Rapport du projet

Analyse des sentiments des tweets relatifs aux prétendants à l'élection présidentielle américaine de 2020

RODRIGUES Mathieu, FANG Zicheng, NAJJAR Ahmad et ANDJELOVIC Lazar

Contents

1 Introduction

Le processus électoral est crucial pour l'ensemble des pays. Les élections présidentielles revêtent une importance particulière, car elles permettent de choisir le dirigeant qui gouvernera le pays pour les prochaines années. À l'heure actuelle, les réseaux sociaux ont une place prépondérante dans notre vie quotidienne et jouent également un rôle de plus en plus important dans les élections présidentielles. Dans ce projet, nous souhaitons examiner si une analyse des sentiments des tweets relatifs aux prétendants à l'élection permettrait d'avoir une idée représentative du résultat final de l'élection. Cette analyse serait particulièrement intéressante, car elle permettrait de comprendre l'opinion publique et ouvrirait de nouvelles perspectives à la science informatique. Les sentiments sont des notions difficiles à appréhender dans un contexte général. Ils varient selon les individus, les cultures, les contextes sociaux, économiques et politiques. De plus, les sentiments peuvent être influencés par des biais cognitifs, des préjugés ou des stéréotypes. Dans le cadre d'une élection présidentielle, l'analyse des sentiments doit donc prendre en compte ces facteurs pour produire des résultats fiables et pertinents. Cette étude permettra également, en tant qu'objectif secondaire, de vérifier s'il existe une corrélation entre les discours sur les réseaux sociaux et les résultats réels des élections présidentielles. Cette corrélation serait intéressante à étudier, car elle permettrait de mesurer l'impact des réseaux sociaux sur l'opinion publique et sur les choix électoraux. Dans notre projet, nous nous sommes intéressés à l'élection présidentielle de 2020 qui opposait Donald Trump à Joe Biden.

2 Travaux connexes

1) "Political Communities on Twitter: Case Study of the 2022 French Presidential Election" (2022) [4]: Cet article explore l'utilisation croissante des médias sociaux dans les campagnes politiques, en se concentrant sur l'élection présidentielle française de 2022. Les plateformes telles que Twitter et Facebook sont devenues des outils importants pour les candidats et les partis politiques pour interagir avec les électeurs et diffuser leurs idées. Les auteurs ont créé un ensemble de données Twitter contenant 1,2 million d'utilisateurs et 62,6 millions de tweets liés à l'élection. Ils ont utilisé des méthodes de détection de communauté pour identifier les groupes de soutien de chaque candidat et ont analysé leur position respective. Enfin, ils ont examiné les tweets offensants et les bots automatiques pour mieux comprendre la stratégie de campagne en ligne de chaque communauté.

2) "Sentiment analysis on twitter data" (2015) [11]: Cet article discute de l'utilisation des réseaux sociaux, en particulier Twitter, pour extraire le sentiment des utilisateurs. Les auteurs expliquent comment l'analyse des données de Twitter peut être utilisée pour déterminer si les tweets sont positifs, négatifs ou neutres. Cela est utile pour les entreprises qui cherchent à évaluer les commentaires sur leurs produits, ou pour les clients qui cherchent à connaître l'opinion des autres avant d'acheter un produit. Les auteurs utilisent des algorithmes d'apprentissage automatique pour classifier le sentiment des tweets en utilisant des données d'entraînement contenant des émoticônes et des acronymes comme étiquettes bruyantes.

3) "Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis" (2018) [6]: Les données Twitter sont utilisées pour prédire les élections présidentielles dans plusieurs pays, y compris en Indonésie. Les auteurs ont utilisé les tweets des candidats à la présidence et les hashtags pertinents pour l'analyse de sentiment et ont élaboré un algorithme pour prédire la polarité du sentiment. Les résultats ont montré que Jokowi est en tête des prévisions actuelles d'élection, ce qui correspond aux résultats de prévision de quatre instituts de sondage en Indonésie. Le Big Data, y compris les réseaux sociaux tels que Twitter, peut être une source précieuse de données pour prédire les résultats électoraux.

4) "Study of Twitter sentiment analysis using machine learning algorithms on Python" (2017)[8]: L'analyse de sentiment sur Twitter est une approche pour extraire les sentiments exprimés par les utilisateurs de Twitter sur différentes occasions. Le format difficile des tweets a rendu le traitement difficile, ce qui a conduit à une augmentation de la recherche dans ce domaine au cours des dernières décennies. Cet article passe en revue quelques articles de recherche sur l'analyse de sentiment sur Twitter, décrivant les méthodologies et modèles utilisés, ainsi qu'une approche généralisée basée sur Python. Twitter reste une plateforme importante pour l'analyse de sentiment en raison de son utilisation généralisée par les utilisateurs pour exprimer leurs opinions et afficher leurs sentiments.

5) "A review of feature extraction in sentiment analysis" (2014)[5]: L'analyse de sentiment vise à déterminer ce que les autres pensent et commentent sur des produits, des services, des politiques et la politique, en utilisant le contenu généré par le public sur des sites d'avis en ligne et des médias sociaux. Les caractéristiques ou aspects du produit ont un rôle important dans l'analyse de sentiment, et l'extraction de caractéristiques devient un domaine actif de recherche. Cet article de revue examine les techniques et approches existantes pour l'extraction de caractéristiques dans l'analyse de sentiment et l'exploration d'opinions, en utilisant un processus de revue systématique de la littérature pour identifier les domaines bien ciblés par les chercheurs et les zones les moins traitées, offrant ainsi des opportunités de recherche pour l'avenir.

6) "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle" (2012)[13]: Cet article présente un système d'analyse en temps réel des sentiments exprimés sur Twitter à l'égard des candidats à l'élection présidentielle américaine de 2012. Twitter est devenu une plateforme centrale pour exprimer des opinions et des points de vue sur les candidats et les partis politiques. L'analyse de sentiments peut aider à explorer la manière dont les événements électoraux affectent l'opinion publique. Le système offre des résultats instantanés et continus, offrant ainsi une nouvelle perspective opportune sur la dynamique du processus électoral et de l'opinion publique pour le public, les médias, les politiciens et les chercheurs.

