MSE | MASTER OF SCIENCE IN ENGINEERING

WEB Mining

Rapport de projet

Yverdon, le 15.06.2023

Sam Corpataux
Gaël Koch
Maël Vial

Table des matières

1. Contexte et objectifs

1.1. Contexte général

Ce projet rentre dans le cadre du cours Master HES-SO MA-WEM. Il a pour but de mettre en pratique les différentes notions de "web scrapping" que nous avons vu en classe pour récupérer des données sur le web. Une fois les données récupérées et nettoyées, elles pourront être utilisées pour entrainer des algorithmes de Machine Learning de sorte à mettre en lumière des statistiques liées à ces données. Un travail de visualisation complètera ce projet. Il présentera les diverses statistiques extraites de nos modèles.

Autrement dit, l'objectif général du projet est donc de récupérer des données depuis internet afin d'en extraire une valeur "business" et de les présenter de façon lisible et compréhensible.

1.2. Choix du sujet

Nous avons tous été confronté à des désagréments (retards, pertes de bagages etc …) ou au contraire, à des bonnes surprises (sur-classement, prix de billets réduits etc…), lors de nos voyages en avion. Nous avons donc trouvé légitime de se poser la question : il y a-t-il une compagnie aérienne qui offre de meilleurs services que les autres ?

Pour répondre à cette interrogation, nous avons donc décider faire de l'analyse de sentiments sur des commentaires de réseaux sociaux, liés à des compagnies aériennes. Plus de détails quant au choix des compagnies aériennes sont expliquées dans le chapitre 4 de ce document.

1.3. Objectifs du projet

Le premier objectif du projet est de récupérer des informations sur les compagnies aériennes depuis internet. Cela se fera à partir d'un réseau social afin de pouvoir réellement obtenir l'avis et les réactions des personnes envers une compagnie.

Le second, sera de faire de l'analyse de sentiment sur ces données récoltées. En effet, il est nécessaire de connaître le sentiment des personnes afin de pouvoir évaluer leurs appréciations envers les compagnies aériennes. Il faudra donc utiliser un modèle de machine learning qui pourra classifier nos données comme étant : positive, négative ou neutre.

Le dernier objectif est de visualiser ces données d'une manière simple, efficace et qui apporte de la valeur. Il sera donc intéressant de mettre en place plusieurs techniques de visualisations afin d'avoir une vision détaillé des données récoltées. De cette façon, nous pourrons déterminer lesquelles ont les meilleurs/les pires services clients. Enfin, il serait intéressant de prendre en compte le facteur "temps" dans ces visualisations, afin de voir l'évolution de l'appréciation de chaque compagnie aérienne.

2. Données

Le choix des données a été plus compliqué à obtenir que prévu et ce pour plusieurs raisons.

2.1. Echec du scrapping sur Twitter

Lorsque nous avons débuté ce projet, nous voulions nous focaliser uniquement sur des tweets. En effet, Twitter est une plateforme largement utilisée mondialement et sur laquelle les utilisateurs n'hésitent pas à partager leur opinion sur tout et n'importe quoi, y compris les compagnies aériennes. Nous avions trouvé un super outil nommé "snscrape" qui nous permettait d'outrepasser la limite imposée par l'API de Twitter, afin de récolter des tweets en très grande quantité. Malheureusement, ce service a été complètement bloqué, littéralement un jour avant que nous commencions à récolter des données (pour plus de détail, veuillez-vous référer à la section "problèmes rencontrés" de notre cahier des charges).

Pour contrecarrer ce problème, nous avons trouvé deux solutions :

- Utiliser un dataset Kaggle pré-annoté
- Faire du scrapping sur Reddit et non sur Twitter

2.2. Dataset de Kaggle

Nous avons trouvé, sur Kaggle, une dataset qui correspond exactement à nos attentes. Il contient des tweets d'utilisateurs concernant 6 compagnies aériennes : American airlines, Delta, Southwest, US Airways, United et Virgin America. Sur ces 6 compagnies, nous n'allons garder uniquement les 4 suivantes : Delta, American airlines, Southwest, United Airlines.

La raison de ce choix est simple, nous avons un deuxième dataset (présenté dans le chapitre suivant) qui contient spécifiquement ces 4 mêmes compagnies. Nous souhaitons garder une cohérence entre nos datasets étant donné que l'un servira à réentraîner notre modèle de machine learning, et l'autre à le tester.

Après l'élimination des deux compagnies non-choisies, il nous reste 11'223 tweets labélisés et utilisables pour l'réentraînement de notre modèle. Chaque tweet est accompagné des informations suivantes : le sentiment, la confiance dans la prédiction du sentiment, l'utilisateur, le texte, la date de création, le nombre de retweet. Il existe encore quelques autres colonnes mais elles sont moins importantes pour ce projet, nous allons donc les laisser de côté.

