Bots, Femmes et Hommes sur Tweeter

Catherine Ikae, Jacques Savoy

Université de Neuchâtel (Suisse) – {Catherine.Ikae, Jacques.Savoy}@unine.ch

Résumé

La communication sur les réseaux sociaux ne correspond ni à l'oral ni à l'écrit mais appartient clairement à un nouveau paradigme. Dans ce cadre, notre communication s'intéresse aux éléments linguistiques permettant d'identifier si un ensemble de tweets a été généré par un *bot* ou par un être humain. Dans ce dernier cas, notre système doit authentifier si ces tweets ont été écrits par un homme ou une femme. Cette étude se base sur 676 000 tweets rédigés en anglais et utilisés lors de la campagne d'évaluation PAN CLEF 2019. La distinction entre êtres humains et machines s'avère plus aisée que celle entre hommes et femmes. Ainsi, le taux de succès s'élève à environ 90 % pour la première tâche. On détecte la présence d'une machine lorsque le rapport Type Token (TTR) s'avère soit peu élevé (inférieur à 20 %) ou soit très élevé (50 % et plus). De même, la densité lexicale présente des valeurs très supérieures aux tweets rédigés par des êtres humains, de l'ordre de 80 % et plus. L'absence de pronoms personnels, peu d'adjectifs signalent également la présence d'un bot. Enfin, les thèmes repris par les machines concernent des domaines particuliers (*job*, *engineer*, *location*, *manager*, *itjob*, *business*, …). Distinguer si un ensemble de tweets a été émis par un homme ou une femme correspond à une tâche plus difficile dont le taux de succès avoisine les 70 %. En analysant ces différences, les hommes utilisent plus fréquemment des articles (*the, an*), certains pronoms (e.g., *he, it, they*), des conjonctions (e.g., *or*), des expressions (e.g., *fucking*), ou certains thèmes (e.g., *brexit*, *game, iot, security*). Les femmes sur-emploient les pronoms personnels (e.g., *me, you, our*) ainsi que *my* ou abordent des thèmes différents (e.g., *love*, *girl*, *social*, *video*). Au niveau des emojis, leur fréquence d'occurrence est plus forte chez les femmes que chez les hommes (88.2/10 000 formes vs. 60.3 for men). Si les hommes optent pour les formes , , , , , les femmes préfèrent les emojis suivants : , , , , , , , .

Abstract

Writing in social networks corresponds to a new type of communication different from the oral or written form. The main objective of this paper is to determine the linguistic aspects allowing us to identify whether a set of tweets has been sent by bot or a human being. In the latter case, the classifier must specify whether these tweets have been written by a male or a female author. This study is grounded on 676,000 tweets written in English and used during the 2019 CLEF PAN evaluation campaign. The classification between bots and humans is easier than between the two genders. For the first task, the accuracy rate can be around 90% while for the second the success rate is clearly lower. When a set of tweets has been sent by a bot, the TTR (Type Token Ratio) is either low (below 20%) or very high (larger than 50%). Similarly, the lexical density is significantly higher when the tweets have been written by a bot (larger than 80%). The low rate of personal pronouns or adjectives are other linguistic aspects that can be used to detect a bot. Finally, bots tend to focus on some particular topics (e.g., *job*, *engineer*, *location*, *manager*, *itjob*, *business*, …). To determine the author's gender is a rather more complex problem in which the best solutions achieved an accuracy rate around 70%. To distinguish between them, men tend to employ more frequently determiners (e.g., *the, an*), some pronouns (e.g., *he, it, they*), conjunctions (e.g., *or*), swear expression (*fucking*), or some specific topics (e.g., *brexit*, *game*, *iot*, *security*). Women over-use some personal pronouns (e.g., *me, you, our*) together with *my* as well as some topics (e.g., *love*, *girl*, *social*, *video*). Emojis occur more often with female writers than males (88.2/10,000 tokens vs. 60.3 for men). Some sequences are more frequently used by men such as , , , , , , and on the other hand, women prefer using , , , , , , , .

Mots-clés : Profilage d'auteur, réseaux sociaux, linguistique de corpus.

Keywords: Author profiling, social network, corpus linguistics.

1. Introduction

Durant la dernière campagne d'évaluation CLEF PAN 2019, les organisateurs ont proposé deux tâches distinctes sur un ensemble de tweets rédigés en langue anglaise. La première consistait à identifier de manière automatique les tweets émis par un bot de ceux rédigés par un être humain. Dans ce dernier cas, une seconde réponse devait être fournie par le système, à savoir si ces tweets étaient écrits par un homme ou une femme. Afin de permettre une détection sur un volume plus conséquent, l'identification ne se basera pas sur un tweet unique mais sur une série de 100 tweets formant un document ou un problème à résoudre.

