Développez une preuve de concept

Projet 9

Introduction

L'engouement actuel pour l'intelligence artificielle est incontestable. Même si les concepts mathématiques sont connus depuis des décennies, les technologies permettant l'usage des modèles les plus complexes deviennent de plus en plus accessibles.
C'est pourquoi j'ai choisi d'axer cette présentation sur un modèle de deep learning : le T5 développé par Google.
Nous verrons que ce modèle est très polyvalent ce qui simplifie grandement sa mise en œuvre.
Nous appliquerons ce modèle à **l'analyse de sentiments**.

La présentation se compose des parties suivantes:

1. **Rappel sur le Natural Language Processing**
2. **Présentation du modèle**
3. **Modélisation**
4. **Déploiement de l'API**
5. **Présentation du tableau de bords**
6. **Note sur l'accessibilité (WCAG)**

Introduction

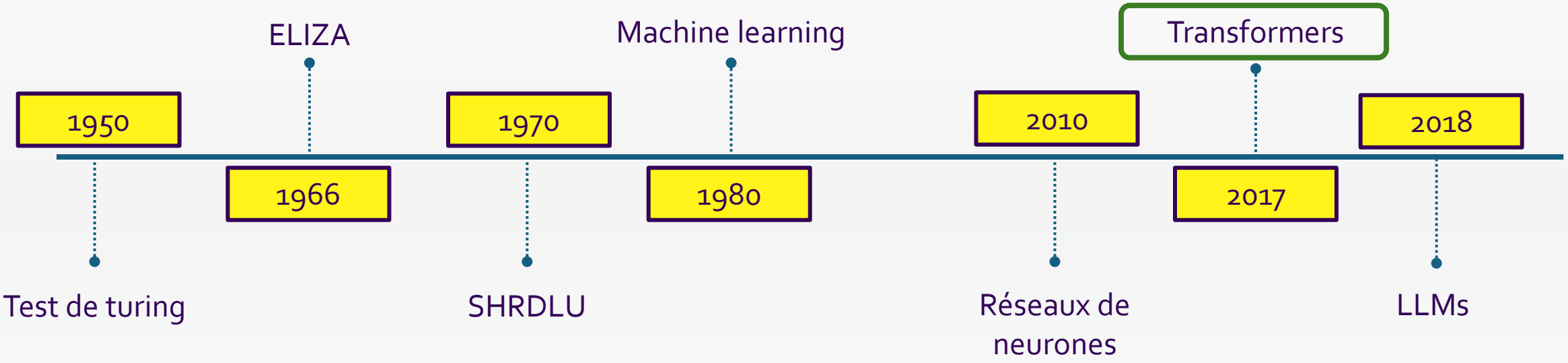
Les enjeux et application de l'analyse de sentiment sont nombreux

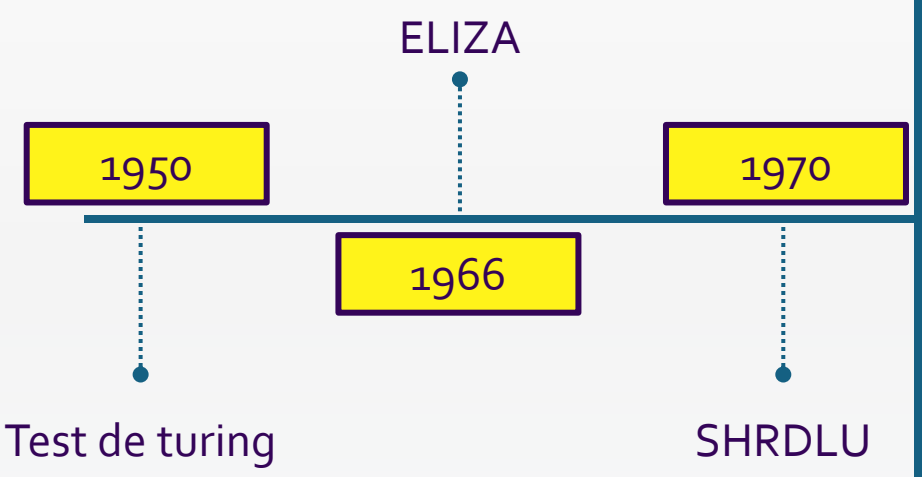
C'est un outil puissant pour mesurer la notoriété d'un produit, d'une marque ou d'une personne au travers des réseaux sociaux par exemple.

