Analyse des sentiments

HMIN232M – Méthodes de la science des données

B. Rima E. Youssef T. Shaqura

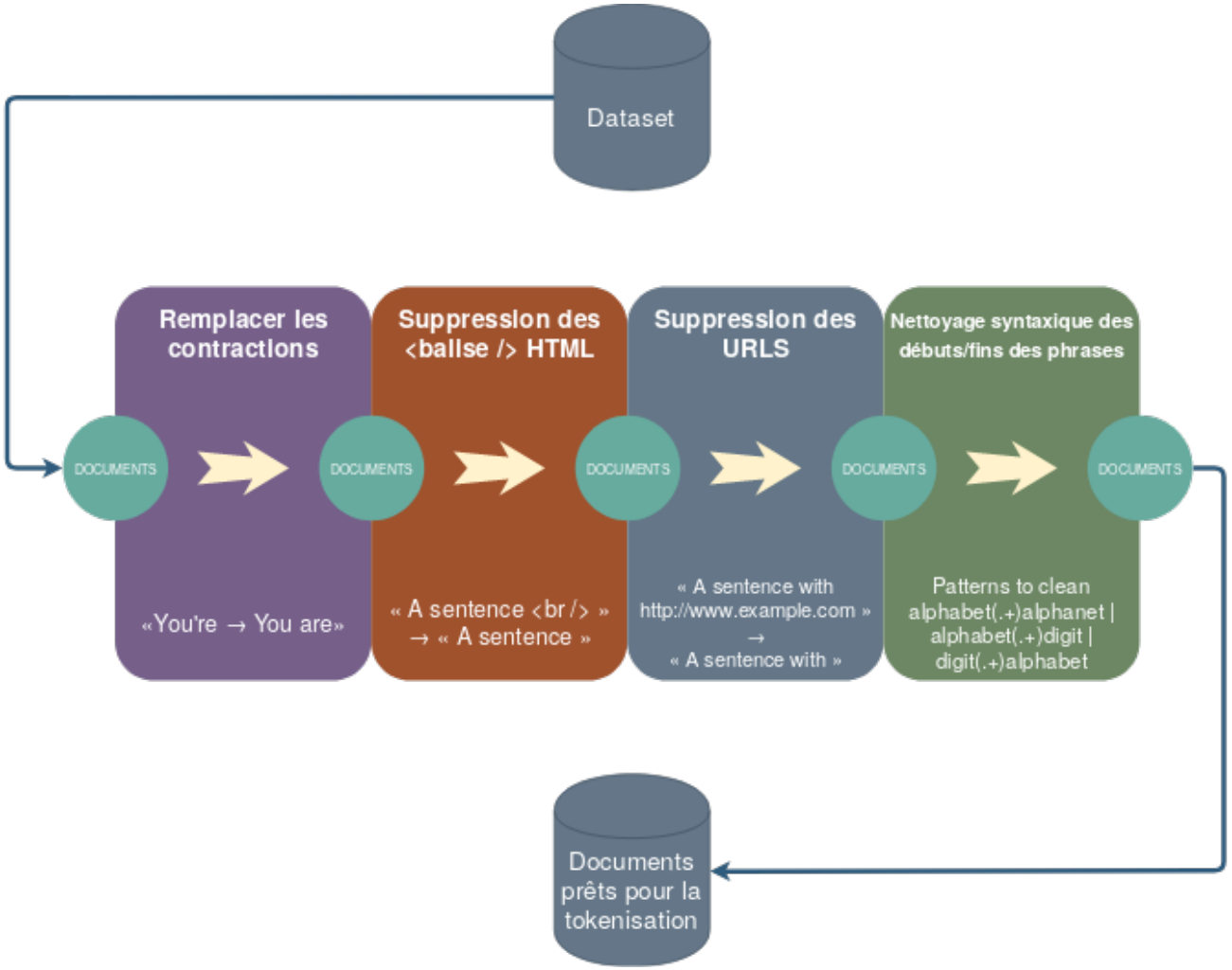
M1 Informatique AIGLE

25 avril 2019

