Blog de la société MIC (Marketing Intelligence Consulting)

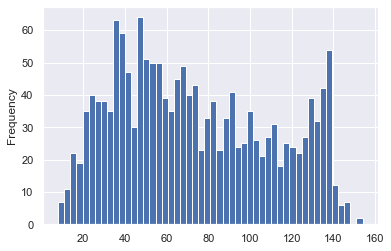
L’objectif du projet est de réaliser pour la société Air Paradis un prototype IA permettant de prédire les sentiments positifs, négatifs ou neutres à partir de tweets.

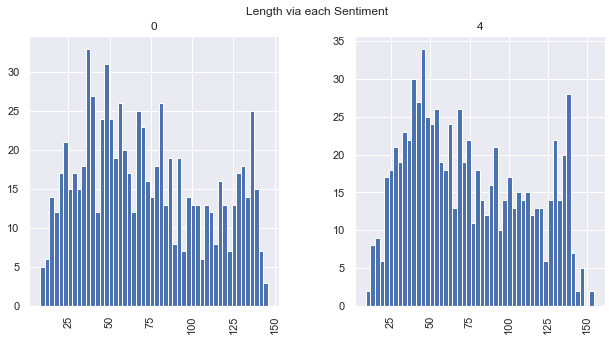
Le jeu de données est composé de seize million de tweets :

Quand la valeur de la variable target est égale à 4 cela correspond à un tweet de sentiment positif.

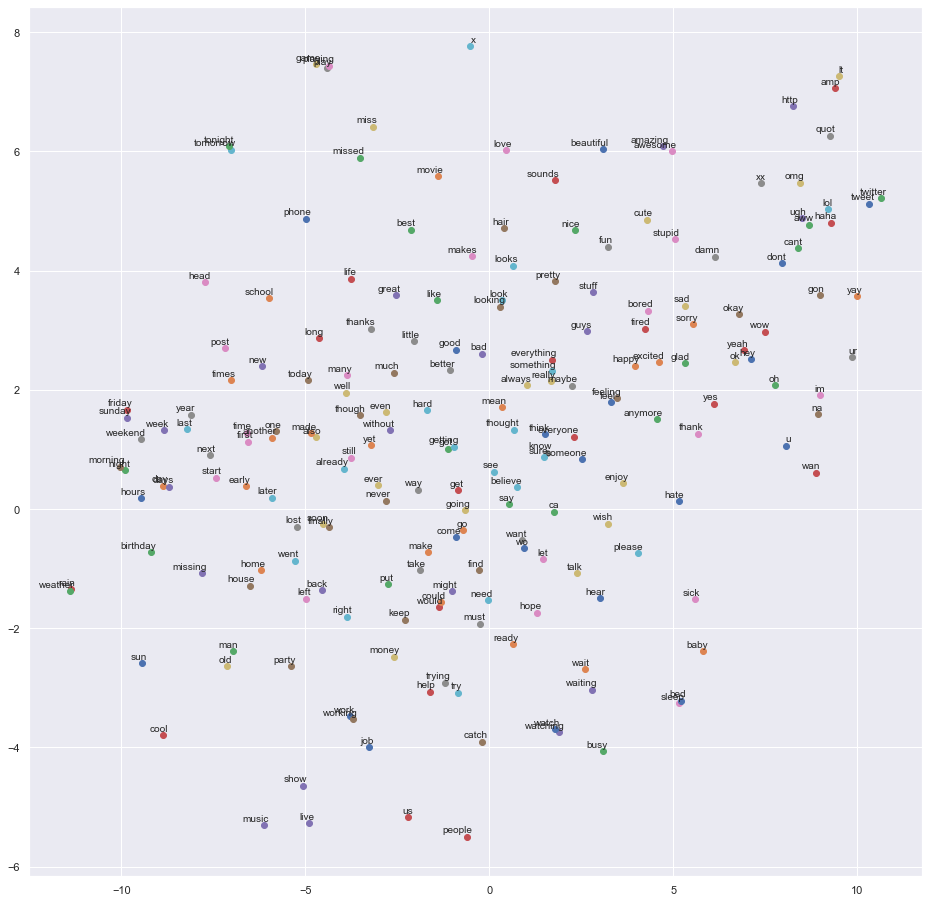
Quand la valeur de la variable target est égale à 2 cela correspond à un tweet de sentiment neutre.

