Compte-rendu d'analyse

François Marès
Merwan Bouvier
Thomas Laurent

8 juin 2021

Résumé

Les près de 3 millions de tweets associés à l'*Internet Research Agency*, une organisation impliquée dans la campagne d'influence russe des élections américaines de 2016, peuvent faire l'objet d'une classification comportementale. Les chercheurs Darren Linvill et Patrick Warren ont proposé une telle classification, comprenant essentiellement les classes *Right Troll*, *Left Troll*, *News Feed* et *Hashtag Gamer*. Nous discutons dans ce document de plusieurs méthodes candidates pour automatiser cette classification avec le plus de précision possible.

Après avoir défini une stratégie de représentation du contenu sémantique des tweets, nous éprouvons plusieurs méthodes pour parvenir à mettre en avant celle capable de distinguer le mieux possibles les comportements principaux des comptes trolls.

1 Introduction

En janvier 2017, la CIA, le FBI et la NSA publient un document conjoint dont la version déclassifiée[3] contient des jugements relatifs aux activités russes en rapport avec les élections présidentielles américaines de 2016. Il affirme qu'une politique d'influence de la démocratie américaine favorable à Trump a été mise en œuvre, et énonce les moyens employés. En particulier, est écrit que la stratégie russe *allie des opérations secrètes de renseignement aux efforts manifestes des agences gouvernementales russes, des médias financés par l'État, des intermédiaires tiers et des utilisateurs rémunérés des médias sociaux ou «trolls».*

En février, l'*Internet Research Agency* (IRA), organisation basée à Saint-Pétersbourg identifiée comme la source des trolls russes, est mise en examen pour ses activités. Le rapport du procureur spécial[4], publié en 2019, conclut à l'absence de collusion entre la Russie et Trump, mais présente de nombreux éléments sur les opérations russes – cependant presque tous les détails, en particulier ceux sur les tweets de l'IRA (p.24), sont noircis.

En juillet 2018, le site web *FiveThirtyEight*, spécialisé dans le journalisme de données, donne accès librement à des fichiers contenant près de 3 millions de tweets associés à l'IRA [2]. Cette base de données a été constituée par deux chercheurs de l'Université de Clemson, Darren Linvill et Patrick Warren, à partir des informations que Twitter a fourni au Congrès américain sur les comptes trolls et leurs agissements entre mai 2015 et novembre 2017. Les chercheurs s'en sont servis pour étudier les méthodes d'influence de la politique américaine par l'IRA [1]. Pour ce faire, il ont classifié les activités des comptes en thèmes, par exemple *Left Troll* ou *Right Troll*.

L'objectif de notre étude est d'automatiser cette classification des tweets en provenance de l'*usine à trolls* russe, grâce à des méthodes d'apprentissage supervisé.

Faciliter l'étude des campagnes de désinformations est nécessaire pour qui souhaiterait prendre des mesures capables de limiter leur efficacité. Cet objectif fait parti de ceux de Twitter : depuis la révélation de l'affaire des trolls russes, l'entreprise a annoncé la suppression de nombreux comptes liés à des opérations d'influence.

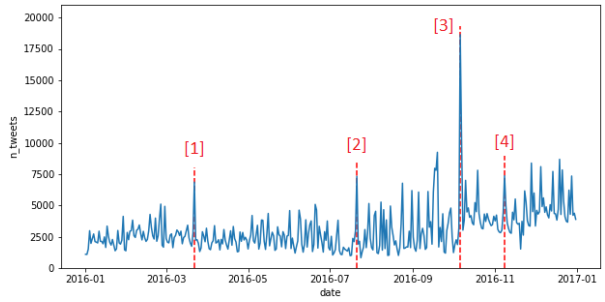
2 Analyse

2.1 Présentation des données

La base de données contient près de 3 000 000 de tweets provenant de seulement 2 848 pseudonymes Twitter. Ils ont été envoyés entre février 2012 et mai 2018 mais la grande majorité a été postée de 2015 à 2017. L'activité des comptes est fortement corrélée avec les évènements médiatiques majeurs liés à la politique aux États-Unis – la vérification du contenu des tweets vient confirmer ce lien. C'est particulièrement remarquable pour l'année 2016 [Figure 1], année d'élection de Do-

nald Trump à la Maison Blanche.

FIGURE 1 – Chronologie des tweets publiés en 2016.