Il est aussi utilisé par exemple en finance afin d'anticiper des mouvements boursiers
https://www.bloomberg.com/professional/insights/data/can-get-edge-trading-news-sentiment-data/

Rappel sur le Natural Language Processing

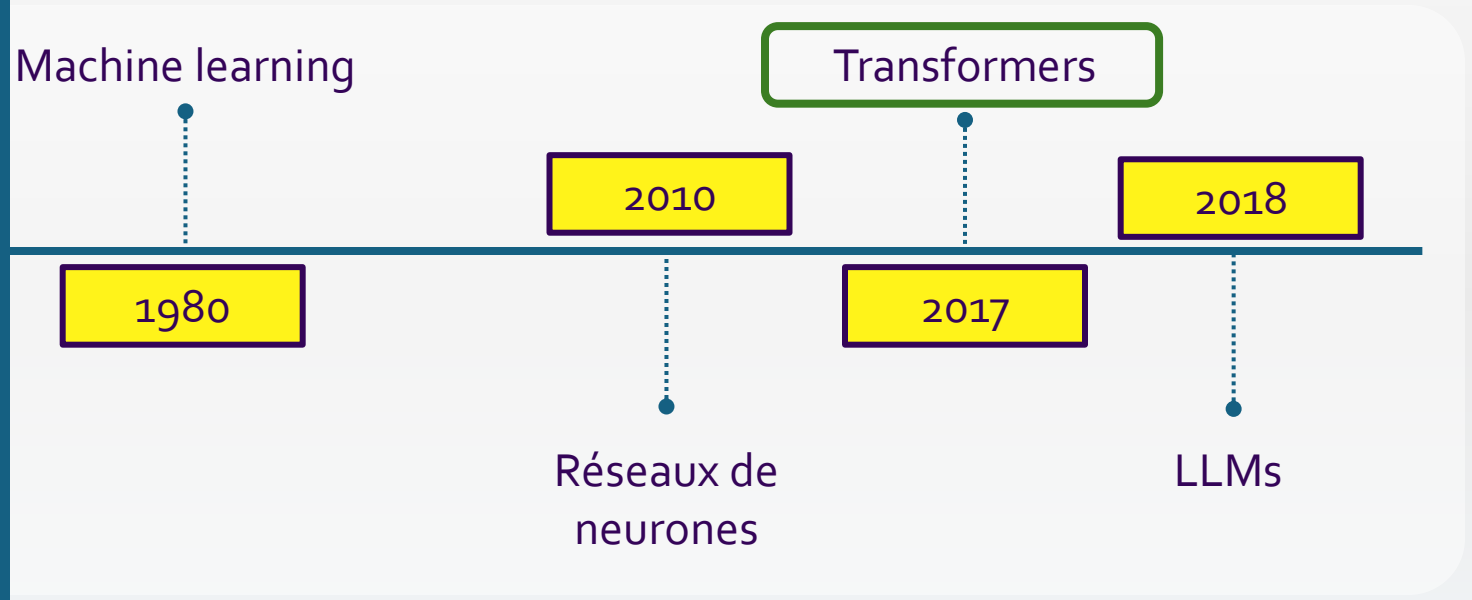
Le NLP est une branche du deep learning qui utilise des réseaux de neurones profonds pour analyser et comprendre le langage humain. Ces modèles apprennent à partir de grandes quantités de données textuelles pour effectuer des tâches telles que la traduction, la classification de texte, et la génération de langage naturel.





Les débuts du traitement automatique du langage naturel (NLP) remontent aux années 1950, lorsque les chercheurs ont commencé à imaginer des machines capables de comprendre le langage humain. Les premières approches étaient **symboliques** et **basées sur des règles :**
« IF …. THEN »

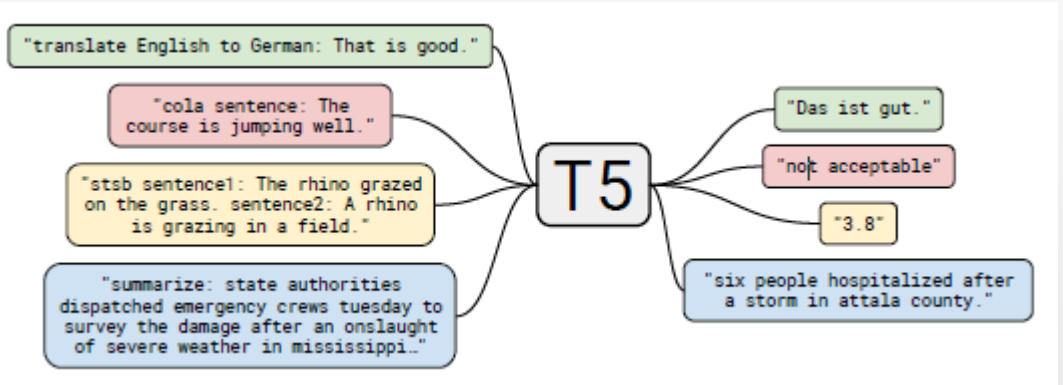
A partir des années 80 arrivent les méthodes statistiques, avec notamment le N-grammes, basées sur la **prédiction de mots** et la **reconnaissance de la parole**, en modélisant la probabilité d'occurrence d'un mot en fonction des précédents.