Quand la valeur de la variable target est égale à 0 cela correspond à un tweet de sentiment négatif.





TSNE :

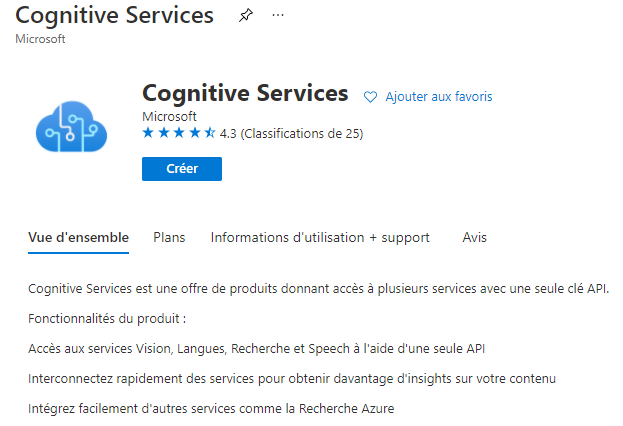


1. Comparaison entre trois approches
   1. Approche “API sur étagère”
   2. Approche “Modèle sur mesure simple”
   3. Approche “Modèle sur mesure avancé”
   4. Approche “API sur étagère”

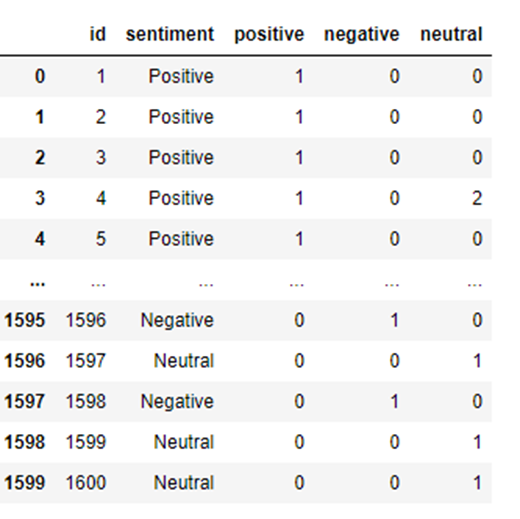
Microsoft Azure est la plate-forme permettant de gérer les applications du cloud de Microsoft. L’API Analyse de texte est un service cloud qui fournit des fonctionnalités de traitement en langage naturel pour l’exploration de texte et l’analyse de texte, notamment l’analyse des sentiments, l’exploration des opinions, l’extraction de phrases clés, la détection de la langue et la reconnaissance des entités nommées.

L’API analyse des sentiments fait partie d’[Azure Cognitive Services](https://docs.microsoft.com/fr-fr/azure/cognitive-services/).

Après avoir créé la ressource Cognitive Services on récupère l’url endpoint et la clé.

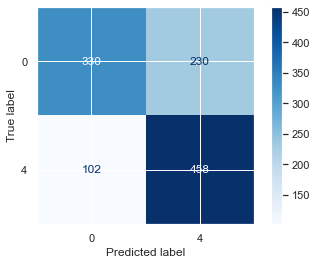


Dans Jupyter notebook nous réalisons la détermination des sentiments de chaque Tweets, l’API d’analyse des sentiments permet de prédire si le sentiment des tweets est soit positif, soit neutre ou soit négatif.

