Documentation en OSINT de l'algorithme de détection

Les bots, outils de désinformation

L'un des principaux enjeux de la désinformation est psychologique. Au-delà de simplement diffuser de fausses informations ou du contenu tronqué, biaisé ou dévié, les phénomènes d'engouement de masse qu'elle peut produire sont particulièrement inquiétants. Sur des plateformes et des réseaux sociaux comme Twitter, un compte influent avec un réseau de *followers* et d'abonnements fourni peut facilement imposer une tendance et manipuler les opinions. Des mouvements populaires naissent, se développent et se polarisent sur Twitter; et il est difficile de croire qu'il s'agisse du fruit du hasard.

En réalité, l'achat de faux abonnés est monnaie courante. Certains utilisateurs du réseau vont chercher à élargir leur zone d'influence en faisant grimper artificiellement leur nombre d'abonnés. Dans d'autres cas, la création de faux comptes sert un but plus élaboré, comme la transmission de messages polarisants, dans tel ou tel contexte d'actualité, en faveur de telle ou telle communauté, à la défense de tels ou tels idéaux. A grande échelle, l'impact est assez conséquent pour alimenter et radicaliser des mouvements sociaux sur le long-terme.

Des comptes "humanisés"

Les comptes ainsi créés, automatisés et régis par des algorithmes mais affichant de faux *personae* pour mieux s'"humaniser", sont de moins en moins facilement repérables. Prenons par exemple le compte de "@FMarlair". "*Je suis une personne sincère mais j'aime vraiment la nature ici c'est la grande France*", indique-t-il en description, avant d'insérer une émoticône de drapeau français.

En plus de relever le manque de cohérence de la phrase, on peut observer trois éléments intéressants. L'utilisation de l'adjectif "sincère", positif, incite le lecteur de la description à la confiance - envers la personne, mais surtout envers le contenu qu'elle publie, aime et partage. On observe également une indication sur ses goûts, par définition humanisante. Enfin, le drapeau et la mention patriotique (voire nationaliste), si elles n'indiquent pas ouvertement une opinion politique tranchée, sont déjà polarisantes.

Inscrit le 26 février 2019 à 3h07 du matin, un horaire pour le moins étrange pour décider de se créer un compte Twitter, "Fabien" a déjà eu le temps de s'abonner à plus de 800 comptes - c'est-à-dire un peu plus de 13 comptes par heure s'il néglige de dormir (ce qui semble être le cas).

Pourtant, sa photo de profil - la seule image de son compte - montre un homme d'âge moyen à la mine bien reposée, aux traits régulier et séduisants, tout droit sorti d'un magazine. D'ailleurs, sa photo est également celle d'un certain Jean-Pierre Dimitri sur Instagram (entre autres) et apparaît sur Pinterest et Webstagram auprès des hashtags #malemodel et #perfectman. *Soit Fabien souffre d'un trouble dissociatif de l'identité, soit celle-ci a été créée de toutes pièces.*

Malgré son activité importante d'abonnement aléatoire, Fabien n'a qu'une cinquantaine d'abonnés. Un faible ratio qui reflète peut-être la qualité de ses publications: celles-ci consistent uniquement en des retweets de publications, pour les trois-quarts émanant du même compte. Aucun tweet, aucune publication rédigée n'apparaissent sur son fil.

Pris individuellement, ces éléments ne sont pas particulièrement probants. En revanche, combinés, l'authenticité du compte est fortement remise en question.

Des caractéristiques communes discrètes

"@FMarlair" est loin d'être le seul compte de ce type. Dans le cadre du hackathon [*de Sciences Po Saint-Germain-en-Laye, en partenariat avec le Ministère de l'Europe et des Affaires Étrangères*], nous avons pu nous rendre compte lors de concertations en équipe que chacun de nous avait rencontré des profils suspects, comprenant des éléments douteux communs.

A première vue, la plupart de ces comptes paraissent "normaux", partageant des contenus variés, affichant des descriptions fournies et publiant parfois plusieurs photos d'une même personne dans des situations différentes. On note donc un clair effort de démonstration d'authenticité dans la création de ces comptes.