[1] Mars 2016, WikiLeaks publie plus de 30 000 courriels et pièces jointes de la messagerie privé d'Hillary Clinton alors qu'elle était secrétaire d'État.

[2] Juillet 2016, Investiture de Donald Trump comme candidat républicain.

[3] Octobre 2016, fuite de milliers de nouveaux mails d'Hillary Clinton.

[4] 8 novembre 2016, Donald Trump est élu quarante-cinquième président des États-Unis.

2.1.1 Catégories de Linvill et Warren

Les tweets et comptes associés ont été classés en 8 groupes comportementales par Linvill et Warren, dont voici une brève description (par ordre décroissant de tweets associés) :

Non English (837,725) Comptes rédigeant le plus souvent dans une autre langue que l'anglais – le plus fréquemment il s'agit du russe.

Right Troll (719,087) Comptes se faisant passer pour des supporters de Donald Trump. Ils publient des tweets aux messages populistes républicains, dénigrent Obama et Hillary Clinton.

News Feed (599,294) Comptes se faisant passer pour des agrégateurs de nouvelles locales officielles partageant les actualités autour de certaines villes américaines.

Left Troll (427,811) En opposition aux *Right Trolls*, cette catégorie se fait passer pour des activistes supporters du parti Démocrate. Ils ont pour but de diviser le parti Démocrate et critiquent pour certains Hillary Clinton.

Hashtag Gamer (241,827) Comptes publiant selon le jeux hashtag – l'utilisateur met un hashtag spécifique et répond à la question impliqué par celui-ci. Le but de cette catégorie est de se faire passer pour des joueurs de jeux vidéo et publier des tweets politiques de temps à autre, notamment à travers les jeux hashtag.

Commercial (122,582) Comptes à buts commerciaux. Leurs tweets décrivent un produit (perte de poids, investissement, etc.) souvent accompagné d'un lien et de mentions vers d'autres comptes twitter.

Unknown (13,905) Comptes inclassables parmi les autres catégories.

Fearmonger (11,140) Comptes propageant des informations sur des événements de crise inventés.

Bien que la catégorie `Non English` soit importante dans les données, pour des raisons de difficultés de traduction nous ignorons cette catégorie dans notre analyse. La catégorie `Unknown`, peu importante en volume sera elle aussi ignorée. Parmi les cinq modèles de comportement – *Right Troll*, *Left Troll*, *News Feed*, *Hashtag Gamer* et *Fearmonger* –, *Fearmonger* fait exception : les tweets des comptes de cette catégorie passent parfois d'une catégorie à une autre. C'est pourquoi nous avons choisi de ne pas en tenir compte pour la suite de notre analyse.

Nous restreignons donc notre problématique à l'automatisation de la classification des tweets en provenance de l'*usine à trolls* russe parmi les quatre catégories *Right Troll*, *Left Troll*, *News Feed* et *Hashtag Gamer*.

2.1.2 Restriction des données d'origine

En plus de ne nous intéresser qu'à quatre catégories, nous retirons de notre analyse tous les tweets rédigés dans une autre langue que l'anglais. Notre étude porte finalement sur les deux-tiers des tweets de la base de données d'origine (1 959 778 tweets) ; par la suite, chaque fois que nous parlerons de données nous feront référence à ce sous-ensemble.

2.1.3 Données d'entraînement et de test

Pour que la comparaison entre les méthodes de classification automatique soit équitable, nous utiliserons toujours les mêmes données d'entraînement et de test. Nous avons séparé aléatoirement nos données en deux jeux : 80% (soit 1 567 823 tweets) pour les données d'entraînement et les 20% restantes (soit 391 955 tweets) pour les données de test.

Sauf précisions contraires, les statistiques évoquées par la suite porterons sur les seules données d'entraînement.

2.1.4 Descripteurs candidats pour la classification

Le jeu de données initial comporte 21 descripteurs pour chaque tweet :

external_author_id	following	new_june_2018
author	followers	alt_external_id
content	updates	tweet_id
region	post_type	article_url
language	account_type	tco1_step1
publish_date	retweet	tco2_step1
harvested_date	account_category	tco3_step1

Cependant, la plupart ne sont pas pertinents pour notre classification.

