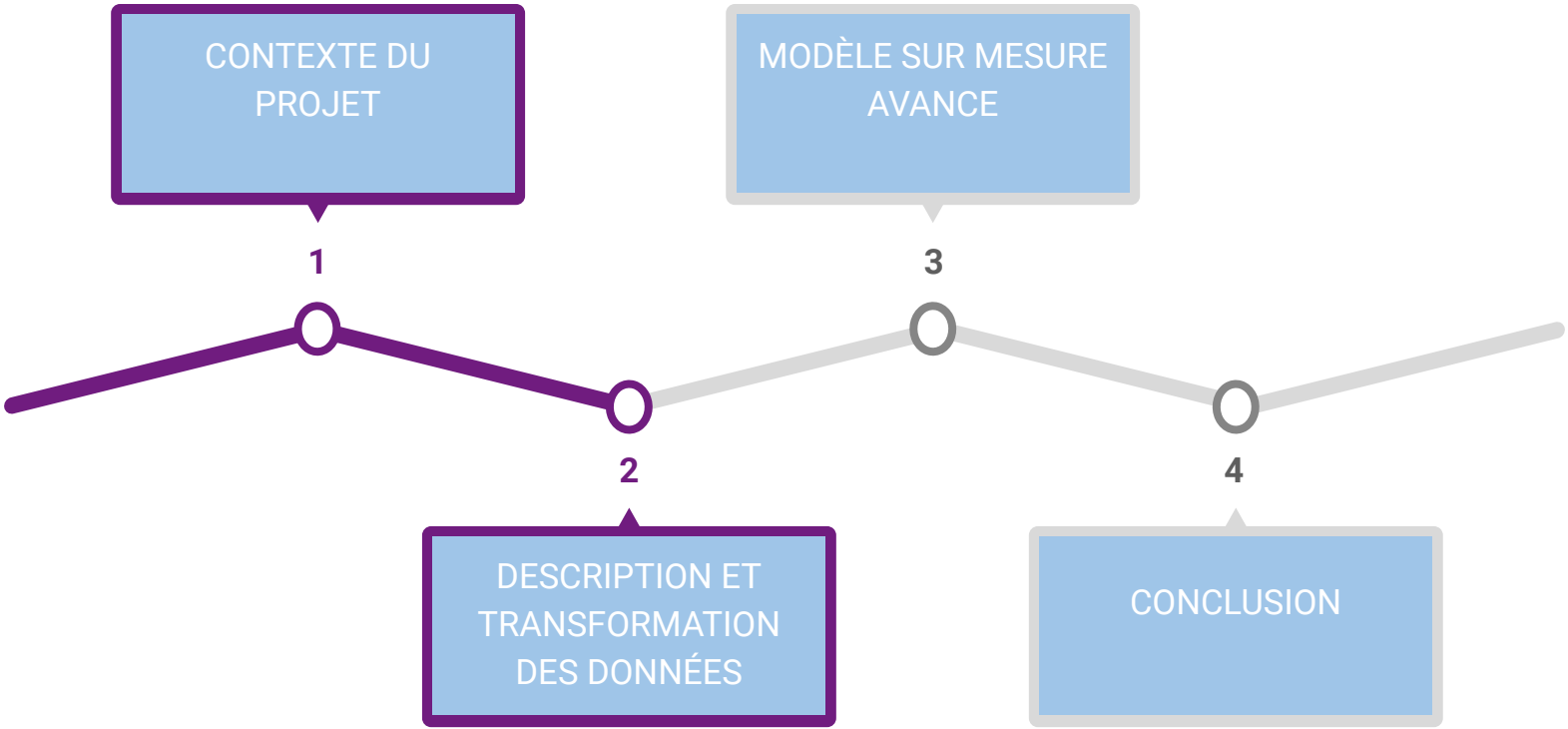
Détectez les Bad Buzz grâce au Deep Learning

Moussa KIBALY

SOMMAIRE


CONTEXTE DU PROJET
1
DESCRIPTION ET TRANSFORMATION DES DONNÉES
2
MODÈLE SUR MESURE AVANCE
3
CONCLUSION
4

1. CONTEXTE DU PROJET

La société Air Paradis souhaite connaître **les sentiments des tweets** postés par leur client sur les réseaux sociaux.