Cependant, la répétition et la combinaison des éléments susmentionnés instille le doute. Parmi ceux-ci, on peut relever: un nom d'utilisateur inhabituel (jeu de mot, nom suivi de nombreux chiffres, noms mal orthographiés, etc) ; peu ou pas de photos personnelles ; une *bio* brève, incohérente ou tout simplement inexistante ; un ratio tweets / retweets extrêmement inégalitaire (en faveur des retweets) ; un contenu polémique ou polarisant, généralement de nature politique et actuelle ; une activité extrêmement importante à des plages horaires inhabituelles ; entre autres.

Par ailleurs, la principale force de Twitter étant l'influence que le réseau peut apporter, ce type de compte va d'abord chercher à augmenter son *following*, en s'abonnant massivement et aléatoirement à des "gros comptes" qui affichent plusieurs milliers d'abonnés. Souvent, ces abonnements ont lieu en nombre important dans un laps de temps très court.

Les indicateurs d'une activité mécanique peuvent se trouver en consultant directement le contenu du compte - par exemple, la répétition de phrases-types du genre "*Coucou, comment ça va?*" , "*Félicitations*!" ou encore "*Je peux te parler en privé?*" adressées à plusieurs comptes, parfois simultanément, est très parlante.
Ils peuvent aussi être mis en valeur par des outils comme les sites *foller.me* ou *accountanalysis.lucahammer.com*, qui fournissent des graphes analysant l'activité des utilisateurs sélectionnés, allant du nombre de tweets moyen horaire aux caractéristiques des contenus les plus partagés.

Compilation et classification

Nous avons constaté l'ensemble de ces caractéristiques communes durant la compilation et la classification de plus de 130 comptes Twitter en fonction de leur apparence, leur contenu, leur ratio abonnés/abonnements et leur rythme d'activité.

En partant d'une base relativement aléatoire d'une trentaine de comptes (mêlant des comptes que l'on savait être "réels" à des comptes jugés suspects trouvés à partir de mots-clés ou de sujets polémiques), nous avons progressivement élargi notre champ de recherche.
Il est rapidement apparu, dans ce processus, qu'il existe des réseaux liant parfois jusqu'à plusieurs dizaines de comptes suspects. Dans la plupart des cas, soit ces comptes se contentent de se suivre mutuellement, soit ils sont pleinement inscrits dans une logique de renforcement mutuel, retweetant, likant et commentant sommairement leurs contenus.
Il est particulièrement classique de trouver une configuration dans laquelle un "gros compte" suit un compte insignifiant, qui a pour seule activité de répéter le contenu publié par le "gros compte". Ce type de compte a été qualifié d'**amplificateur**. Cependant, il n'est pas possible à ce stade de déterminer s'il s'agit d'une activité commandée, achetée par le compte valorisé, ou bien d'une activité gérée par un tiers ou un particulier.