L'identification d'une machine comme auteur d'un document pose un nouveau défi. En effet, les réseaux informatiques sont encombrés de pourriel (courriel non sollicité) et de tweets publicitaires. Éliminer ces informations ayant pas ou peu d'intérêt devrait réduire le trafic sur Internet de l'ordre de 45 % (voir www.statista.com). Par contre, distinguer si un message a été écrit par un homme ou une femme ne constitue pas une nouveauté en soi (Yule, 2010), (Pennebaker, 2011), (Eckert & McConnell-Ginet, 2013), (Schwartz *et al*., 2016). Cependant, étudier cette distinction sur des tweets constitue une tâche que les campagnes d'évaluation abordent depuis quelques années (Potthast *et al.*, 2019).

Dans cet article, ces deux problèmes de classification automatique doivent être résolues non pas avec un système type “boîte noire” mais en tenant compte des aspects linguistiques et des raisons expliquant les décisions proposées. Notre but ne correspond pas à implémenter et à tester différents *classifiers* afin d'atteindre la performance la plus élevée, mais de mieux comprendre les caractéristiques distinctes de chaque catégorie (soit entre bots et êtres humains, ou entre hommes et femmes). De plus, sachant que la communication sur Internet (chat, forum, courriel, réseaux sociaux) a été reconnu comme une nouvelle forme, parfois vue comme intermédiaire entre l'oral et l'écrit (Crystal, 2006), notre démarche doit mettre en lumière les éléments linguistiques propres à twitter.

Afin d'atteindre nos objectifs, nous disposons d'un ensemble de 676 000 tweets rédigés en anglais. En plus, nos analyses permettront de confirmer ou d'infirmer nos connaissances sur les aspects linguistiques proches à chaque genre. Afin de justifier nos conclusions, une procédure automatique d'identification validera la qualité des caractéristiques proposées pouvant déterminer clairement chaque groupe.

2. Travaux reliés

Dans le cadre du profilage, l'identification du genre de l'écrivain correspond à une première étape considérée comme la plus simple. En effet, la classification est binaire et des volumes importants de données peuvent être collectés. Toutefois cette classification automatique s'appuyant sur des différences de style entre hommes et femmes s'avère loin d'être parfaite (Pennebaker, 2011), (Schwartz *et al*., 2016).

Pour distinguer le genre, nous savons que les femmes recourent plus fréquemment aux mots émotionnels (Pennebaker, 2011), (Rangel et Rosso, 2016) ou indiquent leur certitude (e.g., *always, must*). Ces caractéristiques linguistiques s'avèrent plus difficile à déterminer que les pronoms personnels ou les articles définis utilisés plus souvent par les hommes. Afin de faciliter la détection des émotions, des listes de mots ou d'expressions peuvent être établies comme celles incluses dans le système LIWC (*Linguistic Inquiry and Word Count*) (Tausczik & Pennebaker, 2010). D'autres listes de mots peuvent refléter d'autres aspects linguistiques (e.g., les verbes modaux) voire des thèmes particuliers (e.g., avec le système LSD (*Lexicoder Sentiment Dictionary*) (Young & Soroka, 2007)).

Comme on pouvait s'y attendre, certains sujets apparaissent plus fréquemment auprès de l'un des deux genres (e.g., sports, argent *vs.* shopping, amis) (Argamon *et al*., 2008), (Schwartz *et al*., 2013). Basé sur un corpus comprenant 19 320 auteurs de blogs (50 % de chaque genre), la machine s'avère capable d'identifier correctement le genre de l'auteur pour environ 72 % des blogs (longueur moyenne : 7 250 mots) (Argamon *et al*., 2008). En considérant également les thèmes, ce taux de réussite augmente à 76 %. Ainsi, les hommes parlent plus de la technologie (e.g., *game*, software, Linux) tandis que les femmes écrivent plus au sujet des amitiés et des relations sociales (e.g., *love*, *cute*, *mom*).