Présentation du modèle

Développé par Google en 2020, le T5, pour Text-to-Text Transfer Transformer, se distingue des autres modèles NLP par ses applications multiples :

-Classifications
-Analyses de sentiments
-Résumés de textes
-Traduction…



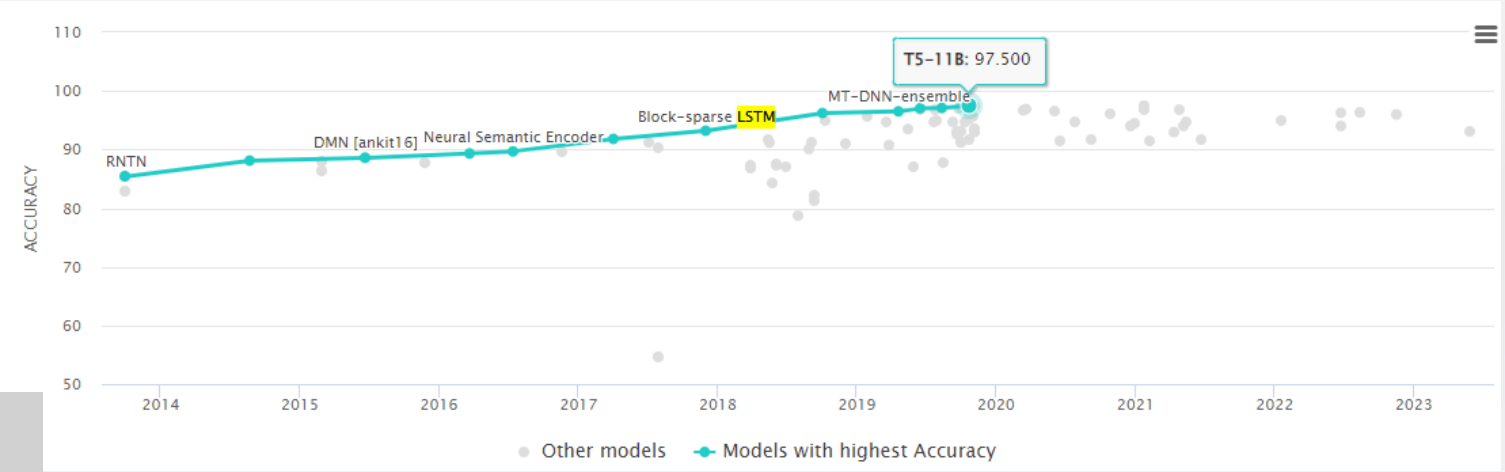
Source : Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Contrairement aux modèles antérieurs qui traitent chaque tâche de NLP différemment (par exemple, BERT pour la compréhension de texte et GPT pour la génération), le T5 formule toutes les tâches sous la forme de texte en entrée et de texte en sortie.

Présentation du modèle

Pourquoi le T5 ?

Ce modèle dans sa version la plus poussée est actuellement une référence en termes de précision sur l'analyse de sentiment d'après le site [Paper with code]



!

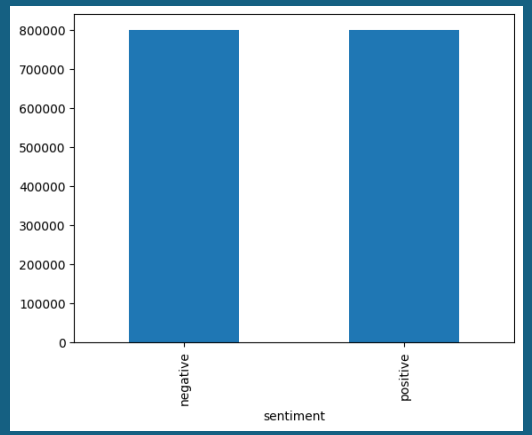
A noter que , à quelques exceptions près, les modèles avec plus de 90% d'accuracy sont des transformers