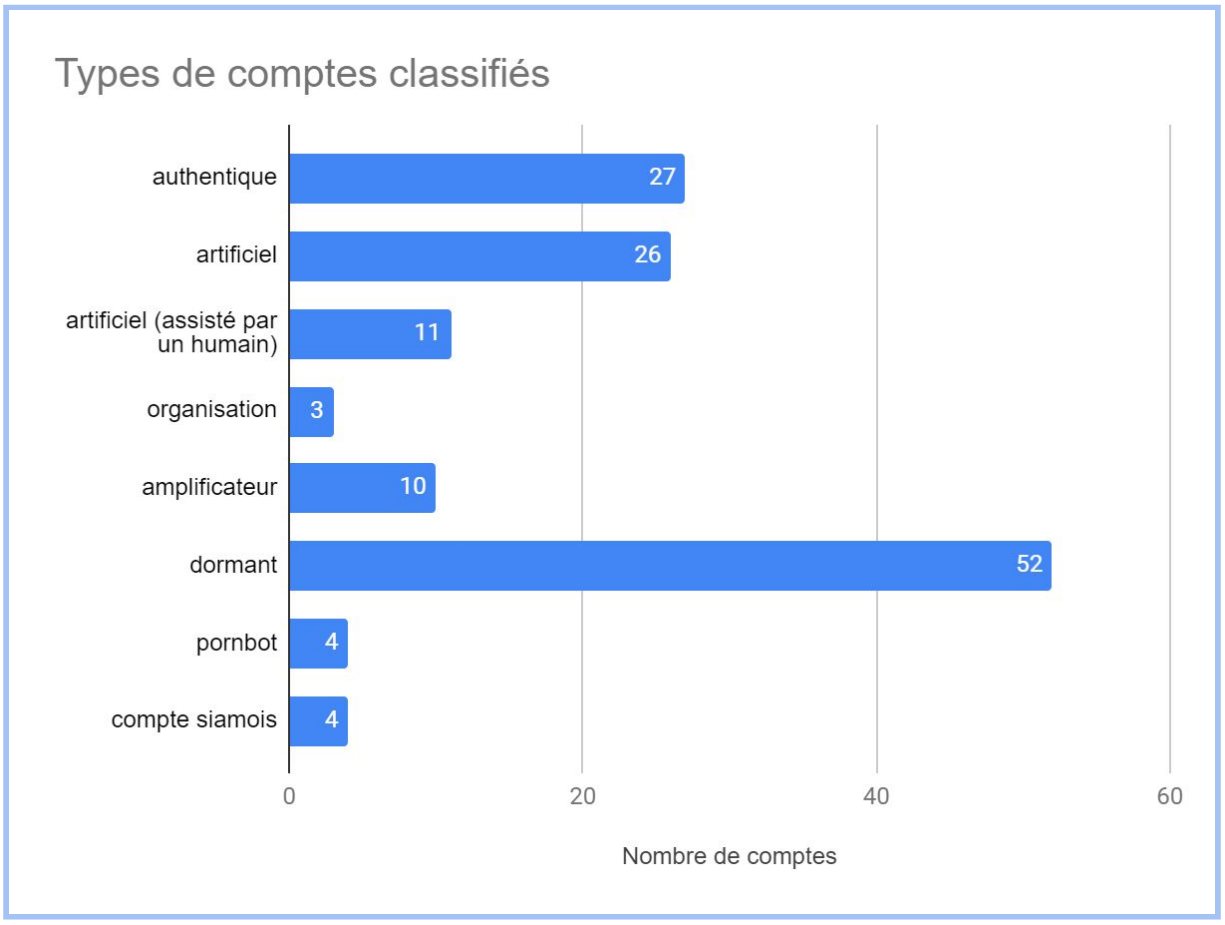
Ce système de réseau nous a permis de débusquer un nombre important de comptes suspects, en partant de comptes influents et polémiques, comme celui du comité français officiel de soutien au premier ministre italien Matteo Salvini @ SalviniFrance, et en cataloguant leurs abonnements.

En allant plus loin, nous avons constaté plusieurs sous-ensembles relativement homogènes parmi ces comptes artificiels. Outre les amplificateurs, les comptes affichant une activité quasi-nulle mais un nombre très importants d'abonnements - dans le but d'acquérir de plus en plus d'abonnés par principe de *follow back* - sont qualifiés de **dormants**. Ceux affichant une photo de profil aguicheuse, avec une activité principale de mention de personnes suivies de messages-types de *flirt* sont désignés comme **séducteurs**.

Ces deux types de comptes deviennent particulièrement intéressants lorsque l'on se penche sur ce qu'ils ont aimé ou ce qu'ils partagent: généralement, il s'agit de contenu polémique, polarisant et / ou politique. Pour les comptes dormants, on constate que les "réveils" (pics d'activité après une longue période d'inactivité de souvent plusieurs mois) ont souvent lieu à des périodes d'instabilité politique ou de naissance de mouvements sociaux. Ayant accumulé plusieurs centaines voire plusieurs milliers de *followers* au fil des mois d'inactivité, l'impact de leurs premières publications est d'autant plus significatif.

D'autres types de compte ont été identifiés comme **pornbots** (contenu à tendance pornographique vendant des services sexuels, mais qui, pour certains, aiment ou partagent soudainement du contenu politique), **comptes siamois** (photos de profil et descriptions similaires mais noms différents) ou encore **organisations** (comptes porte-parole d'entités officielles ou officieuses larges).

