Analyse des sentiments des tweets sur une population importante à travers le temps

Adam DIAKITE
Université de Paris
adam.diakite@etu.u-paris.fr

Amélie WERBREGUE
Université de Paris
amelie.werbregue@etu.u-paris.fr

Aurélie LING
Université Paris
aurelie.ling@etu.u-paris.fr

Avant-propos

L'analyse des sentiments de populations est une notion qui devient de plus en plus importante. En effet, avec l'avènement des réseaux sociaux et de l'ère du big data, grâce auxquels plusieurs centaines de millions d'utilisateurs provenant du monde entier créent de la donnée, il est intéressant de s'intéresser aux différents avis et opinions exprimés par ces personnes sur des domaines variés.
Il existe en effet de nombreuses approches différentes pour analyser les sentiments et l'opinion des utilisateurs de réseaux sociaux. Le regroupement de ceux-ci est souvent fait par utilisateurs de la même région, ou du même pays, mais beaucoup plus rarement par d'autres facteurs comme l'état, la tranche d'âge, l'appartenance politique ou encore par date de publication.
Pour ce faire, il est nécessaire de produire des méthodes s'appliquant sur des populations plus larges et permettant une comparaison et une analyse plus rigoureuse pour peut-être produire des corrélations et des liens de causalité.
Avant d'émettre de telles propositions, il est important d'être à même de prendre compte des différents biais et d'autres facteurs indissociables des données et métadonnées des réseaux sociaux - comme le bruit - pour pouvoir émettre des postulats étant les plus pertinents possible.
En raison des corrélations et des différents informations que nous pouvons tirer des résultats, nous nous sommes demandés si l'opinion exprimée sur les réseaux sociaux est représentative de la réalité et si il était possible de produire des analyses globales grâce à l'analyse syntaxique des sentiments des utilisateurs de twitter du monde entier. Il est enfin fondamental de savoir quelle est la méthode la plus pertinente dans le domaine de l'analyse des sentiments sur Twitter.

En réponse à ces problématiques, nous vous proposons un travail divisé en trois parties.

Il s'agit dans un premier temps de mettre en place plusieurs algorithmes d'analyse numérique des sentiments s'appuyant sur des valeurs quantitatives comme le nombres de *likes,* de *retweets*, ou encore de de *tweets* postés sur les deux candidats. Il s'agit d'une analyse numérique et spatiale car nous nous intéresserons au volume des données.

Ensuite, nous nous intéresserons à une analyse plus poussée des tweets à travers le temps et l'espace pour attester de l'évolution de l'opinion sur un sujet donné, toujours en prenant compte des données spatiales en plus des données temporelles.

I. Introduction

Le problème soulevé dans ce rapport de recherche est l'analyse des sentiments des utilisateurs de twitter. De façon plus précise, nous nous intéresserons à comment l'opinion publique peut varier sur un sujet donné à travers le temps et l'espace. Ainsi, nous avons décidé de travailler et d'étudier différents algorithmes pour pouvoir produire des résultats significatifs et pertinents. Algorithmes étants résumés dans l'avant-propos. L'analyse des sentiments est un domaine important car elle s'applique dans de nombreux domaines : politique, économie, écologie ou encore technologies et est utile pour la prise de décision et la prédiction d'événements. L'extraction de sentiments est très largement utilisée dans plusieurs industries, économiques, de recherche et même sociales, en passant par les industries de l'art et du divertissement. Tandis que de nombreuses études ont été menées pour produire des algorithmes jugeant du sentiment dans un texte brut, nous avons choisi d'étudier l'analyse des tweets car Twitter est un site de

micro-blogging, et que de par sa nature, nous ne pouvions pas nous permettre de choisir les tweets uniquement par type d'utilisateurs particuliers; le site étant très privé à cet égard. L'intérêt de cette recherche est donc multiple : d'une part, il réside dans le fait de pouvoir comprendre les tenants et aboutissants de la recherche sur l'analyse des sentiments par agrégation de contenu, d'une autre part, il nous permet de mieux comprendre différents algorithmes et cheminements prévu à cet effet.

C'est donc en réunissant plus d'1.6 millions de tweets posté du 15 octobre 2020 au 11 novembre 2020 - en pleine période d'élections - que nous avons choisi d'analyser les sentiments des différents utilisateurs, qu'ils soient d'Amérique ou du reste du monde, pour pouvoir établir des analyses plus globales. Pour ce faire et étant donné que les élections ont déjà eu lieu et qu'il s'agit d'une période très courte, mais intense, nous avons choisi de stocker les tweets réunis sur une base de données MySQL, puis de dynamiquement créer des structures de données sur un notebook pour manipuler le contenu de ces bases de données. Cependant, la difficulté de cet exercice réside dans deux points fondamentaux : le volume des données et leur nettoyage.
Dans un premier temps, il faut d'ores et déjà penser à la quantité de données à analyser : ici, il s'agira de plusieurs millions de tweets. Il est donc nécessaire, dans un premier temps, de veiller à l'optimisation des algorithmes pour réduire le plus possible le temps de traitement.
Dans un second temps, il nous semble important d'effectuer un travail de nettoyage des données. En effet, comme précisé plus tôt, les données, et plus précisément les données web sont souvent formatées pour être comprises par des programmes et donc illisibles pour l'homme. De plus, le problème que représente le bruit dans le texte est non négligeable et doit être pris en compte dans l'exploitation des données recueillies. Enfin, la normalisation des données ainsi que leur égalisation est nécessaire pour avoir un jeu de données équitable et pertinent.
Enfin, il est important de mesurer le poids des données corrompues (lignes manquantes dans la base de données, données aberrantes) dans le calcul des différents scores.