Modélisation

Les données

1,6 Millions de Tweets échantillonnés à 500 000 afin de limité le temps d'entraînement.
Les données sont équilibrées : autant de sentiments négatifs que positif. En cas de déséquilibre la classe majoritaire peut être privilégiée lors de la prédiction.

	sentiment	text
0	1	Gonna review again for Geom and read another c...
1	0	@CurlyRockTour I'm missing you in NYC this tim...
2	0	Rain follows me!!! Shopping in the rain = the ...
3	0	Nooo~! I'll miss Jo. #doctorwho #thegreendeath
4	0	Bdubs then Transformers 2 at 12:01 in IMAX. Wh...
...	...	...
29995	1	@lingenla can you bring season 2 into work tom...
29996	0	I hope by july 18 the weather will actually be...
29997	0	@celebritee I wish I had it in me man.. I just...
29998	1	@mrschemdoc He only called because my bro made...
29999	0	Back from Creative Coffee Club, good meeting ...



Tokenisation

```
Original :
 Its official. I'm sickies!

Tokens :
 ['its', 'official', 'i', 'm', 'sickies']

Stopwords removed :
 ['official', 'sickies']

Lem :
 ['official', 'sickies']
```

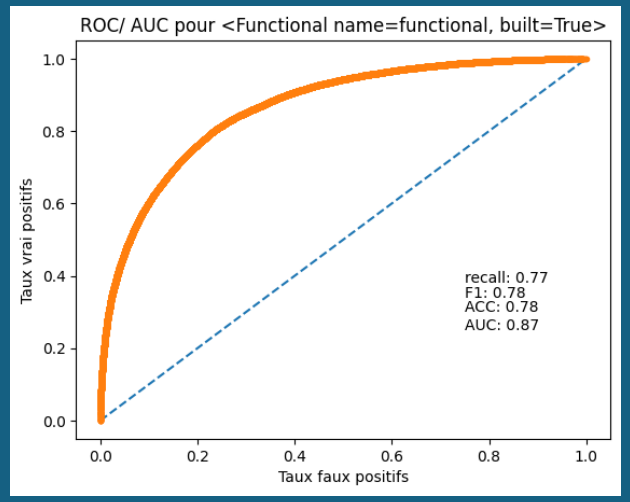
Exemple : « ***sentiment*** *: + texte à analyser* »

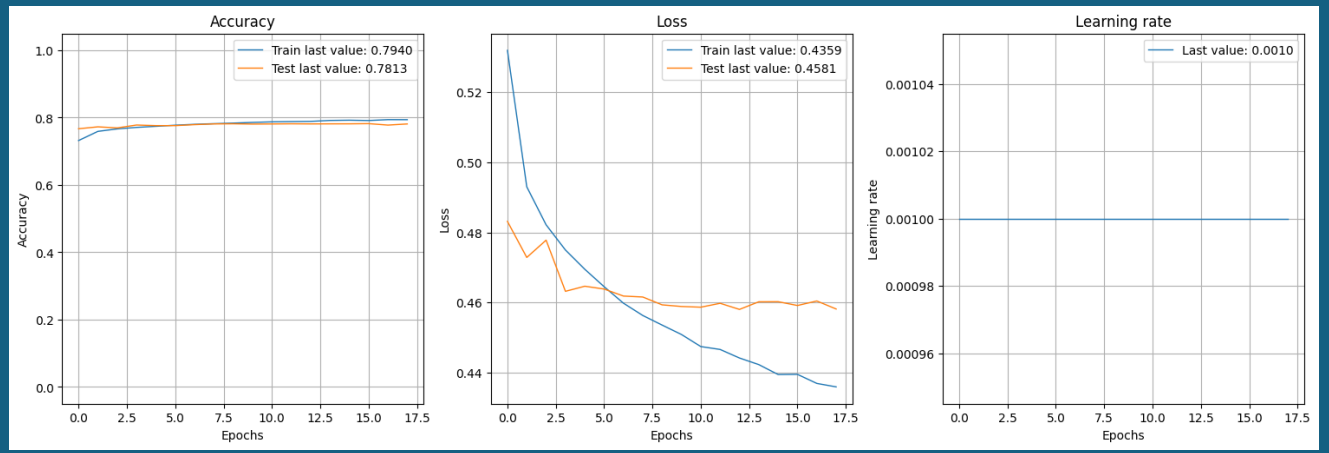
Modélisation

La **Baseline** ou référence est comme son nom l'indique un point de référence qui nous permettra d'évaluer la pertinence de notre nouveau modèle

Le LSTM (Long Short-Term Memory) appartient à la famille des **Recurrent Neural Networks**). Il a marqué une rupture parmi les modèles d'IA par sa capacité de mémorisation et donc d'interprétation du contexte. Il a cependant « la mémoire courte » et ne sait pas traiter les informations en parallèle ce qui limite ses capacités d'apprentissage.