Pour cela, on a développé et déployé un modèle de prédiction du sentiment basé sur des réseaux de neurones profonds, à l'aide du studio **Azure ML de Microsoft**.

2. DESCRIPTION ET TRANSFORMATION DES DONNÉES

On part d'un jeu de données contenant 1 600 000 tweets et sentiments associés de différents utilisateurs de Twitter. Ces données ont été récupérées via une **API de Twitter.**.

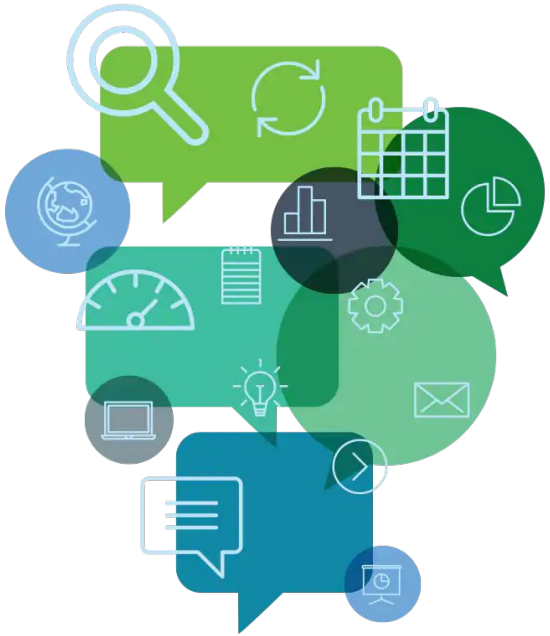
Seules les colonnes "sentiment" et "tweet" sont nécessaires à notre modèle.

La donnée textuelle "tweet" va être pré-traitées avant l'utilisation de nos modèles.

Fichier des tweets et sentiments des utilisateurs de Twitter

	sentiment	id	date	query	user	tweet
159995	1	1689009522	Sun May 03 12:25:02 PDT 2009	NO_QUERY	stephstar20	@mariaruizx lucky.. that is soo great! I wish ...
159996	1	1977823875	Sat May 30 21:47:31 PDT 2009	NO_QUERY	KGWSunrise	BB: I think Directors Paul and Scooter r keepi...
159997	1	1966616026	Fri May 29 18:05:24 PDT 2009	NO_QUERY	tmj4340	Power's back. I really didn't cut the power li...
159998	1	2003659622	Tue Jun 02 06:55:50 PDT 2009	NO_QUERY	jcimagination	Great website for Icon Search...http://www.ico...

PRÉ-TRAITEMENT DES DONNÉES TEXTUELLES



Les commentaires sont nettoyés en supprimant les signes de ponctuation et les nombres, la conversion des mots en minuscule. Les stopwords anglais sont aussi supprimés ainsi que certains stopwords particuliers (a, about, the, my, you, your, ...).

Les commentaires sont normalisés en effectuant une lemmatisation (conversion des verbes à l'infinitif, nom commun au masculin singulier) afin de récupérer les sens des mots. Le stemming qui consiste à récupérer la racine des mots par la suppression des suffixes et préfixes n' a pas été fait.

Tweets avant et après le nettoyage.

(Les signes de ponctuation et caractères spéciaux sont nombreux dans les tweets. Après le nettoyage, les signes de ponctuation et caractères spéciaux ont été supprimés et les mots convertis en minuscule.)

timent	id	date	query	user	tweet	clean_tweet
	1689009522	Sun May 03 12:25:02 PDT 2009	NO_QUERY	stephstar20	@mariaruizx lucky.. that is soo great! I wish ...	mariaruizx lucky that be soo great wish could ...
	1977823875	Sat May 30 21:47:31 PDT 2009	NO_QUERY	KGWSunrise	BB: I think Directors Paul and Scooter r keepi...	bb think director paul scooter r keep thing fr...
	1966616026	Fri May 29 18:05:24 PDT 2009	NO_QUERY	tmj4340	Power's back. I really didn't cut the power li...	power s back really didn t cut power line with...