Depuis 2010, plusieurs questions de stylométrie ont été étudiés dans le cadre des campagnes CLEF PAN (e.g., l'attribution d'auteur, la détection de plagiat, le profilage d'auteur) (Rosso *et al*., 2019). Les tâches sont proposées dans différentes langues mais l'anglais reste le choix le plus fréquent. Dans le cadre de profilage d'auteur, le genre et le groupe d'âge de l'auteur sont les facteurs à déterminer. Parfois, on s'intéresse également à détecter des traits psychologiques (Coppersmith *et al*., 2014), (Boyd & Pennebaker, 2017) ou des signes de dépression (Neuman, 2016), (Guntuku *et al*., 2017). En s'appuyant sur des blogs ou des tweets, les approches les plus performantes recourent à des stratégies de représentation différentes comme les sacs de mots, les *n*-grammes de mots ou de lettres, des séquences de parties du discours ainsi que d'autres indices stylistiques (pourcentage de lettres en majuscules, pourcentage d'émojis, longueur moyenne des mots ou des phrases).

Ainsi en 2014 (Rangel *et al*., 2014), le recours à la régression logistique a permis d'obtenir la meilleure performance tandis qu'en 2015 un modèle SVM (Rangel *et al*., 2015) s'est hissé au premier rang. En 2016 (Rosso *et al*., 2016), la meilleure qualité de réponse a été obtenue par un modèle de régression logistique basé sur une combinaison de mots, *n*-grammes de mots complétés par quelques autres éléments stylistiques. En 2017 (Potthast *et al*., 2017) un modèle SVM linéaire basé sur des uni-grammes et bi-grammes de mots ainsi que des 3-grammes à 5-grammes de lettres a obtenu le score le plus élevé. En 2018 (Rangel *et al*., 2018), un modèle similaire a présenté la performance la plus forte. Finalement en 2019 (Rangel et Rosso 2019), la régression logistique fut la meilleure solution en s'appuyant sur des *n*-grammes de mots et de lettres. Si les performances obtenues ne peuvent pas être directement comparées d'une année à l'autre (les collections de test (et donc la difficulté) ne sont pas similaires), les campagnes d'évaluation CLEF PAN ont mis l'accent sur la prédiction et ont négligé la justification ou la description des données. Les systèmes ayant les meilleures performances s'avèrent très complexes et se basent sur des attributs très nombreux et difficiles à interpréter (e.g., avec des *n*-grammes de lettres). Cet article vise précisément à combler cette lacune en cherchant les éléments linguistiques distinctifs entre les diverses catégories.

De plus, la collection fournie par la campagne d'évaluation CLEF PAN 2019 nous permet d'analyser la détection automatique du genre de l'auteur sous l'hypothèse que ce nombre soit limité à deux. En effet, nous ne savons pas si les personnes LGBT présentent ou non des traits linguistiques distincts. Par contre, la classification entre des textes écrits par des êtres humains ou des bots constitue un nouveau défi. A notre connaissance, les travaux précédents n'ont pas encore étudié les caractéristiques et le degré de difficulté de cette tâche. Elle représente donc une direction de recherche inédite.

3. Le corpus et la méthodologie d'évaluation

Afin de détecter automatiquement si un ensemble de tweets a été envoyé par une machine ou par un être humain, nous disposons d’un corpus généré lors de la campagne d'évaluation CLEF PAN 2019. Dans ce cadre, un document correspond à 100 tweets provenant de la

même source. Le tableau 1 reprend quelques statistiques sur le corpus étudié. Celui-ci comprend 6 760 documents (soit 100 tweets) correspondant à 676 000 tweets. La distribution entre bots et êtres humains est parfaitement équilibrée, de même que la répartition entre hommes et femmes. Nous retrouvons donc 3 380 documents émis par des machines et 1 690 par des hommes et la même valeur pour les femmes (Soit un total de 3 380 pour les êtres humains).

Lors de la campagne d'évaluation CLEF PAN, les participants disposaient de 4 120 documents (rédigés pour 2 060 par des bots, et 2 060 par les êtres humains) pour effectuer la mise au point de leur système. A cet ensemble d'entrainement, on peut ajouter les 2 640 documents (1 320 bots, 1 320 êtres humains) correspondant au jeu d'évaluation utilisé lors de la campagne. Nos évaluations reprennent exactement cette même subdivision.

	Bots	**Hommes / Femmes**
Nombre de documents	3 380	1 690 / 1 690 Humain : 3 380
Nombre de tweets	338 000	169 000 / 169 000 Humain : 338 000
Longueur moyenne (formes) (écart-type)	2 129 (1 248)	2 021 / 2 087 (516 / 524)
Taille du vocabulaire (vocables)	152 719	136 793 / 142 310 Humain : 229 937
Vocables ayant une fréquence supérieure ou égale à dix	21 542	16 633 / 16 431 Humain : 26 825
Nombre de formes	7 205 309	3 416 182 / 3 528 715 Humain : 6 941 517

***Tableau 1** : Quelques statistiques sur notre corpus de tweets.*

En analysant le contenu de chaque document (voir tableau 1), on constate que le nombre moyen de formes par documents s'élève à 2 129 pour les bots, contre 2 021 pour les hommes ou 2 087 pour les femmes. En moyenne, les machines produisent des documents un peu plus long (de l'ordre de 10 %). Par contre, l'écart-type de cette distribution présente également une valeur plus importante (1 248) que celle associée à la distribution des hommes (516) ou des femmes (524).

