Rapport Projet Deep Learning

Auteur: Komlan Jean-Marie Dantodji

Thème:

Analyse de Sentiment des Clients des vols américains à partir des tweets.

Context:

Dans le cadre de l'amélioration du service client que offre des compagnies aériennes américaines, une enquête sur tweeter a été menée afin de connaître les avis des clients qui ont pris des vols américains. L'objectif est de les classer en trois catégories sur la base des tweets qu'ont fait ces clients, le sentiment qu'ils ont eu.

Les données:

La dataset contient des informations sur le tweet du client notamment
Tweet_location: la région de provenance,
Airline: La compagnie prise:,
Negative_reason_gold: La raison négative,
Text: Le tweet effectué,
Airline_sentiment_confidence: la probabilité de confiance que porte le client sur la compagnie.
Negative_reason_confidence: Probabilité de confiance négative sur le vol effectué.

Outils techniques:

Pour réaliser ce projet d'analyse de sentiment des clients des compagnies aériennes, on est amené à utiliser le langage Python. Plusieurs libraisiers ont intervenus telles que:

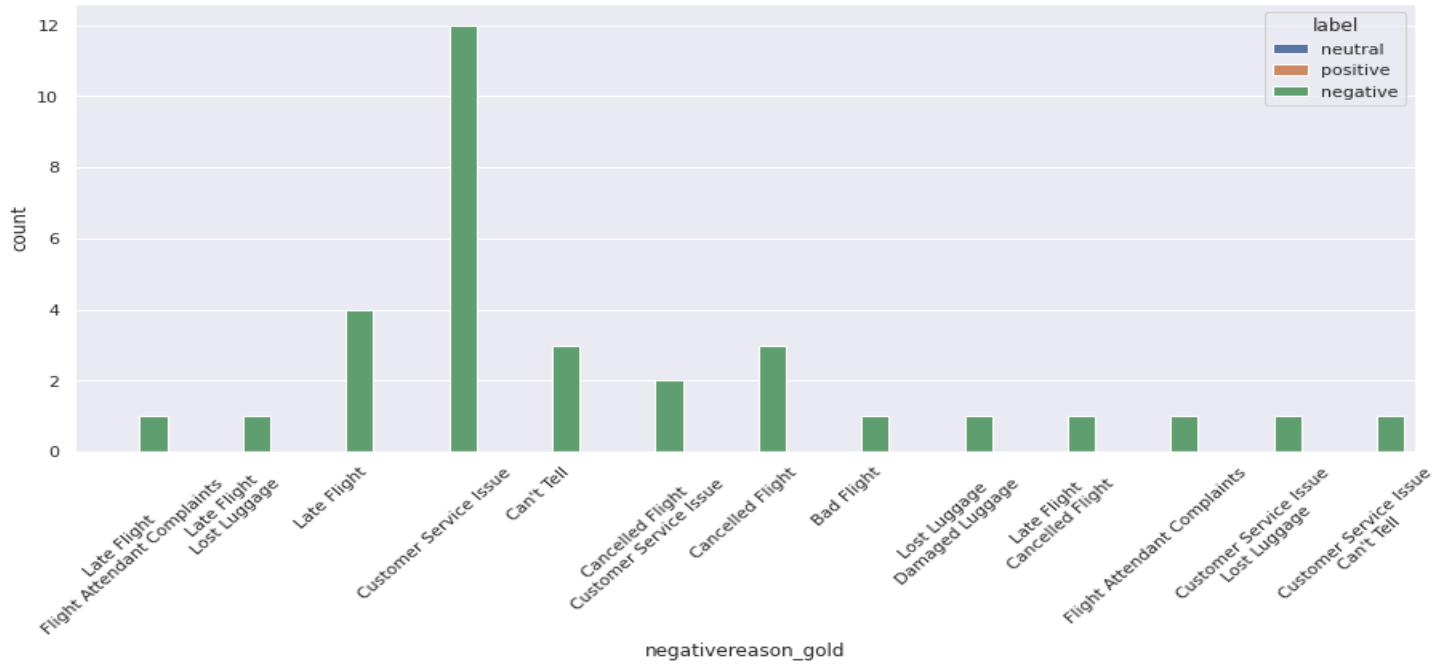
- Seaborn - Matplotlib pour l'analyse des donnée afin de comprendre la donnée,
- NLTK: Pour nettoyer les données textuelles et enlever les caractères indésirables,
- Keras Tokenizer pour encoder le text sous forme numérique afin d'appliquer les modèles de deep learning,
- Keras - Tensor Flow dans la construction des modèles.