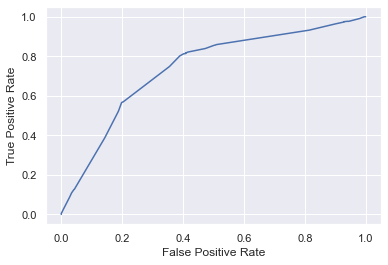


Après une régression logistique nous obtenons une précision de : 70.36%

La matrice de confusion :



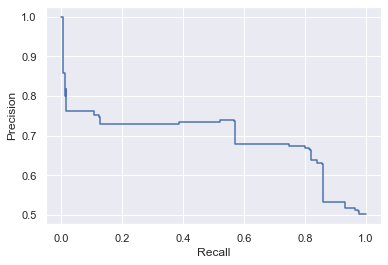
La courbe ROC :



AUC : 0.74

RMSE : 2.76

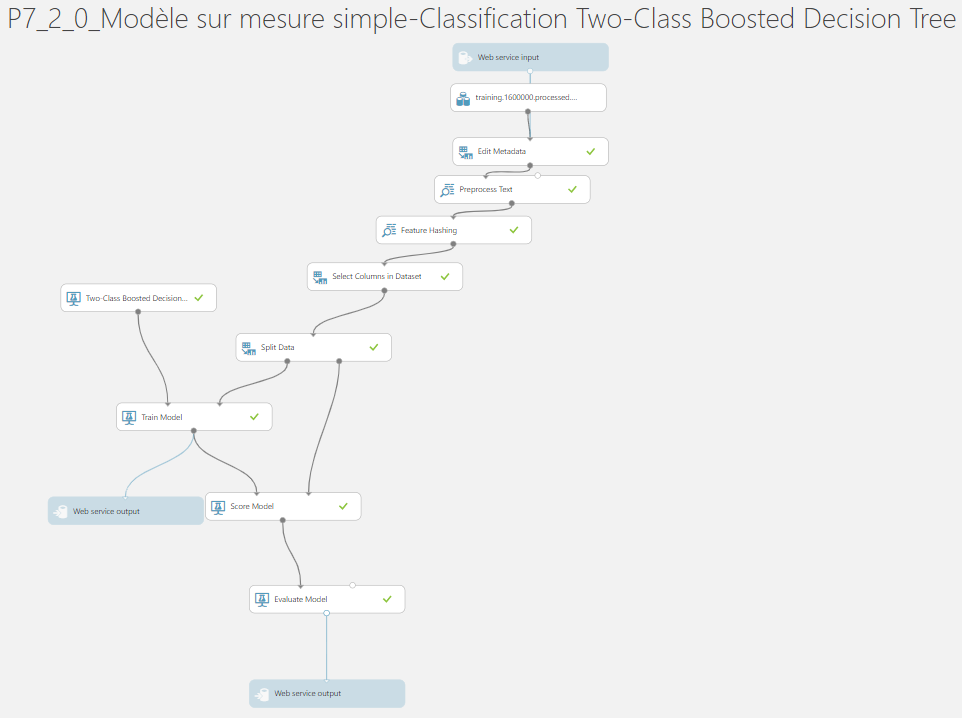
Rappel « Recall » :



* 1. Approche “Modèle sur mesure simple”

Pour cette approche nous utilisons le service Azure Machine Learning Studio classic, c’est une interface graphique drag & drop permettant de développer rapidement un modèle classique pour prédire le sentiment associé à un tweet.

Création du modèle :