Voici un exemple de tweet :

```
tweet_id  :  570301031407624196
airline_sentiment  :  negative
airline_sentiment_confidence  :  1.0
negativereason  :  Bad Flight
negativereason_confidence  :  0.7033
airline  :  Virgin America
airline_sentiment_gold  :  nan
name  :  jnardino
negativereason_gold  :  nan
retweet_count  :  0
text  :  @VirginAmerica it's really aggressive to blast obnoxious "entertainment" in your guests' faces &amp;amp; they have little recourse
tweet_coord  :  nan
tweet_created  :  2015-02-24 11:15:36 -0800
tweet_location  :  nan
user_timezone  :  Pacific Time (US & Canada)
```

Figure 1: Exemple de tweet

Comme expliqué ci-dessus, nous allons utiliser ces données pour réentraîner un modèle de machine afin de pouvoir l'utiliser pour reconnaître les sentiments de nos propres données récupérées, c'est-à-dire le second dataset, basé sur Reddit.

2.3. Scrapping sur Reddit

Notre second dataset contient des données que nous avons récupérées nous-même, directement depuis le réseau social Reddit.

Notre technique est la suivante, nous avons recherché tous les "subreddit importants" liés aux compagnies aériennes présentes dans notre premier dataset (celui de Kaggle). Nous avons été en mesure de récupérer des postes concernant les 4 compagnies aériennes suivantes : Delta, American airlines, Southwest ainsi que United Airlines.

Par "subreddit importants" nous entendons tous ceux qui comportent, au minimum, plusieurs milliers de messages. Cette limite est nécessaire car nous avons besoin de quantité de données considérables pour évaluer correctement un modèles ML.

Une étape de pré-traitement est essentielle car ces postes contiennent un grand nombre d'informations qui nous sont inutiles. Par exemple : le type d'avatar de l'utilisateur, ou si le poste est dans une catégorie spéciale du subreddit etc... En tout, un post est composé de 122 colonnes, mais seul : l'auteur, la date, le subreddit, le titre et le texte du poste nous intéressent réellement. En voici un exemple :

```
subreddit   :  americanairlines
selftext    :  Do you find it annoying when people block the boarding area during boarding.  They announce the
title   :  Blocking the boarding area
author   :  vw2xb
utc_datetime_str   :  2023-04-21 23:48:01
```

Figure 2: Exemple d'un post Reddit après le pré-traitement

Certain de ces posts étaient écrit en langues étrangères (ex : mandarin, japonais etc…). Ils ne seront donc pas pris en compte, étant donné que notre modèle a été entrainé sur des textes écrits en anglais.

Après la phase de pré-traitement, ce second dataset est se compose d'environ 9'000 posts Reddit par compagnie, soit un total d'environ 35'000 posts.

Sur ces 35'000 postes, environ 4'000 appartiennent à des comptes qui sont « deleted ». Il s'agit de posts ayant été écrit avec un compte qui a maintenant été supprimer. Nous perdons donc certaines informations comme l'auteur du post ainsi que son contenu du post. Cependant, nous avons toujours accès au titre du post. Or, bien souvent, le titre d'un post suffit pour savoir si le reste du contenu va être plutôt positif, neutre ou négatif. Il est donc tout à fait envisageable d'utiliser ces données pour faire de l'analyse de sentiment.

2.4. Licences d'utilisation des données

Le premier dataset Kaggle est sous la licence CC BY-NC-SA 4.0. Ce qui nous permet d'utiliser librement les données, tant que ce n'est pas à but lucratif et que le dataset original est cité.

Selon les conditions d'utilisation de Reddit, il est permis de « copier et d'afficher les posts des utilisateurs sans en modifier le contenu, excepté pour le formater selon nos affichages », ce qui est conforme avec notre utilisation des données.

2.5. Première évaluation de tendance

Lors de l'analyse des données, nous avons réalisé un « word cloud » afin de voir ce qui en ressortait et d'avoir quelques premiers « insights » sur nos données. Contrairement à ce que nous pensions, les mots qui ont le plus d'importance semblent être plutôt positifs. Nous avons tout de même des mots négatifs auxquels nous pouvions nous attendre comme « removed », « delay », « cancel » ou encore « issue ».