Traitement des commentaires

Effet de la fonction stemming

get_stemmed_text('adresse immanquable meilleur magret paris ravioles foie gras tomber également personnel agréable superbe sélection')

```
Out[16]: 'adress immanqu meilleur magret paris raviol foi gras tomb égal personnel agréabl superb sélect'
```

Traitement des commentaires

Effet de la fonction de lemmatisation

get_lemmatized_text('adresse immanquable meilleur magret paris ravioles foie gras tomber également personnel agréable superbe sélection')

```
Out[18]: 'adresse immanquable meilleur magret pari raviole foie gras tomber également personnel agréable superbe sélection'
```

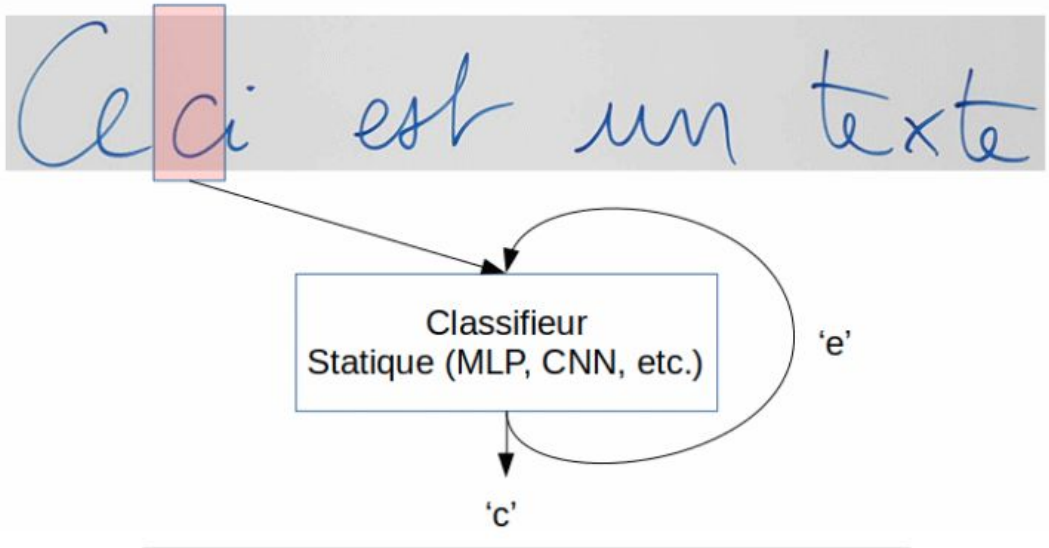
Tokenizer

La **Tokenisation** permet de convertir les mots d'un texte en un ensemble d'entier ou vecteur de longueur fixe

	abord	abordable	abrite	abriter	absolument	absolut	absolute	abîme	ac	accent	...	évidemment	évite	éviter	évolution	évoque	être	île	îlot	œil
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
639	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
640	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
641	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
642	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
643	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0

3. MODÈLE SUR MESURE AVANCE

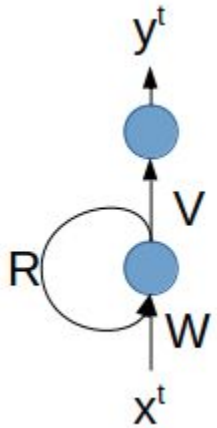
Les réseaux de neurones récurrents (RNN)



Les réseaux de neurones récurrents sont les modèles de réseaux de neurones dédiés au traitement de séquences, c'est-à-dire aux signaux de taille variable. Ils reposent sur 2 principes :

