Documentation de la méthode et des outils

I - Comment trouver des comptes suspects ou inauthentiques?

- ***Par réseau :***

Une première méthode de repérage de comptes suspects ou inauthentiques serait d'explorer les **profils des abonnés et des abonnements de comptes avec un following conséquent** (au minimum > 1K followers) ainsi que de **comptes certifiés** (hommes politiques, célébrités, organisations, partis politiques, etc.).

Dans la recherche de bots, il est logique de s'orienter vers des comptes **d'actualité et d'information** (type Le Monde Histoire, France Inter), des comptes **politisés** (type La France Insoumise) ou des comptes **polémiques** (contenant par exemple des noms comme ceux d'Alain Finkielkraut ou d'Eric Zemmour).

Généralement, les bots ou les comptes de trolls avec une visée d'influence et une portée politique s'abonnent à des comptes **reflétant leurs opinions** ou cherchent à **attirer l'attention** de comptes de ce genre.
On peut également remarquer que ces bots peuvent, dans leur vague d'abonnements, se concentrer sur différents **thèmes** en fonction des moments (il y aura ainsi d'abord un abonnement à 20 comptes de cuisine, puis 20 abonnements à des ligues de protection des animaux...).

Cependant le mécanisme fonctionne également en sens inverse. Certains comptes déjà conséquents ou influents peuvent chercher à renforcer leur **champ d'attention** par le biais de logiciels qui peuvent créer des comptes fictifs s'abonnant à l'acheteur - et auquel l'acheteur s'abonne automatiquement.
Il est donc utile de regarder à la fois les abonnés et les abonnements des "gros comptes".

Dans un second temps, il est intéressant de se pencher sur le réseau de comptes que l'on a **déjà définis comme étant suspects ou robotisés**, qui s'appuient régulièrement sur une méthode de **renforcement mutuel**. Ainsi, ce type de compte va avoir tendance à suivre des comptes similaires et à retweeter leur contenu - et vice-versa - créant des **noeuds d'interactions** utiles à la recherche de bots. Les mêmes contenus, y compris de la désinformation, sont donc partagés par des comptes de type similaire au sein d'un réseau qui se nourrit lui-même.

Ce type de réseau est assez facilement détectable, avec des **éléments similaires** (pas de bannière, photos impersonnelles ou images de banque de données, termes

identiques ou presque évoquant des classifications sociétales ou politiques, logos de type drapeau, croix, orthographe hasardeuse, etc.).

- ***Par sujets :***

Une deuxième méthode de repérage de comptes artificiels serait le ciblage de **sujets types**, en général **polarisants** ou s'inscrivant dans une **actualité polémique** (par exemple, l'affaire du hijab de running de Décathlon).

Pour ce faire, les outils peuvent être une recherche par mots-clés ou par hashtags. On peut alors cibler des **personnalités politiques** (ou des comptes apparentés, ou utilisant des noms de personnalités politiques comme le fait le compte Salvini_France, compte français de soutien au premier ministre italien), des **marques**, des **médias** populaires (RT_France), des mouvements, des communautés, etc. Il est intéressant ici de veiller au vocabulaire employé, qui peut inclure des termes clivants comme par exemple "Français de souche" ou "Juifs de France".

Un compte artificiel peut également se repérer par le retweet de certaines publications polémiques. En effet, un compte artificiel ou automatisé pourra se traduire par de **nombreux retweets,** notamment **par rapport au nombre des tweets ou des réponses**; parfois même sans légende ou bien simplement avec des hashtags.

- ***Par les caractéristiques extérieures :***

Par "caractéristiques extérieures", on entend le nom de l'usager, ses photos, sa bio et tous les éléments accessibles en consultant la "page de garde" du profil.

Sur cette base, une méthode pour repérer des comptes suspects ou artificiels peut être de regarder la date d'inscription de l'utilisateur, en lien avec le nombre de tweets, et le ratio nombre d'abonnements/abonnés.