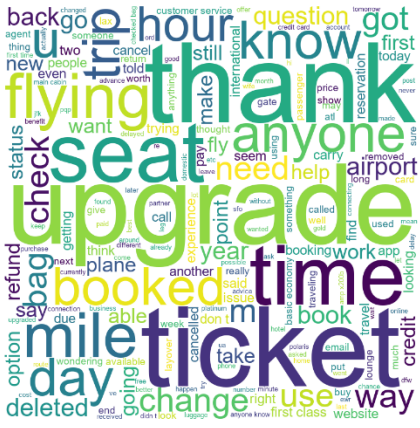

Figure 3: Exemple d'un word cloud

3. Etat de l'art

Pour faire de l'analyse de sentiment sur des tweets, nous nous sommes basés sur le travail de Daniel Loureiro et de ses collègues. Ils ont notamment écrit un article très intéressant, nommé « *TimeLMs: Diachronic Language Models from Twitter* » qui présente une approche innovante pour améliorer les performances des modèles de langage basés sur les données de Twitter. Les auteurs ont développé des modèles de langage diachroniques appelés TimeLMs, qui sont spécialement conçus pour traiter les données de Twitter au fil du temps.

Expérience menée :

L'expérience menée dans cette étude consiste à entraîner des modèles de langage sur des données de Twitter collectées sur une période de plusieurs années. Les auteurs ont utilisé une stratégie d'apprentissage continu pour adapter les modèles aux changements dans le langage et les tendances sur Twitter. Ils ont également examiné les effets temporels sur les performances des modèles dans diverses tâches de traitement du langage naturel en aval.

Résultats :

Les résultats de l'étude montrent que l'utilisation d'une stratégie d'apprentissage continu améliore les performances des modèles de langage basés sur Twitter. Les auteurs ont également constaté que les modèles pré-entraînés sur des données de domaine général ne sont pas toujours la meilleure option pour les tâches spécifiques à Twitter. Dans certains cas, il est plus efficace de partir de zéro ou d'utiliser un modèle plus petit et spécifique au domaine.

Conclusions :

Les conclusions de cette étude remettent en question l'hypothèse dominante selon laquelle la préformation spécifique au domaine peut bénéficier de l'utilisation de modèles de langage de domaine général. Les auteurs suggèrent que, pour certaines tâches et domaines, il est préférable d'utiliser des modèles de langage spécifiques au domaine ou de commencer avec un modèle

non-pré-entraîné. Cette recherche souligne également l'importance de prendre en compte les effets temporels lors de l'évaluation des performances des modèles de langage, en particulier pour les données de Twitter, qui évoluent rapidement.

Source : https://arxiv.org/pdf/2202.03829.pdf

4. Conception

Notre projet a pour but de répertorier les avis des clients de multiples compagnies aériennes. Dans ce chapitre on va voir quelques cas d'utilisation dans lesquels ils seraient intéressant d'utiliser notre site web et nous allons également présenter la structure générale du projet.

4.1. Cas d'utilisation

Un premier cas dans lequel notre site web pourrait être intéressant à utiliser est le cas où une personne souhaite partir en voyage et elle se soucie du confort de son voyage. Un trajet en avion est source d'angoisse pour de nombreuses personnes, qui ne se sentent pas à l'aise en vol ou même à l'aéroport. Afin d'éviter tout désagrément, il serait intéressant de pouvoir récolter des avis sur les précédents clients de plusieurs compagnies aériennes afin de voir quelle compagnie offre les meilleures prestations. Sur notre site, nous pouvons facilement distinguer les postes négatifs des postes positifs, de cette façon on peut se faire une première idée de la qualité de service qu'offre la compagnie que l'on a choisi. Par exemple, si tout le monde parle en mal d'une compagnie, on peut s'attendre à ce que notre avion ait plus de chance d'être en retard ou que le service à bord de l'engin soit de moins bonne qualité.

Un deuxième cas dans lequel on pourrait imaginer utiliser notre site serait dans le cas où l'on a un intérêt lucratif en jeu (ex : on possède des actions d'une compagnie aérienne) et que l'on souhaite surveiller la bonne santé de notre investissement. Concrètement, si l'image d'une compagnie est salie à cause d'un incident, cela se ressentira dans les tweets et les autres postes que les clients vont laisser sur les réseaux. Grâce à notre site, on pourrait donc savoir si l'image de l'entreprise se dégrade et si notre investissement est voué à chuter.

4.2. Architecture générale du projet

Pour créer nos visualisations, nous utilisons la librairie Python "Plotly". Cette librairie est assez populaire et nous permet de créer une large variété de graphiques très simplement et rapidement.

Nous allons également utiliser Python afin de récolter nos données, mais également pour faire la partie classification des sentiments. Python est le langage de programmation de référence pour faire du machine learning et c'est pourquoi nous avons choisi de l'utiliser.

