Dans cet article, nous allons comparer les trois approches que nous avons essayé afin de prédire le sentiment d'un tweet : l'API sur étagère, le modèle sur mesure simple et le modèle sur mesure avancé.

Pour l'API sur étagère, nous avons utilisé l'API de Microsoft Azure pour l'analyse de sentiments. La fonctionnalité d'analyse des sentiments fournit des étiquettes de sentiment (par exemple « négatif », « neutre » et « positif ») en fonction du score de confiance le plus élevé trouvé par le service au niveau de la phrase et du document. Cette fonctionnalité retourne également des scores de confiance compris entre 0 et 1 pour chaque document, ainsi que les phrases qu'il contient pour les sentiments positif, neutre et négatif. Pour des raisons de coût, nous n'avons pas testé ce modèle sur toute la base que nous avions à notre disposition mais uniquement sur un ensemble de tests de quelques milliers d'exemples, représentatifs de l'ensemble des données.

Ensuite concernant le modèle sur mesure simple, nous avons pris en main le Concepteur/Designer de Microsoft nous offrant une interface par glisser-déposer utilisée pour entraîner et déployer des modèles dans Azure Machine Learning. Pour ce qui est du modèle utilisé ici, il s'agit d'une simple régression logistique.

Enfin, pour le modèle sur mesure avancé, nous avons utilisé la librairie Keras. Ici, nous avons essayé 2 modèles différents. Un premier sans couche LSTM (Long Short-Term Memory) et un second avec. La couche LSTM est une couche qui possède une mémoire interne appelée cellule. La cellule permet de maintenir un état aussi longtemps que nécessaire. Cette cellule consiste en une valeur numérique que le réseau peut piloter en fonction des situations. Pour chacun de ces deux modèles, nous avons tenté 4 approches différentes :

- Deux premières avec des données ayant subies un pré-traitement différent :
    - La « lemmatisation » est un processus qui 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 de ne conserver que le sens des mots utilisés dans le corpus.
    - La racinisation 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.

- Et deux autres avec deux word embeddings différents :
    - Word2vec, qui est un groupe de modèles associés qui sont utilisés pour produire des incorporations de mots. L'idée derrière Word2Vec est analogue à dire : "montre-moi tes amis, et je te dirai qui tu es" car l'hypothèse principale est que la signification d'un mot peut être déduite par ses voisins. Si nous avons 2 mots qui ont des voisins très similaires, alors ces mots ont probablement un sens assez similaire ou sont au moins liés.
    - GloVe (Global Vectors for Word Representation), qui vise à capturer la relation entre les mots de l'ensemble du corpus. Il entraîne un modèle basé sur le nombre global de cooccurrences de mots, des statistiques globales et utilise l'erreur quadratique moyenne comme fonction de perte. L'incorporation de mots générés avec un tel modèle préserve les relations et les similitudes entre les mots.

Pour évaluer la qualité de nos modèles, nous avons utilisé le F1-score, la précision et le recall comme métriques d'évaluation. Le F1-Score est une moyenne harmonique entre le recall (le nombre de positifs bien prédit (Vrai Positif) divisé par l'ensemble des positifs (Vrai Positif + Faux Négatif)) et la précision (le nombre de positifs bien prédit (Vrai Positif) divisé par l'ensemble des positifs prédit (Vrai Positif + Faux Positif)).

Le recall permet de savoir le pourcentage de positifs bien prédit par notre modèle. Plus il est élevé, plus le modèle de Machine Learning maximise le nombre de Vrai Positif. Mais cela ne veut pas dire que le modèle ne se trompe pas. Quand le recall est haut, cela veut plutôt dire qu'il ne ratera aucun positif. Néanmoins cela ne donne aucune information sur sa qualité de prédiction sur les négatifs.

La précision est assez similaire au recall, elle permet de connaître le nombre de prédictions positives bien effectuées. Plus elle est élevée, plus le modèle de Machine Learning minimise le nombre de Faux Positif. Quand la précision est haute, cela veut dire que la majorité des prédictions positives du modèle sont des positifs bien prédit.

Bien qu'ils soient utiles, ni la précision ni le recall ne permettent d'évaluer entièrement un modèle de Machine Learning. Séparément c'est deux métrique sont inutiles :

- si le modèle prédit tout le temps « positif », le recall sera élevé
- Au contraire, si le modèle ne prédit jamais « positif », la précision sera élevée.

Outre l'API sur étagère ayant obtenu un un recall de 79,5%, une précision de 69,5% et un F1 score de 74%, le modèle avec les meilleurs résultats est le modèle sur mesure avancé utilisant un word embedding Glove et une couche LSTM. Ce modèle a obtenu 71,1% de recall, 65,6% de précision et 68,2% de F1-Score.