Projet final

Analyse du degré de polarisation des tweets de politiciens aux États-Unis

Laurent Alsène-Racicot, Juliette Beaulieu-Lépine et
Simon-Olivier Laperrière

Rapport présenté à
Guy Wolf

Le 30 avril 2021

Table des matières

1 Introduction

Avec la montée en popularité des réseaux sociaux, plusieurs chercheurs s'entendent pour dire que la société n'a jamais été aussi divisée. La polarisation des opinions est devenue un problème majeur. Constamment exposé à du contenu aligné à notre propre opinion, des bulles médiatiques se forment sur nos plateformes sociales, telles Instagram, Facebook et Twitter, et renforcent nos convictions.

L'année 2020 à été particulièrement chargée. Le nouveau Coronavirus a soulevé de nouveaux enjeux éthiques et sociaux et la polarisation s'est fait remarquée à de nombreux égards. Plusieurs personnes ont pris des positions radicales par rapport aux thèmes tels que l'accès aux soins de santé, le dépistage massif, et la vaccination. Les dirigeants des pays autour du monde ont pris des approches différentes pour faire face à cette crise.

Au milieu de tout ces débats, les États-Unis ont mené une campagne présidentielle. Les discours ont atteint une polarisation sans précédent qui a été particulièrement remarquée sur la plateforme Twitter, qui permet à ses utilisateurs d'échanger en 140 caractères. Le phénomène a atteint des proportions telles que la plateforme a jugé bon de bannir le compte de l'ex-président Donald Trump, lequel s'est vu accusé à maintes reprises de partager de fausses informations. Encore aujourd'hui, l'analyse de la polarisation dans les réseaux sociaux renseigne sur le climat politique et social.

2 Objectifs

Nous proposons de mesurer la polarisation présente dans des tweets de politiciens des États-Unis en les classifiant selon leur polarité : positif ou négatif. Il s'agit d'une tâche d'apprentissage supervisé et de traitement des langues naturelles. Différentes méthodes de classification sont utilisées pour analyser les données : classificateur de Bayes naïf, machine à vecteur de support, arbres de décision et réseau de neurones. Pour chacune de ces méthodes, les différents paramètres sont optimisés pour avoir une efficacité maximale. Il est ainsi possible de comparer les méthodes afin de déterminer laquelle est la plus appropriée pour cette tâche.

3 Description des données utilisées

3.1 Ensemble d'entraînement

Les différents modèles sont entraînés en utilisant un jeu de données de tweets étiquetés selon leur degré de polarisation. Plusieurs jeux de données de la sorte ont été produits dans le cadre de compétitions Kaggle et sont accessibles sur le web. Dans notre cas, nous utilisons la banque de données sentiment140 : `https://www.kaggle.com/kazanova/sentiment140`. Celle-ci contient 1.6 millions de tweets étiquetés en fonction de leur polarisation. L'étiquette 0 a été attribuée aux tweets négatifs et 4 aux tweets positifs. Afin de réduire la taille immense de cet ensemble de données, nous avons conservé 100 000 tweets sélectionnés de façon aléatoire.

3.2 Ensemble d'analyse

Nous testons les modèles entraînés sur un ensemble de tweets de politiciens. Plus particulièrement, nous avons sélectionné huit démocrates et huit républicains. Nous avons collecté leur tweets en ayant recours à la librairie `tweeply`, nous permettant de nous connecter à l'API twitter. À l'aide de celle-ci, nous recueillons au plus 200 tweets de la dernière semaine (limite imposée par Twitter) pour chaque politicien.

4 Méthodologie

4.1 Pré-traitement des tweets

Les tweets recueillis sont des chaînes contenant au plus 140 caractères. Nous les pré-traitons, d'une façon propre à une tâche de traitement des langues naturelles. Plus spécifiquement, nous commençons par utiliser une expression régulière (regex) afin d'extraire les mots contenus dans chaque tweets d'après un patron de recherche. En particulier, nous considérons les mots constitués exclusivement de caractères alphabétiques et de taille arbitraire. Ce choix d'expression régulière se base sur la prémisse que les caractères numériques, spéciaux ainsi que les hyperliens ne rajoutent pas d'information pertinente à la connotation du tweet. Par la suite, nous enlevons les mots d'arrêts (*stop words*), qui correspondent aux mots à usage courant au sein d'une langue naturelle de type adverbes, pronoms ou mots de liaisons.

Ces mots n'offrent pas de valeur ajoutée à la tâche de classification. Nous utilisons la liste de mots d'arrêts de l'anglais disponible dans la librairie `nltk`. À la fin de cette étape, nous obtenons donc une liste ne contenant que des mots significatifs pour chaque tweet.