Par exemple, un compte créé dans les deux derniers mois, avec plus de mille abonnements et beaucoup moins d'abonnés (près d'⅓), avec presque aucune activité (tweets ou retweets) peut être considéré comme un **compte dormant**, c'est-à-dire un compte cherchant à attirer l'attention d'utilisateurs, en général ciblés pour appartenir à une telle communauté, afin de par la suite, "se réveiller" et poster des contenus visés.
A l'inverse, certains comptes ont une activité quasiment exclusive de retweet, généralement d'un seul type de publication, portant sur un sujet précis, ou encore d'une série de comptes spécifiques; il s'agit alors de **comptes amplificateurs** dont la visée peut-être de booster l'activité de certains comptes affiliés, ou bien de polariser leur audience.

Il est également intéressant de prêter attention aux divers **signes**, caractères ou émoticônes apparaissant dans la *bio* (ex: chiffres, croix, drapeaux, smilies, etc), voire dans le nom du compte (qui par ailleurs incluent souvent des jeux de mots pour les comptes

trolls ou automatisés), qui peuvent aider à **identifier** une communauté ou un type de compte. De même, un français incorrect ou un biais international dans la façon de s'exprimer peuvent être révélateurs d'inauthenticité. A ce titre, la légende peut également aider à distinguer des comptes pilotés à **l'étranger**, dont une légende en français par exemple, pourrait comporter des fautes ou des erreurs de syntaxe (par exemple, la bio d'une personne le déclarait "homme d'action Sociales", la majuscule traduisant une origine anglo-saxonne: Social actions man).

Des indices peuvent aussi apparaître quant au **choix de la photo** de profil et de l'image de bannière. Il est utile de vérifier si la photo est référencée, privée, personnelle ou non. Il arrive de distinguer un compte où les personnes sur les photos de profils ne sont pas les mêmes dans le temps, suite à une nouvelle publication de photo.

Par ailleurs, on trouve régulièrement des photos de personnalités publiques, d'événements marquants, de drapeaux, ou bien des images incluant un **signe ostentatoire** d'appartenance communautaire (tatouage, signe religieux, etc…) ou bien avec un **objet** particulier, comme une arme.

Il est notamment aisé d'identifier des comptes avec des photos de profil aguicheuses, cherchant à obtenir davantage d'abonnés, pour par la suite peut-être publier du contenu polarisant; ou bien constituant simplement des comptes pornographiques. C'est par exemple le cas de certains **pornbots**, qui utilisent généralement des images d'individus réels pour attirer des followers, mais sur lesquels il est intéressant de se pencher lorsqu'ils commencent à liker et partager du contenu politique ou polémique.

Il est possible également de repérer des comptes ayant la même photo de profil, mais également la même date de création, les mêmes termes en *bio*, le même type de nom ou bien de même nationalité, indiquant une **création massive de comptes par un acteur unique**.

II - Comment classifier les comptes une fois trouvés?

Certaines caractéristiques communes permettent d'identifier les bots, mais ceux-ci ne semblent pas toujours servir les mêmes objectifs.

- ***En fonction du rythme***

Tout d'abord, une **fréquence élevée** de post (supérieure à 200 par jour environ) apparaît peu naturelle. S'il pourrait s'agir de réels individus derrière ces rythmes, cette activité pourrait difficilement concerner une personne isolée qui tweeterait simplement pour exprimer son opinion.
La fréquence des tweets de chaque compte Twitter peut être analysée grâce à des sites tels que foller.me ou accountanalysis, qui permettent aussi de voir la **proportion de**

tweets, retweets et likes parmi les tweets du compte. Un nombre élevé de retweets et likes peut aussi démontrer une **activité mécanique et moins personnelle** que la production de tweets.