II. Recherches liées

L'analyse des sentiments est aujourd'hui très explorée et étudiée. Elle est utilisée dans divers domaines. Un travail de recherche effectué en amont nous a permis de mieux comprendre l'environnement dans lequel nous allions nous plonger.

Dans [9], la mission d'analyse de sentiments sur Twitter organisée par SemEval, l'atelier international sur l'évaluation sémantique, est présentée. Grâce à cela, nous avons pu comprendre les grands points de ce sujet puisqu'on peut donc y retrouver les méthodes importantes qui ont été utilisées lors de cet évènement. Par exemple, pour assigner une polarité à un tweet, il est possible d'utiliser le texte dans son ensemble, des ensembles de mots ou encore des lexiques de mots porteurs de sentiments. Nous pouvons retrouver comme « features », des n-grammes, des clusters de mots mais aussi les caractères spécifiques à Twitter comme les hashtags ou les tags. De plus, certains aspects sont à prendre en compte pour l'évaluation, comme le contexte ou le sarcasme. Enfin, différentes méthodes de Machine Learning peuvent être appliquées pour classifier les données, comme les *Support Vector Machines*.

L'article [6], nous présente une méthode permettant de prédire les prix futurs des actions, sur un temps donné, en se basant sur l'analyse des données des réseaux sociaux. D'une part, ce sont des tweets en anglais qui sont récupérés et pré-traités, de l'autre, des données à propos des actions. Les sentiments des tweets sont décidés grâce à SentiWordNet. Les tweets subjectifs sont d'abord sélectionnés et ensuite classés selon leur polarité. Grâce à la classification des sentiments avec la méthode Naive Bayes et les données sur des anciens prix, il est montré dans cet article qu'il est possible de prédire des prix futurs sur certain temps.

L'analyse des sentiments a été utilisée pour connaître l'opinion des passagers de transports en Indonésie sur la transmission du virus du Covid-19, comme le décrit [5]. Les données sont prétraitées dans le but d'obtenir des mots-clés parmi les tweets. Pour faire cela, celles-ci sont nettoyées, le texte est mis en minuscule, elles sont filtrées et les mots sont séparés. Les méthodes de *Naive Bayes* et d'arbre de décision ont été comparées pour la classification. L'approche Naive Bayes est intéressante et importante car elle permet de réduire la complexité sur des larges données. Les chercheurs ont réussi à montrer qu'elle

était plus efficace qu'un arbre de décision puisqu'elle donnait une précision égale à 73,82% contre 58,24%.

Dans [7] un outil d'analyse de sentiments a été développé. D'abord, une extraction, tokenisation et nettoyage des données (en enlevant les mots non importants) sont réalisées. Des monogrammes (*unigrams)* ou des bigrammes (*bigrams)* sont pris en compte, c'est-à-dire un ou deux mots, puis les *Part Of Speech* (POS) tags, c'est-à-dire le type des mots, si c'est un verbe, un sujet, un adjectif… Ces différentes méthodes sont comparées grâce à la méthode de classification *Naïve Bayes*. Soit les monogrammes sont utilisés seuls soit avec les POS tags, de même pour les bigrammes. Il a été conclu que l'analyse était plus précise en choisissant les monogrammes sans les POS tags. De plus, il est précisé que la différence entre un tweet positif et négatif est plus ambiguë que celle entre un tweet subjectif et objectif. Enfin, des données sur un temps court peuvent mener à des erreurs car les opinions changent selon les périodes.

Dans [8], nous avons une analyse des sentiments des tweets concernant l'élection présidentielle de 2011 à Singapour. Le but, ici, est de calculer le score de sentiment pour chaque candidat et de quantifier les données, la difficulté étant la différence de sens entre l'anglais américain ou britannique, les mots n'ont pas la même signification. Ce qui est intéressant dans cet article, c'est qu'il y a, en plus du traitement de base, une correction des données qui est faite. En effet, d'autres données sont récupérées en plus, comme l'âge ou l'appareil utilisé pour obtenir des informations plus précises. Malgré une marge assez petite, les premiers candidats ont pu être distingués.