Puisque, par ses tweets, c'est en réalité un compte que nous voulons classifier, tous les descripteurs qui permettent de l'identifier avec certitude sont à exclure (*external_author_id*, *author*, *alt_external_id*). Nous aurions pu garder le nom des comptes, pour exploiter le lien sémantique qui existe parfois avec le thème principal de ce dernier – en particulier pour les comptes *NewsFeed* –, mais nous n'avons pas voulu entrer dans une telle analyse.

Le descripteur *region* est presque identique pour tous les tweets (États-Unis) puisque nous avons sélectionné les tweets en anglais.

La date de publication (*publish_date*) n'est pas retenue puisque nous désirons que notre classifieur ne se limite pas à la période des tweets d'apprentissage.

Les métadonnées techniques telles que la date de collecte du tweet, un booléen indiquant si le tweet vient de l'extension de la base de données fournie au Congrès, l'identifiant du tweet ou encore le lien vers celui-ci (respectivement *harvested_date*, *new_june_2018*, *tweet_id* et *article_url*) ne sont pas retenus pour notre classifieur.

Les trois liens présents dans un tweet (*tco1/2/3_step1*) ne sont pas non plus directement retenus ; étant donnée, d'une part, la difficulté d'analyse des liens inconnus des données d'entrainement, et d'autre part la faible probabilité qu'un ancien lien soit utilisé par un autre tweet ultérieurement. Si notre objectif avait été de nous limiter à la classification de tweets sur une courte période, ou si nous voulions prendre en compte le maximum d'informations, nous pourrions ajouter par exemple quatre nouvelles variables. Chacune serait le score des liens du tweet relatif à chaque catégorie (comme nous l'avons fait pour l'analyse sémantique des tweets, cf. 2.1.5), défini par exemple comme la somme des fréquences d'utilisations des trois liens – où des sites référencés – parmi les données d'apprentissages de la catégorie.

Cependant, on observe que le nombre de liens est un bon discriminant pour les classes *Hashtag Gamers* et *News Feed*. Les *Hashtag Gamers* publient des tweets avec une fréquence de présence de lien beaucoup plus faible (0.26), tandis que les tweets avec 3 liens proviennent plus largement des comptes *News Feed*. Une variable contenant le nombre de liens peut être utilisée par notre classifieur lorsqu'il inclura des variables qualitatives.

Le type du tweet (décrit par *post_type* et *retweet*) pourrait être retenu pour notre classification automatique, en particulier lorsque nous utiliseront des arbres de décision. Jusque là, cette variable qualitative n'est pas prise en compte.

Mis à part le contenu du tweet, les trois variables restantes *following*, *followers* et *updates* sont quantitatives. Elles pourraient facilement être utilisées comme entrées de notre classifieur, mais nous nous heurterions alors à un biais important. En effet les données qui sont à notre disposition ne proviennent que d'environ 3000 comptes différents. Si un tweet de l'ensemble de test proviennent d'un compte représenté dans les données d'entraînement il est très probable que le classifieur fasse ce lien : ces variables prendraient le rôle d'identifiants du compte. On peut illustrer ce problème en mettant en œuvre la méthode des k plus proches voisins : pour $k = 3$ par exemple on obtient une précision de 98% si le partage des tweets en données d'apprentissage et de test est indifférent aux comptes. Un moyen d'éviter ce problème serait de rendre qualitatives ces variables.

Finalement, c'est à partir du contenu des tweets que nous avons choisi de les classifiers dans un premier temps, les autres descripteurs candidats étant qualitatifs [Table 1]. La partie suivante (Sec. 2.1.5) présente la transformation de ce descripteur très complexe en huit descripteurs quantitatifs – que nous appellerons *scores*.

descripteur d'origine	descripteurs dérivées	type
content	8 scores	quantitatif
tco1/2/3_step1	nombre de liens	qualitatif
post_type		qualitatif

Table 1 – Descripteurs candidats pour la classification.

2.1.5 Nouveaux descripteurs quantitatifs

Naturellement, un humain qui voit un tweet dont il ne connait pas l'origine juge du thème de celui-ci à partir de son contenu textuel. Souvent, l'identification de quelques mots clef suffit. Les hashtags, en particulier, sont utilisés a escient pour maximiser l'efficacité de cette analyse ; le temps qui y est consacré par le lecteur étant presque instantané.

Nous voulons donc construire des descripteurs qui quantifient l'orientation des mots clef d'un teweet, soit une crédence en son appartenance à chaque classe selon ses mots et hashtags utilisés, en traitant à part les hashtags.