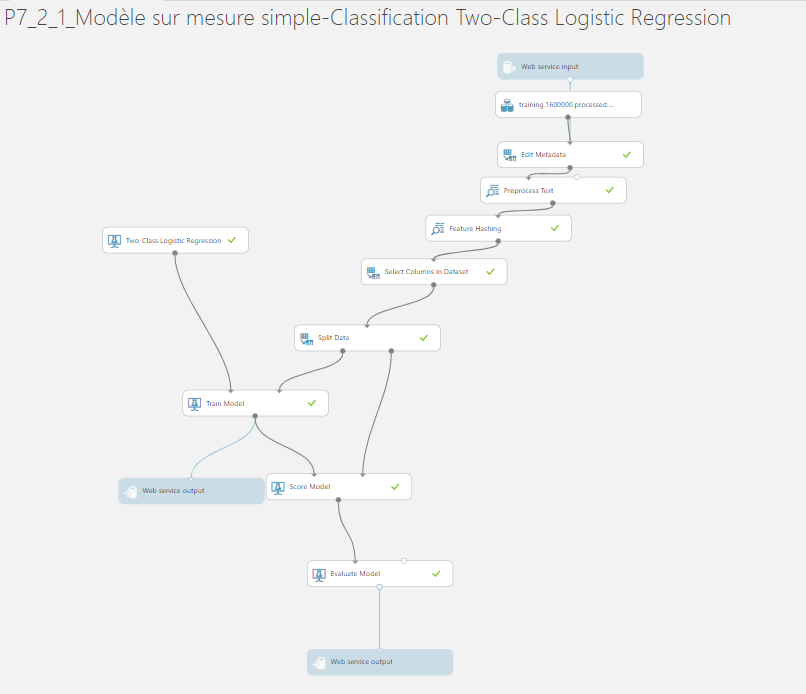
Après nettoyage du jeu de données, split du jeu de données, définition du type de modèle et entrainement du modèle.

Evaluation du modèle :



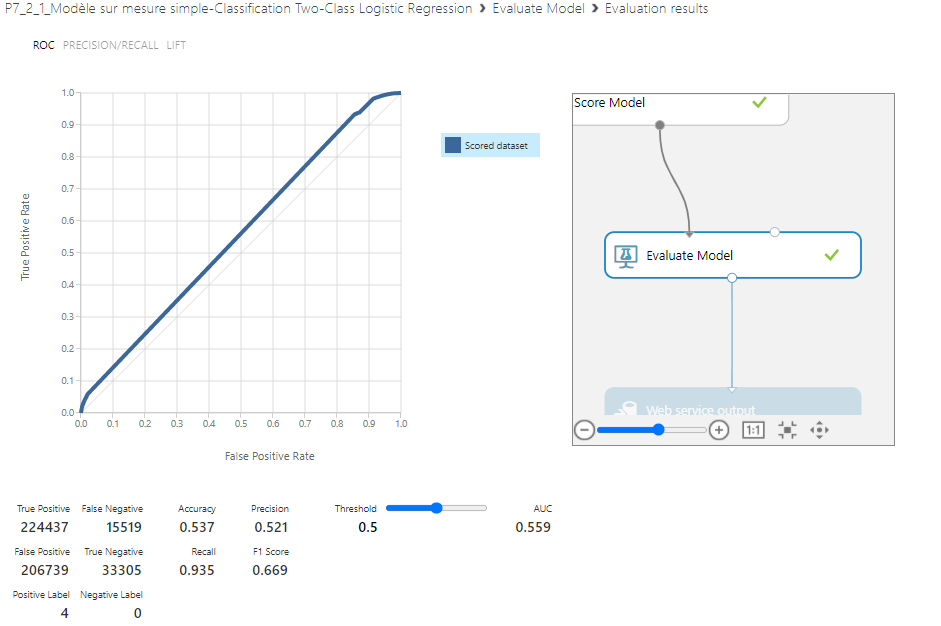


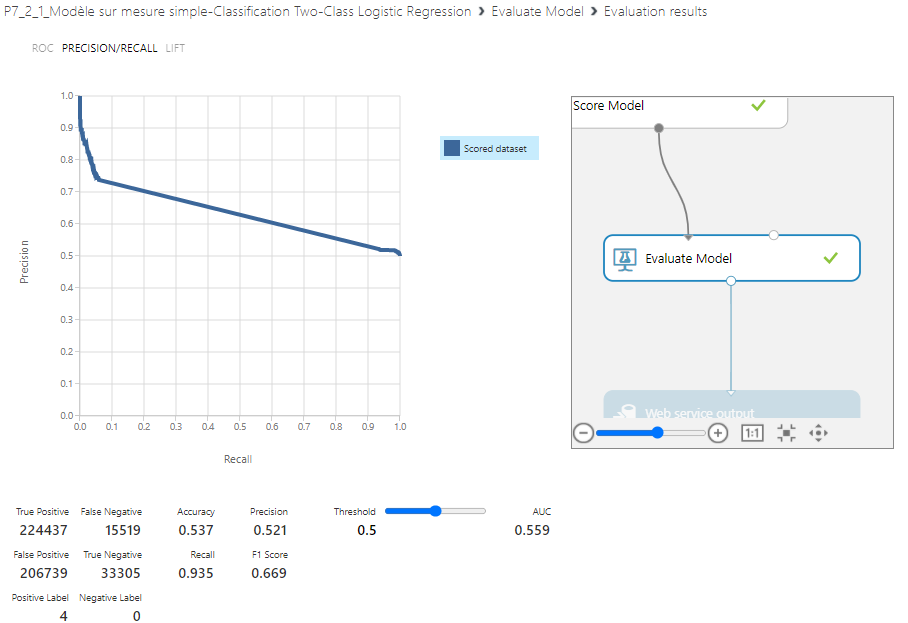
Création du modèle :