Au niveau du vocabulaire, les machines présentent une richesse lexicale moindre (152 719 vocables ou mots distincts) contre 229 937 pour les humains et ceci pour le même nombre de documents. Entre les deux genres, les femmes présentent un plus grand nombre de vocables que les hommes (142 310 vs. 136 793). En comptant le nombre de formes, des volumes similaires se retrouvent chez les bots (7 205 309) que chez les humains (6 941 517).

Lors de notre prétraitement, nous considérons les signes de ponctuation comme des formes, de même que pour les émojis. Ainsi depuis la séquence “Paul’s books!!!”, nous en tirons la suite {paul ’ s book !!!}. Les lettres en majuscules ont été remplacées par leur équivalent en minuscules. La répétition des signes de ponctuations ou des emojis est considérée comme un seul vocable. Les hyperliens (e.g., http:/t.co/OLtFu5aPrm) sont remplacés par le vocable "hyperlink". Enfin, nous avons appliqué un enracineur léger (le S-stemmer de (Harman, 1981)) afin de supprimer le pluriel des noms. En effet, la solution proposée par Porter s'avère trop agressive, réduisant par exemple le mot “organization” à “organ”.

Afin d'illustrer nos deux problèmes de classification, le tableau 2 illustre par quelques exemples de tweets pour les trois catégories de notre corpus.

Catégorie	**Tweet**
Bot	Lead Security Engineer: Lead Security Engineer REQ# 1800002V Cygnacom Solutions is currently looking for a Senior PKI Engineer/Architect to join our Entrust consulting services team at our Washington, DC area location. https://t.co/gnEp7cIblc
Bot	RT @MikeQuindazzi: 5 #autonomous cooking solutions. #ai #robot chefs, #smart ovens, #robotics, #iot https://t.co/WG5jwLkEC3
Bot	President Obama Again Pushes Congress to Act on Student Loans: President Barack Obama, with Education... http://t.co/RjHazw8u
Bot	20:3 And cast him down: deliver my people shall do no wrong, do no service.
Femme	Thoroughly enjoying folding blankets whilst listening to the deadly new single from @spiesboys https://t.co/oakQqpqxVH
Femme	Enjoying today's weather since it's supposed to start getting into the 30s tomorrow #100happydays http://t.co/OLtFu5aPrm
Femme	@little_magpie1 Definitely!! I hope yours feel better soon!
Femme	Day 112: Ummmmmm whaaaaattttt!!!! @mindykaling and I are one step closer to being bffs! … https://t.co/ltBbyOPuCC
Homme	RT @piagnone: where are you on the big cow political compass? https://t.co/d8yWXVSZk3
Homme	U know when u're watching a hockey game of 2 teams u dont cheer for & u dont really care. Then all of a sudden a team scores and you cringe.
Homme	RT @kkellett10: Chilling with the boys! Sunday funday, cottage, beers & sports! @mike_heslin @eric_forsyth @markymark0987 http://t.c ...
Homme	@RosemaryMacCabe Yeah, tall and buff... your standard self-centred bland flakes! Ahh thanks Rosemary xx

***Tableau 2** : Quelques exemples de tweets avec leur catégorie.*

4. Vocabulaire caractéristique

Face à un vocabulaire volumineux (e.g., 152 719 vocables apparaissent dans des tweets émis par des bots), une analyse complète s'avère difficile et pouvant conduire à des résultats quelque peu décevant (par exemple, en mettant en lumière une caractéristique présente uniquement dans un ou deux tweets). Dans ce but, les mots possédant une fréquence d'occurrence inférieure à dix ont été supprimés. En effet, le style d'une catégorie ne devrait pas dépendre d'éléments rares. De plus, notre étude doit se focaliser sur les éléments stylistiques qui sont à la fois *frequent and ubiquitous* comme le soulignent Biber & Conrad (2009). Avec cette première sélection, on remarque que la taille du vocabulaire se réduit de 90 % (voir tableau 1). Par exemple, le vocabulaire usité par les hommes passe de 136 793 vocables à 16 633.