7) "A Bi-level assessment of Twitter in predicting the results of an election: Delhi Assembly Elections 2020" (2022)[12]: Cet article étudie l'utilisation de Twitter pour évaluer les résultats des élections de l'Assemblée de Delhi en 2020. Les auteurs analysent la corrélation entre les activités des différents candidats et partis sur Twitter, les mentions et les sentiments des électeurs envers un parti, et les résultats électoraux réels. Ils constatent que le nombre d'abonnés et les réponses aux tweets des candidats sont des indicateurs utiles pour prédire les résultats des élections. Cependant, le nombre de tweets mentionnant un parti et le sentiment des électeurs envers ce parti ne sont pas alignés sur les résultats électoraux. Les auteurs utilisent également des modèles d'apprentissage automatique pour prédire les résultats des élections, avec des résultats prometteurs en utilisant les fonctionnalités de tweets.

8) "A Semi-Supervised Approach to Sentiment Analysis of Tweets during the 2022 Philippine Presidential Election" (2022)[10]: Cet article présente une méthode semi-supervisée pour l'analyse de sentiment de tweets en anglais et en tagalog. Les auteurs ont utilisé des tweets de la campagne présidentielle des Philippines pour entraîner un classificateur de base en utilisant des techniques de traitement du langage naturel. Les tweets ont été annotés en trois polarités: positif, neutre et négatif, et ont été auto-entraînés avec un classificateur Multinomial Naïve Bayes avec 30% de données non étiquetées. Les résultats ont montré une précision de 84,83%, ce qui dépasse les autres études utilisant des données Twitter des Philippines.

9) "Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections" (2018)[9]: Cet article décrit un système d'analyse en temps réel des sentiments exprimés sur Twitter envers les candidats à l'élection présidentielle américaine de 2012. Twitter est devenu un site central pour exprimer les opinions politiques et les événements émergents sont souvent suivis par une explosion de tweets, offrant une opportunité unique pour mesurer la relation entre le sentiment public et les événements électoraux. Le système offre une analyse instantanée et continue du sentiment de l'ensemble du trafic Twitter relatif à l'élection, offrant une nouvelle perspective opportune sur la dynamique du processus électoral et de l'opinion publique pour le public, les médias, les politiciens et les chercheurs.

10) "UK general election 2017: A Twitter analysis" (2017)[7]: Ce rapport analyse les publications sur les réseaux sociaux liées à l'élection générale britannique de 2017. Il montre que le sentiment en faveur du Parti travailliste domine la conversation sur Twitter et que le Parti national écossais est également bien représenté. La question du Brexit est populaire dans les hashtags et le débat sur Twitter est influencé par des événements externes et les médias. Les médias sociaux ont prédit un changement d'opinion publique

plus tôt que les sources de médias traditionnels.

11) "Location-based sentiment analyses and visualization of Twitter election data" (2020)[14]: Cet article utilise des analyses de sentiment basées sur les données de localisation de Twitter pour étudier les élections présidentielles américaines de 2016 et les élections générales britanniques de 2017. Ils découvrent que le sentiment des tweets basé sur la localisation est cohérent avec les résultats des élections dans les deux cas, et qu'il y a des tendances similaires dans le sentiment envers les candidats et les partis politiques indépendamment de la méthodologie adoptée pour la collecte des données.

Ces articles présentent une similarité dans leur domaine d'application qui est l'analyse des sentiments à partir des tweets sur Twitter. Ils ont tous recours à des techniques de traitement de langage naturel pour extraire et classifier le sentiment véhiculé dans les tweets. Cependant, ils diffèrent par leur contexte d'application, leur langue, leur méthode de classification et leurs résultats. Par exemple, certains articles se concentrent sur l'élection présidentielle dans un pays particulier tandis que d'autres s'intéressent aux élections législatives. Certains articles utilisent des méthodes de classification supervisées tandis que d'autres recourent à des méthodes semi-supervisées. Enfin, certains articles montrent une forte corrélation entre les sentiments exprimés sur Twitter et les résultats électoraux, tandis que d'autres ne montrent pas de corrélation significative. Malgré ces différences, tous les articles démontrent la pertinence de l'analyse des sentiments sur Twitter pour évaluer l'opinion publique et prévoir les résultats électoraux.

3 Problématique

Dans ce projet, notre objectif principal était de résoudre le problème de l'analyse de l'opinion publique lors d'une élection présidentielle à travers l'analyse des sentiments exprimés sur le réseau social Twitter. L'importance des réseaux sociaux dans la vie quotidienne de la population et leur influence croissante dans les choix électoraux font d'eux un domaine de recherche privilégié pour étudier l'opinion publique.

Dans ce contexte, notre projet vise à répondre à la question suivante : est-il possible d'obtenir une estimation de l'opinion publique et des résultats des élections présidentielles à partir de l'analyse des sentiments des tweets relatifs aux prétendants à l'élection présidentielle de 2020 ? La question est importante, car les choix électoraux ont un impact majeur sur la vie politique, économique et sociale du pays. Par conséquent, notre projet a cherché à développer une méthodologie d'analyse des sentiments qui prenait en compte les spécificités de l'élection présidentielle américaine de 2020.

Pour atteindre notre objectif, nous avons choisi de nous concentrer sur l'élection présidentielle de 2020, qui a opposé le président sortant, Donald Trump, au candidat démocrate Joe Biden.