Au final, nous aurons donc un serveur web, que l'on pourra lancer localement et qui contiendra nos différents graphiques interactifs. De plus, les scripts pour faire la classification des sentiments seront disponibles. Ces graphiques seront, entre-autres, une visualisation au fil du temps des sentiments par compagnies, cette même analyse mais pour chaque compagnie individuellement, un clustering des sentiments afin de représenter la proportions des sentiments, un wordcloud par compagnie, des statistiques, des barres charts des sentiments par années, etc.

Voici un schéma qui résume l'architecture qui va être mise en place

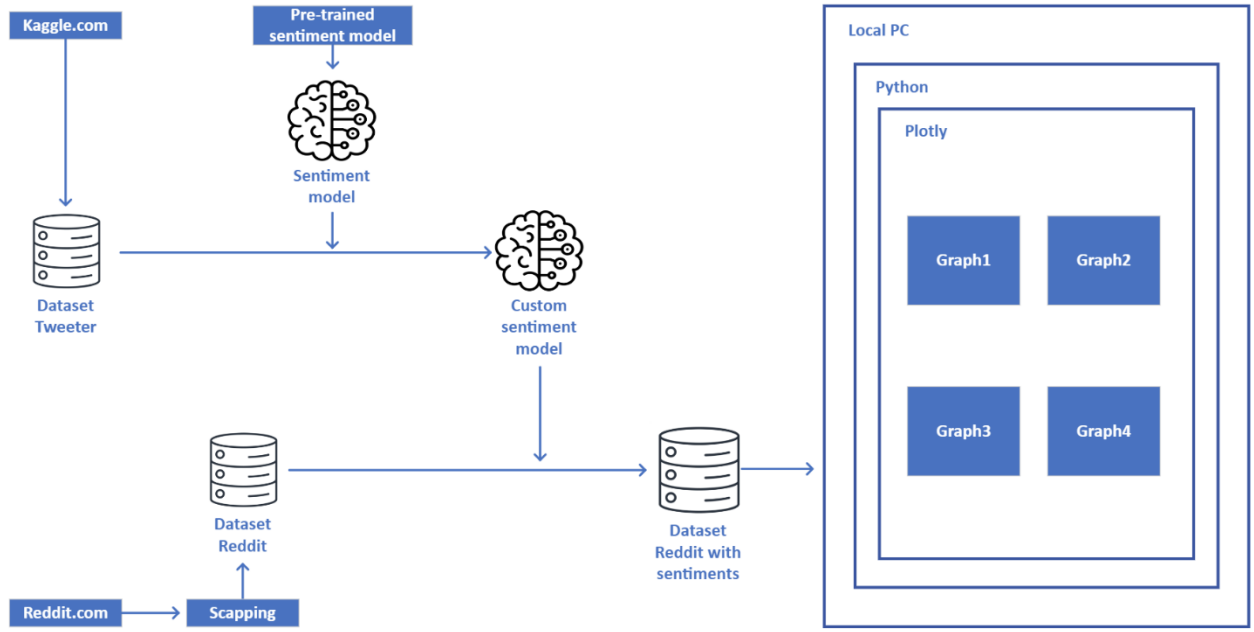

Figure 4: Schéma résumant l'architecture du projet

5.Fonctionnalité

Pour représenter au mieux les différentes caractéristiques de nos données, nous avons créé plusieurs visualisations sur notre site web.

5.1. Evolution temporelle

La première page de notre site contient des graphiques qui présentent l'évolution temporelle des tweets en fonction de chaque compagnie.

Ces graphiques nous permettent de voir s'il y a des tendances qui changent au fil du temps. Autrement dit, on peut facilement savoir si une événement marquant est arrivé à une compagnie aérienne.

Voici le premier graphique :

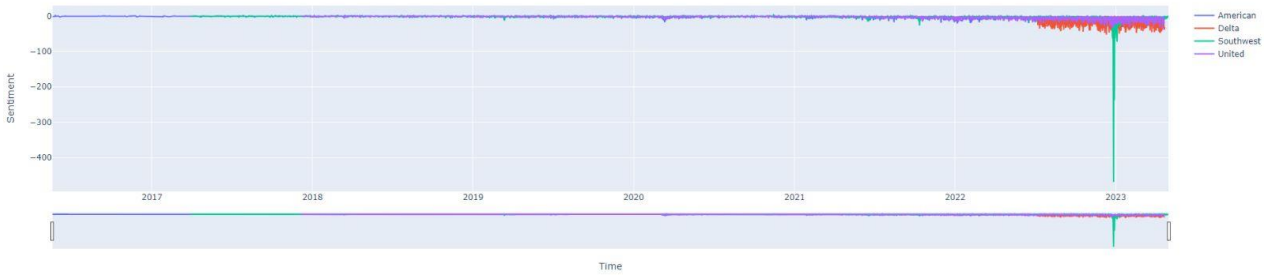

Figure 5: Graphique présentant l'évolution de la tendance sur les compagnies aériennes

Ce graphique présente l'évolution de l'appréciation des tweets, en fonction de la compagnie. Pour chaque jour, on prend le nombre de tweets positifs et on y soustrait le nombre de tweets négatifs. Comme on peut le voir sur l'axe Y, les tendances n'arrivent quasiment jamais en positifs, autrement dit, il y a toujours plus de tweets négatifs que de tweets positifs. On remarque aussi

un grand pic en fin 2022 pour la compagnie Southwest (en vert). La compagnie a fait face à une crise assez importante, ce qui explique l'explosion du nombre de tweets négatifs.