4.2 Construction des vecteurs d'attributs

Les algorithmes utilisés pour classifier les tweets nécessitent des entrées de taille constante. Ainsi, nous construisons un vecteur d'attributs de taille fixe (3000 attributs) pour chacun des tweets. Nous utilisons trois types de vecteurs d'attributs, tous construits avec la librairie `scikit-learn`. En premier lieu, nous dénombrons le nombre d'occurrences des mots dans chaque tweets (*count vectorizer*). Puis, à partir du nombre d'occurrences, nous construisons un nouveau vecteur d'attributs correspondant à la fréquence des mots dans chaque tweet (*term-frequency*). Finalement, pour le dernier vecteur d'attribut, nous ajustons l'échelle du dernier vecteur de façon à ce que la fréquence des mots apparaissant dans plusieurs tweets de l'ensemble d'entraînement soit diminuée (*term fequency times inverse document frequency*). De cette façon, la pondération des mots jugés impertinents est moindre.

4.3 Algorithmes utilisés

Pour ce projet, quatre algorithmes de classification ont été testés et évalués en fonction de leur taux de précision. Le plus performant a par la suite été utilisé pour analyser les tweets des politiciens. Nous avons tout d'abord considéré des algorithmes simples, tels le classificateur de Bayes naïf, SVM et un arbre de décision. Ces méthodes ont l'avantage d'être en général peu coûteuses en temps (à l'exception de SVM) et offrent parfois des résultats de haute précision. Par la suite, nous avons utilisé une méthode plus complexe, soit un réseau de neurones perceptron multi-couches, afin de vérifier si une augmentation de capacité améliorerait nos résultats.

4.3.1 Classificateur "Naïf Bayes"

Le classificateur de Bayes naïf est un algorithme de classification probabiliste, qui se base sur l'hypothèse que les attributs des observations sont indépendants entre-eux. Il classifie un exemple (tweet) en retournant la classe (polarité) qui maximise la fonction de vraisemblance (*likelihood*), c'est-à-dire

la probabilité conditionnelle qu'un tweet donné appartienne à la dite classe étant donné ses caractéristiques. En vertu du théorème de Bayes, cette probabilité se décompose en plusieurs termes, soient la *prior* qui correspond à la probabilité d'appartenance à une classe à priori, puis la probabilité d'observer les caractéristiques étant donné l'appartenance à une certaine classe. En ajoutant l'hypothèse naïve que les caractéristiques sont indépendantes entreselles, cette probabilité conditionnelle peut à son tour être décomposée en plusieurs termes, soit les probabilités d'observer chaque caractéristique étant donné l'appartenance à une classe. L'algorithme approxime toutes ces valeurs en utilisant les fréquences d'appartenance dans l'ensemble d'entraînement.

4.3.2 SVM

L'idée centrale dernière l'algorithme de machine à vecteurs de support est de trouver le meilleur hyperplan qui divise un ensemble de données étiquetées en deux classes. Par meilleur, il est entendu que l'hyperplan propose la plus grande marge entre les données et lui-même. Pour ce faire, il est nécessaire de résoudre un système d'optimisation quadratique. Lorsque les données ne sont pas linéairement séparables, on peut utiliser une astuce de noyau pour représenter les données dans un espace plus grands dans lequel il est souhaité qu'elles deviennent linéairement séparables. Dans notre cas, nous avons tester trois différents noyau communs, soit le noyau linéaire, le noyau polynomial et le noyau RBF. Celui qui s'est révélé le plus performant pour nos données est le noyau RBF et il a donc été utilisé en conséquence.

4.3.3 Arbres de décision

L'arbre de décision est un algorithme qui effectue des coupes dans les données, attributs par attributs, et détermine ainsi récursivement des frontières de décision. Une fois la phase d'entraînement complétée, l'espace est divisée en différentes région et à chacune d'entre elles correspond à une classification des exemples qu'elle contient. Pour sélectionner quel attribut est optimal pour effectuer la prochaine coupe dans la phase d'entraînement, nous avons utilisé l'index de Gini comme mesure d'impureté.