4 Expériences et solutions

4.1 Dataset

Le dataset (Figure 1) dont nous disposons comporte une multitude d'informations relatives à des tweets publiés entre le 15 octobre 2020 et le 6 novembre 2020. Les colonnes du dataset contiennent des données telles que la date et l'heure de création du tweet, l'identifiant unique du tweet, le texte intégral du tweet, le nombre de likes et de retweets, ainsi que l'outil utilisé pour publier le tweet. De plus, les informations relatives aux utilisateurs sont également incluses, telles que l'identifiant de l'utilisateur, le nom d'utilisateur, le nom d'écran, une brève description personnelle, la date d'adhésion, le nombre de followers, la localisation donnée sur le profil de l'utilisateur, la latitude et la longitude de cette localisation, ainsi que des informations sur la ville, le pays, l'État et le code d'État.

Ces données ont été collectées à des intervalles irréguliers allant de une à plusieurs secondes à l'aide de Snsscrape et de l'API Twitter. Le dataset comprend des tweets associés aux hashtags #DonaldTrump et #Trump pour les tweets du compte de Donald Trump, ainsi que des hashtags #JoeBiden et #Biden pour les tweets du compte de Joe Biden. Ces données ont été séparées en deux fichiers CSV différents. Il convient de souligner que les données ont été obtenues sur internet et que nous ne sommes pas les créateurs originaux de ces données. En effet, nous avons réussi à obtenir cette énorme quantité d'information grâce à une personne ayant travailler sur une autre analyse et qui a généreusement partage ce dataset en ligne.

	tweet_id	likes	retweet_count	user_id	user_followers_count	lat	long	likes_norm	retweet_norm
count	4.998280e+05	499828.000000	499828.000000	4.998280e+05	4.998280e+05	234867.000000	234867.000000	499828.000000	499828.000000
mean	1.321588e+18	9.401198	2.294263	4.593279e+17	2.684420e+04	35.584230	-50.817607	0.318651	0.069670
std	2.436775e+15	447.252176	111.271005	5.594152e+17	4.018437e+05	17.014470	65.128743	13.438163	2.148067
min	1.316529e+18	0.000000	0.000000	2.654000e+03	0.000000e+00	-79.406307	-161.755833	0.000000	0.000000
25%	1.319467e+18	0.000000	0.000000	2.241751e+08	7.900000e+01	32.776272	-99.512099	0.000000	0.000000
50%	1.322267e+18	0.000000	0.000000	2.540326e+09	4.710000e+02	39.783730	-77.036558	0.000000	0.000000
75%	1.323860e+18	2.000000	0.000000	1.099413e+18	2.285000e+03	43.120257	1.888334	0.050750	0.000000
max	1.324502e+18	165702.000000	63473.000000	1.324492e+18	8.241710e+07	90.000000	179.048837	5475.323927	605.701123

Figure 1: DataSet (Biden)

4.2 Cloud

Le stockage de données en nuage est une solution populaire pour stocker les données de tweets collectées en raison de sa flexibilité et de son évolutivité. Les services de stockage en nuage tels qu'Amazon S3, Google Cloud Storage et Microsoft Azure Storage offrent des options de stockage économiques et à grande échelle pour les données de tweets.

Ces services de stockage en nuage offrent des fonctionnalités telles que la sécurité, la disponibilité, la redondance et la sauvegarde automatique, ce qui garantit la sécurité et la disponibilité de vos données. De plus, ils offrent une évolutivité horizontale, ce qui signifie que vous pouvez augmenter ou réduire facilement la capacité de stockage en fonction de vos besoins.

Voici quelques-uns des avantages du stockage de données en nuage pour les données de tweets:

Coût: Les services de stockage en nuage sont souvent plus économiques que les solutions de stockage traditionnelles car ils ne nécessitent pas d'investissement initial important pour l'achat de matériel.

Flexibilité: Les services de stockage en nuage offrent une grande flexibilité, vous permettant de stocker des données de manière temporaire ou permanente et de les accéder de n'importe où dans le monde.

Évolutivité: Les services de stockage en nuage peuvent facilement évoluer pour répondre aux besoins en matière de stockage des données de tweets en constante évolution.

Sécurité: Les services de stockage en nuage offrent des mesures de sécurité avancées telles que la cryptographie et les contrôles d'accès pour protéger vos données contre les menaces de sécurité.

Bien que nous ayons considéré l'utilisation de Microsoft Azure Storage pour stocker nos données de tweets, nous avons finalement décidé de ne pas mettre en place cette solution en raison de ses tarifs. Cependant, nous reconnaissons que le stockage en cloud avec Microsoft Azure Storage est une solution efficace pour les data scientists qui travaillent avec de grandes quantités de données de tweets. Si nous avions décidé d'utiliser Microsoft Azure Storage pour stocker nos données de tweets, nous aurions bénéficié de plusieurs avantages en termes de performance et de praticité d'utilisation.

Tout d'abord, Azure Storage est une solution qui permet aux utilisateurs de stocker et de gérer des volumes de données importants. La plateforme est capable de gérer des quantités massives de données de tweets, ce qui nous aurait permis de stocker toutes nos données de manière centralisée et de les rendre facilement accessibles à notre équipe de data scientists.

De plus, Azure Storage stocke les données sur des serveurs partout dans le monde, ce qui signifie que les données auraient été stockées sur plusieurs centres de données pour garantir une disponibilité maximale en cas de défaillance d'un centre de données.

En termes de praticité, Azure Storage est, en théorie, facile à utiliser et à configurer, avec une interface intuitive et des outils de gestion de données.

En fin de compte, l'utilisation de Microsoft Azure Storage aurait pu être une solution pratique et performante pour stocker nos données de tweets, mis le problème tarifaire de côté.

4.3 Descriptions et Explications des résultats

Le graphique présenté (Figure 2) est une visualisation informative des valeurs nulles dans les datasets de tweets concernant les candidats à la présidence américaine, Joe Biden et Donald Trump. Le graphique a été généré à partir d'un code Python qui a créé deux DataFrames, chacun contenant le nombre de valeurs nulles pour chaque variable dans les datasets respectifs. Les lignes présentant des valeurs nulles ont ensuite été sélectionnées pour être représentées dans le graphique.