Ensuite, nous avons calculé le score Z pour chaque vocable (Muller, 1992). Si cette valeur est supérieure à trois, le terme est sur-employé pour la catégorie étudiée (par exemple, les bots, les hommes ou les femmes). Pour déterminer sa valeur, on compare le nombre d'occurrence du vocable t_i dans la catégorie (la variable a dans l'équation 1) avec celle qui serait attendue selon une répartition uniforme entre toutes les catégories. Cette dernière valeur se calcule

(voir équation 1) en tenant compte de sa probabilité d'occurrence définit par la fréquence relative (*rtf*) de ce terme dans le corpus ($rtf_i = tf_i / n$, avec n le nombre total de formes dans le corpus).

$$\text{Z score}(t_j) = \frac{\left(a - rtf_i \cdot n'\right)}{\sqrt{¿¿¿}} \quad (1)$$

dans laquelle la variable n' le nombre de formes dans la catégorie étudiée.

Le tableau 3 indique les termes ayant les scores Z les plus élevés pour les différentes catégories. En se focalisant sur les bots, on constate que les thèmes abordés constituent une source d'attributs significatifs pour cette classe comme *job*, *hiring*, *engineer*, ou *developer*. De plus, le mot *urllink* signale que cette catégorie se caractérise par l'emploi récurrent de liens hypertexte (e.g., https://t.co/WG5jwLkEC3). Enfin, les symboles de ponctuation s'avèrent également des indices pertinents (e.g., **- – /** ou **,**).

Rang	**Bots**	**Humains**	**Hommes**	**Femmes**
1	129.49 –	410.88 @	44.85 **.**	28.29 #
2	114.14 **-**	217.93 rt	28.03 the	25.31 my
3	112.50 job	126.58 i	22.29 **..**	24.10 so
4	94.48 urllink	122.66 _	21.73 •	23.03 **…**
5	90.27 developer	86.91 this	20.99 iot	22.60 mailonline
6	86.21 engineer	83.53 my	20.56 he	22.35 me
7	83.51 /	80.46 **!**	19.63 that	22.14 rt
8	79.70 software	74.62 '	19.05 golfbreak	21.51
9	71.89 **,**	62.98 s	17.44 game	20.86 thank
10	65.35 hiring	61.62 on	16.96 they	20.51 love

***Tableau 3** : Les termes les plus caractéristiques selon les différentes catégories.*

Pour les humains, le symbole @ (e.g., @rogerfederer) et le vocable *rt* indiquent la présence fréquente de mentions et de retweets. Comme autre éléments linguistiques caractéristiques, on peut indiquer l'usage du génitif anglais ('s) ou le pronom *I* (je/moi).

Si l'on distingue entre les hommes et les femmes, on découvre que les hashtags (#) ou les retweets (*rt*) sont plus fréquents dans la catégorie féminine. Cette dernière se caractérise également par le pronom *me* ou *my*, les relations sociales (*thank*) ou un thème (*love,*). Plus loin dans la liste, on retrouve les autres pronoms personnels comme *I*, *you* (*your*), *us*, *we* (*our*), *her* ou *she*. Les relations humaines sont également présente via les termes *x*, *xx* ou *xxx* (bisous), le *hi*, ou *please*, voire le *hashtag #girlstreamer*. Les femmes sont également sensibles à la technologie (e.g., *mailonline*) ou via les mots *apps*, ou *video*. On les dit moins sensible à l'argent (Argamon *et al*., 2009) mais les tweets indiquent que les mots *account* et *credit* sont sur-employés par les femmes.

Les hommes possèdent des sujets de conversation distincts (*game*, *iot* (Internet of Things), *golf, playlist, stanleycup, marijuana, theatre*). Au niveau linguistique, on retrouve l'usage prépondérant des articles (*the*, *that*) ou des conjonctions (e.g., *but, or*) mais, ce qui un peu surprenant, les pronoms personnels *he*, *they*, *him*, ou *it*. Leur langage ne manque pas de couleurs (e.g., *fucking*) et ils recourent souvent à la négation (e.g., *not*) contrairement à une étude précédente (Pennebaker, 2011).

La distinction entre bots et êtres humains peut également s'opérer avec d'autres variables stylométriques comme celles reliés à la richesse lexicale. Ainsi, le rapport TTR (Type Token Ratio) permet de mesurer une telle richesse lexicale pour un document. Ce rapport se calcule

en comptant le nombre de vocables (taille du vocabulaire) divisé par le nombre de formes (or longueur du document). Le résultat varie entre 0 (ou 0 %) et 1.0 (ou 100 %). Une valeur élevée indique que l'auteur emploie un vocabulaire riche et diversifié ou qu'il aborde des thèmes fort variés nécessitant des termes différents. A l'inverse, une valeur faible indique des répétitions des expressions afin, souvent, d'insister sur quelques points jugés importants. Toutefois, ce rapport TTR s'avère peu stable lors que la longueur du texte s'accroit (Baayen, 2008). Dans notre cas, les textes demeurent dans des tailles comparables et cet inconvénient peut être ignoré.