Ce site nous indique aussi les **heures** à laquelle le compte produit le plus de contenu; si la production de contenu se **concentre** à certaines heures de la journée, et si ces plages se trouvent au milieu de la **nuit** du fuseau horaire français, alors que le compte se présente comme appartenant à un.e français.e, un doute peut être émis sur l'authenticité du compte.
C'est surtout l'accumulation de ces facteurs et leur répétition sur plusieurs comptes qui permettent de dégager une dynamique générale.

- ***En fonction du contenu:***

1 / De la présentation

La **présentation** du compte est un élément crucial dans la détermination de son authenticité. Par exemple, des comptes automatisés auront généralement une **description** (*bio*) sommaire, avec des **qualificatifs valorisants** et des **termes simples**, rassurants ; voire des phrases d'accroche positives ou **aguicheuses**.

En effet, une méthode classique pour augmenter son nombre d'abonnés est l'utilisation de phrases ou d'images attrayantes, appelant à la rencontre ou à suivre l'auteur.

Pour ce qui est **des photos et des images utilisées,** la question se pose de savoir si elles sont personnelles, ou s'il s'agit de photos de personnalités, voire si elles sont simplement issues d'une banque de données. Une simple recherche inversée permet de le déterminer. Bien évidemment cet indicateur n'est pas suffisant, mais il fait partie d'un faisceau d'indices probants dans la détermination de l'authenticité des comptes.

La présence de **liens** dans la bio est également importante; si ces liens sont morts ou redirigent vers des pages introuvables, il est probable que le compte ne soit pas authentique.

Enfin, un dernier élément que l'on peut trouver sur la page de présentation d'un compte est son nombre d'abonnés et son nombre d'abonnements. Un **ratio** indiquant au moins trois fois le nombre d'abonnements par rapport au nombre d'abonnés peut être considéré comme suspect. En effet, il peut sous-entendre que l'individu s'est massivement abonné à des comptes aléatoires dans l'espoir d'être suivi en retour et ainsi booster sa popularité et sa zone d'influence, ce que cherchent à faire les bots.

2 / Des publications

Le contenu des publications en elles-mêmes et pas seulement des *bios* est également un indicateur parlant. Le type de contenu (re)tweeté, dans le cas de comptes artificiels, est généralement **polémique** ou **polarisant**. Qu'il s'agisse d'actualités, d'intox,

de liens vers des articles de presse accompagnés ou non d'opinions, ces posts visent à **cliver** leur public pour créer ou renforcer des communautés d'opinion souvent radicalement opposées - et manipulables.

La **répétition systématique** de questions types ("Coucou, comment ça va?" ; "Je peux te parler en privé?") ou de phrases accrocheuses est un fort indicateur d'inauthenticité. Généralement, ce type de phrase n'est pas seulement tweeté mais s'accompagne de **tags de "gros comptes"**, choisis aléatoirement, avec pour but d'éveiller leur intérêt et de les inciter à suivre le bot.

Il est également de crucial de prêter attention à la qualité de l'activité du compte. S'agit-il presque uniquement de retweets, de partage de liens sans commentaire, ou encore de copiés-collés sans âme? Ou s'agit-il plutôt de tweets relativement longs, argumentés, avec un effort d'écriture et une orthographe correcte? Ces éléments permettent d'appuyer la distinction entre un compte réellement humain et un compte automatisé.

Un **volume** particulièrement important de (re)tweets (plusieurs centaines voire plusieurs milliers), surtout dans un laps de temps très court - si la personne s'est inscrite très récemment ou si elle tweete très (trop?) régulièrement dans la journée - indique aussi une action mécanique, inhumaine.

Pour que les messages soient transmis de manière innocente et discrète, on peut recourir à la création de personae. Ainsi, un bot pourra avoir une description fournie précisant ses opinions politiques, sa nationalité ou encore ses goûts, dans le but de **l'humaniser** et de permettre au public de **s'identifier** personnellement, pour mieux être influencé.
Cela passe par l'inclusion de **photos de personnes réelles**, qui à première vue paraissent authentiques, ou encore par le **partage de publications très populaires** sans lien particulier entre elles pour indiquer des centres d'intérêt variés.