LSTM
+ Glove embendding

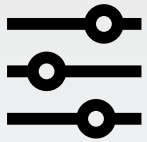




Modélisation

Entraînement du T5

Corpus C4 (Colossal Clean Crawled Corpus) : Pour entraîner T5, Google a utilisé un jeu de données très vaste appelé **C4,** qui est dérivé d'une grande partie du contenu disponible sur le web, mais avec un nettoyage et un filtrage intensif. Ce jeu de données contient environ **750 Go de texte**.



Le fine-tuning est l'étape qui permet au modèle de se spécialiser. Dans notre cas d'application qu'est l'analyse de sentiment cela se fait grâce aux tweets et à leurs attributs prédéfinis.

Modélisation

L'évaluation

T_5

Précision	Rappel	F1	AUC	Entraînement
0,78	0,77	0,78	0,87	~1h30

LSTM

Précision	Rappel	F1	AUC	Entraînement
0,85	0,86	0,85	N/a	~3h

Accuracy/précision
proportion de prédictions correctes

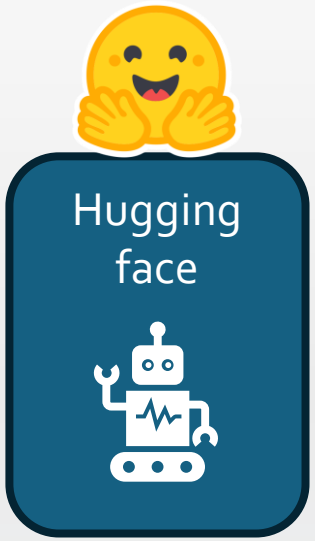
Recall/rappel
proportion de vrais positifs

F1
Equilibre entre rappel et précision

AUC
aire entre la droite représentant la baseline (hasard) et la courb ROC

Entraînement durée nécessaire pour la phase d'entraînement

Déploiement

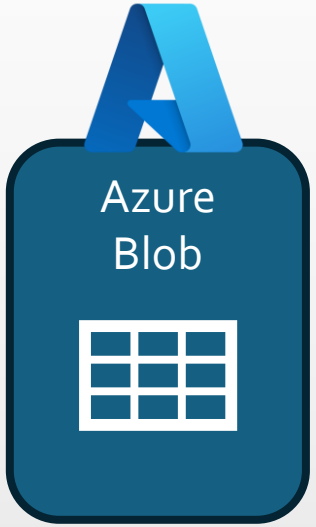


Hugging face est une plateforme collaborative qui héberge notamment des modèles pré-entraîné. Nous l'avons utilisé pour stocker notre modèle T5 entraîné sur nos Tweets.

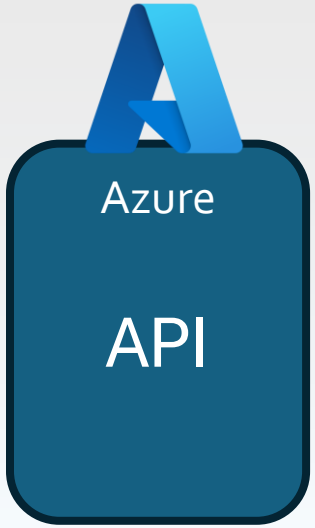
Déploiement

Les codes de l'**API** ainsi que de notre **interface Streamlit** sont stockés et déployé grâce à **Github**