Sommaire

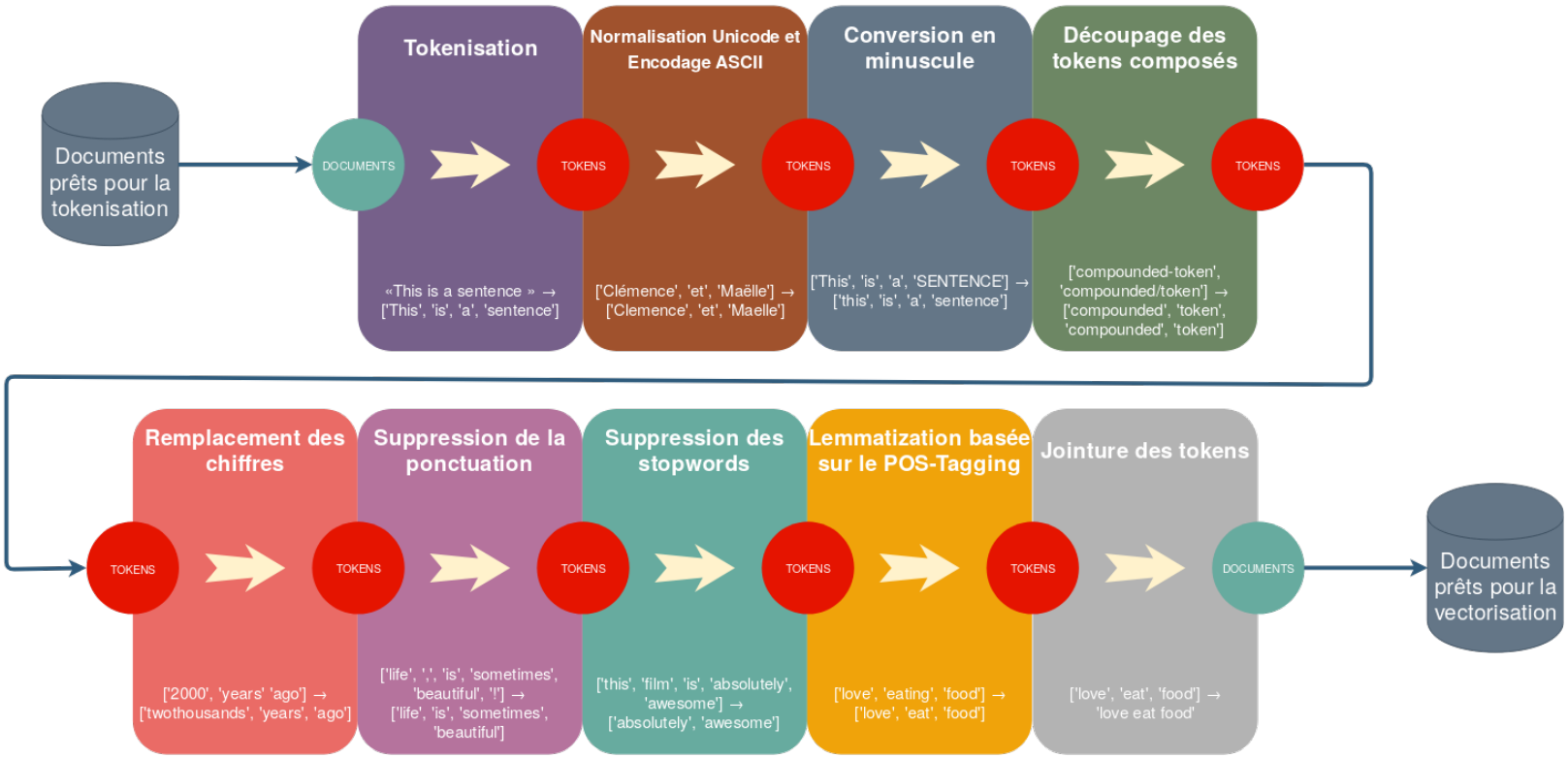
Préparation à la tokenization

Pré-traitements



Tokenisation et normalisation

Pré-traitements



WordCloud

Visualisation des données