Dans [3], un dictionnaire de sentiment est appliqué pour classifier les tweets. Celui-ci comporte des mots annotés comme positifs ou négatifs. Les tweets sont nettoyés et les mots sont séparés. Pour chaque mot, on compare avec les mots dans le dictionnaire, s'il est positif on augmente le score de positivité, de même s'il est négatif. En fonction du score que l'on obtient, cela donnera un pourcentage de sentiment qui permettra de décider de la polarité. Cela a permis aux chercheurs de comparer les scores obtenus sur différents sujets quotidiens, comme la politique, les fake news ou les films. La plupart du temps, nous retrouvons des scores bien contrastés entre positif et négatif mais avec une forte proportion de neutres.

L'article [2] présente une méthode d'analyse des sentiments avec une représentation en réseau multicouche hétérogène des tweets. Ce dernier est décomposé en plusieurs couches, hashtags, mots-clés et mentions. D'abord, une marche aléatoire est appliquée, ensuite un traitement d'intégration des nœuds et enfin un algorithme de Deep Learning. Pour chaque point, différentes approches sont utilisées et comparées. Les données récupérées sont des tweets en anglais et hindi pour différents sujets. L'étude de cet article se révèle efficace, plus particulièrement avec un marche aléatoire biaisée dite consciente de la centralité et une expansion des nœuds orientée sur le sentiment.

Mais les outils d'analyse des sentiments peuvent parfois se révéler inconsistants, ce qui est expliqué dans [4]. En effet, il est montré qu'avec des textes sémantiquement similaires nous pouvons retrouver une polarité différente. C'est ce qu'on peut appeler des exemples contradictoires. En prenant en compte deux documents avec la même sémantique, sont évalués, *l'intra-tool inconsistency,* qui correspond à une analyse inégale de ces documents pour un même outil et *l'inter-tool inconsistency*, qui correspond à une analyse inégale de ces documents pour deux outils différents. Cette étude est appliquée sur six outils d'analyse de sentiments importants et populaires dans le domaine. Il est alors montré qu'en résolvant ces inconsistances, il est possible d'obtenir des outils plus efficaces.

Enfin, dans [1], une recherche est faite sur la corrélation des sentiments entre différents groupes démographiques. En effet, les opinions et sentiments ne sont pas toujours les mêmes pour différents groupes de personnes ou dans différentes zones géographiques. Un objectif de cet article est de mesurer la similarité entre plusieurs groupes sur des intervalles de temps donnés afin de retrouver les groupes avec les mêmes opinions. Cette mesure est effectuée grâce au coefficient de corrélation.

III. Formulation du problème

Le problème auquel nous nous sommes intéressé était de savoir comment manipuler des données de masse pour extraire des opinions globales. En d'autres termes, il s'agit d'émettre trois classes d'appartenances à partir de millions de tweets. Il est donc très important d'effectuer un travail de classification préliminaire en amont du traitement des résultats.

Nous voulions être capable de savoir s'il était possible de prévoir l'issue d'élections et si les avis exprimés sur les réseaux sociaux étaient représentatifs de l'issue de ces dernières. Pour ce faire, nous avons choisi d'utiliser un algorithme de classification des sentiment dans les textes : Sentiment Intensity Analyzer permettant d'attribuer un score d'intensité aux tweets pour ensuite nous permettre d'établir des classes dans lesquelles placer les différents tweets, afin d'ensuite établir des graphes explicatifs.
De plus, nous avons essayé plusieurs autres algorithmes comme la méthode K-means pour la classification des tweets en différentes classes afin de calculer les différents scores de précision et les taux de spécificité et sensibilité. Nous avons alors pu voir quelle est la meilleure méthode de clustering pour établir des classes de tweets, sans utiliser d'algorithmes de classification pré-conçu.
Pour l'algorithme des k-means, nous avons choisi d'établir k à 2 pour les trois classes suivantes : positif, et neutre afin d'établir des k-moyennes représentant ces deux classes.
Le fonctionnement de K-means est le suivant : à partir d'un ensemble composé de n points, nous devons répartir ces n points (x1, x2, x3, x4… xn) en k classes différentes avec une distance la plus petite possible entre chaque point d'un même cluster S. La formule suivante décrit le fonctionnement de K-Means :

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

IV. Solution

Nous allons maintenant vous présenter notre solution à ce problème. Pour l'élaborer, nous avons utilisé le langage Python sur Jupyter Notebook afin d'obtenir des résultats facilement explicables et démontrant des corrélations intéressantes. Il s'agit en premier lieu de récupérer les données nécessaires au traitement.