4.3.4 Réseau de neurones perceptron multi-couches

Le réseau de neurones perceptron multi-couches (MLP) est en fait un simple réseau de neurones artificiels (ANN), utilisant des couches cachées

contenant un nombre de neurones pré-déterminés. Le réseau est entraîné en ajustant les poids des arêtes inter-neurones, par l'algorithme de rétro-propagation du gradient. Cet algorithme utilise une fonction d'erreur qui agit comme une métrique entre le résultat obtenu lors de la classification d'un exemple et le résultat souhaité, soit l'étiquette exact de l'exemple. Le gradient de cette fonction est par la suite calculé et propagé vers l'arrière au sein du réseau pour déterminer l'ajustement à effectuer de façon à réduire l'erreur. L'architecture du réseau que nous avons utilisé est de 4 couches cachées, contenant respectivement 30, 25, 20 et 15 neurones. L'algorithme a été entraîné durant 200 époques.

5 Résultats

5.1 Fiabilité des résultats

Les modèles entraînés ont été testés sur la même base de données. Le tableau suivant présente la précision des différentes méthodes.

Modèle	Précision
Bayes naïf	0.75
SVM	0.76
Arbres de décision	0.68
Réseau de neurones MLP	0.69

TABLE 1 – Précision des modèles selon la méthode utilisée

On constate que les algorithmes classifient les tweets de test mieux que le hasard, mais qu'ils sont loin d'être parfaits. L'algorithme SVM est celui qui performe le mieux avec une précision de 76% ; en conséquence, c'est ce dernier qui est utilisé dans l'analyse de la polarisation. Il reste qu'il va falloir prendre l'analyse qui suit avec un certain recul, car près d'un tweet sur quatre est mal classifié. Toutefois, les erreurs de classifications se font dans les deux sens avec autant de probabilité, de sorte que bien que les tweets individuels puissent être mal classifiés, les structures globales devraient être assez proches de la réalité.

Le niveau de précision assez faible de nos algorithmes peut être expliqué par divers facteurs, mais la complexité du langage naturel est probablement le

facteur le plus important. Chaque langue comprend plusieurs subtilités, telles que des expressions, des styles de ponctuation et du sarcasme. Pour un être humain, ces styles sont facilement détectables, mais ils sont perdus dans le texte écrit et pré-traité. L'utilisation de mots négatifs ou de double négation peut aussi confondre un algorithme, aussi précis soit-il. Les algorithmes sont donc efficaces, mais ils ne sont pas parfaits.

5.2 Vue d'ensemble des tweets politiques

Les tweets proviennent de politiciens des États-Unis, autant républicains que démocrates. Les nuages de mots suivants font ressortir les mots les plus utilisés par chacun des deux partis, après avoir effectué le pré-traitement.

FIGURE 1 – Nuage des mots les plus fréquents chez les démocrates

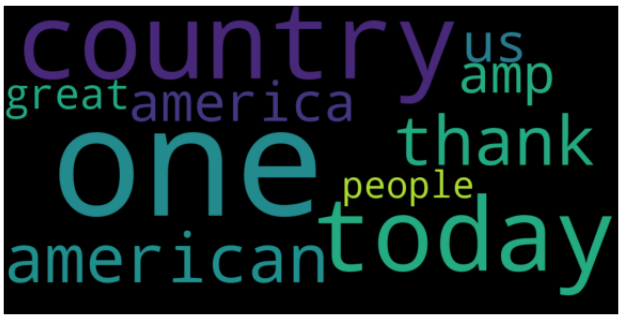


FIGURE 2 – Nuage des mots les plus fréquents chez les républicains

On observe tout d'abord que certains mots sont présents dans les deux nuages de points, tel par exemple «american». Évidemment, étant donné

que les tweets traitent en majeure partie de politique américaine, il était fort probable que des mots soient récurrents au sein des deux partis. De plus, on remarque la présence du mot «amp», qui correspond en fait au caractère spécial &. Par ailleurs, nous pouvons remarquer que les deux mots les plus fréquents chez les démocrates sont «help» et «people». L'égalité et la justice sociale étant au coeur des intérêts des démocrates, il est intéressant de remarquer que les mots les plus fréquemment utilisés font ressortir ces valeurs. Les deux mots les plus fréquemment utilisés par les républicains sont quant à eux «country» et «one». Ces mots sont beaucoup moins chargés de sens que ceux du parti démocrate. D'ailleurs, ils apparaissent aussi dans le nuage de point de ce parti.

Ci-dessous sont présentés à titre d'exemples trois tweets ayant été catégorisés « positif » et trois tweets ayant été catégorisés « négatif ».

Positif : BarackObama - *Last year, I sat down with my good friend Bruce @Springsteen for a long and meaningful conversation that touched on so much of what we're all dealing with these days. I'm excited to share it with you over the next few weeks :* {*hyperlien*}.