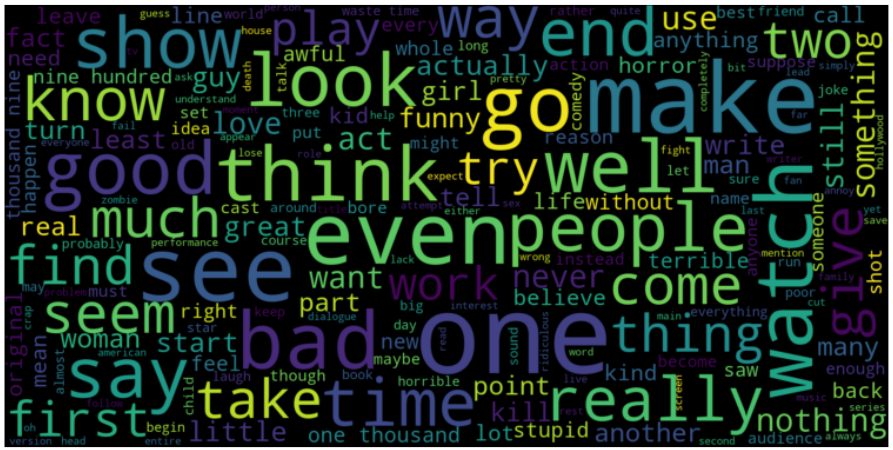
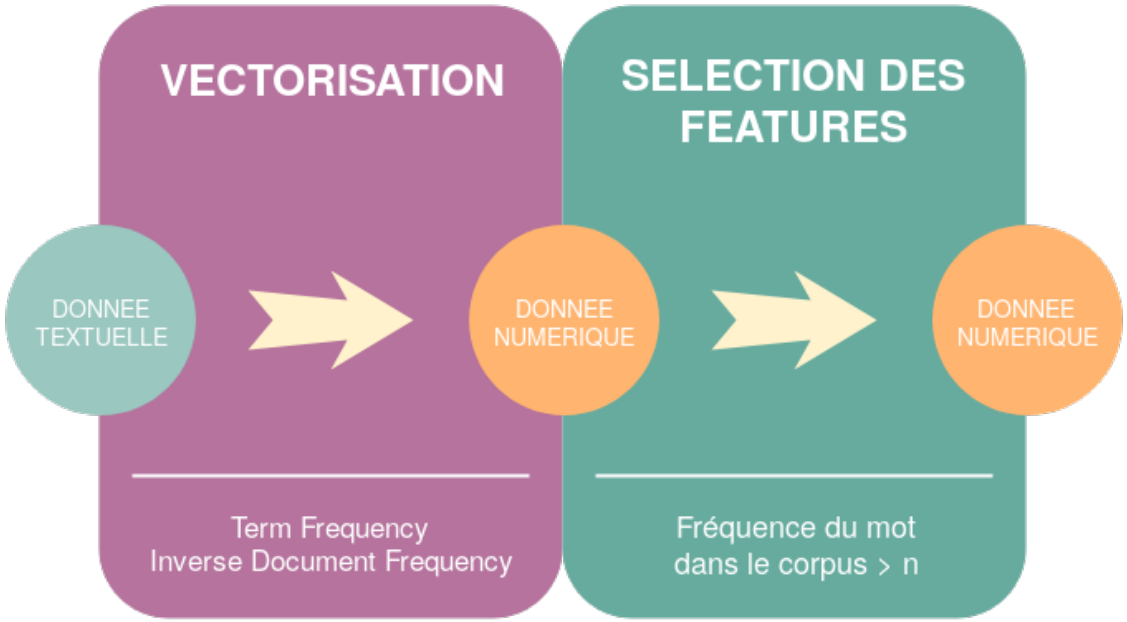


Figure – Les mots les plus fréquents dans les avis négatifs

On peut s'attendre à ...

