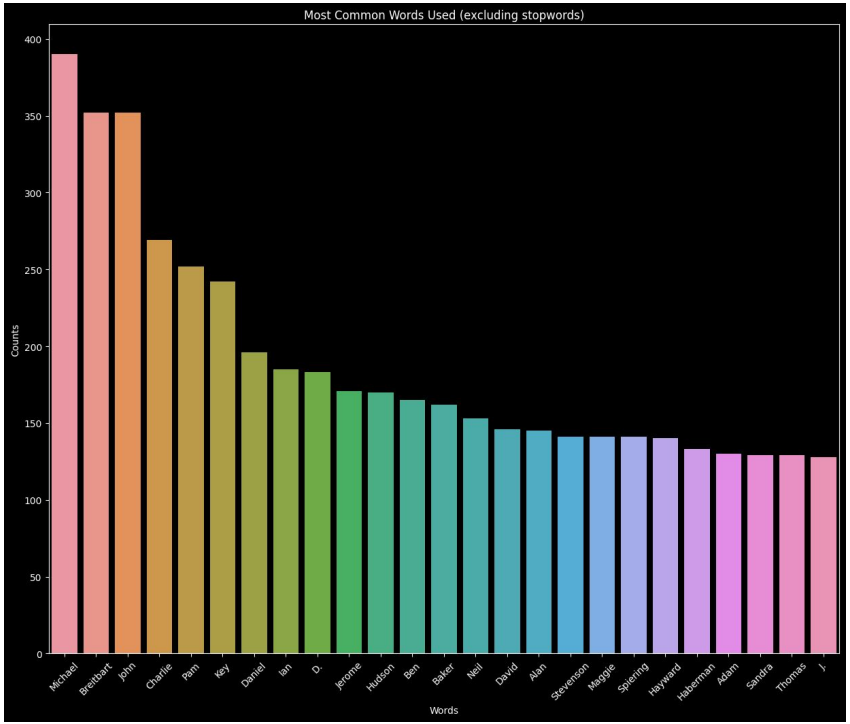
Comment différencier Fake news des vraies news?