Déploiement

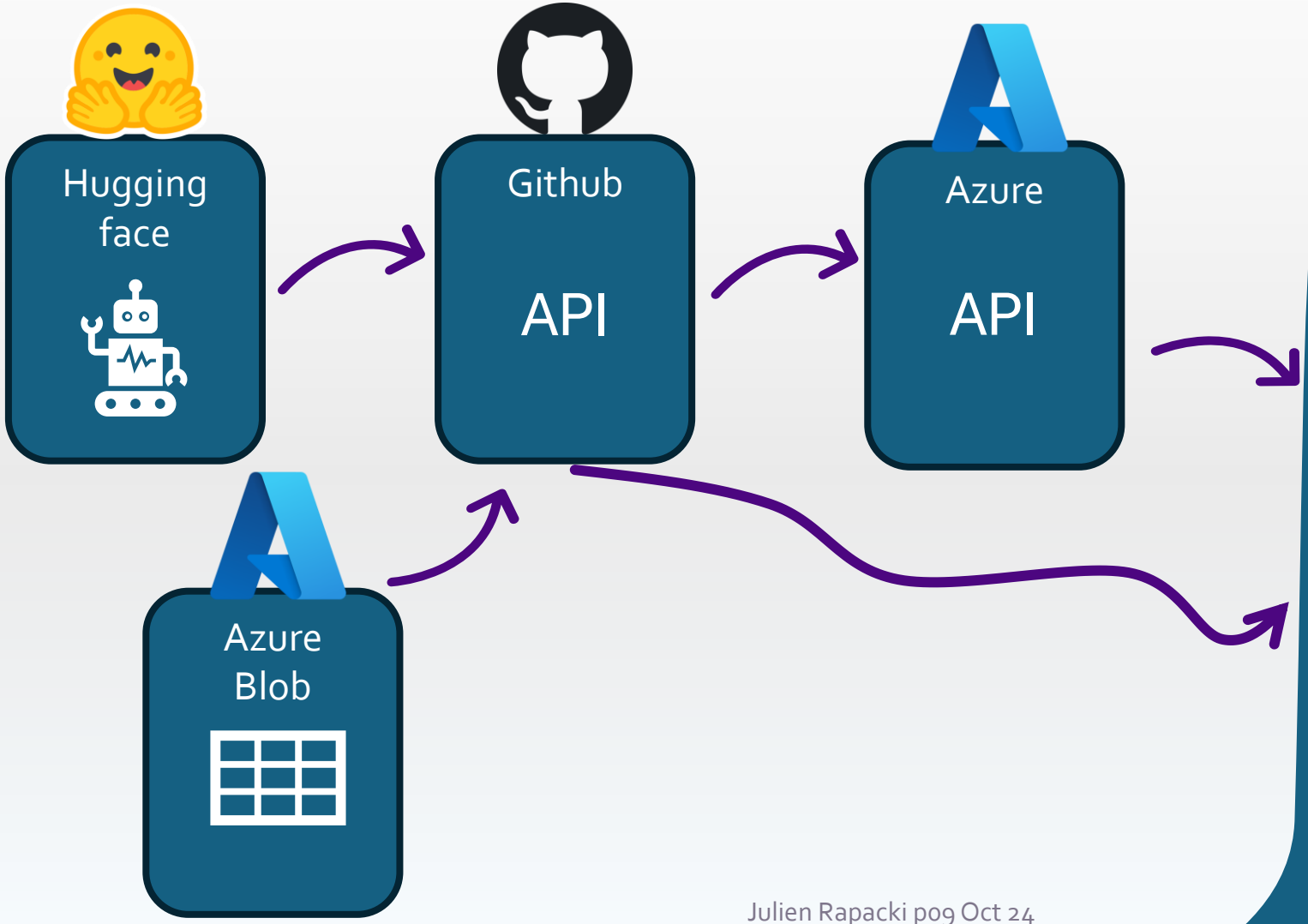


Azure Blob est un service de stockage cloud qui nous permet de stocker nos données sources (Tweets pour l'entraînement). Les données sont lues directement par notre application Streamlit.



L'API est déployée via Github Action sur Azure Web Apps et sera interrogée via l'interface Streamlit