Positif : JoeBiden - *Heading back to Georgia with @DrBiden next Thursday to mark 100 days of our administration and thank folks for helping us build back better as a nation. See you soon !*

Positif : SpeakerPelosi - *With @POTUS Biden's leadership, @HouseDemocrats are making historic progress #ForThePeople – lowering the cost of health care, delivering bigger paychecks by building our infrastructure and ensuring our government works for the public interest, not dark money special interests.* {*hyperlien*}

Négatif : BarackObama - *With COVID-19 cases reaching an all-time high this week, we've got to continue to do our part to protect one another. This pandemic is far from over and your actions can help save lives.* {*hyperlien*}

Négatif : AOC - *I don't care what Cruz said at CPAC, but I do care that it appears Texas was just a layover stop for him between Cancun and Orlando to drop a pack of water into someone's trunk and abandon his constituents again as they get slammed with $16,000 electrical bills.* {*hyperlien*}

Négatif : HillaryClinton - *Want to help take on the climate crisis ? I encourage you to tune in to the Greenlight Climate Festival tomorrow. It's two days of conversations about the opportunities we all have to be a part of solutions.* {*hyperlien*} {*hyperlien*}

Le premier tweet est clairement bien classifié, avec beaucoup de mots connotés positivement, comme « good ». Les deux tweets suivants sont bien classifiés. Le quatrième tweet, quant à lui, est plus complexe à classifer. En effet, la première partie est négative, avec l'annonce de la montée du nombre de cas de la pandémie de COVID-19. Par contre, la seconde partie est un message d'espoir, et aurait pu être classifiée positive. Le 5e tweet est clairement négatif est a bien été classifié. Le dernier tweet, bien que ressemblant au 4e, de par le fait qu'il traite d'un sujet négatif en termes positifs, est néanmoins mal classifié, la part positive étant bien plus importante que la part négative.

5.3 Polarisation démocrates/républicains

Les modèles ont été utilisés pour classifier les tweets de politiciens des États-Unis. Nous les avons répartis selon leur parti politique. Les résultats sont présentés dans le tableau 2.

On observe que tous les utilisateurs n'ont pas la même fréquence d'utilisation. La tendance générale est d'exprimer des tweets positifs, et ce, quel que soit le parti. Les deux politiciens les plus positifs sont les républicains Arnold Schwarzenegger et Mike Pence avec un pourcentage de 84% de tweets positifs, alors que le politicien le plus négatif est le démocrate Bernie Sanders avec seulement 43% de tweets positifs - le seul politicien à écrire plus de tweets négatifs que de tweets positifs. En allant scruter le compte Twitter de Bernie Sanders plus en profondeur, nous remarquons que la plupart de ses tweets soulignent des enjeux concernant les inégalités économiques entre les différentes classes sociales, l'accès aux soins de santé et la crise environnementale. Ceci explique la connotation fortement négative de ses tweets. Au contraire, Mike Pence a tendance à féliciter publiquement les bons coups de son parti sur la plateforme. Ses tweets se veulent donc beaucoup plus positifs.

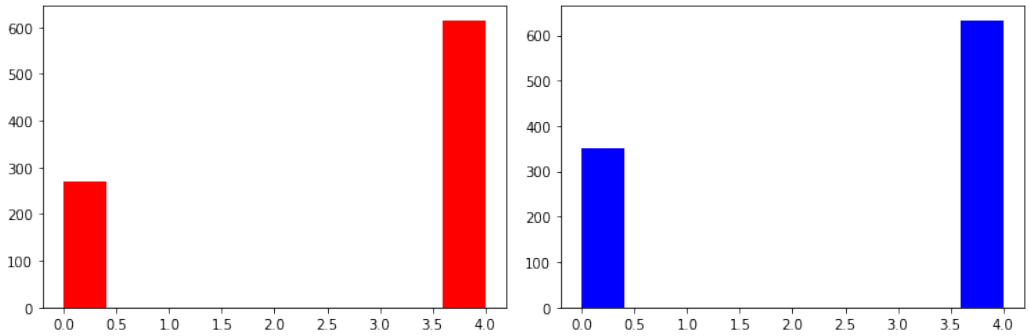
Afin de mieux visualiser la polarisation des politiciens au sein d'un même