Nous avons également la possibilité de voir le nombre total de tweets tweeter par jour et par compagnie. Le graphique ci-dessous montre l'évolution des tweets lié à la compagnie Southwest :

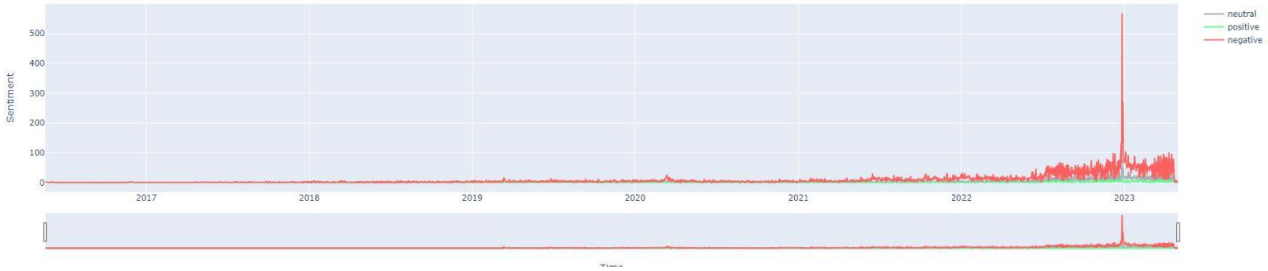

Figure 6: Graphique présentant le nombre de tweets par sentiment

On remarque à nouveau cet immense pic en fin 2022, il correspond évidemment à la crise citée précédemment et comme on peut le voir, le pic est rouge donc les tweets étaient bel et bien négatifs.

Enfin, nous avons un troisième graphique sur cette page qui nous montre le nombre total de tweet effectué chaque jour :

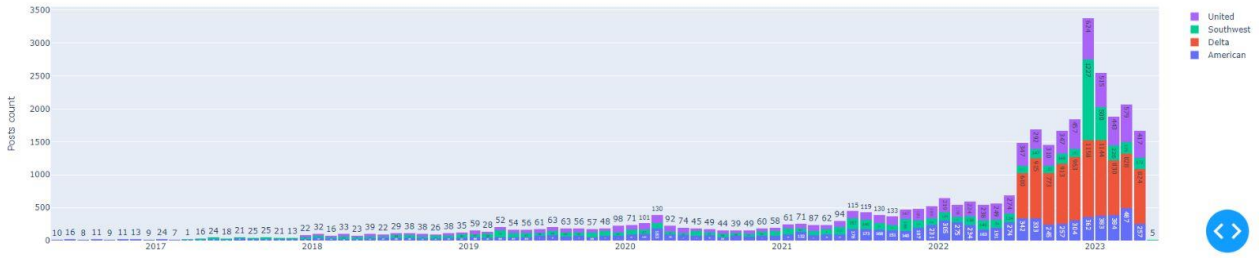

Figure 7 7: Graphique en bar présentant le nombre de tweets par jour par compagnie

Les barcharts empilées nous permettent facilement de distinguer quelle compagnie est la plus populaire pour une période donnée.

5.2. Répartition des classes

La deuxième page de notre site contient des pie charts qui permettent la visualisation de la répartitions des classes de chaque compagnie. Autrement dit, cela nous permet de voir le taux de tweets positifs, neutres et négatifs, pour chaque compagnie.

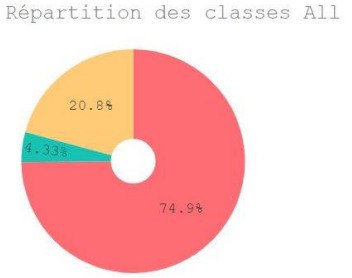

Figure 8: Graphique pie chart présentant la répartition des classes

Nous avons un graphique par compagnie et avons également ajouté un graphique représentant la mise en commun de toutes les compagnies, le voici :

5.3. Diagramme de Sankey

Sur la troisième page, nous affichons un diagramme de Sankey.