Pour former les data sets, il a fallu utiliser snscrape et l'API Twitter pour streamer les tweets par *hashtag*. C'est à dire que grâce à l'outil snscrape, il est possible de streamer les tweets d'une date *x* à une date *y* qui comprennent le mot clé *k*. Dans cette recherche, nous avons choisi d'effectuer deux commandes snscrape : une pour les tweets concernant Donald Trump et une pour les tweets concernant Joe Biden, à chaque fois du 15 octobre au 8 novembre 2020, période d'effervescence pour les élections américaines.
Grâce à ces commandes, nous avons pu indiquer au scraper, qui effectue des commandes automatiquement via l'API twitter et un compte, quelles données entrer et quand les rechercher.
De cette commande a pu résulter deux fichiers .csv, comprenant 21 colonnes chacun pour plusieurs centaines de milliers de lignes.
Les colonnes représentent à chaque fois des informations que nous avons jugé intéressantes pour l'analyse : *created_at* qui montre sous le format datetime la date de création du tweet, *tweet_id* qui sert de clé primaire unique et qui permet d'identifier chaque tweet, *tweet*, qui correspond au texte brut posté par l'utilisateur, *likes* et *retweets* qui équivalent respectivement au nombres de "j'aime" et de "retweet" (partages) reçus par le tweet précédemment posté, *source* qui décrit la provenance du tweet (Ios, android, ordinateur...), *user_id* qui définit l'utilisateur ayant posté, *user_name* son nom d'affichage, *user_description* qui est la description que l'utilisateur fait de lui-même, *user_join_date* qui est la date de création du compte de l'utilisateur, encore une fois en format datetime, *user_location* qui est la localisation que l'utilisateur entre lui même dans son profil, *lat et long*, respectivement la latitude et la longitude de la localisation de l'utilisateur et enfin, plusieurs données géospatiales comme *country, city, state, state_code* et *continent.* Comme nous l'étudierons plus tard dans cet article, ces types de données nous seront très utiles, cependant, il faudra effectuer avant toute chose un travail de nettoyage de la base de données.

1. Nettoyage des données

Il est d'abord nécessaire de nettoyer les données. Pour cela, nous avons appliqué plusieurs méthodes que nous avons pu voir dans différents articles.
Premièrement, nous ne sélectionnons que les tweets qui contiennent des mots-clés comme "Trump" ou "Biden" par exemple. Ensuite, nous enlevons les URL, les mentions, les émoticônes, les ponctuations et les hashtags pour avoir plus de précision dans l'analyse des données. (Data cleaning) [5][7]
Enfin, il faut détecter la langue écrite des tweets, pour pouvoir la qualifier d'"unknown" si elle n'est pas précisée, afin de ne pas utiliser ces tweets lors des résultats finaux.

2. Clustering des données

Pour le clustering des données, nous utilisons différents regroupements. Dans un premier temps, nous réalisons des regroupements simples afin de visualiser les données. Nous séparons les tweets en fonction des mots-clés qu'ils contiennent c'est-à-dire, ici, "Trump" et "Biden" et ensuite, selon la temporalité ou la localisation.
Dans un second temps, nous réalisons des regroupements selon plusieurs critères. Par exemple, nous allons regarder en fonction des appareils ou plateformes des utilisateurs (portable, site web) pour chaque pays ou encore en fonction de chaque langue pour chaque pays. Cela nous permettra d'identifier les types de personnes et à quel point elles sont liées à l'élection présidentielle.

3. Score du sentiment

Pour analyser les sentiments sur les données, nous savons que chaque tweet peut avoir des critères différents (sarcasme, subjectivité, etc). [9]
[7] Nous avons exploité les classes de python qui permettent d'analyser les sentiments des tweets (nltk.sentiment).
Tout d'abord, nous avons analysé les tweets filtrés en appliquant, d'un côté, la subjectivité et de l'autre, la polarisation. Ensuite, nous avons appliqué sur la polarisation des tweets une analyse globale, afin de déterminer si les tweets sont positifs, négatifs ou neutres. Nous avons alors pu visualiser les résultats obtenus sous formes de différents diagrammes afin de mieux les comprendre.

4. Analyse spatiale et temporelle

L'analyse de la géolocalisation des données, permet de voir si les élections des Etats-Unis sont très suivies dans le monde et différencier les provenances des tweets. Pour visualiser les tweets, nous avons utilisé tweet Loc [10]. Nous avons pu avoir une vue globale des données. Puis nous examinerons, selon les pays, le contenu des tweets. Enfin, nous pouvons aussi obtenir des informations en fonction du temps, en heure ou en en jour, ce qui nous servira à voir s'il y a eu des changements de sentiments et d'opinions au cours d'une période.

5. Comparaison des résultats

Pour évaluer les résultats obtenus avec l'analyse des sentiments, nous avons utilisé une méthode Naive Bayes avec la classe nltk.classify de python. Comme dans cet article [5] , qui regroupe les modèles puis calcule les probabilités de chaque classe en se basant sur la division des mots dans un document. Dans notre cas, nous allons exploiter cela sur des tweets et non des documents. Nous avons également utilisé l'algorithme K-Means, qui permet de résoudre les problèmes de classification et de régression.