Le graphique est composé de deux sous-graphiques qui permettent une comparaison directe entre les datasets de Biden et de Trump. Les axes des sous-graphiques présentent le nombre de valeurs nulles sur l'axe y et les noms des variables sur l'axe x. Les barres dans le graphique sont colorées en bleu pour le dataset de Biden et en rouge pour celui de Trump, afin de faciliter la distinction entre les deux, ce choix de couleurs va être utiliser tout au long de l'exploration du dataset.

Il est intéressant de noter que les variables "Source" et "Username" ne présentent pas de valeurs nulles dans les datasets de tweets sur Joe Biden et Donald Trump, ce qui peut être considéré comme un point positif pour l'analyse ultérieure de ces données.

Par ailleurs, en examinant le graphique représentant les valeurs nulles pour les autres variables dans les deux datasets, on constate des similitudes dans les données manquantes. En effet, plusieurs variables présentent un nombre important de valeurs nulles dans les deux datasets, telles que "user_description", "user_location", "geo" et "place". Cette observation peut suggérer que ces variables ne sont pas collectées de manière systématique ou que les utilisateurs choisissent de ne pas les renseigner lorsqu'ils publient des tweets.

Cependant, il convient de noter que l'absence de données dans ces variables ne signifie pas nécessairement qu'elles ne sont pas pertinentes pour l'analyse. En effet, certaines variables telles que "user_description" et "user_location" peuvent fournir des informations importantes sur l'auteur du tweet et sur le contexte de la publication, ce qui peut influencer la compréhension des données. Par conséquent, il est important de prendre en compte les valeurs manquantes lors de l'analyse des données pour s'assurer que les résultats obtenus sont cohérents et fiables.

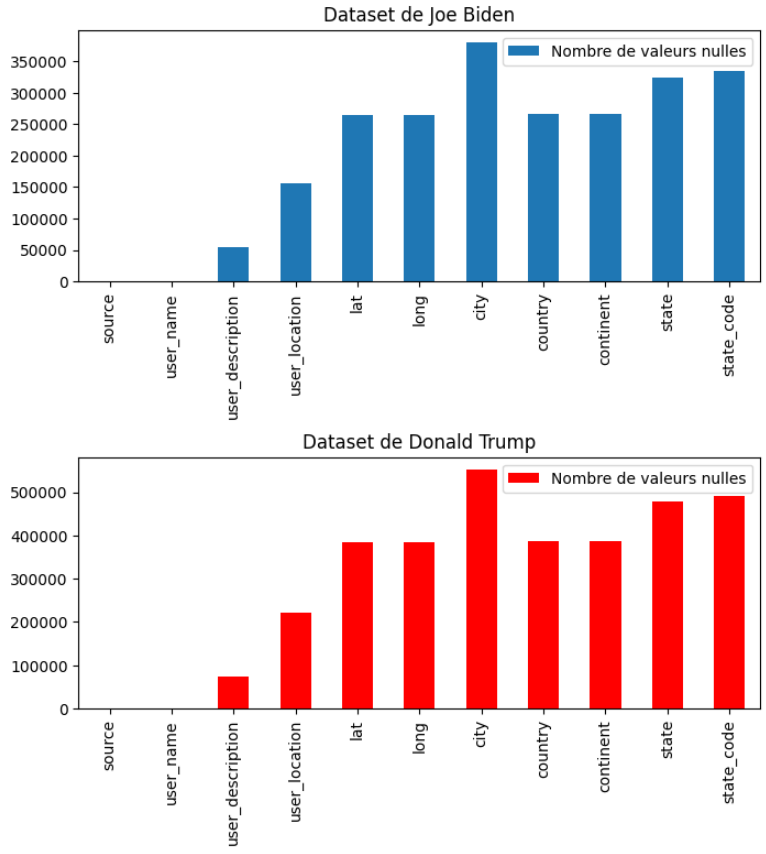

Figure 2: Représentation des valeurs nulles du DataSet

Nous avons ensuite représenté le nombre de likes (Figure 3) et de retweets (Figure 5) par tweet pour les comptes Twitter de Joe Biden et de Donald Trump. Le premier sous-graphique représente les données de "likes" pour les tweets de Joe Biden tandis que le deuxième sous-graphique représente les données de "likes" pour les tweets de Donald Trump. La plage de valeurs pour les axes des ordonnées et des abscisses des deux graphiques sont également définie de manière à garantir une comparaison pertinente des données pour les deux comptes. Pour ce qui est des deux autres sous-graphes on a le troisième sous-graphique de la figure qui représente les données de "retweets" pour les tweets de Joe Biden et le quatrième sous-graphique représente les données de "retweets" pour les tweets de Donald Trump.

En observant attentivement la visualisation produite par le code, on peut immédiatement remarquer que les tweets concernant Joe Biden ont tendance à avoir un nombre de likes et retweet plus élevé que ceux de Donald Trump. En effet, le premier sous-graphique de la figure montre une plus grande dispersion des valeurs de "likes" pour les tweets de Joe Biden, avec plusieurs tweets ayant un nombre de likes et retweet très élevé. Cela peut être interprété comme une indication que les tweets de Joe Biden ont une plus grande probabilité de devenir viraux et de figurer en tête des tendances sur Twitter.

En revanche, les tweets concernant Donald Trump présentent une répartition des valeurs de "likes" et "retweets" plus resserrée, avec une plage de valeurs de "likes" et "retweets" moins importantes. Cela suggère que les tweets de Donald Trump ont moins de chances de devenir viraux et d'attirer l'attention des utilisateurs de Twitter.