Des auteurs bien différents

Fake

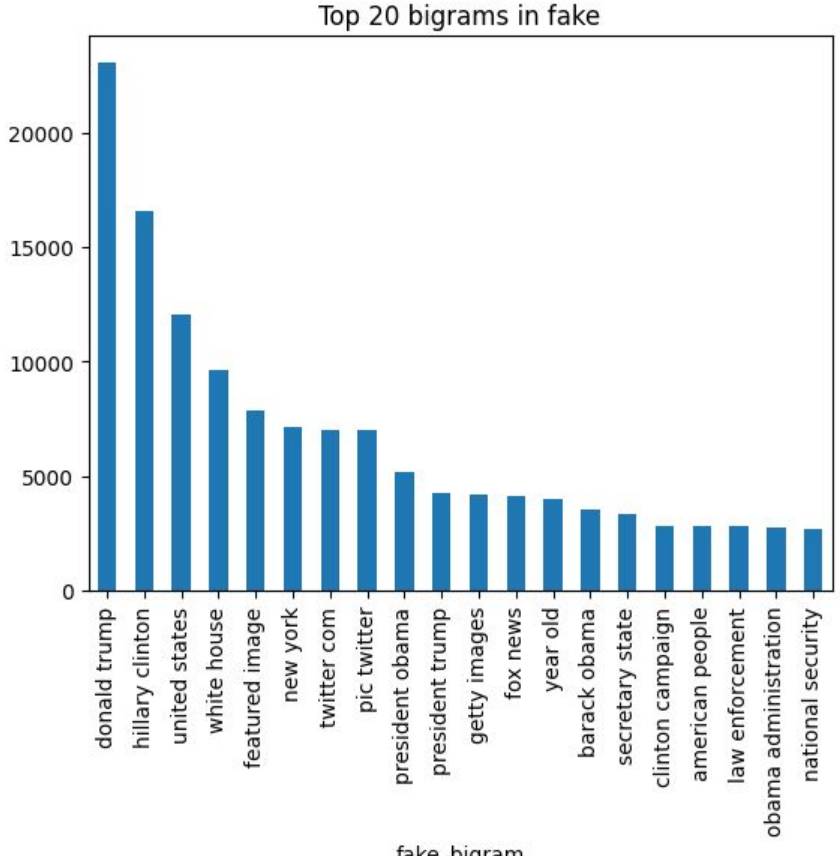


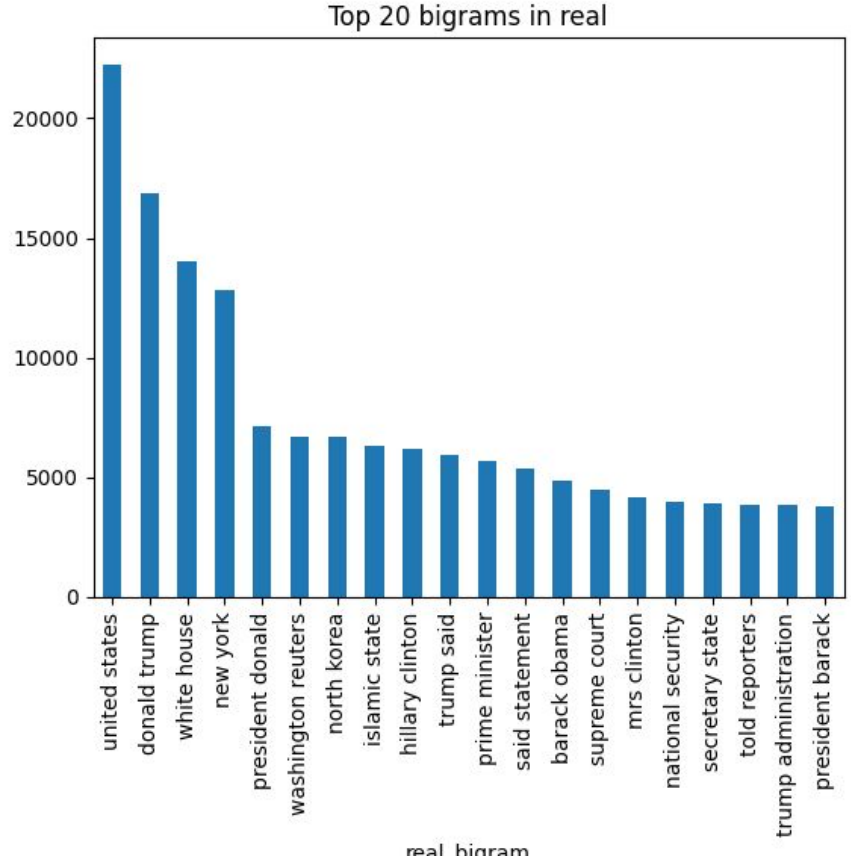
Real



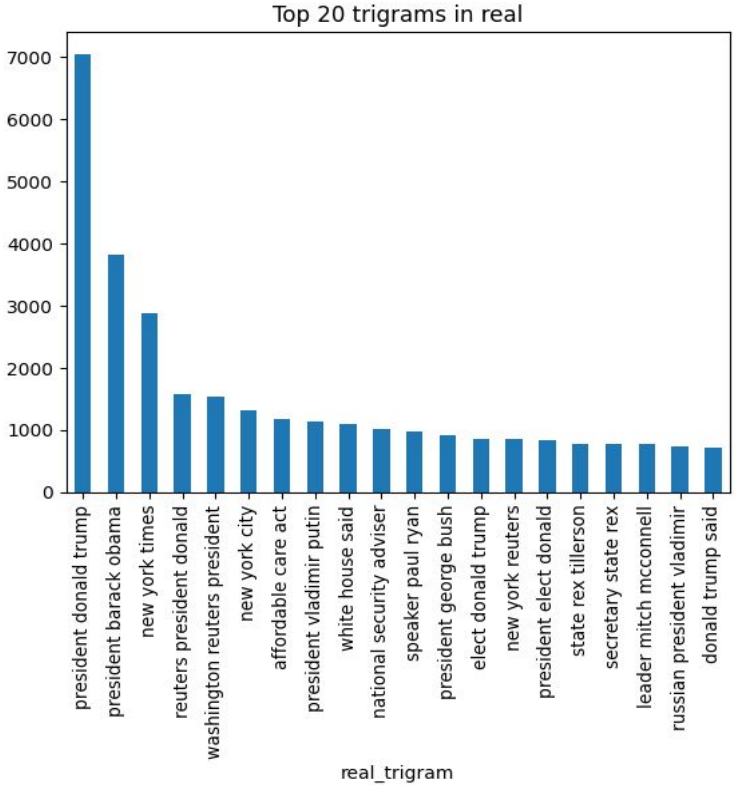
Etude de la fréquence des mots

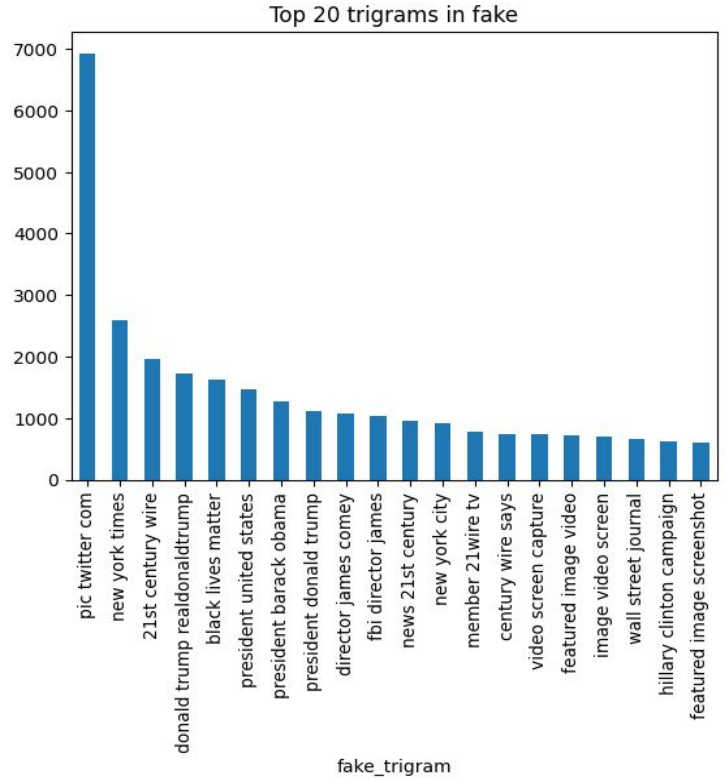
Les bigrams


Top 20 bigrams in fake
20000
15000
10000
5000
0
donald trump
hillary clinton
united states
white house
featured image
new york
twitter com
pic twitter
president obama
president trump
getty images
fox news
year old
barack obama
secretary state
clinton campaign
american people
law enforcement
obama administration
national security