Ce diagramme, également connu sous le nom de diagramme de flux ou de diagramme d'énergie, est un type de visualisation qui illustre le flux de quantités entre différentes entités ou catégories. On l'utilise pour représenter des données de manière visuellement attrayante et intuitive.

Il se compose de blocs rectangulaires ou de colonnes verticales qui représentent les différentes catégories ou entités, telles que des pays, des produits ou des processus. Ces blocs sont reliés par des flèches ou des lignes de différentes largeurs, qui représentent le flux ou le mouvement des quantités d'une catégorie à une autre.

La largeur des flèches dans le diagramme de Sankey représente la quantité de flux ou de mouvement entre les catégories. Plus la flèche est large, plus la quantité est grande. De cette manière, le diagramme permet de visualiser les proportions relatives des différentes catégories et les transferts de quantités entre elles.

Ce diagramme est donc particulièrement utile pour avoir un aperçu des flux et des transferts de quantités entre différentes catégories de nos tweets.

Figure 9: Diagramme de Sankey présentant la répartition des données

Comme on peut le voir sur le graphique ci-dessus, les tweets sont plutôt bien répartis entre les diverses compagnies aériennes. On remarque également qu'il y a une majorité de tweets négatifs.

5.4. Page de statistiques

Enfin, nous avons décidé d'ajouter une dernière page à notre site web, pour afficher des statistiques plus précises sur nos tweets. En voici un exemple :

First post date : 05.20.2016

Last post date : 05.01.2023

Number of posts : 34028

Number of positive posts : 1473 (4.33%)

Number of negative posts : 25473 (74.86%)

Number of neutral posts : 7082 (20.81%)

Random post : Favorite lounge food hacks? [removed] --> negative

Number of users : 19698

Figure 10: Image des statistiques sur les tweets

Nous affichons également le contenu d'un tweet aléatoire afin de voir à quoi ressemble un tweet positif et/ou négatif.

Cette page de statistique contient également un « word cloud » qui nous permet assez rapidement de voir les mots les plus couramment utilisé dans les tweets analysés.

Figure 11: Exemple de word cloud

Les mots « thanks », « refund » ou encore « upgrade » semblent plutôt positifs. En revanche les mots « time » et « delete » ont plutôt une connotation négative.

5.5. Switch entre les datasets

Comme expliqué dans le chapitre 2 de ce document, nous avons dû utiliser deux datasets pour réaliser ce projet. L'un a été télécharger directement depuis Kaggle, il s'agit de tweets déjà annotés, et l'autre contient des postes Reddit que nous avons scrapper nous-même. Nous avons pris la décision de travailler avec ces deux datasets, car nous n'étions pas certains d'arriver à de bons résultats en faisant du sentiment analysis sur les postes Reddit. Au final, les résultats sont plutôt satisfaisants, c'est pourquoi nous avons la possibilité de switcher entre ces deux datasets.

Nous pouvons donc définir, sur chacune des pages de notre site web, si nous voulons affichés des statistiques liés à notre base de données de tweets ou si nous préférons analyser les postes Reddit que nous avons scrapper.

6.Techniques, algorithmes et outils utilisés

Pour ce qui est du dataset Kaggle, nous n'avons pas eu de travail spécifique à faire dessus, les sentiments étant déjà labellisé. Cependant, ce n'était pas le cas des données venant de Reddit.

6.1. Algorithme de ML utilisé

Afin de pouvoir labelliser les données de Reddit dans le même format que les données de Kaggle, nous avons dû mettre en place un modèle de Machine Learning permettant de catégoriser les messages en négatif, neutre ou encore positif.

Afin de réaliser cela, nous avons utilisé Huggingface. En effet, huggingface est un outil complet permettant de faire facilement de l'entraînement de modèle existant, avec les fonctions comme la tokenization nécessaire pour le Natural Language Processing.

En effet, il est nécessaire d'utiliser le tokenizer prévu par le modèle pour faire le pre-processing des données, cela afin de convertir le texte en features voulues par le modèle.

Deux modèles ont ensuite été testé afin de pouvoir comparer les performances. Pour cela nous avons aussi pris en compte ce qui est ressorti de l'état de l'art.

6.1.1. DistilBERT

Le premier modèle est une version allégée d'un modèle BERT classique, introduit par ce papier. Il a l'avantage d'être plus rapide et moins gourmand à entraîner qu'un modèle BERT classique. La version de DistilBERT a été pré-entraîné de la même manière que BERT, c'est-à-dire avec du texte brute anglais, d'une manière non-supervisée. Le fait d'utiliser une version allégée de BERT nous a permis de pouvoir l'entraîner en un temps raisonnable, avec bien sûr la nécessité de le faire sur un GPU.