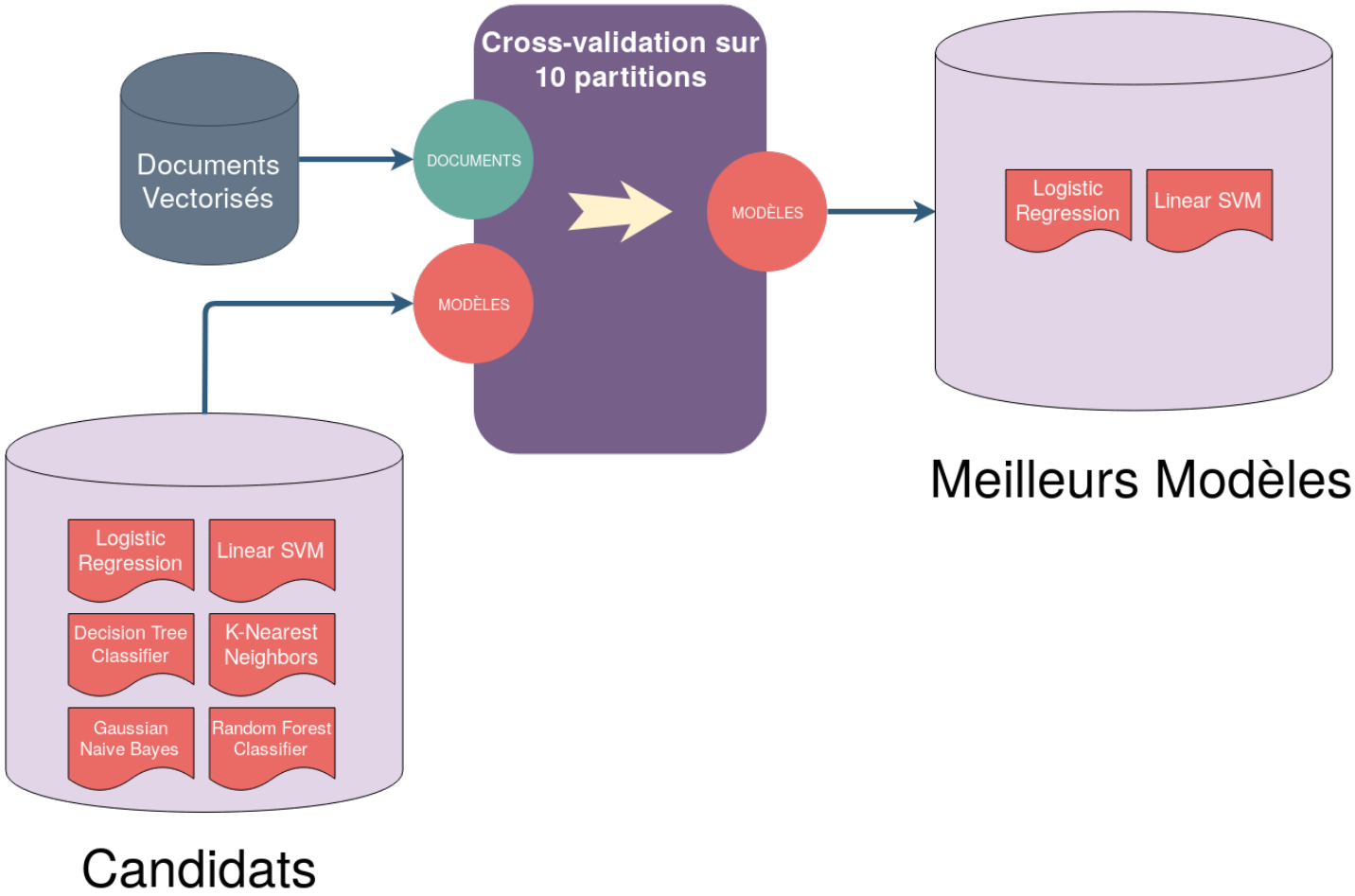
- Beaucoup d'ironie
- Phrases à polarités différentes dans les avis