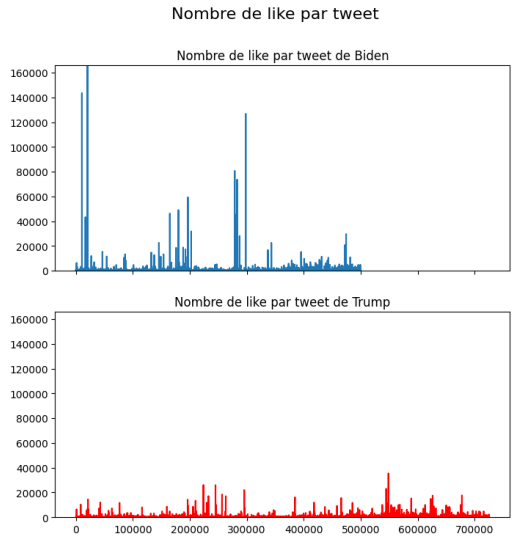

Figure 3: Nombre de likes par tweet

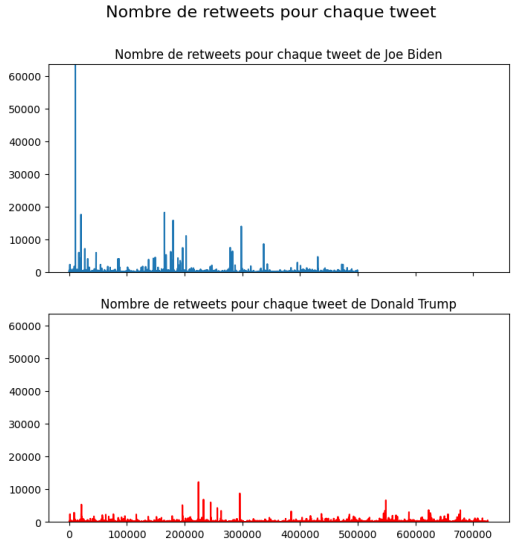

Figure 4: Nombre de retweets par tweet

La somme cumulée des "likes" (Figure 5) ou des "retweets" (Figure 6) vient confirmer la tendance générale observée précédemment, selon laquelle les tweets de Joe Biden ont tendance à être plus populaires que ceux de Donald Trump. En effet, on peut voir que la courbe bleue, représentant la somme cumulée des "likes" ou des "retweets" pour les tweets de Joe Biden, a une pente plus élevée que la courbe rouge, représentant la somme cumulée des "likes" ou des "retweets" pour les tweets de Donald Trump.

Il est également intéressant de noter que les valeurs de la somme cumulée des "likes" ou des "retweets" pour chaque compte augmentent de manière significative à la fin des graphiques. Cela est sûrement dû au fait que la date de l'élection présidentielle américaine de 2020 qui se rapprochait.

En somme, on peut emmètre l'hypothèse que ces graphes suggèrent une certaine loyauté envers Joe Biden, avec des tweets le concernant ayant tendance à générer plus d'engagement que ceux concernant Donald Trump. Bien que le nombre de tweets sur Donald Trump soit plus élevé dans le jeu de données, cela ne semble pas être un facteur décisif pour la popularité sur les réseaux sociaux. On peut donc en déduire que Biden a un soutien fidèle de ses partisans, qui rivalise avec celui de Trump malgré une présence en nombre de tweets plus faible.

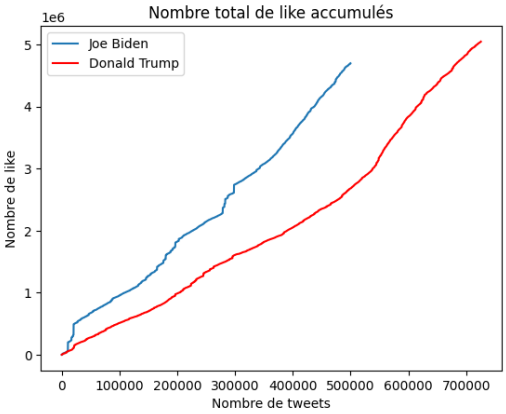

Figure 5: Likes accumulés

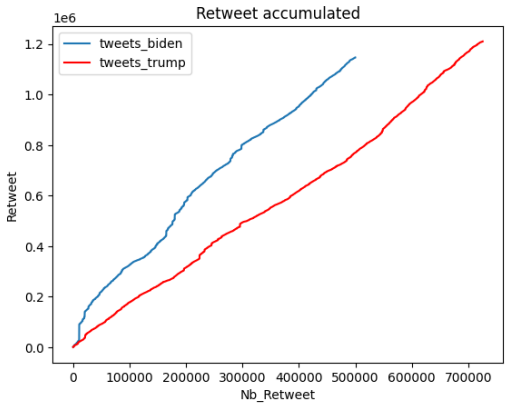

Figure 6: Retweets accumulés

Nous avons observé dans notre ensemble de données que les tweets provenaient de différents pays. Par conséquent, nous avons choisi de visualiser la provenance des tweets des 10 pays les plus fréquents en représentant cela dans un graphique avec en abscisse ces pays et en ordonnée le nombre de tweets (Figure 7). La couleur verte du graphique est utilisé pour les graphiques génériques.

Nous avons remarqué que les États-Unis occupaient une place importante, ce qui est logique car les élections se déroulent chez eux. Cependant, l'Angleterre, l'Allemagne, la France, l'Inde et le Canada ont presque le même nombre de tweets. Cela indique que ces pays sont très intéressés par les élections présidentielles américaines, soit en raison d'une grande diaspora américaine dans leur pays, soit parce que les gens de ces pays sont intéressés par les élections américaines.

À ce stade, nous avons décidé de travailler uniquement sur les tweets provenant des États-Unis, car les tweets des autres pays ne sont pas dans la même langue, ce qui compliquerait l'utilisation de VADER pour l'analyse des sentiments.

Nous avons également examiné les types d'appareils utilisés par les utilisateurs pour publier des tweets. Notre graphique présente les différents types d'appareils en abscisse et le nombre de tweets publiés par ces appareils en ordonnée (Figure 8). Nous avons constaté que les tweets ont été majoritairement publiés à partir d'un ordinateur, suivi de l'iPhone, puis d'Android. Ces trois appareils sont les plus populaires dans le monde et donc les plus utilisés pour publier des tweets.