Dans la distinction entre des tweets émis par un bot ou rédigés par des humains, la mesure de la richesse lexicale peut s'avérer utile. Ainsi, la figure 1 expose la distribution des valeurs TTR pour les deux catégories. On remarque aisément que des valeurs inférieures à 0,2 signalent la présence d'un bot. A l'autre extrémité, des valeurs supérieures à 0,5 correspondent plus à une machine qu'à un être humain.

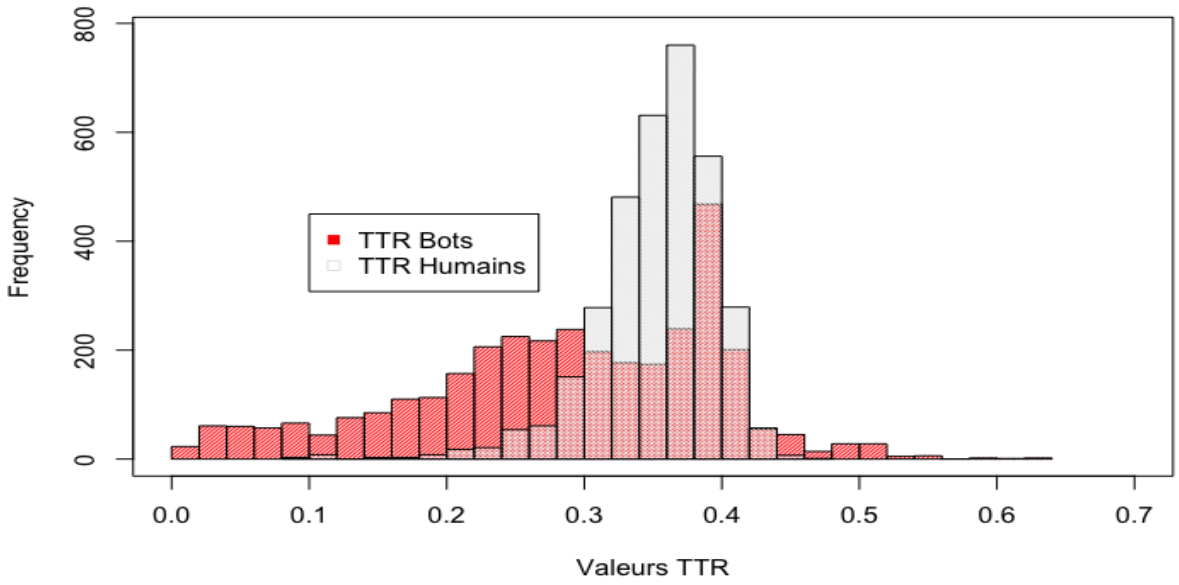

Figure 1 *: Rapport TTR (Type Token Ratio) pour les tweets émis pas des bots ou des humains.*

Pour être plus précis, la moyenne TTR pour les bots s'élève à 0,287 (écart-type : 0,11) tandis que pour les humains cette moyenne est de 0,353 (écart-type : 0,04). Dans ces deux distributions, on rencontre 685 documents émis par des bots ayant une valeur TTR inférieure à 0,2 mais seulement 25 documents rédigés par des êtres humains. En fixant le seuil à 0,46, on observe 88 documents rédigés par des machines contre un seul par un humain.