Vectorisation et sélection des features



Principe

Cross-validation



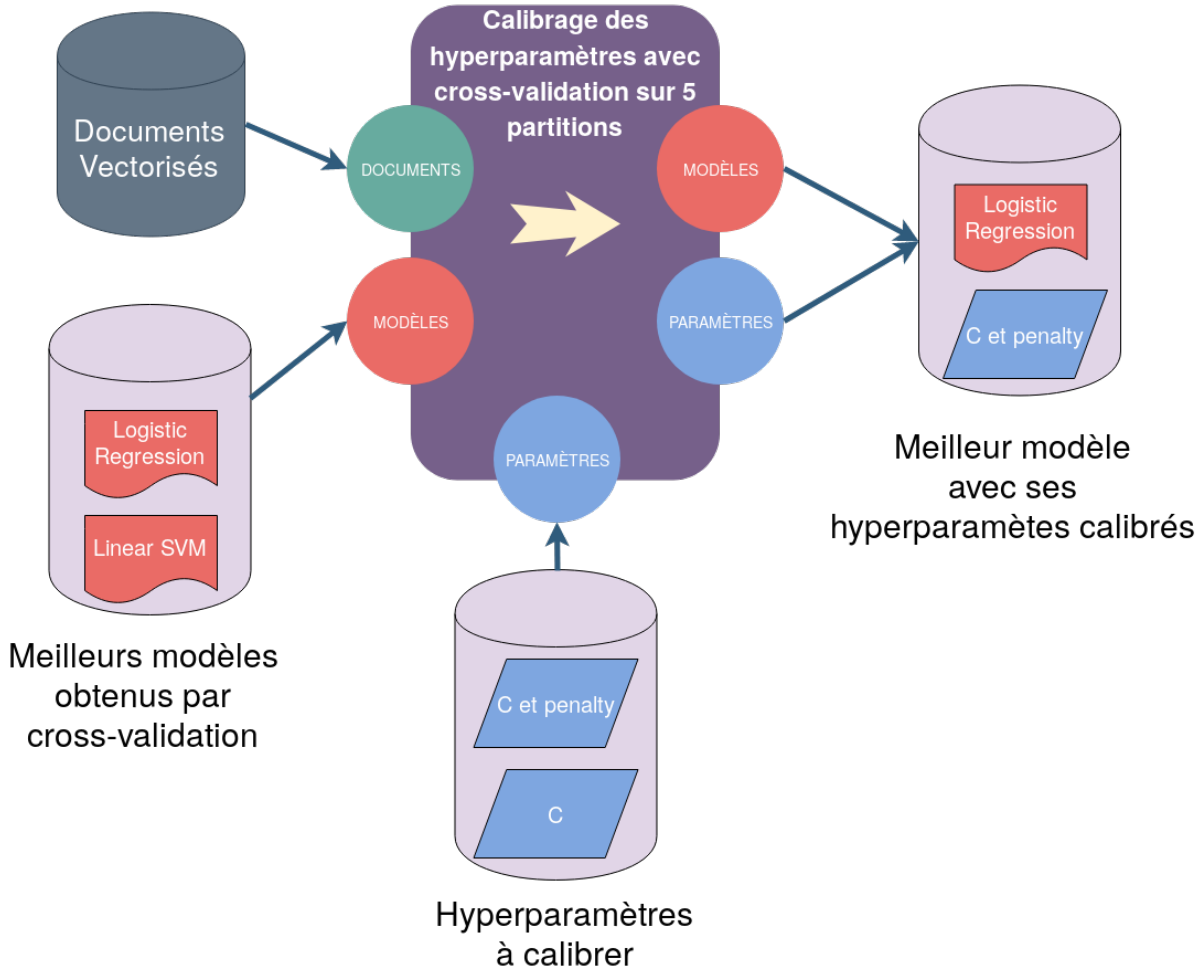
Résultats de la cross-validation

Cross-validation

Modèle	$\overline{score}$	σ
LinearSVC	92%	1%
SGDClassifier	92%	1%
LogisticRegression	91%	0.8%
GaussianNB	84%	1%
RandomForestClassifier	81%	1%
KNeighborsClassifier	79%	1%
DecisionTreeClassifier	75%	0.8%