Après nettoyage du jeu de données, split du jeu de données, définition du type de modèle et entrainement du modèle.

Evaluation du modèle :





* 1. Approche “Modèle sur mesure avancé”

Nous allons utiliser Tensorflow keras et python pour réaliser l’approche “Modèle sur mesure avancé” :

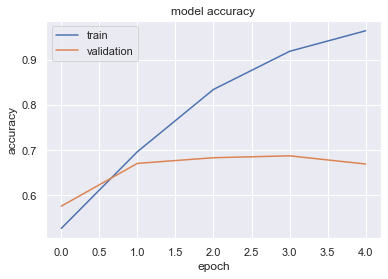
Etape de nettoyage des données :

Le nettoyage de données consiste à enlever les URLs, les tags HTML, la ponctuation et les stop words qui sont des mots qui n'apportent pas ou peu d'information ensuite nous choisirons entre le stemming et la lemmatisation.

Le processus de « lemmatisation » consiste à représenter les mots (sous leur forme canonique. Par exemple pour un verbe, ce sera son infinitif. Pour un nom, son masculin singulier. L'idée étant encore une fois de ne conserver que le sens des mots utilisés dans le corpus.

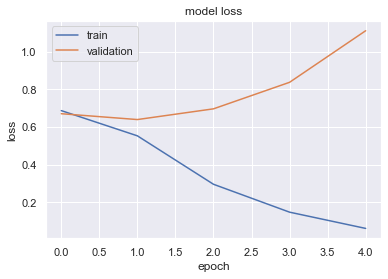
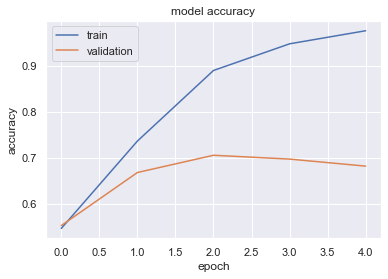
Il existe un autre processus qui exerce une fonction similaire qui s'appelle la racinisation (ou stemming en anglais). Cela consiste à ne conserver que la racine des mots étudiés. L'idée étant de supprimer les suffixes, préfixes et autres des mots afin de ne conserver que leur origine. C'est un procédé plus simple que la lemmatisation et plus rapide à effectuer puisqu'on tronque les mots essentiellement contrairement à la lemmatisation qui nécessite d'utiliser un dictionnaire.

Stemming :

Accuracy: 66.94%

Lemmatisation



Accuracy: 68.19%

On obtient de meilleurs résultats avec la lemmatisation car on obtient au model accuracy une correspondance avec la validation ainsi que l’accuracy soit meilleur avec lemmatisation.

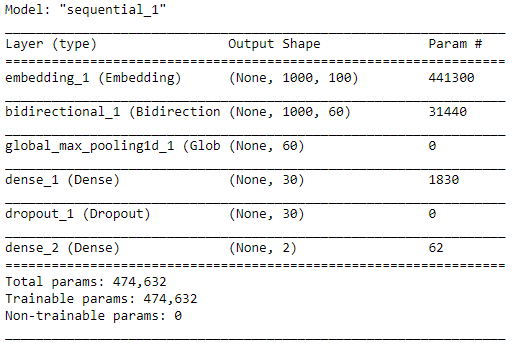
Deux méthodes plongement de mots, word embedding : Dans ces méthodes chaque mot est représenté par un vecteur de nombres réels.