Problématique:

Dans ce projet il s'agira de classifier les sentiments des clients en trois groupes: Sentiment positif, Sentiment neutre et Sentiment négatif suite à leur voyages aux abord des vols des compagnies américaines. On implémentera des algorithmes de Deep learning en considérant les données numériques (ex: probabilités de satisfaction, …) et des données textuelles (les tweets des clients).

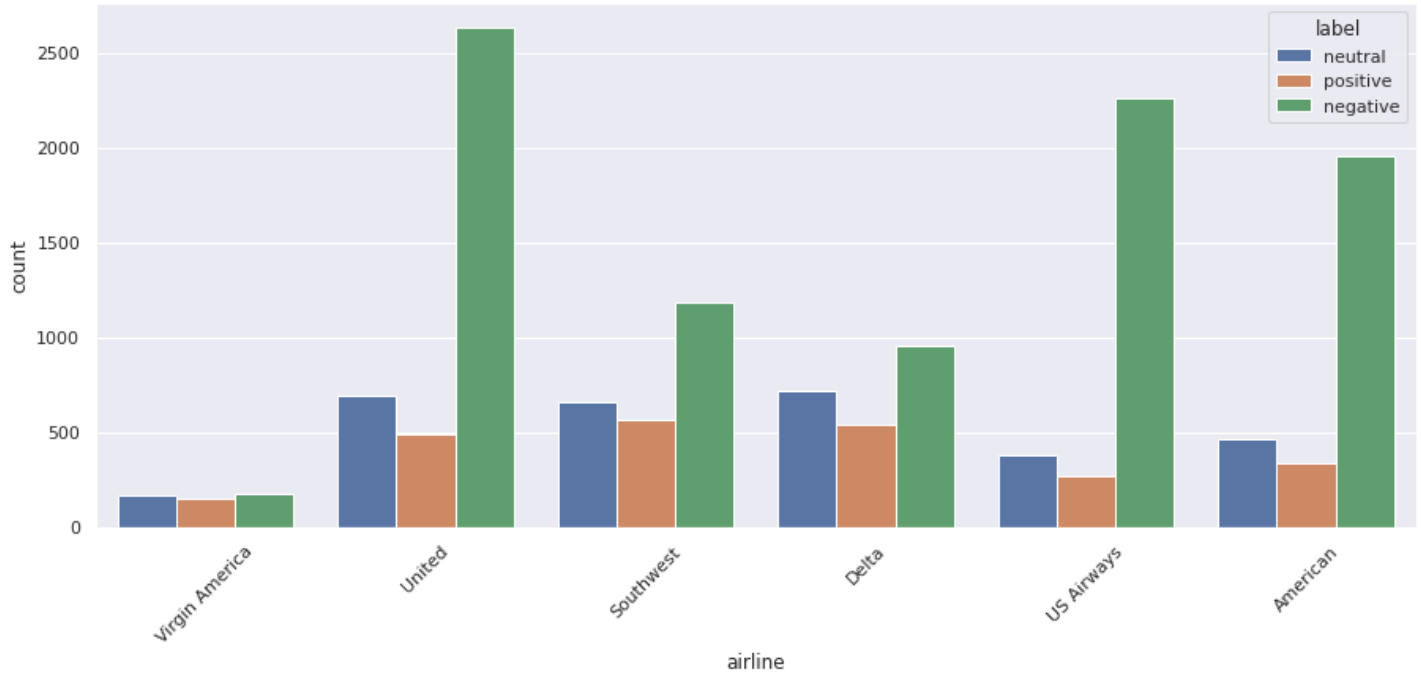
Analyse des données:

1. Part des clients en fonction de leur raison de sentiments:



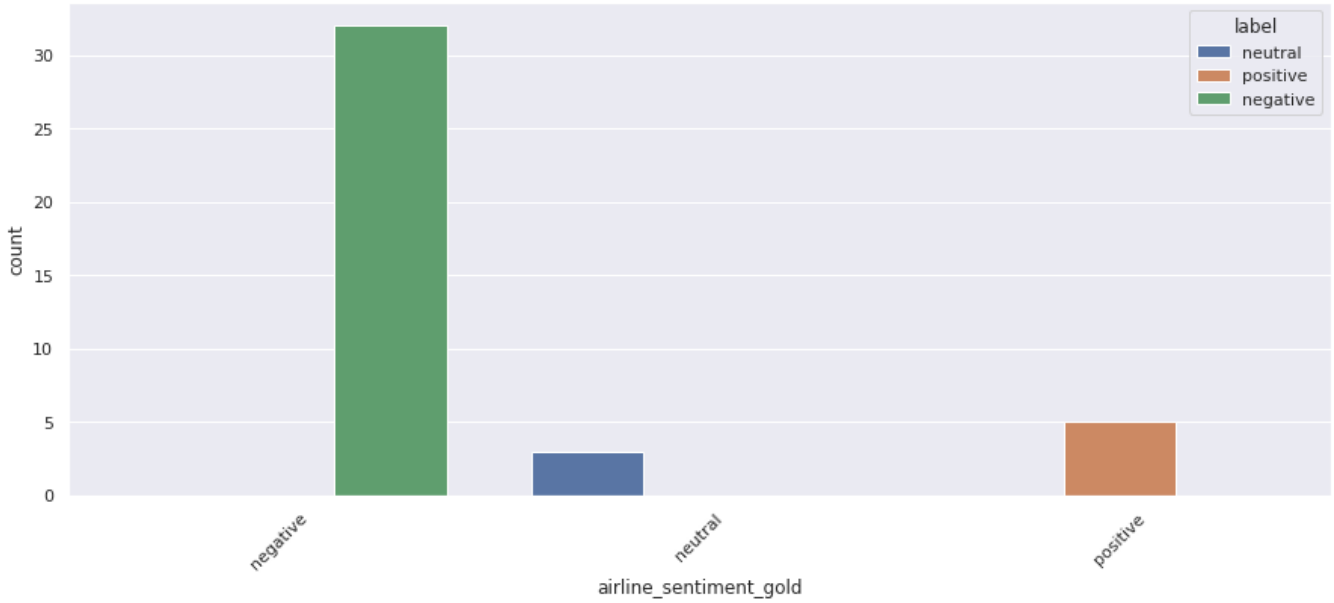
On remarque que dans la majeur des cas, le problème de service client explique le sentiment négatif des clients.

2. Part des clients en fonction des compagnies:



On remarque dans la globalité que toutes les compagnies ont des problèmes avec leur voyage. La compagnie United enregistre plus d'avis négatifs sur ses vols que les autres.

3. Part de categories de sentiments:



Etat de l'art des modeles:

1. Neural Network (Dense)

Dans ce model, on a empilé des couches denses qui sont connectés à tous les neurones de la couche précédente. Chaque couche effectue une multiplication matricielle avec le vecteur d'entrée.

2. LSTM : Long Short-Term Memory (RNN)

Ce type de modèle est une récurrence des réseaux de neurones classiques. Il permet de persister les informations anciennes dans le processus d'apprentissage. Ce que ne font pas les autres modèles classiques.

3. Convolutional Neural Network (CNN)

Les réseaux de neurones à convolutions sont basés sur l'application de matrice de convolution permettant de réduire la taille de la donnée à traiter. Il est généralement applicable sur des images 2D, mais il est possible de l'appliquer sur un vecteur 1D comme dans ce cas pratique.