Principe

Calibrage des hyperparamètres



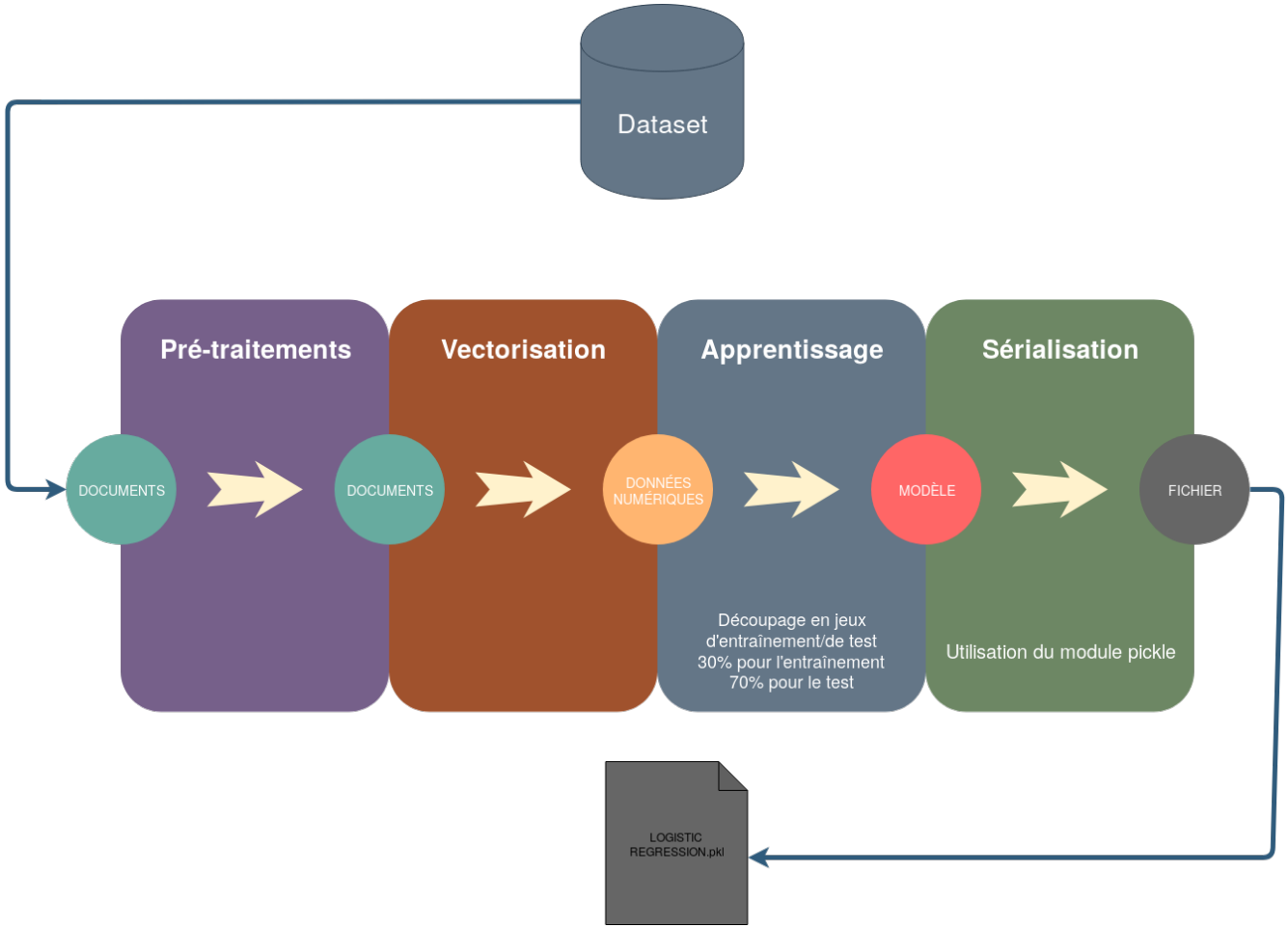
Résultats du calibrage

Calibrage des hyperparamètres

Modèle	$\overline{score}$	**Meilleurs calibrages**
LogisticRegression	90%	C = 11.288 ; penalty = L_2
LinearSVC	90%	C = 1

Pipeline pour Logistic Regression

Création des pipelines



Résultats pour le dataset du challenge

Création des pipelines

Accuracy : 89%

Temps pour effectuer la prédiction $\approx$ 41 secondes

Matrice de confusion :

$$\begin{pmatrix} 1770 & 230 \\ 190 & 1810 \end{pmatrix}$$

	Precision	**Recall**	**F1-score**	**Support**
-1	90%	89%	89%	2000
1	89%	91%	90%	2000
Micro avg	90%	90%	90%	4000
Macro avg	90%	90%	89%	4000
Weighted avg	90%	90%	89%	4000