Il est important de noter que l'utilisation d'appareils mobiles pour publier des tweets est en constante augmentation, notamment grâce à la popularité croissante des smartphones. Il est donc probable que la part de tweets publiés à partir d'appareils mobiles augmentera dans les années à venir.

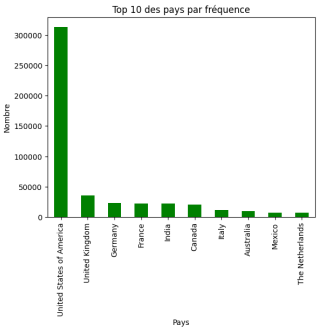

Figure 7: Les top 10 des pays par nombre de tweets

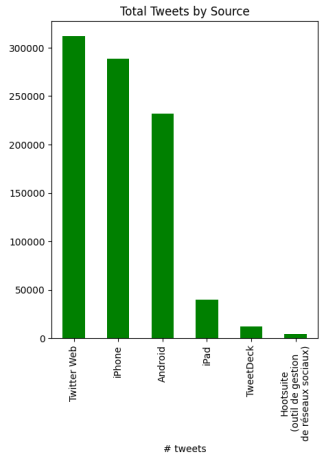

Figure 8: Nombre des tweets par source

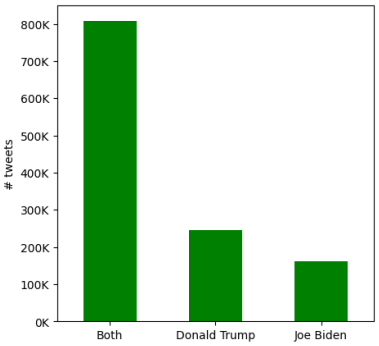

Figure 9: La repartition des mentions des candidats dans les tweets.

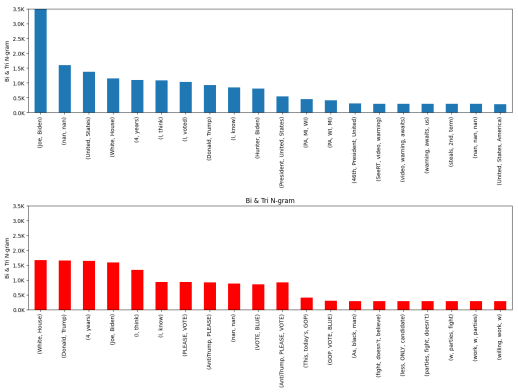

Figure 10: Les bi et tri-grammes

Le graphique que nous avons présenté illustre la répartition des mentions des candidats dans les tweets (Figure 9), à la fois séparément et ensemble. Il montre clairement que les deux candidats sont souvent mentionnés ensemble dans les tweets, tandis que Donald Trump est le candidat le plus mentionné individuellement.

Cette tendance peut être interprétée comme un signe que les utilisateurs comparent souvent les deux candidats dans leurs tweets. Il est également intéressant de noter que la forte prévalence de mentions de Donald Trump dans les tweets individuels peut refléter une focalisation de l'attention sur lui en tant que personnalité publique controversée.

Nous nous sommes intéressés aux termes les plus représentatifs des tweets collectés en fonction des candidats, en utilisant des bi et tri-grammes pour analyser les tweets de Joe Biden et de Donald Trump. Le graphique comporte deux sous-graphiques, l'un pour les tweets de Joe Biden et l'autre pour les tweets de Donald Trump (Figure 10). Les bi et tri-grammes les plus fréquents sont représentés par des barres plus hautes.

L'analyse des termes les plus fréquents dans les tweets concernant Joe Biden et Donald Trump révèle des différences notables. Pour Joe Biden, les termes les plus courants comprennent "Joe Biden", "United States", "White House", "4 years", "I think", "I voted", "Donald Trump" et "I know". Ces termes reflètent les préoccupations liées à sa candidature à la présidence, notamment sa politique, son programme et sa rivalité avec Donald Trump. En revanche, les termes les plus fréquents dans les tweets concernant Donald Trump incluent "White House", "Donald Trump", "4 years", "Joe Biden", "i think", "i know", "PLEASE VOTE", "AntiTrump", "PLEASE" et "VOTE BLUE". Ces termes révèlent une préoccupation particulière pour la campagne électorale en cours, y compris l'appel au vote, la mobilisation contre Trump, mais aussi la mention fréquente du nom de Joe Biden, presque aussi récurrent que le nom de Trump.

Après avoir parcouru toutes ces données, nous pouvons avoir une idée préliminaire de l'opinion publique avant même de commencer l'analyse de sentiment. Nous pouvons conclure que les tweets concernant Biden ont moins de volume, mais ont tendance à devenir viraux contrairement à ceux de Trump. De plus, les termes les plus fréquemment utilisés dans les tweets concernant Biden sont généralement bienveillants, tandis que ceux pour Trump sont plutôt mitigés. Cela pourrait indiquer que Biden bénéficie d'une image plus positive et que Trump doit faire face à une opinion publique plus divisée. À ce stade de l'étude une victoire du candidat Joe Biden fait sens. On peut en plus s'appuyer sur les données "réel" que sont les sondages nationaux, datant d'avant les résultats, afin de faire une corrélation avec nos observations. Ils ont montré que Biden était en tête, mais la course était serrée dans certains États clés. De plus, en raison du vote indirect, le candidat avec le plus de voix populaires ne gagne pas automatiquement l'élection présidentielle, ce qui ajoute une incertitude. Cependant, pour avoir une analyse plus approfondie, nous devons plonger dans l'analyse de sentiment pour obtenir une image plus précise et complète de l'opinion publique.

4.4 VADER-Analyse des sentiments

Dans notre DataFrame, nous avons choisi de récupérer uniquement les tweets des utilisateurs américains. Cela s'explique par le fait que nous souhaitons nous concentrer sur l'opinion américaine et uniquement les tweets en anglais car c'est une élection américaine.