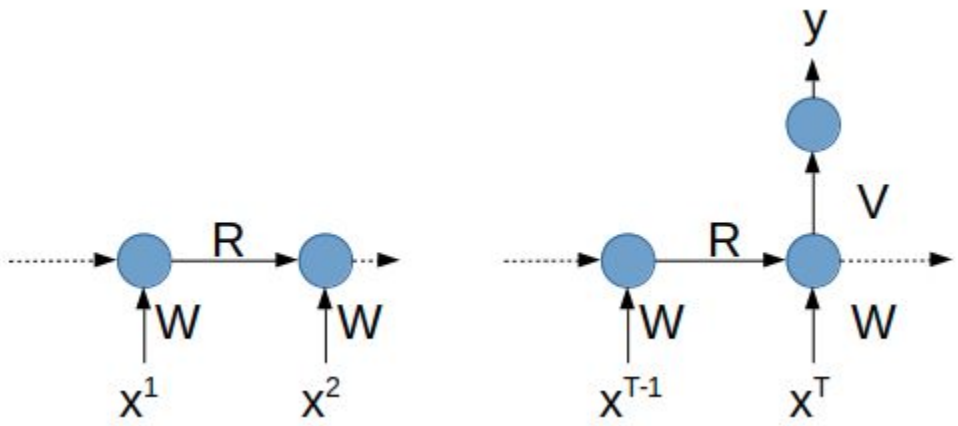
- principe de la fenêtre glissante permettant de traiter les signaux de taille variable
- ils utilisent des connexions récurrentes permettant d'analyser la partie du signal passée

Représentation simplifiée d'un RNN



Le RNN simple est constitué d'une couche récurrente et d'une couche dense. La couche récurrente permet au modèle de prendre une décision avec la mémoire de la décision précédente, qui elle-même dépend d'une décision précédente. On peut considérer le réseau a une mémoire infinie.

RNN pour la classification de séquence

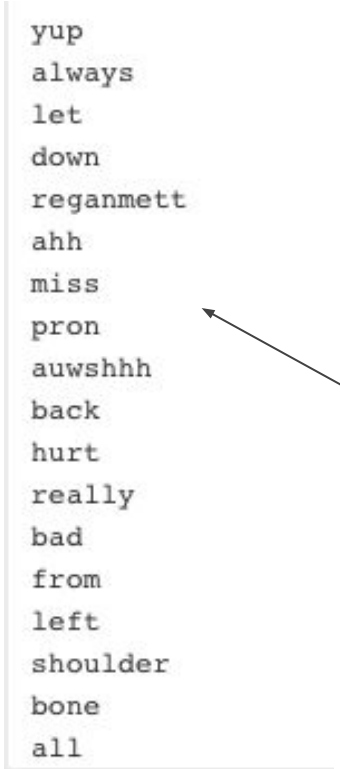


L'analyse de sentiment d'un texte passe par un mode de fonctionnement particulier du RNN permettant la classification de séquence.

La sortie y est produite lorsque toute la séquence a été lue.

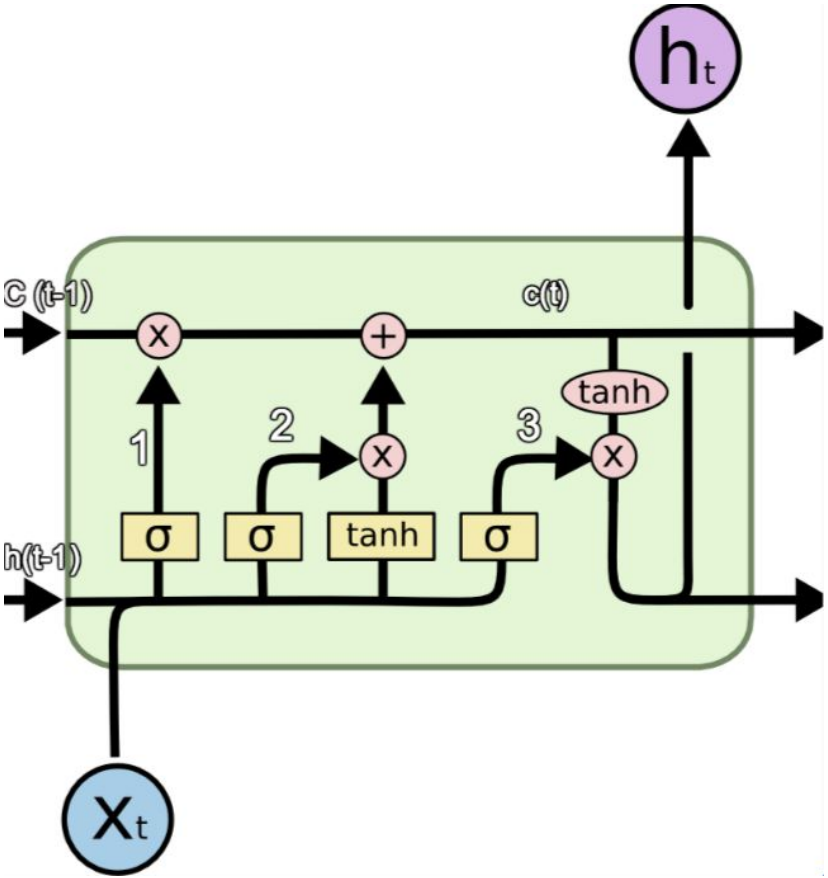
Méthode **Word2Vec sur tous les commentaires**