V. Expérimentation

1. Hardware et données

Pour mener à bien notre étude, nous avons utilisé les notebook Jupyter ainsi qu'une base de données MySQL.
Les raisons suivantes nous ont poussé à utiliser MySQL plutôt qu'un autre SGBD :

- MySQL est le système de gestion de base de données offrant la meilleure performance.
- Projet Open Source et donc accessible pour la recherche. l'accessibilité est un élément important dans notre contexte
- Connectivité : plusieurs clients peuvent se connecter en même temps sur une database. Travaillant en groupe, c'est une donnée importante pour tester nos résultats simultanément.
- Utilisation de SQL possible grâce à l'intégration du langage. Cela nous a permis une meilleure exploration de nos données pour mieux les appréhender ensuite dans les résultats.
- Enfin, il s'agit du système de gestion de base de données le plus populaire, ce qui est utile en cas de problème quelconque.

De plus, nous avons utilisé un notebook Jupyter car ceci nous permettait de travailler en parallèle et d'avoir des résultats très descriptifs et nous permettant de les analyser rapidement.

Notre dataset est composé de 2,470,710 tweets postés entre la période du 15 octobre 2020 au 8 novembre 2020 comprenant les mots clé ou *hashtag* "Joe Biden" ou "Donald Trump". Ces tweets sont recueillis indépendamment et sans aucun filtrage (celui-ci s'est effectué après).

Ensuite, nous nous intéresserons au traitement empirique des données et à des explications à ce niveau. Nous avons tout d'abord choisi d'effectuer un travail préliminaire de valeurs quantitatives. Pour ce faire, nous avons trié les tweets par provenance (Pays) puis compté pour chaque pays puis classé. Comme on peut le voir, nous voulions effectuer une analyse globale et cela a été possible car presque chaque continent est représenté dans le top 5 des plus gros émetteurs de tweets, comme on peut le voir dans la figure 1 suivante :

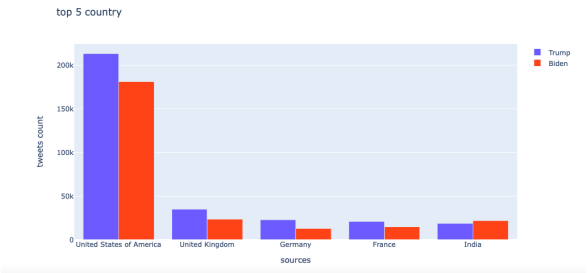

Figure 1 : Nombre de tweets postés par les 5 plus grands pays émetteurs de tweets

En revanche, la base de données comporte de nombreuses valeurs nulles, c'est-à-dire non renseignée par l'utilisateur qui a posté le tweet. Nous l'avons remarqué plus en détail, avec la figure 2 ci-dessous. En effet, beaucoup de tweets n'ont pas d'informations sur certains ou tous les critères de localisation.

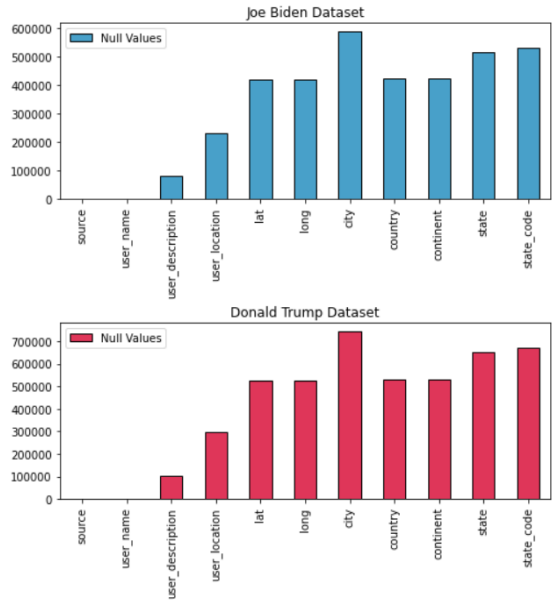

Figure 2 : Nombre de tweets n'ayant pas de valeurs pour chaque catégorie de la base de données

2. Visualisation des données

2.1. Catégorisation

Dans un premier temps nous avons catégorisé les données. En effet, La base de données nous donne accès à différentes informations sur les tweets. Grâce à celles-ci, nous pouvons regrouper les tweets selon certaines catégories afin de mieux visualiser les données que nous possédons et permettre une meilleure analyse.

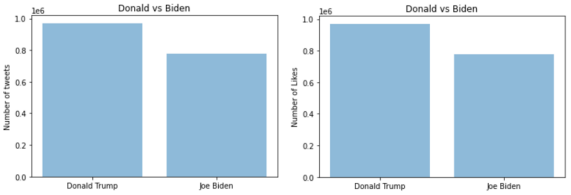

Figure 3 : A gauche, nombre de *tweets* pour chaque candidat, à droite, nombre de *likes* pour chaque candidat

Nous avons d'abord regardé combien de tweets concernaient chaque candidat ainsi que le nombre de likes. Nous pouvons voir sur la figure 3 que pour les likes ou pour les tweets, nous avons un nombre plus important lorsqu'il s'agit de Donald Trump.