Après avoir trié les tweets, nous avons entamé la partie nettoyage des données. Nous avons constaté que les tweets contenaient non seulement du texte, mais également des hashtags, des liens web, des images, des émojis et une forte ponctuation. Il est possible que des citations et des bouts de texte non anglophone se glissent également dans les tweets. En outre,

la présence d'argot et de jargon a également nécessité notre attention. Contrairement à certaines études qui ont choisi d'ignorer ces spécificités et de se concentrer uniquement sur le texte, nous avons décidé de prendre en compte les émojis et l'argot, car ils peuvent fournir de nombreuses informations importantes.

Nous avons utilisé VADER, un outil d'analyse de sentiments spécialement conçu pour les messages sur les réseaux sociaux. VADER prend en compte la négation, l'usage excessif de ponctuation, l'argot, les émojis et les acronymes pour calculer un score de sentiment. Nous avons supprimé les hashtags, les liens et les caractères bruyants qui n'apportent rien de plus et n'influencent pas le calcul du score de sentiment. Ensuite, nous avons itéré sur chaque texte du DataFrame et avons appliqué le calcul du score de sentiment de VADER, puis nous les avons rangés dans une liste. Les résultats ont été stockés sous la forme d'un dictionnaire tel que celui-ci : 'pos': 0.746, 'compound': 0.8316, 'neu': 0.254, 'neg': 0.0.

Le calcul du score de sentiment comprend quatre catégories, parmi lesquelles figure le "compound". Ce score est calculé par la somme du score de chaque mot dans le texte donné et normalisé entre -1 (le plus négatif) et 1 (le plus positif). Les trois autres catégories représentent le ratio de la proportion de texte considéré comme positif, négatif ou neutre. Ainsi, pour décider de la polarité d'un texte, nous nous sommes basés sur le score "compound". Si le score est supérieur à 0,05, le texte est considéré comme positif et nous avons représenté cela par 1. S'il est inférieur à -0,05, le texte est considéré comme négatif, représenté par -1. Sinon, le texte est considéré comme neutre, représenté par 0. Nous avons créé un tableau des scores associés à chaque message pour chaque texte analysé.

Algorithm 1: Analyse des sentiments avec VADER.

```
text1 ← Charger le jeu de données contenant les tweets de Biden
text2 ← Charger le jeu de données contenant les tweets de Trump
text1 ← text1 avec pays == 'Etats-Unis'
text2 ← text2 avec pays == 'Etats-Unis'
Pour text1 et text2 faire:
    text ← text nettoyé (caractères spéciaux, texte en minuscules, stopwords)
words1 ← liste des mots de text1
words2 ← liste des mots de text2
sentiment ← Instancier l'analyseur de sentiment VADER
sentiment_biden ← sentiment(words1)['compound']
sentiment_trump ← sentiment(words2)['compound']
sentiment_biden ← Seuillage des valeurs (-1, 0 et 1)
sentiment_trump ← Seuillage des valeurs (-1, 0 et 1)
```

Ici nous avons deux graphiques (Figure 11 et 12) représentant l'évolution du sentiment lié aux candidats Biden et Trump au fil des tweets . Chacun des graphiques est composé d'une courbe représentant la valeur de sentiment pour chaque tweet, ainsi qu'une courbe représentant la moyenne mobile de cette valeur sur 7 tweets consécutifs. Les axes y des deux graphiques ont été étiquetés avec des valeurs numériques correspondant à la polarité des sentiments (négatif, neutre, positif). De plus, les points de données de la courbe ont été marqués par des points individuels pour une meilleure visualisation de l'évolution du sentiment.

Lorsqu'on observe les graphiques présentés, il est clair que les sentiments exprimés envers Joe Biden et Donald Trump ont évolué de manière significative au fil des tweets. En ce qui concerne les tweets relatifs à Biden, on peut noter une certaine stabilité dans les sentiments exprimés avec une section très marquée par une forte positivité. À l'inverse, la courbe représentant les sentiments exprimés à l'égard de Trump est extrêmement chaotique et instable, avec des variations fréquentes et significatives. Les sentiments exprimés envers ces deux candidats sont donc très différents. Ces graphiques offrent un aperçu fascinant de la manière dont les opinions politiques et les sentiments peuvent évoluer et fluctuer sur les réseaux sociaux.

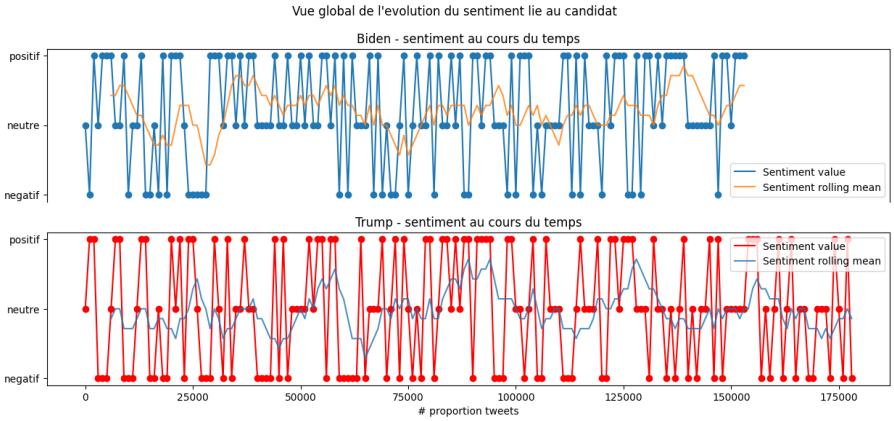

Figure 11: Analyse des sentiments.