Pour chaque catégorie, nous avons construit une liste des 100 mots les plus fréquemment employés parmi les tweets de l'ensemble d'entraînement, qui ne font pas partie des mots les plus communs de la langue anglaise. L'occurrence de chaque mot est ensuite remplacée par la proportion qu'elle représente par rapport au nombre total d'occurrences dans la liste. La TABLE 2 donne pour deux classes les dix premiers mots de ces classements.

	`RightTroll`		`LeftTroll`	
1	trump	0.1025	black	0.0523
2	obama	0.0340	trump	0.0466
3	just	0.0324	just	0.0317
4	hillary	0.0279	white	0.0260
5	breaking	0.0276	police	0.0213
6	video	0.0218	love	0.0195
7	president	0.0197	after	0.0163
8	clinton	0.0179	need	0.0161
9	realdonaldtrump	0.0168	today	0.0153
10	america	0.0165	women	0.0142

TABLE 2 – 10 mots les plus employés par les classes *Right Troll* et *Left Troll*.

Le score de chaque tweet selon une classe donnée est ensuite déterminé comme la somme des proportions de l'intersection entre les mots qu'il contient et les 100 mots les plus employés par cette catégorie. L'opération est identique pour les hashtags.

La distribution des scores est très similaire entre ceux d'un même type sur des classes différentes, mais la distribution est très différente entre les scores sur les mots ou sur les hashtags. La TABLE 3 donne la moyenne, l'écart-type, les quartiles et le maximum des scores des tweets pour la classe *Right Troll*.

On remarque que les scores sur les hashtags sont très faibles et souvent nuls, comme le confirme la TABLE 4, qui donne le pourcentage de tweets avec un score nul

	Mots	Hashtags
count	1 567 823	1 567 823
mean	0.0184	0.0027
std	0.0344	0.0078
min	0	0
25%	0	0
50%	0.0064	0
75%	0.0179	0
max	1.7439	0.2978

TABLE 3 – Statistiques basiques sur les scores pour la classe *Right Troll*.

pour chacun des huit descripteurs. Une raison commune à la faiblesse des scores sur les hashtags ainsi que de telles proportions de scores nuls est le peu de hashtags contenus dans chaque tweet, et le choix de ne prendre en compte que les 100 hashtags les plus fréquemment utilisés par les catégories. Pour le cas des hashtags des tweets de catégorie *Hashtag Gamer*, les tweets utilisent des hashtags très variés, d'où les scores observés. Ce descripteur ne servira à discriminer la classe que de seulement quelques tweets – en cas de besoin de réduire la dimension des données d'entrée de modèles, nous pourrons ignorer ce score.

	Mots (%)	Hashtags (%)
RightTroll	43	83
LeftTroll	43	86
NewsFeed	52	84
HashtagGamer	50	99

TABLE 4 – Pourcentage des tweets au score nul pour chaque type de score.

2.1.6 Limite des scores définis

Plus embêtant, 323 588 tweets (soit 20% des données) ont tous leurs scores nuls, parce qu'ils s'éloignent des tweets typiques de leurs catégories – ou même ne contiennent aucune information sémantique trahissant leur appartenance à un groupe comportemental.

Ces dernières remarques nous font anticiper que nos classifieurs sur les tweets n'auront surement pas une grande précision : une partie des tweets ne sont pas caractéristiques d'une catégorie comportementale, et nos scores ne rendent pas compte de toute l'information sémantique contenue dans un tweet.

Un moyen de pallier ce problème est, comme nous le faisons à la fin de l'analyse, de préférer comme mesure

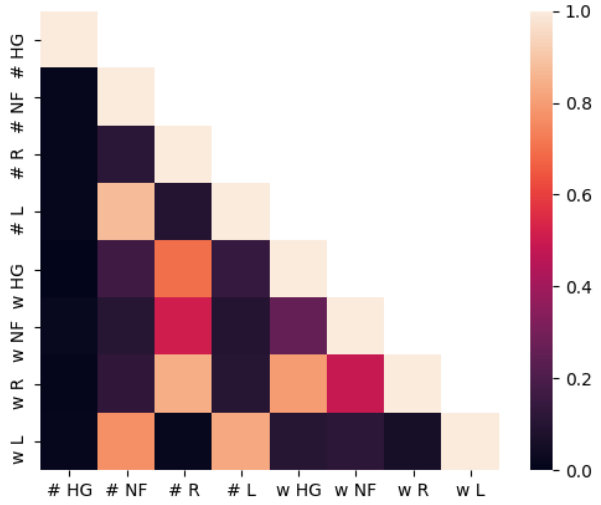
de précision celle concernant la prédiction de la catégorie des comptes (par le vote des prédictions sur ses tweets) plutôt que de celle des tweets isolés.