fake_bigram


Top 20 bigrams in real
20000
15000
10000
5000
0
united states
donald trump
white house
new york
president donald
washington reuters
north korea
islamic state
hillary clinton
trump said
prime minister
said statement
barack obama
supreme court
mrs clinton
national security
secretary state
told reporters
trump administration
president barack
real_bigram

Les trigrams





Feature Transformation :

- Elimination des stop_words
- Stemmatisation (PorterStemmer)
- Vectorization (TFIDVectorizer)

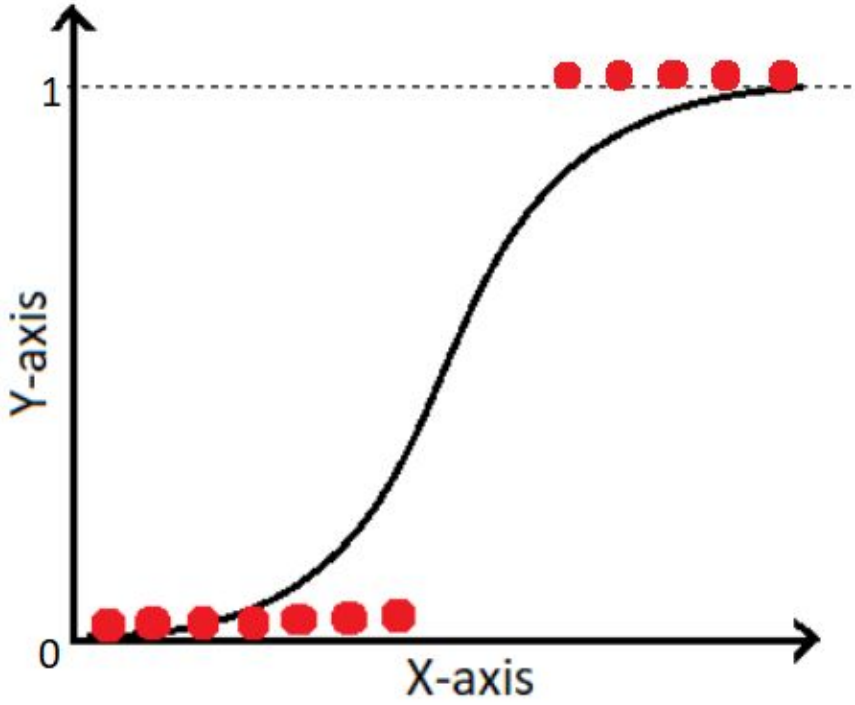
Natural Language Processing with Python!