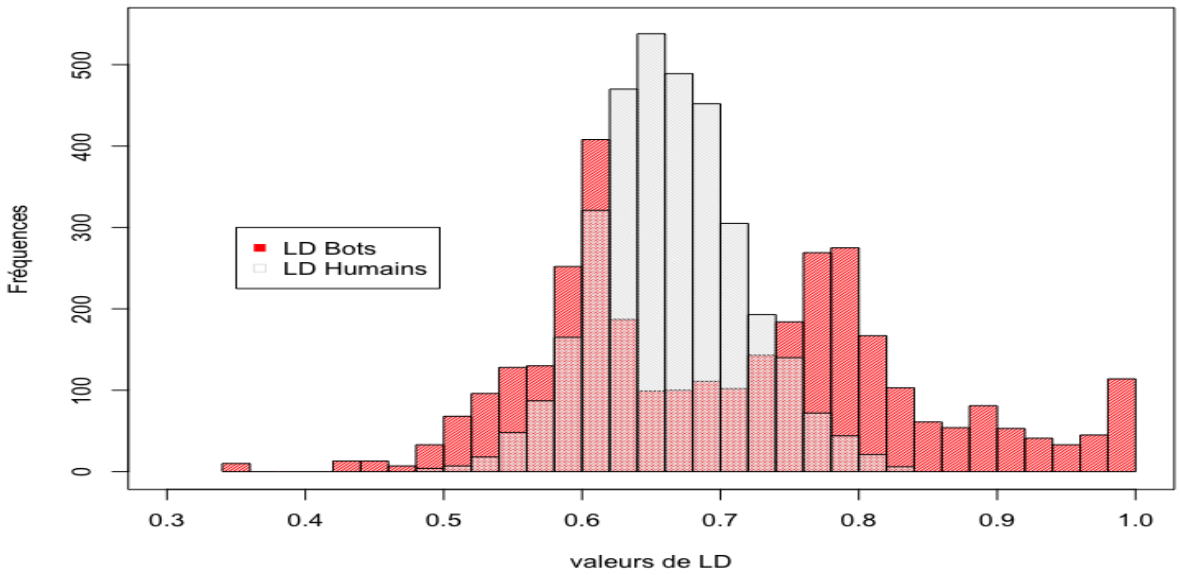

***Figure 2** : Densité lexicale pour les tweets émis pas des bots ou des humains.*

Comme deuxième mesure, nous avons analysé la densité lexicale mesurée par le pourcentage de noms, d'adjectifs, d'adverbes et de verbes dans un texte (sans tenir compte de la ponctuation, des emojis et autres symboles (#, @, etc.)). Pour les bots, cette mesure possède une moyenne de 0,708 (écart-type : 0,129) tandis pour les humains la moyenne s'élève à 0,664 (écart-type :0,053). Les histogrammes de ces deux distributions sont repris dans la figure 2. A nouveau, les valeurs extrêmes indiquent la présence d'un bot. Ainsi, on retrouve 484 documents émis par des bots ayant une densité lexicale supérieure à 0,84 mais aucun par des êtres humains. De même, avec une limite à 0,5, on n'observe que quatre ensembles de tweets écrits par des humains mais 76 par des bots.

Rang	**Bots**	**Hommes**	**Femmes**
1	2 531	1 320	1 806
2	2 156	626	688
3	2 154	472	632
4	2 019	321	530
5	2 017	259	529
6	1 997	220	524
7	1 619	218	469
8	1 009	209	469
9	953	208	332
10	890	200	274

***Tableau 4** : Les emojis les plus fréquent par catégories.*

La distribution des emojis par catégorie révèle également une nette différence entre les bots et les êtres humains. Les tweets émis par une machine contiennent plus d'emojis d'indications (le doigt pointé dans une direction, e.g.,), des objets (e.g., ,), des animaux (e.g.,), des places (e.g.,) ou des drapeaux (e.g.,)[1]. Pour les humains, ce sont les emojis de visage qui occupent les rangs les plus hauts, avec une préférence assez nette pour l'émoji . Ce dernier correspond aussi à la séquence la plus fréquente avec trois emojis de ce type. On notera que ces emojis signalent une émotion qui est le plus souvent positive (sauf avec ou

1 Dans ce cas, une ambiguïté demeure entre la dénomination d'un pays ou la croix rouge.

). Comme autre distinction entre les deux genres, les femmes introduisent plus d'emojis dans leurs tweets soit 8,82 ‰ contre 6,03 ‰ pour les hommes.

5. Classification automatique

Le problème posé lors de la campagne d'évaluation CLEF PAN 2019 s'articule en premier lieu sur l'identification des tweets écrits par des bots ou des humains. Dans un second temps, le système doit distinguer les tweets rédigés par des hommes de ceux envoyés par des femmes. Comme notre analyse portait sur les termes ayant des scores Z élevés, nous avons repris ces mots ou expressions comme attributs pour une classification automatique. Afin de démontrer qu'un nombre restreint de termes suffisent à obtenir une performance similaire à une approche basée sur un grand nombre d'attributs, nous avons comparé les taux de succès avec 200 ou 2 000 termes. Ces derniers sont choisis en fonction de leur score Z, en prenant soin d'extraire le même nombre pour chaque catégorie. Ainsi, pour un total de 200 attributs choisis, 100 reflètent l'écriture des bots et 100 celle des êtres humains (ou 100 le style masculin et 100 le féminin lorsque que le but est de distinguer les deux genres).