2.2. Données spatiales

Grâce aux différentes données des utilisateurs à propos de leur localisation, nous pouvons comparer le nombre de tweets par candidat selon leur provenance. Par exemple, dans la figure 1, nous pouvons voir le nombre de tweets pour les six pays qui en comportent le plus. Comme attendu, nous retrouvons bien une quantité de tweets plus importante pour les Etats-Unis, avec toujours plus de tweets concernant Donald Trump.

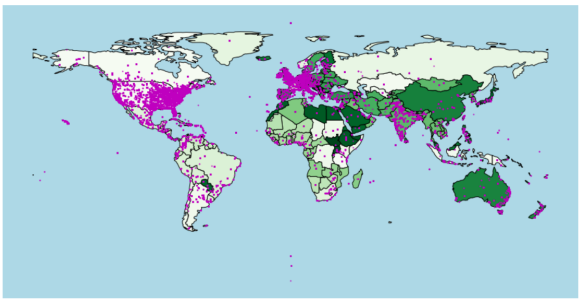

Figure 4 : Localisation des tweets (un tweet est représenté par un point violet sur la carte)

Nous avons aussi comparé selon les différents continents, les différentes villes et les différentes régions ou états du monde. De tout évidence, les localisations les plus fréquentes se situent en Amérique et aux Etats-Unis. Nous retrouvons cela avec la figure 4 où nous pouvons voir d'où proviennent les tweets.

2.3. Données temporelles

Nous disposons également de données temporelles à propos du tweet, c'est-à-dire la date et l'heure à laquelle il a été posté.

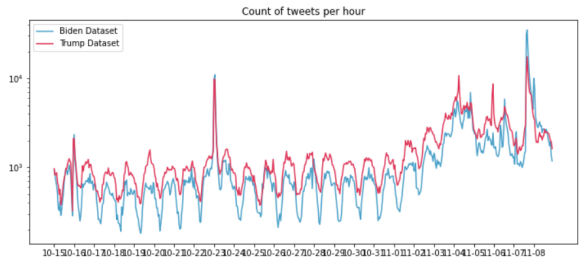

Figure 5 : Nombre de tweets par heure pour chaque candidat du 15 Octobre 2020 au 08 Novembre 2020

Comme représenté sur la figure 5, le nombre de tweets est principalement en augmentation à partir de novembre, c'est-à-dire en fin de période d'élection. La haute proportion retrouvée est cohérente puisqu'il s'agit des dates de fin d'élection, lors de l'annonce du vainqueur.

3. Analyse des sentiments simple

Dans un second temps, nous avons pu classer les tweets selon le sentiment qui a été retourné, *positif* *négatif* ou *neutre* pour chaque candidat.

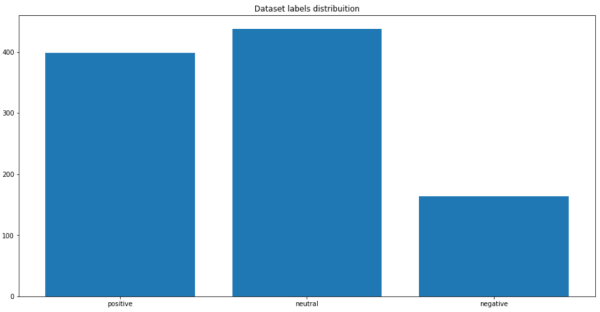

Figure 6 : Nombre de tweets concernant Donald Trump par sentiment *positive*, *negative* ou *neutral*

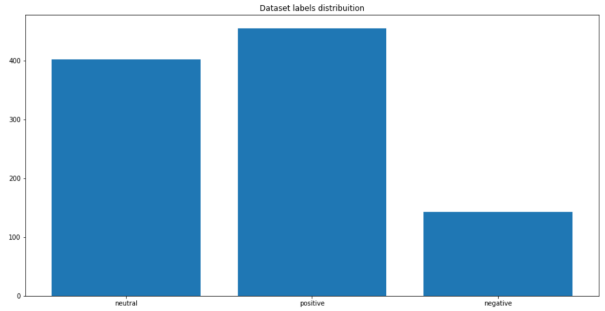

Figure 7 : Nombre de tweets concernant Joe Biden par sentiment *positive*, *negative* ou *neutral*

Comme nous pouvons le voir dans la figure 6, qui montre le nombre de tweets par sentiment pour le candidat Donald Trump, on retrouve beaucoup plus de tweets neutres et positifs que négatifs. C'est également le cas pour la figure 7, qui concerne le candidat Joe Biden. En comparant les deux graphiques, nous remarquons qu'il y a légèrement plus de tweets positifs pour Joe Biden.

Figure 8 : Pourcentage de tweets selon l'analyse du sentiment, à gauche pour les tweets concernants Trump, à gauche pour ceux concernant Biden

Plus précisément, lorsque nous regardons les pourcentages obtenus, que nous pouvons retrouver sur la figure 8, nous voyons qu'il y a plus de tweets analysés positifs pour Biden, avec une marge de 5,7%. En adéquation, nous retrouvons plus de tweets négatifs pour Trump, avec une marge de 2,1%. Nous remarquons tout de même que dans les deux cas, il y a une forte proportion de tweets analysés comme neutres.