Résultats pour le dataset IMDB

Création des pipelines

Accuracy : 85%

Temps pour effectuer la prédiction $\approx$ 106 secondes

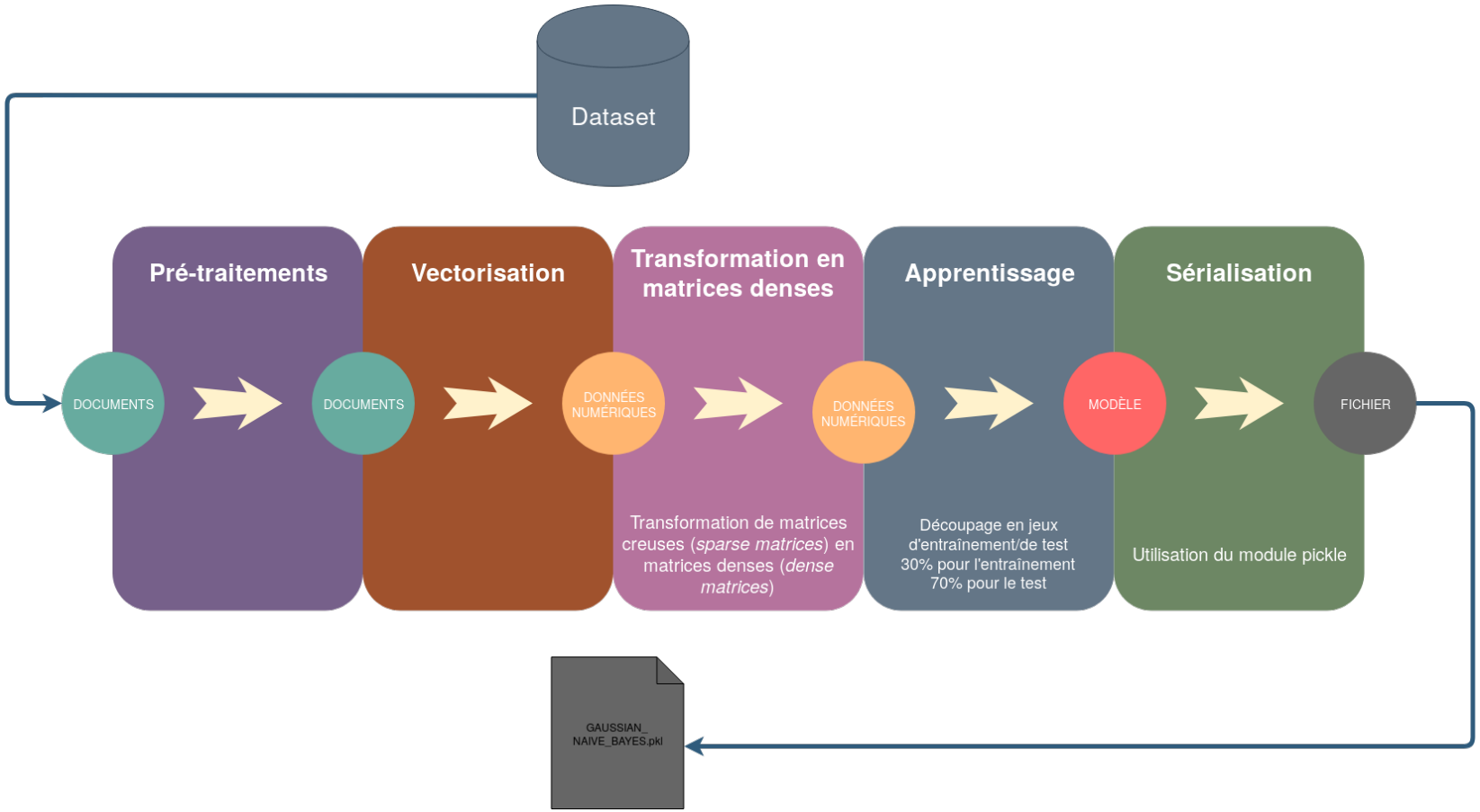
Matrice de confusion :

$$\begin{pmatrix} 4107 & 893 \\ 602 & 4398 \end{pmatrix}$$

	Precision	**Recall**	**F1-score**	**Support**
-1	87%	82%	85%	5000
1	83%	88%	85%	5000
Micro avg	85%	85%	85%	10000
Macro avg	85%	85%	85%	10000
Weighted avg	85%	85%	85%	10000

Pipeline pour Gaussian Naive Bayes

Création des pipelines



Résultats pour le dataset du challenge

Création des pipelines

Accuracy : 84%

Temps pour effectuer la prédiction $\approx$ 42 secondes

Matrice de confusion :

$$\begin{pmatrix} 1666 & 334 \\ 290 & 1710 \end{pmatrix}$$

	Precision	**Recall**	**F1-score**	**Support**
-1	85%	83%	84%	2000
1	84%	85%	85%	2000
Micro avg	84%	84%	84%	4000
Macro avg	84%	84%	84%	4000
Weighted avg	84%	84%	84%	4000