Parti	Nom	Tweets +	Tweets -	% Tweets +
Démocrate	Barack Obama	120	41	75
Démocrate	Bernie Sanders	80	103	43
Démocrate	Kamala Harris	35	15	70
Démocrate	Joe Biden	28	15	65
Démocrate	Bill Cliton	116	48	70
Démocrate	Hillary Clinton	101	46	69
Démocrate	Nancy Pelosi	101	55	65
Démocrate	Alexandria Ocasio-Cortez	52	29	64
Républicain	Arnold Schwarzenegger	77	15	84
Républicain	Mike Pence	73	14	84
Républicain	Ted Cruz	61	40	60
Républicain	Rick Santorum	80	50	62
Républicain	Ben Carson	75	19	80
Républicain	Lindsey Graham	58	45	56
Républicain	Bobby Jindal	119	67	64
Républicain	Jeb Bush	72	21	77

TABLE 2 – Polarité de différents politiciens

parti, nous avons comptabiliser le nombre total de tweets à connotation négative et positive. Les deux histogrammes suivants présentent la polarité par parti (0 = négatif, 4 = positif). Nous remarquons que le parti démocrate a une tendance légèrement plus négative que le parti républicain. Ceci nous indique que l'exemple de Bernie Sanders précédemment mentionné peut fort probablement être généralisé pour le reste des politiciens. Ainsi, l'utilisation de la plateforme Twitter par les démocrates pour souligner certains enjeux politiques polarisants peut être à la source du pessimisme observé.

FIGURE 3 – Polarisation au sein des partis républicain et démocrate



(a) Polarité dans le parti républicain (b) Polarité dans le parti démocrate

6 Conclusion

En conclusion, bien que l'analyse de sentiment ne nous donne pas d'information rigoureuse sur le contenu des tweets, notre recherche nous permet de constater que certains politiciens ont des tendances plus polarisées, optimisme ou pessimisme, que d'autres. Une analyse sociologique plus approfondie permettrait d'expliquer les raisons de ces phénomènes, mais nous nous contentons de remarquer que cette différence est remarquable entre le parti républicain, qui semble légèrement plus positif que le parti démocrate, mais aussi, au sein des politiciens des ces partis.

Il serait d'ailleurs intéressant d'approfondir davantage cette analyse de sentiments en l'appliquant à des thèmes de l'actualité précis. Par exemple, en 2020, le mot clé #Coronavirus a été le deuxième dièse le plus utilisé sur Twitter. En étudiant si ton usage par le gouvernement était plus souvent lié à un sentiment positif ou négatif, il serait possible d'associer les différents partis politiques, ou leur représentants, à une opinion favorable ou défavorable face à certaines mesures. D'autres thèmes comme l'avortement et les armes à feu risqueraient de fortement refléter les convictions des différents partis américains et pourraient aussi mener à des conclusions intéressantes.

En bref, l'analyse de la polarité et l'analyse sentimentale nous donne beaucoup d'information sur les politiques d'un gouvernement. Ce pouvoir d'in-

fluence vient toutefois avec des responsabilités compte tenue de la porté de certaines personnalités publiques. Visiblement, lors de la campagne américaine 2020, le président américain a dépassé les limites, ce qui lui a valu d'être banni de Twitter.

De nos jours, il est important d'être conscient de l'existence de cette forte polarisation de l'opinion et de s'y en protéger en diversifiant volontairement notre bulle médiatique. Ceci est atteignable en y incluant des opinons vairées et différentes de celles notre entourage.

7 Contribution des membres de l'équipe

Comme prévu, nous avons travaillé en étroite collaboration dans chaque étape du projet. Comme Simon-Olivier s'est principalement occupé de la partie extraction et pré-traitement des données, il a aussi rempli cette partie du rapport. Pour les algorithmes, nous en avons chacun implémenté et décrit un. Une fois les premiers résultats obtenus, Simon-Olivier et Laurent ont travaillé à améliorer la précisons des algorithmes pendant que Juliette commençait à rédiger le rapport. Laurent s'est occupé de produire les statistiques et les graphiques des résultats qui ont ensuite été analysés par Simon-Olivier et Juliette. Le rapport a donc été rédigé de manière collaborative du début à la fin.

8 Références

[1] Kazanova, M. M. (2017, September 13). Sentiment140 dataset with 1.6 million tweets. Kaggle. https ://www.kaggle.com/kazanova/sentiment140

[2] Yu, P. (2020, January 6). How to Access Twitter's API using Tweepy - Towards Data Science. Medium. https ://towardsdatascience.com/how-to-access-twitters-api-using-tweepy-5a13a206683b

[3] Hebbar, N. (2021, January 8). Twitter Sentiment Analysis Using Python for Complete Beginners. Medium. https ://medium.com/swlh/tweet-sentiment-analysis-using-python-for-complete-beginners-4aeb4456040