4. Analyse des sentiments avec la localisation et la temporalité

Dans un dernier temps, nous avons pris en compte la localisation et la temporalité des tweets pour comparer l'analyse des sentiments obtenus.

Avec la figure 9, nous pouvons voir que pour le candidat Joe Biden, nous obtenons en grande partie

des scores de sentiments positifs, seulement 5 états ont un score négatif lors des 14 premiers jours et un seul pour les 14 derniers jours. Pour le candidat Donald Trump, nous remarquons que les scores de sentiments sont assez variables selon les états. En revanche, il y a très peu de différences entre les 14 premiers et 14 derniers jours.

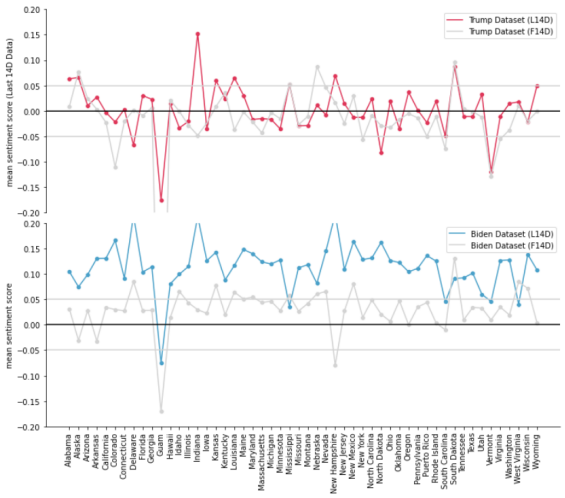

Figure 9 : Moyenne des scores pour chaque candidat, pour les 14 premiers jours (en gris) et pour les 14 derniers jours (en bleu pour Joe Biden et en rouge pour Donald Trump), pour chaque état des Etats-Unis

Nous remarquons facilement qu'il est possible d'obtenir plus de détails et d'informations par rapport aux sentiments des tweets en regardant les localisations et la temporalité.

VI. Analyse des résultats et tests de différentes méthodes

Pour juger plus précisément notre méthode de classification des tweets, il nous a semblé intéressant de mettre en place plusieurs tests, utilisant d'autres algorithmes dont certains énoncés dans les articles que nous avons utilisés pour notre recherche.

1. Méthode d'Elbow pour déterminer le nombre de clusters

Le nombre de cluster était déjà déterminé pour notre analyse, mais nous voulions voir si l'algorithme pouvais détecter un nombre optimal de clusters, c'est à dire voir si il était possible d'avoir des clusters pour des opinions plus précises allant par exemple de "très négatif" à "très positif" en passant par "neutre" et "moyennement positif" et "moyennement négatif".

En expérimentant avec la méthode d'Elbow, qui consiste à voir où un "coude" apparaît sur le graphe pour déterminer le nombre de clusters pertinents nous avons trouvé qu'il était judicieux d'utiliser uniquement 2 clusters, comme on peut le voir sur la figure 10 ci-dessous.

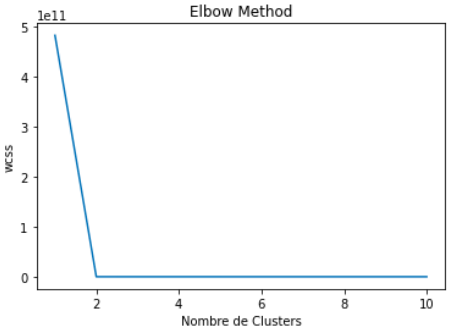

Figure 10 : détermination du nombre de clusters à utiliser pour le clustering K-means.

2. Clustering simple

Avant de se lancer sur le clustering K-means, nous avons choisi d'essayer un clustering "simple" pour voir la proportion de tweets positifs et négatifs pour le candidat Donald Trump. Pour ce faire, nous avons créé deux classes d'appartenance : une pour les tweets positifs et une pour les tweets négatifs.
Nous avions au préalable instancié deux tableaux pour chacun des types de tweets, et, en fonction du pos_score (score de positivité d'un tweet) et du neg_score (score de négativité d'un tweet), nous les avons placés dans le tableau posi[] ou negat[].
Enfin, nous avons créé des *dataframes* pour chacun des tableaux pour les comparer et mesurer leur volume.

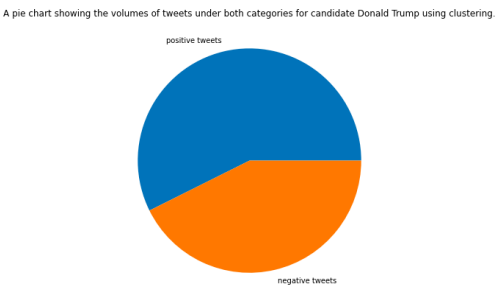

figure 11. clustering de données sur l'ensemble des tweets sur Donald Trump

Comme il est remarquable sur la figure ci-dessus, la proportion de tweets considérés comme négatifs et

positifs correspond à ce que l'on peut examiner lors de la classification plus haut, en utilisant un algorithme d'analyse des sentiments.