Nous comparons la méthode Word2Vec et la méthode GloVe

Avec GloVe on obtient de meilleur accuracy.

Après le nettoyage du jeu de données, split du jeu de données, nous définissons le type de modèle et entrainement du modèle.

Nous réalisons un modèle LSTM bidirectionnel :



Ce modèle utilisant la lemmatisation avec LSTM Bidirectional avec la méthode GloVe on obtient une Accuracy: 68.19%

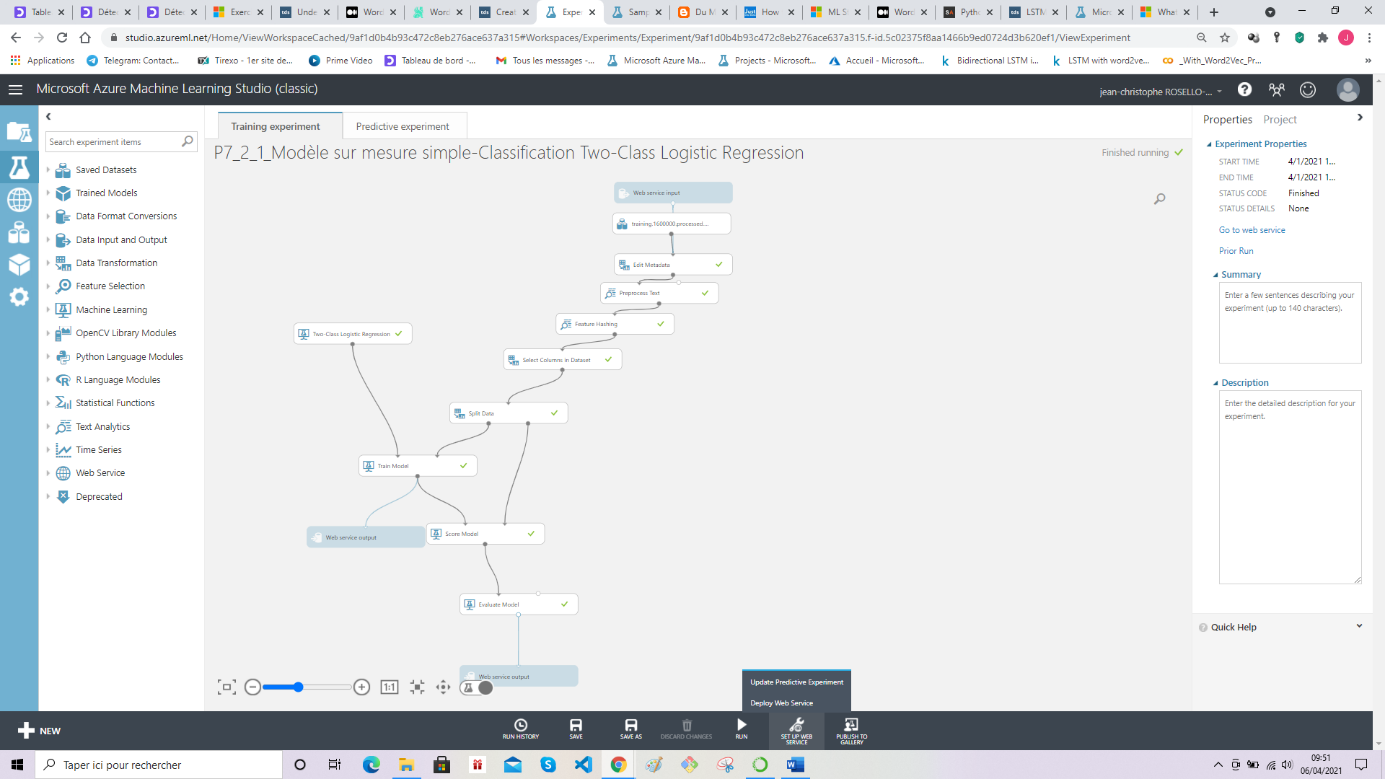
1. Déploiement des modèles de chaque approche :
   1. Approche “API sur étagère”

Le modèle est déjà déployé avec cognitive services ressource de Microsoft Azure.