Stop Words

["This", "is", "a", "test"]

✓ X X ✓

La régression logistique



Model : Logistic Regression (Author)

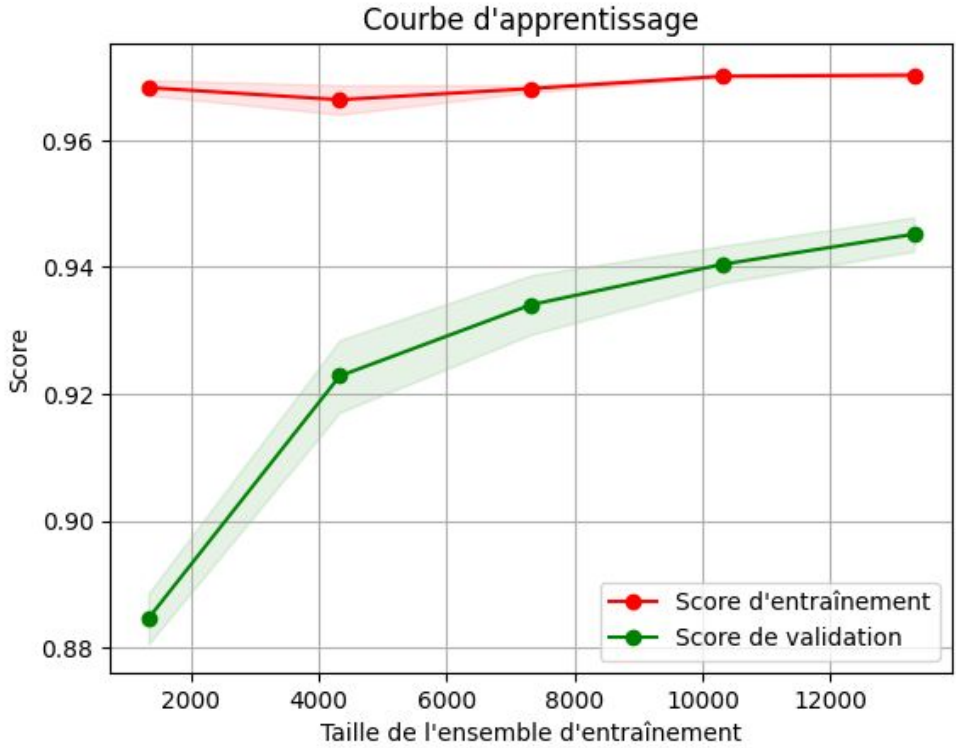
Accuracy score: 0.9865985576923076

	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)

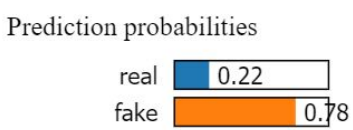
Accuracy score: 0.9401442307692308

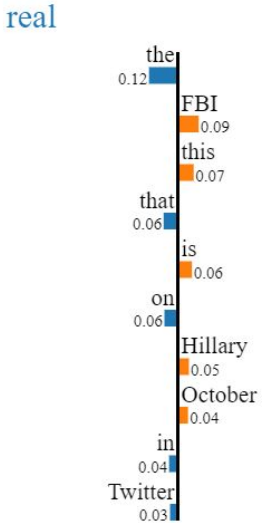
	precision	recall	f1-score	support
Real	0.94	0.94	0.94	2077
Fake	0.94	0.94	0.94	2083
accuracy			0.94	4160
macro avg	0.94	0.94	0.94	4160
weighted avg	0.94	0.94	0.94	4160