4. Gated Recurrent Units (GRU)

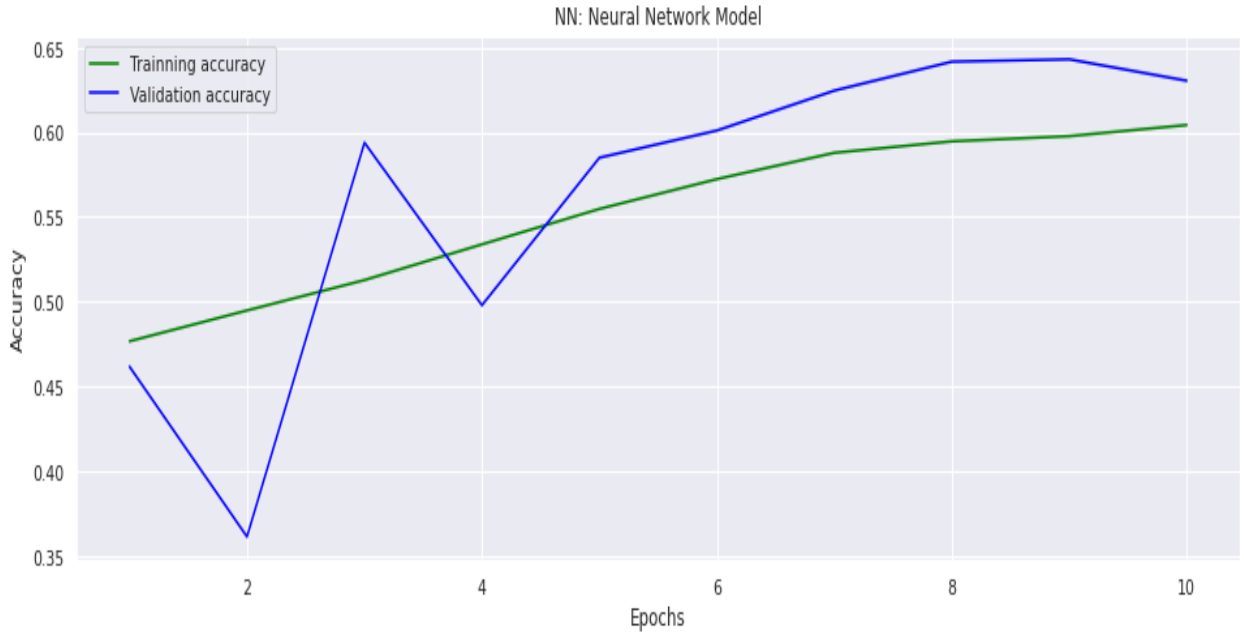
GRU est un modèle de réseau de neurone récurrent qui a la particularité de contrôler quand la couche cachée doit être mise à jour ou réinitialisée.

Application des modèles:

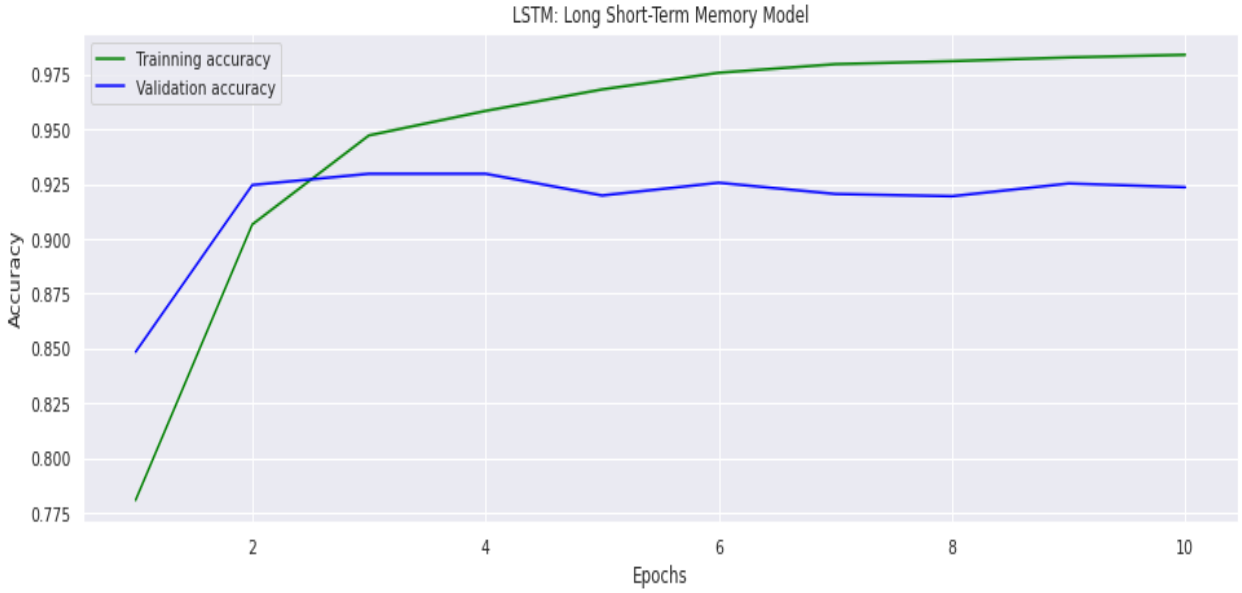
- Variables des modèles:
  Nombre de neurones dans la couche cachée: 100
  Nombre de neurone dans la couche de sortie: 3
  Taille de Batch: 32
  Nombre d'époques: 10