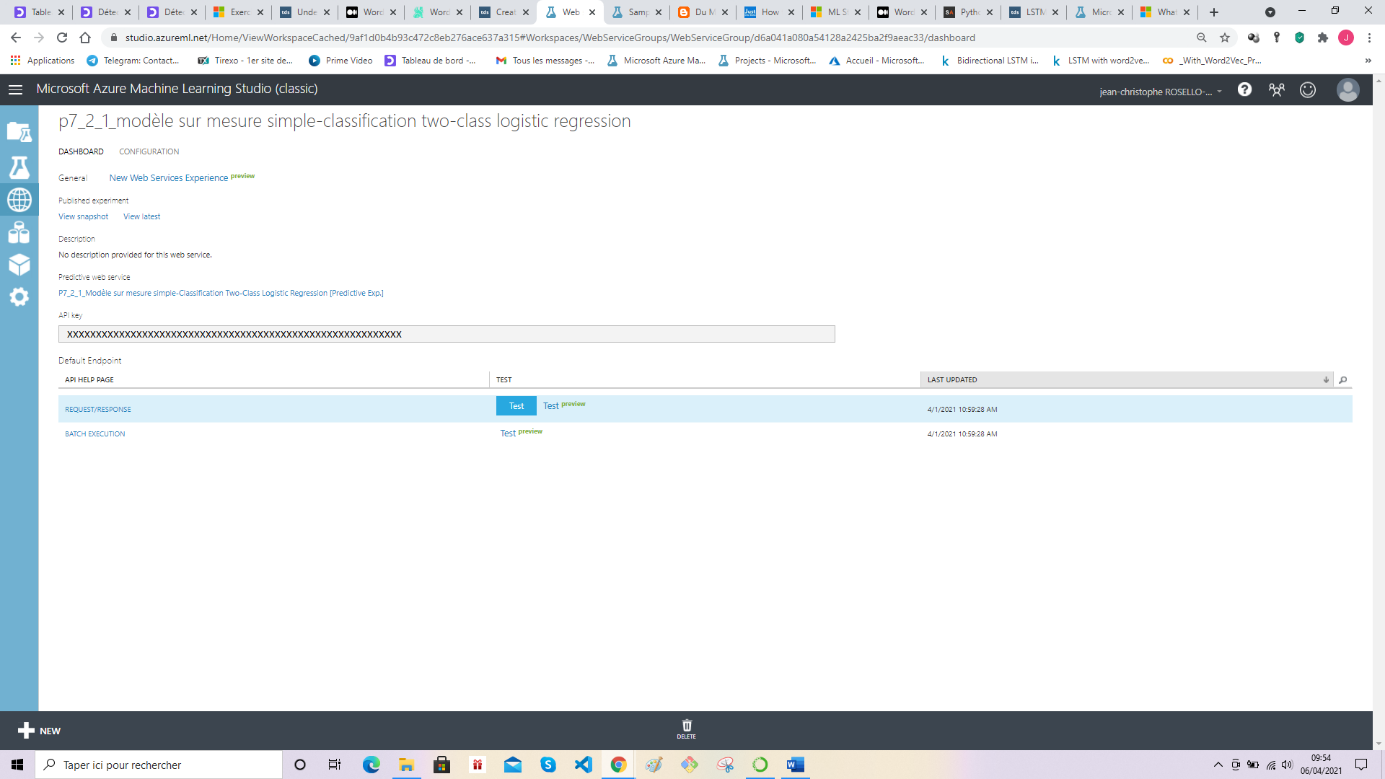
* 1. Approche “Modèle sur mesure simple”

Après l’entrainement du modèle dans l’onglet set up web service on clique sur deploy web service pour déployer l'API. Ce qui génère un endpoint et une clé.