Explicabilité du modèle: lime

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

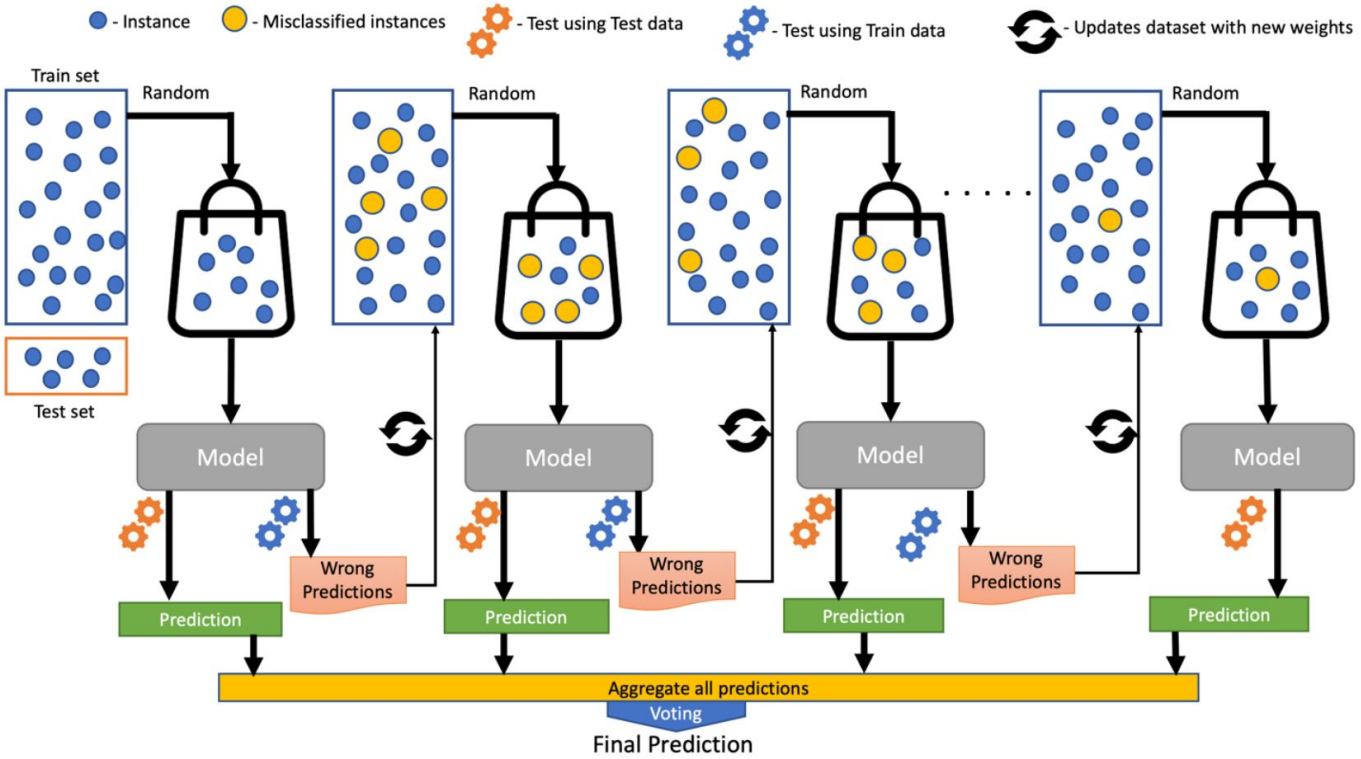




Text with highlighted words

House Dem Aide: We Didn't Even See Comey's Letter Until Jason Chaffetz Tweeted It By Darrell Lucus 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