3. Clustering K-Means

Pour aller plus loin avec notre analyse et classification, nous avons tenté d'expérimenter un algorithme de classification K Means pour l'analyse des sentiments en utilisant les clusters déjà établis précédemment.

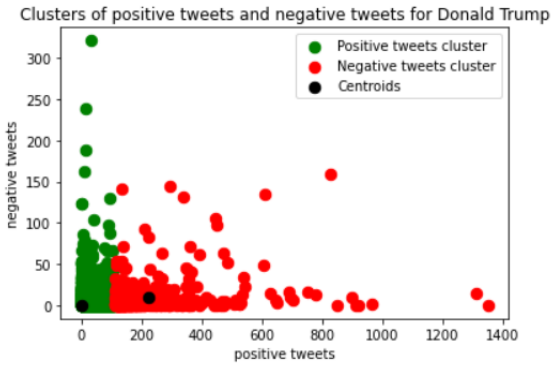

Figure 12 : Algorithme de classification K-Means pour l'analyse des sentiments des tweets à propos de Donald Trump

Le graphique ci-dessus montre que les utilisateurs se découpent en deux classes distinctes qui peuvent parfois être très proches en raison de la nature textuelle des données à analyser. La frontière semble très fine entre positif et négatif.

Conclusion

Le problème que nous nous étions posés en début d'étude était le suivant avait plusieurs facettes : nous nous demandions d'une part quelles étaient les meilleurs moyens de classifier les sentiments des utilisateurs sur un site de micro-blogging comme twitter, mais également si cette classification serait comparable à la réalité.
Nous avons pu tester plusieurs méthodes de classification : clustering, k-means clustering ou encore Sentiment Intensity Analyzer pour pouvoir les comparer.

Il a été important de prendre plusieurs éléments en compte comme le volume des données, le bruit, les différentes langues et localisation ainsi que le formatage des données afin d'obtenir des conteneurs de données corrects.
La solution que nous proposons répond au problème suivant : il est bel et bien pertinent d'effectuer des analyses globales s'appliquant sur le monde entier, et il existe des moyens plus ou moins intéressants de mettre ceci en place. La plupart des études mises en place jusqu'à présent s'intéressent à un pays, une région ou encore un continent. Le challenge de cette étude était également d'essayer de traiter un nombre massif de données provenant de sources, individus, pays, utilisateurs différents.

Bibliographie

[1] T. Palpanas, S. Amer-Yahia et M. Tsytarau. Efficient Sentiment Correlation for Large-scale Demographics. 2014 http://helios.mi.parisdescartes.fr/~themisp/publications/sigmod13-correlations.pdf

[2] L. Singh, A. Mitra, S.Singh. Sentiment Analysis of Tweets using Heterogeneous Multi-layer Network Representation and Embedding. 2020 https://www.aclweb.org/anthology/2020.emnlp-main.718.pdf

[3]H. Bagheri, M.J Islam. Sentiment analysis of twitter data. 2017 https://arxiv.org/pdf/1711.10377.pdf

[4] W. Kouadri, M. Ouziri, S. Benbernou, I. Ben Amor, K. Echihabi, T. Palpanas, W. Kouadri. Quality of Sentiment Analysis Tools: The Reasons of Inconsistency. 2020 http://www.vldb.org/pvldb/vol14/p668-kouadri.pdf

[5] Intania Cahya Sari, Yova Ruldeviyani. Sentiment Analysis of the Covid-19 Virus Infection in Indonesian Public Transportation on Twitter Data: A Case Study of Commuter Line Passengers. 2020 https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9255531

[6] Michal Skuza, Andzzej Romanowski, Sentiment Analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction, 2015 https://annals-csis.org/proceedings/2015/pliks/230.pdf

[7] James Spencer, Gulden Uchyigit, Sentimentor : Sentiment Analysis of Twitter Data, 2012 http://ceur-ws.org/Vol-917/SDAD2012_6_Spencer.pdf

[8] Murphy Choy,Michelle L.F. Cheong,Ma Nang Laik,Koo Ping Shung, A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census, 2011 https://arxiv.org/pdf/1108.5520.pdf

[9] Preslav Nakov, Semantic Sentiment Analysis of Twitter Data ,2017 https://www.researchgate.net/publication/318166816_Semantic_Sentiment_Analysis_of_Twitter_Data

[10] Pavlos Paraskevopoulos, Giovanni Pellegrini, Themis Palpanas, TweeLoc: A System for Geolocalizing Tweets at Fine-Grain,http://helios.mi.parisdescartes.fr/~themisp/tweeloc/