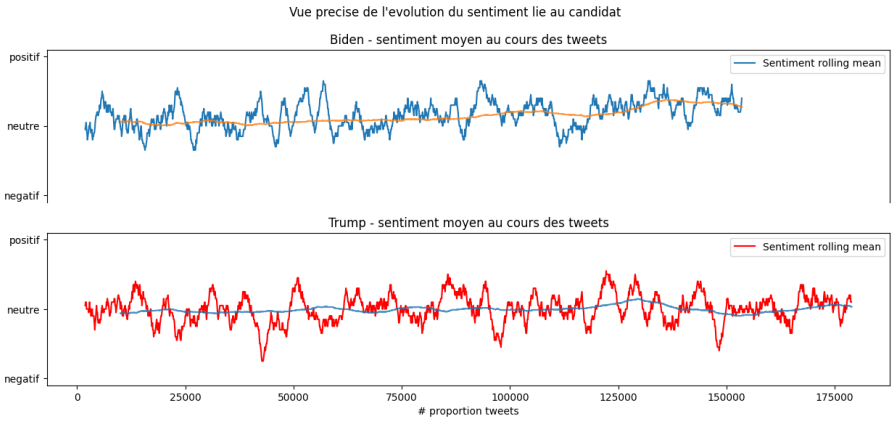

Figure 12: Analyse des sentiments précis.

5 Conclusion

En conclusion, notre étude suggère que Joe Biden avait une longueur d'avance sur Donald Trump lors de l'élection présidentielle de 2020. En effet, les indicateurs tels que la viralité des tweets, la loyauté de la communauté et la nature moins controversée de sa campagne ont clairement favorisé Biden. De plus, la mauvaise gestion de la crise de la COVID-19 par Trump au cours de son premier mandat a également contribué à une image négative de sa présidence. Bien que l'analyse de sentiment avec Vader ait permis de recueillir des informations utiles, elle comporte des limites telles que l'incapacité à détecter l'ironie dans les tweets. Cependant, nos résultats préliminaires sont en ligne avec les résultats finaux de l'élection, ce qui suggère que l'utilisation de l'analyse de sentiment peut fournir des informations précieuses pour comprendre l'opinion publique et aider à prédire les résultats des élections.

Une autres limite de notre analyse, est que nous nous sommes concentrés sur l'analyse de contenu partagé sur une seule plateforme, Twitter. Il est important de noter que Twitter n'est pas représentatif de l'ensemble de la population, car seulement 42% de ses utilisateurs américains ont entre 12 et 34 ans[1], tandis que la tranche d'âge des votants en 2020 était majoritairement composée de personnes âgées de 35 à 64 ans[2] . Par conséquent, notre analyse ne reflète pas nécessairement l'opinion publique dans son ensemble, mais plutôt celle des utilisateurs de Twitter. Une étude faite par "Statista Research Department" et intitulée "Exit polls of the 2020 Presidential Election in the United States on November 3, 2020, share of votes by age" montre que 62% des 18-29ans ont voté pour Joe Biden[3]. En ce qui concerne les autres tranches d'âge, les résultats sont plus serrés, ceci est bien en lien avec nos résultat obtenue en supposant que les 18-29 ans sont majoritaire sur Twitter. Malgré ces limites, notre étude a fourni des indications utiles sur l'opinion publique avant l'élection présidentielle de 2020, et notre analyse de sentiment avec Vader a montré des tendances cohérentes avec les résultats de l'élection.

References

[1] https://www.blogdumoderateur.com/chiffres-twitter/.

[2] https://www.census.gov/library/stories/2021/04/record-high-turnout-in-2020-general-election.html#.

[3] https://www.statista.com/statistics/1184426/presidential-election-exit-polls-share-votes-age-us/.

[4] Hadi Abdine, Yanzhu Guo, Virgile Rennard, and Michalis Vazirgiannis. Political communities on twitter: Case study of the 2022 french presidential election. *arXiv preprint arXiv:2204.07436*, 2022.

[5] Muhammad Zubair Asghar, Aurangzeb Khan, Shakeel Ahmad, and Fazal Masud Kundi. A review of feature extraction in sentiment analysis. *Journal of Basic and Applied Scientific Research*, 4(3):181–186, 2014.

[6] Widodo Budiharto and Meiliana Meiliana. Prediction and analysis of indonesia presidential election from twitter using sentiment analysis. *Journal of Big data*, 5(1):1–10, 2018.

[7] Laura Cram, Clare Llewellyn, Robin Hill, and Walid Magdy. Uk general election 2017: A twitter analysis. *arXiv preprint arXiv:1706.02271*, 2017.

[8] Bhumika Gupta, Monika Negi, Kanika Vishwakarma, Goldi Rawat, Priyanka Badhani, and B Tech. Study of twitter sentiment analysis using machine learning algorithms on python. *International Journal of Computer Applications*, 165(9):29–34, 2017.

[9] Ema Kušen and Mark Strembeck. Politics, sentiments, and misinformation: An analysis of the twitter discussion on the 2016 austrian presidential elections. *Online Social Networks and Media*, 5:37–50, 2018.

[10] Julio Jerison E Macrohon, Charlyn Nayve Villavicencio, X Alphonse Inbaraj, and Jyh-Horng Jeng. A semi-supervised approach to sentiment analysis of tweets during the 2022 philippine presidential election. *Information*, 13(10):484, 2022.

[11] Varsha Sahayak, Vijaya Shete, and Apashabi Pathan. Sentiment analysis on twitter data. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, 2(1):178–183, 2015.

[12] Maneet Singh, SRS Iyengar, Akrati Saxena, and Rishemjit Kaur. A bi-level assessment of twitter in predicting the results of an election: Delhi assembly elections 2020. *arXiv preprint arXiv:2204.08746*, 2022.

[13] Hao Wang, Doğan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 system demonstrations*, pages 115–120, 2012.

[14] Ussama Yaqub, Nitesh Sharma, Rachit Pabreja, Soon Ae Chun, Vijayalakshmi Atluri, and Jaideep Vaidya. Location-based sentiment analyses and visualization of twitter election data. *Digital Government: Research and Practice*, 1(2):1–19, 2020.