2.1.7 Corrélations des scores

Le coefficient de Spearman permet de décrire à quel point la dépendance entre deux variables est monotone[1]. Plus une telle relation entre les variables est marquée, plus le coefficient se rapproche de 1 en valeur absolue.

Les coefficient de Spearman des huit variables score sont représentés sur la FIGURE 2.

FIGURE 2 – Matrice de corrélations de Spearman (valeurs absolues).



Pour nos descripteurs, la monotonie des dépendances n'est pas un indicateur suffisant pour juger de leur capacité de discrimination, mais une forte monotonie de dépendances est mauvais signe. En effet, si deux scores évoluent l'un par rapport à l'autre toujours de la même manière – par exemple sont proportionnelles si le coefficient de Pearson est proche de 1 en valeur absolue –, alors il y a un risque que les deux variables soient de mauvais discriminants entre les deux catégories associées. Les scores associés aux classes *News Feed* et surtout *Hashtag Gamers* sont les plus monotonement corrélées aux autres classes, on peut s'attendre à des difficultés de classifications les concernant.

À l'opposé, il est bon que le coefficient de corrélation entre le score sur les mots et sur les hashtags pour une même catégorie soit élevé. C'est le cas pour les classes *Left Troll* et *Right Troll*.

1. Nous aurions aussi pu choisir le coefficient des rangs de Kendall, moins sensible aux nombreuses valeurs nulles de nos scores, mais les résultats sont très proches.

2.2 Apprentissages supervisés

2.2.1 Classifications triviales à battre

Commençons par évoquer les méthodes de classifications triviales qui serviront de référence pour interpréter les résultats des méthodes plus avancées.

Sans apprentissage, la probabilité à priori qu'un tweet appartienne à une parmi quatre classes est de 0.25. Un classifieur naïf pourrait tirer au hasard la classe de chaque tweet [*N.1*].

Ou bien, comme la répartition parmi les classes n'est pas égale dans les données, tirer au hasard selon les probabilités à postériori d'appartenir à une classe, selon les proportions des données d'entraînement [*N.2*].

Enfin, une troisième méthode simple est de prendre pour prédiction la catégorie associée au meilleur score parmi les huit définis [*S.0*], ou seulement les scores sur les mots [*Sw.0*], ou sur les hashtags [*S#.0*] – lorsque tous les scores sont nuls, la prédiction est selon *N.2*.

On obtient les précisions, par classe et totale, données dans la TABLE 5.

Méthode	Gamer	Left	News	Right
N.1	0.25	0.25	0.25	0.25
		total accuracy : 0.25		
N.2	0.13	0.21	0.30	0.36
		total accuracy : 0.28		
S#.0	0.14	0.22	0.60	0.41
		total accuracy : 0.39		
Sw.0	0.42	0.29	0.67	0.49
		total accuracy : 0.49		
S.0	0.43	0.30	0.77	0.50
		total accuracy : 0.53		

TABLE 5 – Précision des classifieurs simples.

La méthode [*S.0*] donne déjà de bons résultats – on sait qu'il est difficile de faire des prédictions supérieurs à 70% comme nous l'avons déjà vu –, surtout pour la catégories *NewsFeed*.

Nous cherchons désormais des classifieurs plus performants, ou avec une précision répartie différemment parmi les classes.

2.2.2 Analyses discriminantes

Le modèle le plus simple que nous avons expérimenté est l'analyse discriminante linéaire [*ADL*], ou quadratique [*ADQ*]. Sous l'hypothèse d'une loi normale multidimensionnelle, ces deux méthodes font des estimation des paramètres de l'expression de la formule de Bayes et dont on déduit les règles de décision des classes.

En isolant un groupe de validation à partir des données d'entraînement, nous avons énuméré puis testé toutes les combinaisons de variables scores qu'il est possible d'utiliser pour l'ajustement des paramètres des modèles. Nous avons ainsi mesuré que les performances de classification étaient maximales en utilisant les quatre variables de score associées au mots.