Déploiement de l’expérience :



Api déployée dans « web service » :



* 1. Approche “Modèle sur mesure avancé”

Pour cette approche nous utilisons le service Azure Machine Learning pour développer un modèle basé sur des réseaux de neurones profonds pour prédire le sentiment associé à un tweet.

Nous téléchargeons le fichier config.json de notre workspace que nous avons créé sous Azure.



Trois étapes à réaliser pour le déploiement :

1.Enregistrement des modèles :

2. Création du script d'entrée score.py, créant l'environnement et définissant une configuration d’inférence.

Le script charge les modèles dans init().Le run() utilise les modèles pour les prédictions.

L'environnement contient toutes les bibliothèques utiles aux modèles.

La configuration d'inférence conjugue les script et l'environnement.



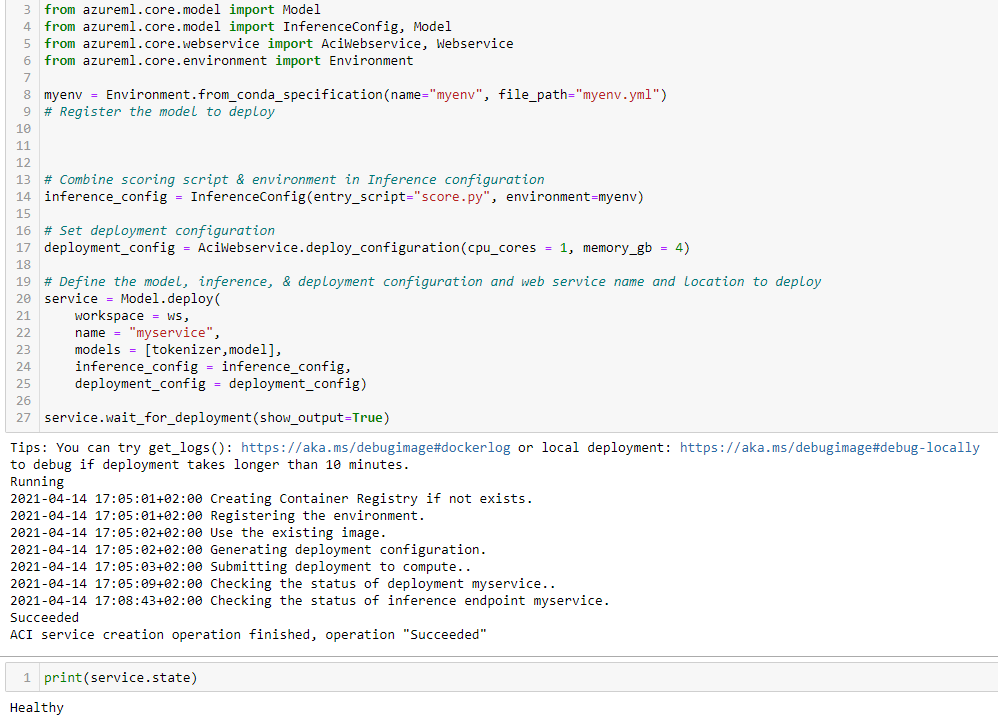
3.Déploiement du modèle :

Nous déployons l'API avec Aci.

Configuration de déploiement :



Déploiement du service :



1. Avantages et inconvénients des approches et performances

3.1 Approche “API sur étagère”

C'est l’approche la plus rapide et plus simple à réaliser, mais n'est pas modifiable, cette méthode a un coût non négligeable.

3.2 Approche “Modèle sur mesure simple”

C’est une approche simple à réaliser mais il faut connaître la méthode et les blocs pour la construction des expérimentations.

L’avantage est que le déploiement du modèle est possible dans Excel.

3.3 Approche “Modèle sur mesure avancé”

C’est l’approche qui est totalement modifiable et modulable. C'est cette approche qui est choisi pour une maintenabilité dans la durée et les modifications de paramètres.

3.4 Tableau récapitulatif des performances