Déploiement



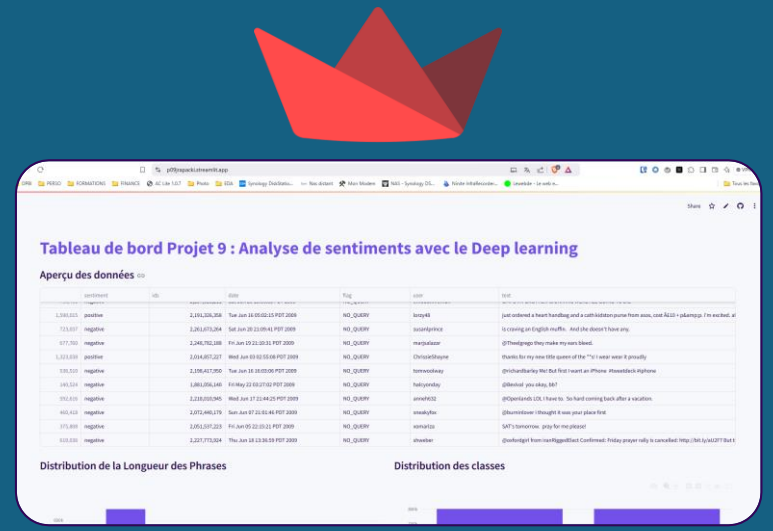


Tableau de bord

Echantillons de données dans un dataframe

- échantillon aléatoire de 50 tweets

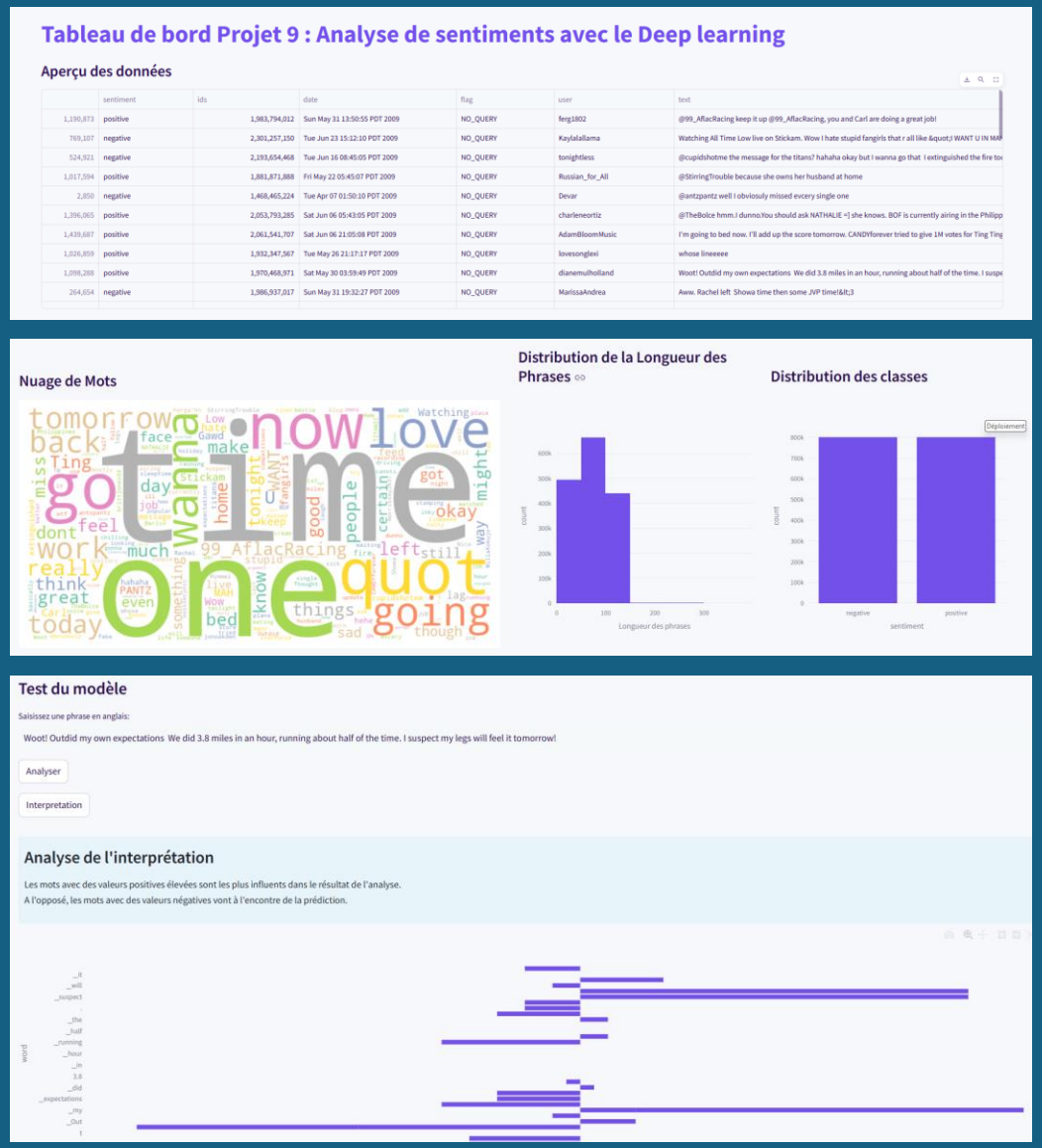
Analyses exploratoires

- Nuage de mots
- Distribution de la longueur des phrases et modalités

Test et analyse de l'interprétation du modèle

- Interprétation du résultat avec l'importance locale des mots de la phrase saisie