Dans le cas de l'analyse discriminante, le recours à l'ACP n'est pas nécessaire aux vues de la simplicité du modèle : il n'y avait pas de raison de réduire la dimension de l'ensemble d'entraînement. La TABLE 6 donne les performances des analyses

Méthode	Gamer	Left	News	Right
ADL	0.06	0.20	0.51	0.78
		total accuracy : 0.49		
ADQ	0.33	0.14	0.85	0.29
		total accuracy : 0.43		

TABLE 6 – Précisions des classifieurs simples.

Ces méthodes ne présentent ici pas de résultats satisfaisants. De plus, les modèles linéaire et quadratique présentent une forte dispersion des précisions calculées par classe : les tweets dont la classe de vérité fait partie des deux classes *Right Troll* et *News Feed* sont relativement bien classifiés et les tweets dont la classe de vérité fait partie des deux classes *Left Troll* et *Hashtag Gamer* sont relativement mal classifiés.

2.2.3 Méthode des K plus proches voisins

La méthode des K plus proches voisins (KPP) consiste à affecter au tweet des données de test la classe la plus représentée parmi celles de ses K plus proches voisins dans l'ensemble d'apprentissage.

Puisque nous avons choisi la distance euclidienne comme mesure de proximité, les variables de score avec les valeurs les plus importantes (les scores sur les mots) ont une plus grande importance. Les autres scores avec beaucoup de valeurs faibles ne sont décisives que dans les cas où les scores sur les mots sont courants (0 notamment).

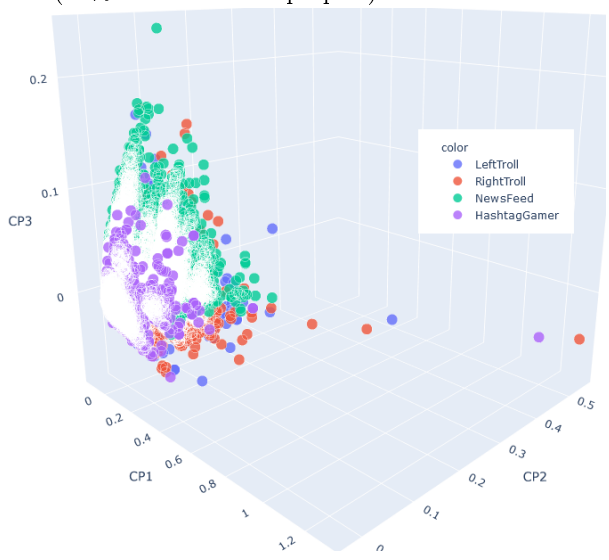
Le problème que nous avons très vite rencontré est que le nombre de dimensions (8) est trop important pour mettre en œuvre la méthode avec nos deux millions de données de test. La recherche d'un K optimal s'annonce ainsi difficile.

Analyse en Composante Principales L'objectif d'une Analyse en Composantes Principales (ACP) est d'obtenir une représentation fidèle de notre nuage de points sur un espace de faible dimension. L'ACP nous assure une solution optimale pour maximiser l'inertie expliquée dans le nouvel espace.

Cependant, la maximisation de l'inertie n'est pas un gage de performances pour notre méthode des KPP, et des scores associés à une inertie faible peuvent être en réalité cruciaux pour déterminer la bonne classe d'un tweet.

Pour comparer les résultats avec et sans ACP nous travaillons avec un sous-ensemble aléatoire des données de 200 000 tweets. On observe la répartition des tweets selon leur catégorie dans l'espace de trois dimensions FIGURE 3.

FIGURE 3 – 200 000 tweets selon les trois axes d'une ACP (93% de variance expliquée).

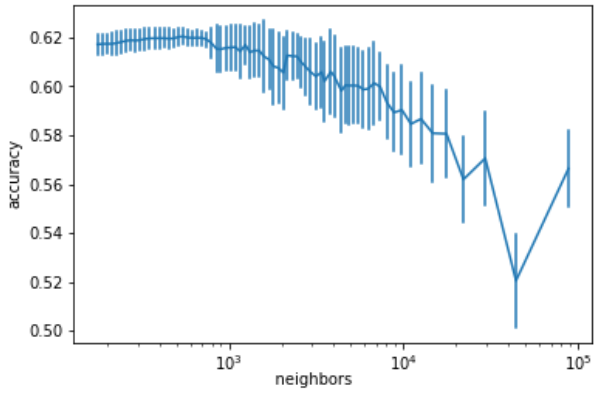


Les tweets sont bien regroupés selon leur catégorie quand on s'éloigne de l'origine, mais pour les scores faibles, proches de l'origine, c'est près de la moitié des tweets qui sont regroupés sans que l'on puisse distinguer des groupes par catégories.