Cette méthode permet de transformer un texte en vecteur dense numérique de dimension inférieure. Elle regroupe les mots par similarité sémantique et syntaxique.



Mots les plus utilisés des commentaires quelque soit le rating. L'adjectif bon, les mots restaurant, plat sont souvent utilisés

RNN à mémoire interne

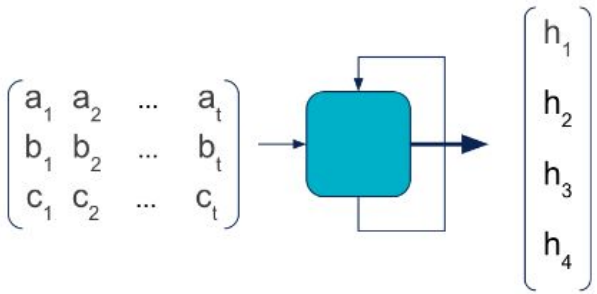


Ces RNN utilisent des cellules LSTM permettant de maintenir une mémoire à long terme.

La cellule possède 3 portes d'entrée considérées comme des vannes :

- porte d'entrée (modification ou non du contenu de la cellule)
- porte d'oubli (remise à 0 ou non du contenu de la cellule)
- porte de sortie (influence ou non du contenu de la cellule sur la sortie du neurone)