Afin de réentraîner ce modèle pour notre utilisation, le dataset de Kaggle a été importé, préparé puis séparé en train et test (25%/75%). Une fois cela fait, l'entraînement a été effectué et un f1_score de 84% a été obtenus. Ce résultat est déjà tout à fait satisfaisant au vu de la quantité de données d'entraînement à disposition, tout en sachant que le sentiment analysis est une tâche pouvant être complexe. Cela s'explique aussi par l'ambiguïté qu'il peut y avoir dans la classification de certains tweets. En effet, il est parfois compliqué même pour un humain de savoir dans quel classe se trouve certains tweets, qui peuvent être du second degré ou tout simplement à cheval entre deux classes.

6.1.2. Twitter-roBERTa-base

Le second modèle qui été entraîné est un modèle que nous avons découvert durant la réalisation de l'état de l'art. En effet, il est basé sur le paper dont nous parlons plus haut, « TimeLMs: Diachronic Language Models from Twitter ». Ce modèle est basé sur RoBERTa, qui a la même architecture que BERT. Ce modèle a cette fois déjà été entraîné à faire de la classification comme nous le souhaitons, sur 124M de tweets.

Pour le réentraînement, le même processus que pour le modèle précédent a été utilisé. Cela nous a permis d'obtenir des résultats légèrement meilleurs, avec un f1_score de 88%. Ces résultats sont sans grandes surprises meilleurs que ceux du modèle précédent, étant déjà pré-entraîné à faire de la classification sur une grande quantité de tweets. Ce modèle est donc choisis pour labelliser les données de Reddit.

6.1.3. Inférence

Une fois le modèle entraîné, nous avons fait de l'inférence sur nos données Reddit afin de les labelliser. Cela a permis finalement d'obtenir un dataset de la même forme que celui. Provenant de Kaggle, avec le sentiment pour chacun des posts Reddit.

6.1.4. Résultats

Finalement, nous avons pu obtenir comme prévu un dataset de posts Reddit labelisé en fonction des sentiments. Il est cependant complexe de définir exactement la précision de la labellisation sur ces données, au vu du fait qu'elles ne sont pas labellisées. Nous avons des résultats tout à fait satisfaisant sur le set de données provenant de Kaggle, mais il ne nous est pas possible de le démontrer sur le dataset Reddit autrement qu'en regardant des posts et en contrôlant manuellement la labellisation. Nous pouvons tout de même faire l'hypothèse que les résultats seraient proches, au vu de la similitude entre un post Reddit et un tweet et des quelques contrôles que nous avons fait manuellement.

Voici un exemple d'utilisation du modèle avec trois post Reddit en entrée, un par catégorie :

```
sentiment_model([
"I flew United last month and the experience was AWESOME!",
"is flight 587 from DFW to ORD currently on-time?",
"@united my luggage has been broken!! #youcouldntmakethis up #brokenwheel"
])
```

Et la sortie que nous obtenons, qui sont dans ce cas tout à fait correct :

```
[{'label': 'positive', 'score': 0.9289741516113281},
 {'label': 'neutral', 'score': 0.8915866613388062},
 {'label': 'negative', 'score': 0.9768679141998291}]
```

7. Planification et organisation

Voici un diagramme de GANTT qui représente notre planification et répartition du travail pour ce projet :