Nous avons ensuite mis en œuvre l'algorithme KNN pour K allant de 1 à 100, avec différentes divisions aléa-

toires des données, pour trouver le paramètre optimal du nombre de voisins. La précision du modèle en fonction de K est représentée sur la FIGURE 4.

FIGURE 4 – Précision de la méthode KNN après une ACP pour 200 000 tweets. K-optimal de 23 voisins.



Le résultat est plutôt bon (62% de précision), mais lorsque l'on augmente le nombre de données le classifieur perd en qualité : on tombe à 56% de précision pour un K optimal de 90 voisins. On peut cependant poser en conjecture que le K optimal est bien plus grand en réalité, mais le temps de calcul devient trop important pour que nous le cherchions.

Pour tenter de diminuer le temps de calculs, nous avons mis en œuvre l'algorithme des *K plus proches prototypes*. Cette variante de la méthode des *K plus proches voisins* comporte une phase d'apprentissage constituée du calcul des prototypes qui résument les individus d'apprentissage dans chaque classe. Cependant, les performance sont restées un peu en dessous de celles des *K plus proches voisins*.

Enfin, sans ACP et sur un jeu réduit des données la méthode des KNN nous a donné de bons résultats comme on pouvait s'y attendre. La TABLE 7 qui suit résume les résultats des méthodes KPP :

Tweets	ACP	K-opt	Précision au K-opt
200 000	3D	23	0.62
1 000 000	3D	90	0.56
200 000	non	172	0.63

TABLE 7 – Précisions des méthodes KPP.

Finalement, les méthodes de K plus proches voisins sont trop gourmandes en calculs pour répondre à notre problématique. Nous en feront un meilleur usage en fin d'analyse quand nous modifieront légèrement celle-ci en nous intéressant à la classification des comptes plutôt que des tweets, ce qui réduira considérablement la taille des données d'entrée.

2.2.4 Arbres de décision et forêts

Les méthodes à base d'arbres binaires consistent à partitionner de manière récursive l'espace des caractéristiques en régions homogènes au sens de notre classification.

Pour augmenter la robustesse de cette classification, on applique la technique du *bagging* qui consiste à faire voter plusieurs classifieurs entraînés sur des sous-ensembles de données différents. Pour maximiser la diversité de ces classifieurs, on passe directement par la méthode des forêts aléatoire. En effet, les descripteurs utilisés pour la construction de chaque arbre sont tirés au sort (on en tire le nombre optimal, soit $\sqrt{9} = 3$).

On teste les classifieurs suivants, leurs résultats sont présentés dans la TABLE 8. Comme nous l'avions proposé (TAB. 1), on peut ajouter à nos variables scores le nombre de liens présents dans le tweet. Cette nouvelle dimension permet d'augmenter légèrement les performances.

[*RF.1*] 50 classifieurs, 50 nœuds terminaux, variables : 8 scores.

[*RF.2*] 50 classifieurs, 50 nœuds terminaux, variables : 8 scores et le nombre de liens.

[*RF.3*] 50 classifieurs, 100 nœuds terminaux, variables : 8 scores et le nombre de liens.

[*RF.4*] 100 classifieurs, 100 nœuds terminaux, variables : 8 scores et le nombre de liens.

Méthode	Gamer	Left	News	Right
RF.1	0.05	0.20	0.74	0.84
		total accuracy : 0.58		
RF.2	0.57	0.14	0.84	0.76
		total accuracy : 0.62		
RF.3	0.58	0.21	0.84	0.73
		total accuracy : 0.63		
RF.4	0.57	0.20	0.84	0.74
		total accuracy : 0.63		

TABLE 8 – Précisions des forêts aléatoires.

Les méthodes de forêts aléatoires donnent de bons résultats tout en n'étant pas trop gourmands en calculs. De plus, ils permettent l'ajout de descripteurs qualitatifs, sans craindre un trop grand nombre de dimensions. Cependant, la précision est encore très inégale pour les différentes classes.