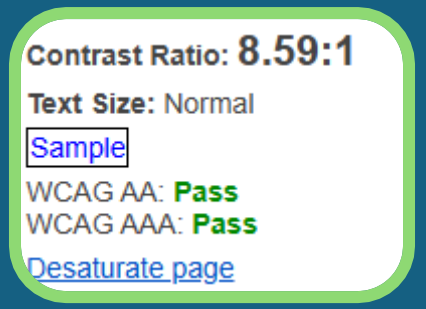
https://iap9jrapacki.streamlit.app/

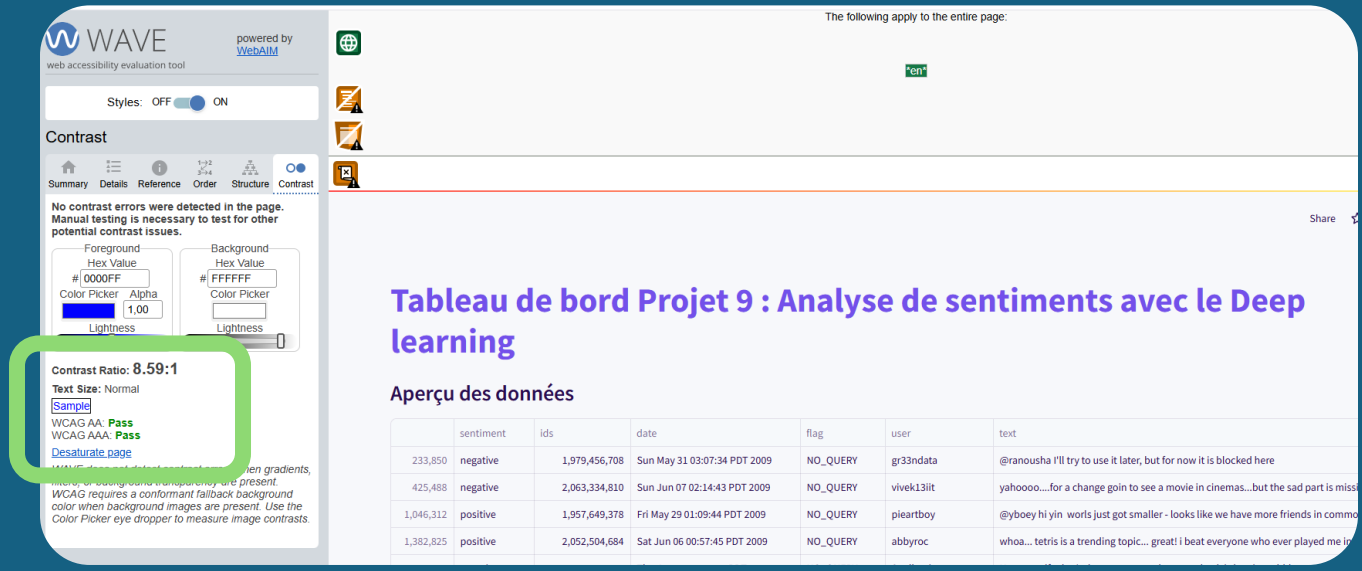


Accessibilité

La vérification de la conformité WCAG niveau A est effectuée à l'aide d'une extension du navigateur web appelée WAVE

- ✓ Contenu non textuel
- ✓ Utilisation de la couleur
- ✓ Contraste
- ✓ Dimensionnement du texte
- ✓ Titre de page





Conclusion

Le modèle T5 est très complet :

il est entraîné sur un corpus très vaste
il s'adapte à plusieurs cas d'usages (analyse de sentiments, résumés de textes….)
il est capable d'appréhender le contexte global de grande ampleur d'un texte
il peut donc distinguer les nuances subtiles du langage

Il faut cependant noter qu'à l'heure actuelle, et cela est confirmé par une publication de 2023 intitulée *Sentiment Analysis in the Era of Large Language Models: A Reality Check* les LLMs progressent à grande vitesse et surpassent le T5 même dans l'analyse de sentiment.

Dans le cadre de ce POC une version standard et gratuite d'Azure web a été utilisée et montre quelques limites quant à la durée d'inférence et son interprétation, en particulier pour les phrases de 10 tokens et plus.