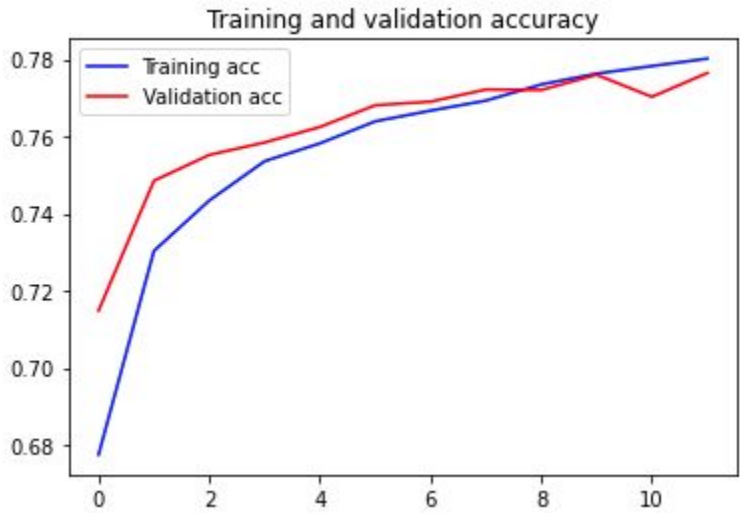
Architecture du modèle récurrent

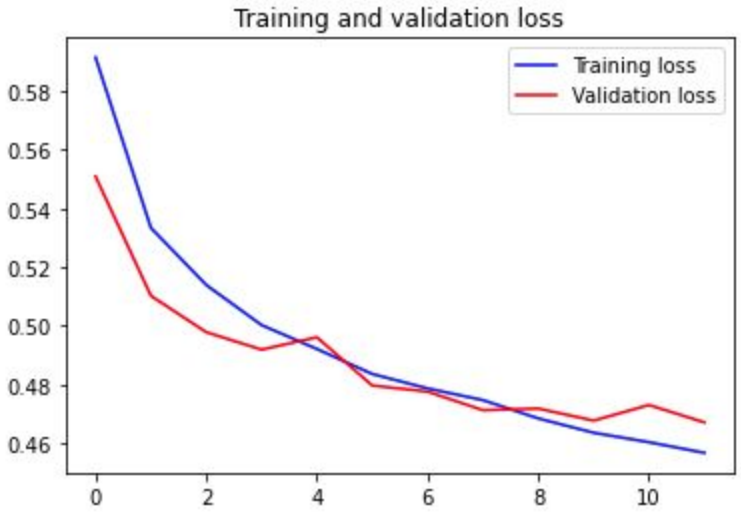


- Embedding Layer : Layer responsable de la conversion des tokens en représentation vectorielle généré par Word2Vec
- Bidirectional: Traitement bidirectional du texte. Cela signifie que le contexte des tweets est traité de gauche à droite et de droite à gauche
- LSTM: Long Short Term Memory, C'est un variant du RNN contenant une cellule à mémoire interne pour apprendre le contexte à long terme des mots plutôt que les mots voisins effectué par le RNN classique.
- Conv1D: couche convolutionnel 1D
- GlobalMaxPool1D: Réduction de la dimension en entrée en prenant le maximum pour chaque Dimension
- Dense : couche dense contenant 2 neurones et la fonction sigmoid donnant

Performance du modèle

Accuracy et fonction de perte

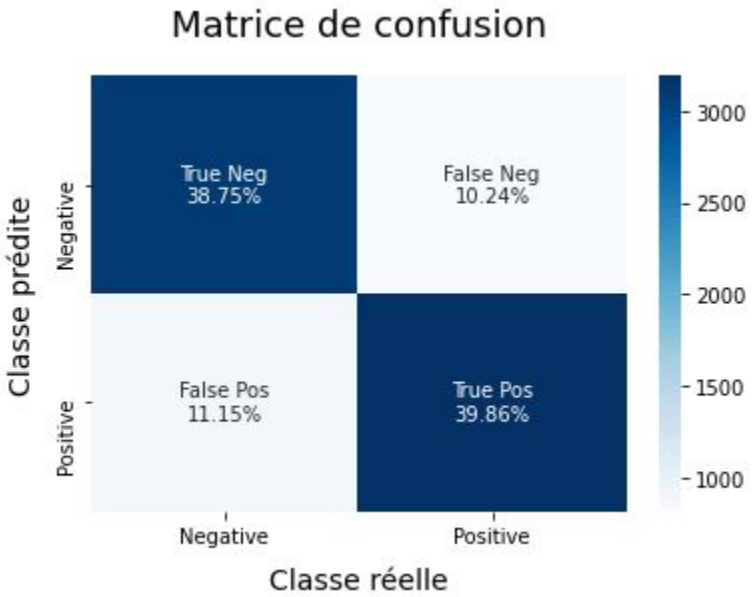




L'analyse de la performance par le biaise de l' **accuracy** et la fonction de perte (**loss**) nous montre une augmentation de la première et une diminution de la seconde en fonction des itérations (**epoch**).

Le modèle s'améliore

Matrice de confusion



Le heatmap à gauche permet d'évaluer notre modèle à maximiser les vrais négatifs (spécificité). En effet, il est primordial pour Air Paradis de connaître **les tweets vraiment négatifs** afin d'améliorer son image auprès des clients.

Rapport de la classification

	precision	recall	f1-score	support
0	0.79	0.78	0.78	3992
1	0.78	0.80	0.79	4008
accuracy			0.79	8000
macro avg	0.79	0.79	0.79	8000
weighted avg	0.79	0.79	0.79	8000

Le rapport de classification permet d'évaluer notre modèle de classification du sentiment en mesurant notamment le **f1-score**..

4. CONCLUSION

Modèle fonctionnel de prédiction du sentiment

Le modèle RNN implémenté et déployé dans Azure Machine Learning est appelé à travers une application web utilisant le framework **Streamlit.**

Ainsi, nous pouvons analyser le sentiment d'un texte en saisissant un texte via l'interface Streamlit puis en cliquant sur **Prédire**.