2.3 Classification des comptes

Les parties précédentes présentait la classification des tweets pris individuellement à partir de leur contenu. Comme nous l'avons vu en fin de présentation des variables scores, cette démarche n'est pas entièrement satisfaisante [Sec. 2.1.6]. La variabilité du contenu des tweets d'une même classe – y compris pour un même compte – ainsi que parfois l'absence d'informations sur la classe adéquate nous empêchent de toujours identifier le comportement à associer à un tweets.

Une approche plus stable est de classifier les auteurs des tweets plutôt que les tweets eux-mêmes. Chacun des auteurs est associé à un comportement principal, c'est cette classification que nous pouvons faire plus précisément.

Deux approches ont été envisagées (la table 9 donne leurs précisions) :

— travailler directement en donnant des scores aux auteurs, après la concaténation du texte de tous leurs tweets.

— associer à l'auteur la classification majoritaire de ses tweets.

La première option donne plus d'importance aux tweets très caractéristiques d'une catégorie et à ceux avec beaucoup de contenu. Nous l'avons adoptée avec la méthode des *K plus proches voisins*, sans ACP et sur touts les comptes des données [*CKPP*].

La seconde donne autant de poids à chaque tweets, qu'il soit caractéristique d'une classe comportementale ou inclassable. Nous l'avons adoptée avec la méthode des forêts aléatoires [*CFA*].

Méthode	Gamer	Left	News	Right
CFA	0.96	0.08	0.83	0.72
		total accuracy : 0.61		
CKPP	0.55	0.65	0.75	0.98
		total accuracy : 0.84		

Table 9 – Précisions des méthodes appliqués aux comptes.

Le résultats sont ceux espérés pour la méthode *CKPP*, qui a notamment une très bonne répartition de sa précision entre les classes (une moyenne de 0.73).

La méthode *CFA*, elle, prédit très mal l'appartenance du compte à la classe *Left Troll*. Ce résultat n'est pas étonnant puisque cette classe a été la plus difficile à distinguer pour toutes les méthodes, elle a donc peu de chance d'être la classe majoritaire des tweets d'un compte.

Avec les variables scores que nous avons défini la solution du regroupement des informations sémantiques de tous les tweets d'un même compte est certainement la meilleur solution.

3 Conclusion

Notre objectif dans cette analyse était de parvenir à classifier automatiquement les tweets de l'IRA dans différentes catégories comportementales.

Après avoir pris en main le jeu de données, nous l'avons enrichi en créant de nouvelles variables quantitatives à partir du contenu textuel des tweets afin de pouvoir mettre à l'épreuve différents algorithmes de classification.

Dans un premier temps, nous avons cherché à classifier les tweets pris séparément. Nous avons obtenu des précisions globalement satisfaisantes, en particulier avec les forêts aléatoires, mais nos résultats étaient limités par le fait que les tweets sont très variables et peuvent parfois être impossibles à classifier à partir de leur contenu, et ce même pour un opérateur humain.

Nous avons donc expérimenté une seconde démarche consistant à regrouper les tweets par auteur, ce qui palliait aux limites précédemment observées. Cette fois-ci, les variables quantitatives créées étaient dans la grande majorité du temps très pertinentes, et celles-ci nous ont permis, via les méthodes déjà employées, d'obtenir des résultats de classification très satisfaisants.

À travers notre classification, nous n'avons pas apporté de nouvelle information par rapport au jeu de données initial. En effet, chacun des tweets présent dans le jeu de données a été catégorisé par Linvill et Warren. Cependant, dans le cas hypothétique où de nouveaux troll tweets associés à de nouveaux comptes troll liés avec l'IRA étaient récupérés, notre solution de classification par apprentissage supervisé à partir de mesures de caractères sémantiques associées aux comptes serait un moyen rapide et pertinent de classification automatique.

Références

[1] Darren L. Linvill and Patrick L. Warren (Juin 2018). Troll Factories : The Internet Research Agency and State-Sponsored Agenda Building. *Disponible en ligne à cette URL.*

[2] Oliver Roeder (Jul. 2018). Why We're Sharing 3 Million Russian Troll Tweets. Article publié par FiveThirtyEight, *disponible en ligne à cette URL.*

[3] Office of the Director of National Intelligence (Jan. 2017). Assessing Russian Activities and Intentions in Recent US Elections (ICA 2017-01D). *Available online at this URL.*

[4] Special Counsel Robert S. Mueller (March 2019). Report On The Investigation Into Russian Interference In The 2016 Presidential Election Volume I of II. *Available online at this URL.*