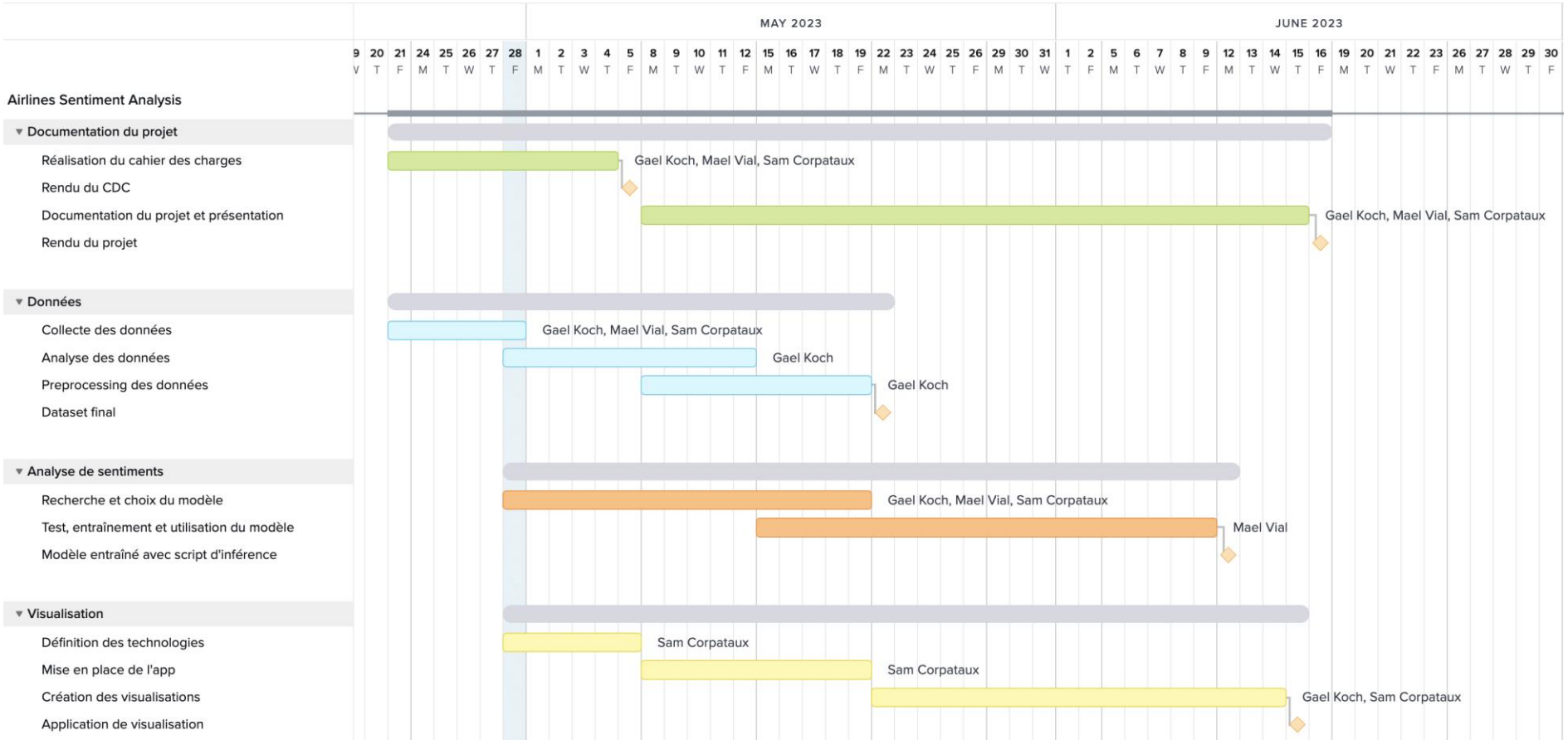

Figure 12 : Diagramme de GANTT du projet

8. Conclusion

En conclusion, notre projet s'est déroulé de manière satisfaisante, marqué par une excellente entente au sein de notre groupe. Cependant, un obstacle imprévu est venu perturber notre progression lorsque notre outil de scrapping de Twitter a été bloqué. Cette situation s'est malheureusement produite moins de 24 heures avant le lancement de notre opération de scrapping, ce qui a été un coup de malchance indéniable.

Malgré cet inconvénient majeur, nous avons su réagir rapidement et trouver une solution intermédiaire en nous tournant vers les publications Reddit et les datasets disponibles sur Kaggle. Cette alternative a permis de compenser la perte de données provenant de Twitter et a assuré la poursuite de notre projet dans des conditions raisonnables.

En dehors de cet incident, le reste du projet s'est déroulé sans accroc. Nous avons réussi à mettre en pratique les connaissances acquises en classe, ce qui nous a donné une occasion précieuse d'appliquer nos compétences techniques et de renforcer notre compréhension des concepts abordés.

En somme, malgré l'obstacle inattendu que nous avons rencontré, nous sommes satisfaits des résultats obtenus grâce à notre bonne entente et à notre capacité à trouver une solution alternative. Ce projet nous a permis de consolider notre apprentissage et de développer nos compétences, tout en nous confrontant à des défis réels propres au monde de la collecte et de l'analyse de données.

Pour les travaux futurs, il serait opportun d'explorer de nouvelles perspectives afin d'améliorer encore davantage notre projet. Une première piste consisterait à essayer d'autres modèles de ML, tels que les réseaux de neurones profonds ou les algorithmes d'apprentissage par renforcement, pour obtenir des résultats plus précis et améliorer les performances globales de notre système d'analyse des sentiments. De plus, afin d'évaluer les performances réelles de notre modèle sur les données Reddit, il serait nécessaire de labelliser les posts Reddit et de les intégrer à notre ensemble de données. Cela nous permettrait de mieux comprendre les spécificités de cette plateforme et d'ajuster notre modèle en conséquence, pour assurer une analyse des sentiments plus pertinente et adaptée aux différentes sources de données. Ces travaux futurs seraient une continuation naturelle de notre projet, ouvrant ainsi de nouvelles perspectives d'amélioration et de développement.