Résultats pour le dataset IMDB

Création des pipelines

Accuracy : 77%

Temps pour effectuer la prédiction $\approx$ 106 secondes

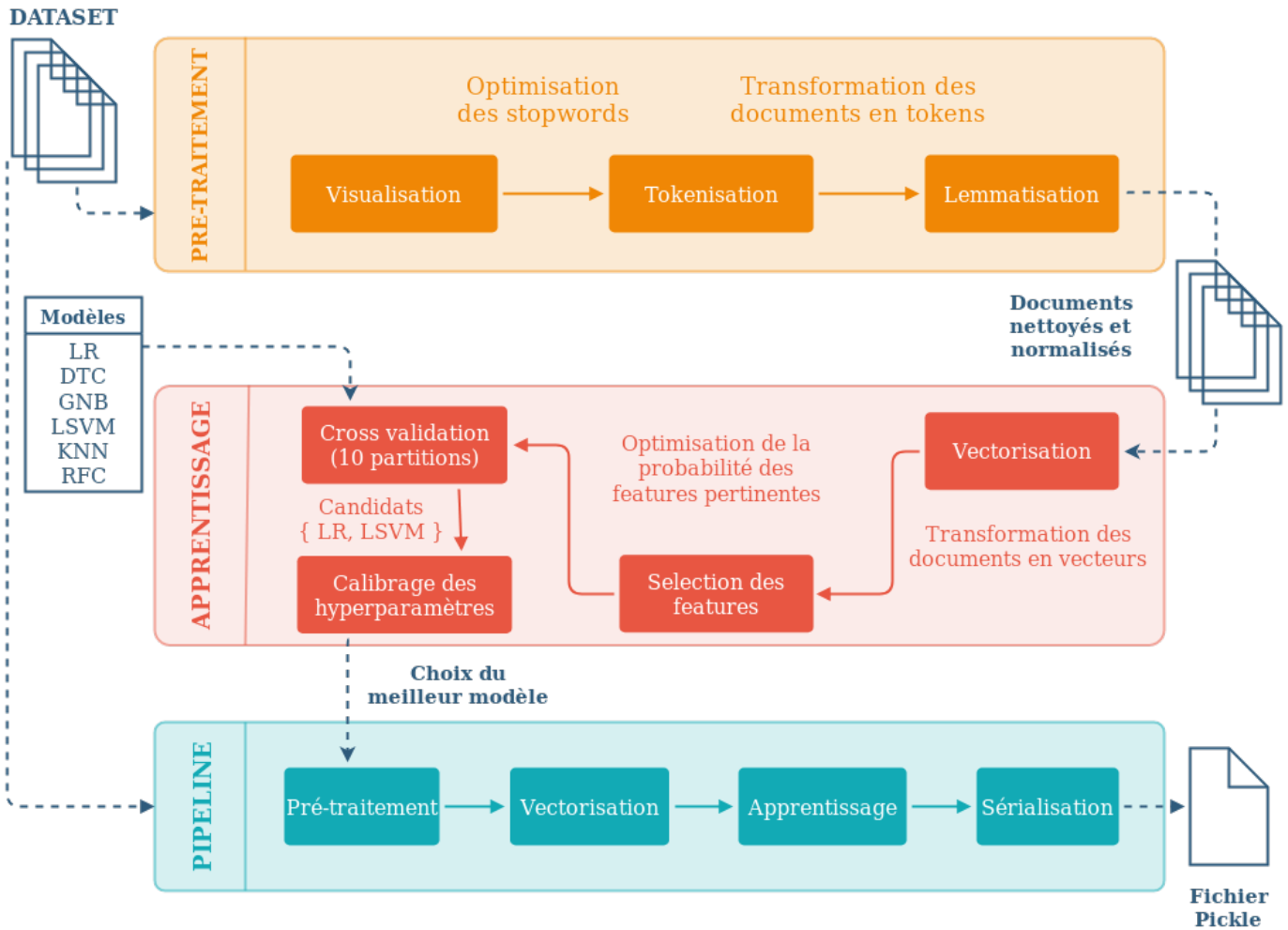
Matrice de confusion :

$$\begin{pmatrix} 3634 & 1366 \\ 914 & 4086 \end{pmatrix}$$

	Precision	**Recall**	**F1-score**	**Support**
-1	80%	73%	76%	5000
1	75%	82%	78%	5000
Micro avg	77%	77%	77%	10000
Macro avg	77%	77%	77%	10000
Weighted avg	77%	77%	77%	10000

Schéma globale de nos traitements

Conclusion



Perspectives

Conclusion

- Named Entity Recognition (NER)
- Traitement des ponctuations (?, !, ..., etc)
- SentiWordNet

Example

I love being cheated on !