Sur la base de ces classifications, nous avons pu établir une représentation graphique relativement précise de la répartition de ces catégories parmi l'échantillon étudié:



Le projet d'algorithme de détection

Sur la base de ce travail de recherche et de classification, nous avons été à même d'identifier trois éléments-clés retrouvés combinés dans la quasi-intégralité des

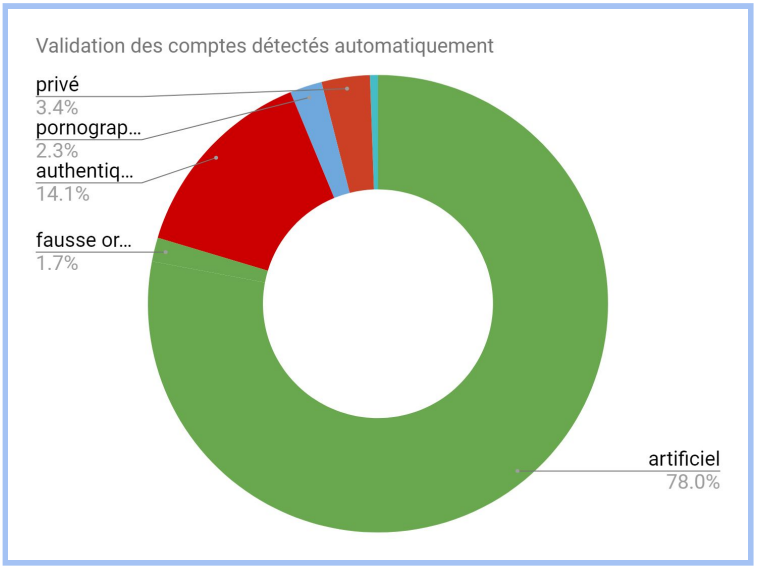
comptes jugés inauthentiques, à savoir: un faible volume d'activité sur Twitter (tweets, retweets et publications uniquement) ; un laps de temps d'existence court ; un ratio abonnés / abonnements particulièrement déséquilibré en faveur des abonnements.

En effet, une telle combinaison indique que pour une activité récente et faible - voire inexistante - ce type de compte attire tout de même des *followers* (ce qui peut s'expliquer par un nombre très élevé d'abonnements, reposant sur le principe du *follow back* mentionné plus haut ; ou par l'abonnement de comptes influents, qui auraient acheté des comptes artificiels via des sites webs spécialisés pour gonfler leur nombre d'abonnés, et se seraient alors automatiquement abonnés à leurs *followers* fictifs nouvellement acquis.)

Un premier algorithme de détection de faux comptes, comptes suspects et *bots* potentiels a donc été développé sur la base de ces trois critères, avec pour objectif principal de long-terme de contrer les dynamiques de désinformation une fois perfectionné.
Le choix a été fait de sélectionner les comptes ayant publié moins de 15 tweets, avec un ratio abonnements/abonnés supérieur à 3 et une date de création remontant au dernier mois au maximum.

En l'espace d'une nuit, l'algorithme a détecté 1708 comptes suspects correspondant aux critères énoncés. Après l'analyse de 10% de ces comptes par notre équipe, dans un processus de vérification, seuls 17.8% ont été jugés authentiques contre 78% de comptes artificiels. Il est important de préciser que près de 4% des comptes trouvés par l'algorithme sont en privé, et donc non-vérifiables.
Cette première étape aboutit donc à un outil au taux de fiabilité respectable de 78%.

Si des modifications et des précisions sont bien évidemment à apporter pour obtenir un taux de fiabilité satisfaisant sur le temps long, on peut considérer que la mise en oeuvre de cet outil est en bonne voie et peut réellement constituer un instrument efficace de lutte contre la désinformation sur Twitter, éventuellement adaptable à un plus grand nombre de plateformes et réseaux sociaux.


Validation des comptes détectés automatiquement
privé
3.4%
pornograp...
2.3%
authentiq...
14.1%
fausse or...
1.7%
artificiel
78.0%