1. Neural Network
   Dans ce modèle, j'ai empilé quatre couches Dense avec des activations RELU. Ensuite à la sortie, une couche de Dense avec une activation de probabilité

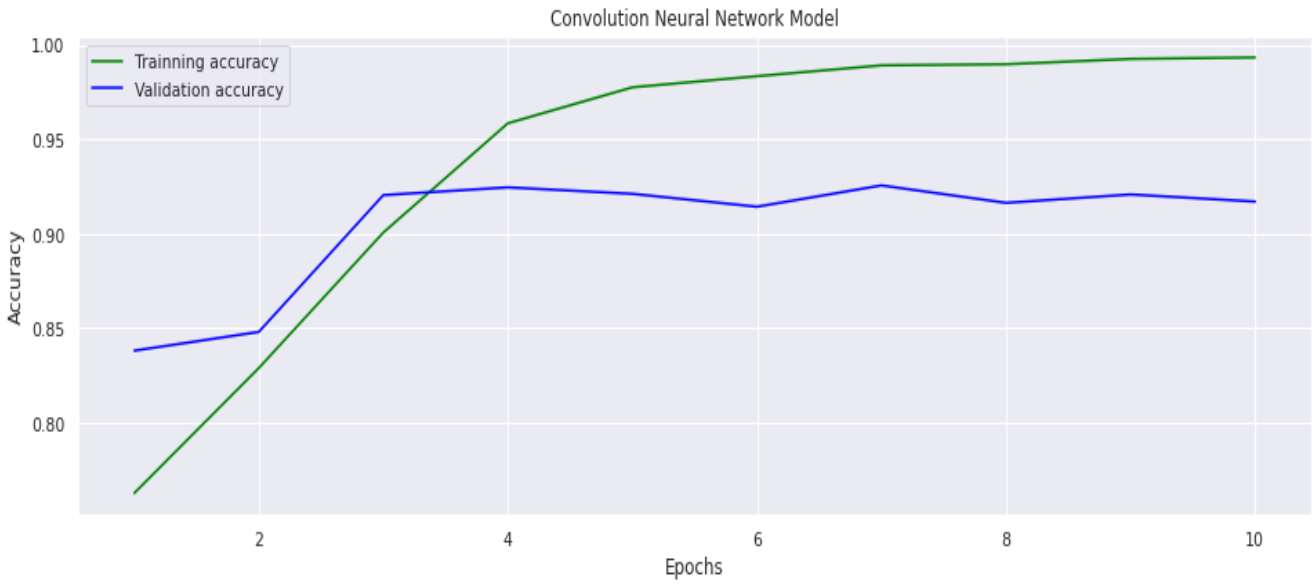
SOFTMAX.



2. Long Short-Term Memory:



3. Convolution Neural Network (CNN)



4. Gated Recurrent Units (GRU)

Analyse des résultats:

Tableau récapitulatif des performances des modèles:

Model	Time	Nombre d'époque	Accuracy Training	Accuracy Validation	Lost Training	Lost Validation
NN	11 s	10	0.6043	0.6305	0.9963	0.9680
LSTM	350 s	10	0.9839	0.9235	0.0468	0.2778
CNN	100 s	10	0.9933	0.9170	0.0212	0.3339
GRU	202 s	10	0.9892	0.9242	0.0314	0.2518

Conclusion:

Après application de ces quatres modèles, on a remarqué que le modèle du CNN a le meilleur accuracy à 0.99 à l'apprentissage avec 0.021 comme perte. Il a de plus un temps d'exécution meilleur que LSTM et GRU.