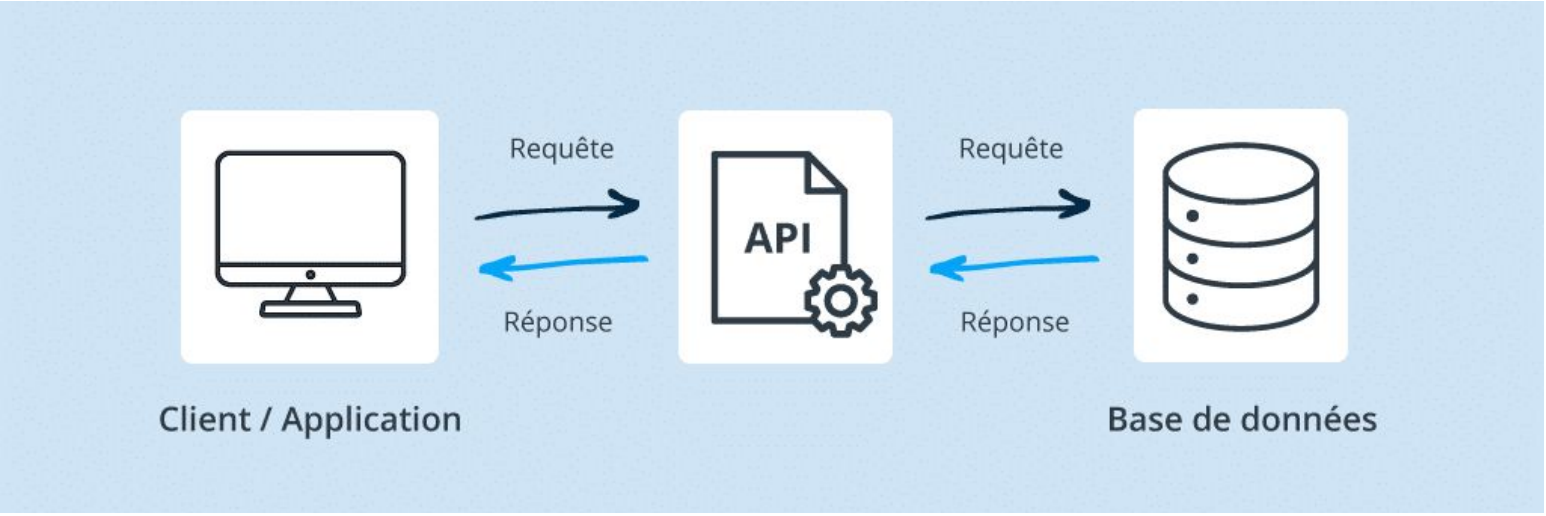

- Instance
- Misclassified instances
- Test using Test data
- Test using Train data
- Updates dataset with new weights
Train set
Random
Random
Random
Random
Test set
Model
Model
Model
Model
Wrong Predictions
Wrong Predictions
Wrong Predictions
Prediction
Prediction
Prediction
Prediction
Aggregate all predictions
Voting
Final Prediction

Model : XGBoost (Text)

Accuracy score: 0.9024038461538462

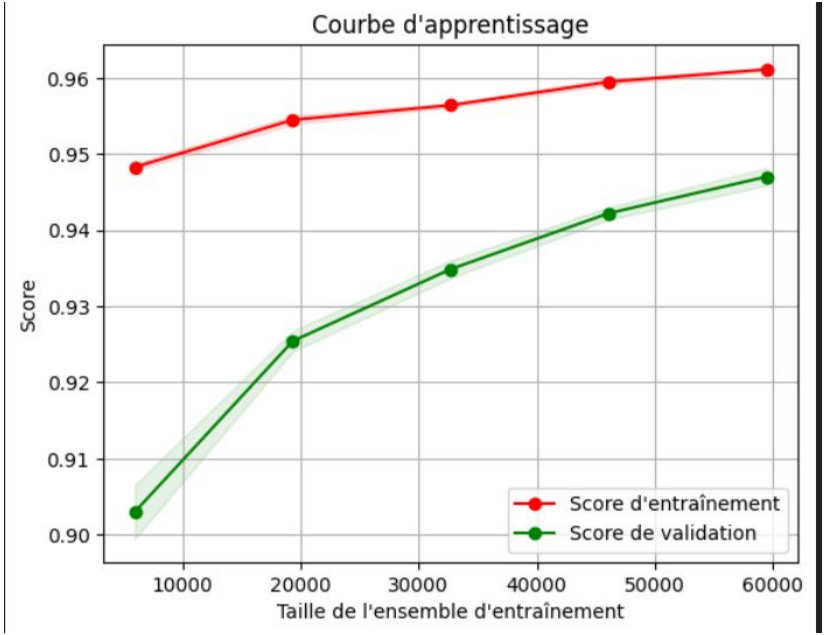
	precision	recall	f1-score	support
Real	0.90	0.90	0.90	2077
Fake	0.90	0.90	0.90	2083
accuracy			0.90	4160
macro avg	0.90	0.90	0.90	4160
weighted avg	0.90	0.90	0.90	4160

L'application



Réentrainement : 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