Cette analyse de performance s'appuie donc sur des documents contenant 100 tweets et ayant une longueur moyenne d'environ 2 100 mots. Chaque document est représenté par un vecteur dont la longueur correspond au nombre d'attribut retenu (indiqué par la variable *m*) . Pour chacun, la valeur numérique correspond à la fréquence absolue dans le document multipliée par le score Z du terme. Par exemple si l'on doit distinguer les hommes des femmes et en se basant sur le tableau 3, si le terme *the* étant présent cinq fois dans une série de tweets, son poids sera égale à 28,03 x 5.

L'objectif poursuivi n'est pas d'atteindre la meilleure performance mais de savoir si une réduction importante du nombre d'attributs (e.g., limité à $m = 200$) permet d'obtenir un taux de réussite comparable à un modèle possédant un grand nombre d'attributs (par exemple, $m = 2\,000$).

Comme classifier, nous avons retenu l'approche des *k* plus proches voisins (*k*-NN) avec des valeurs pour *k* variant entre 1 et 7 (voir tableau 5). Comme mesure de distance, nous avons choisi la fonction Manhattan (définie dans l'équation 2) et celle de Canberra (voir équation 3) (les deux appartenant à la famille L^1). Dans ces formules, les vecteurs sont symbolisés par les lettres majuscules A et B tandis que chaque composante est représentée par a_i ou b_i.

$$Distance_{Manhattan}(A, B) = \sum_{i=1}^{m} |a_i - b_i| \quad (2)$$

$$Distance_{Canberra}(A, B) = \sum_{i=1}^{m} \frac{|a_i - b_i|}{(a_i + b_i)} \quad (3)$$

La fonction de Manhattan est souvent proposée dans des études stylométriques (e.g., le modèle Delta de Burrows (2002)). Toutefois, si elle correspond à un choix raisonnable, elle n'apporte pas souvent la meilleure performance. En revanche, la fonction Canberra permet souvent d'accroitre le taux de succès (Kocher & Savoy, 2017).

Modèle	Nombre de termes	Bots vs Humains	Hommes vs. Femmes
k-NN, *k*=3, Manhattan	**2 000**	**88,82**	**61,36**
k-NN, *k*=3, Manhattan	200	90,30	59,85
k-NN, *k*=5, Manhattan	**2 000**	**89,39**	**61,36**
k-NN, *k*=5, Manhattan	200	90,45	60,75
k-NN, *k*=3, Canberra	**2 000**	**93,00**	**59,55**
k-NN, *k*=3, Canberra	200	92,65	65,99
k-NN, *k*=5, Canberra	**2 000**	**93,22**	**62,27**
k-NN, *k*=5, Canberra	200	93,83	69,02
Régression logistique	**2 000**	**84,69**	**72,05**
	200	87,23	76,43

***Tableau 5** : Les taux de succès en fonction du nombre de termes.*

Les taux de réussite reportés dans le tableau 5 ont été obtenu avec 2 640 documents pour la classification entre bots et humains. Pour la distinction entre hommes et femmes, le jeu de test comprenait 1 320 documents. Le jeu d'entraînement contenait 2 060 documents émis par des bots et 1 030 par des hommes et le même nombre pour les femmes.

En conclusion, lorsque le système dispose de 2 000 attributs les performances obtenues ne présentent pas de manière systématique une valeur plus élevée que celles atteintes avec un nombre restreint d'attributs (200 dans notre étude). Bien que le modèle *k*-NN soit sensible à la présence d'attributs non pertinents (Efron & Hastie, 2016), ceux retenus (même dans le cas où m = 2 000) possédaient toujours un score Z positif et supérieur à trois. En recourant à la régression logistique, la même conclusion peut être tiré.

6. Conclusion

Notre analyse porte sur les éléments stylistiques présents dans les tweets permettant de différencier ceux émis par une machine de ceux rédigés par un être humain. Dans ce dernier cas, nous devons également distinguer entre les tweets écrits par un homme ou une femme. Créé lors de la campagne d'évaluation CLEF PAN 2019, le corpus étudié comprend 676 000 tweets rédigés en anglais.

Les traits distinctifs d'une machine correspondent à des valeurs TTR soit très forte (supérieure à 0,5) soit très faible (inférieure à 0,2) (voir figure 1). La machine a donc tendance soit à répéter des messages similaires (TTR très faible) soit à générer un ensemble très disparate d'information (TTR élevé). De même, les valeurs de densité lexicale (LD) peuvent apparaître comme très fortes (supérieures à 0,84) soit très faibles et inférieures à 0,5 (voir figure 2). La machine ne construit pas souvent des phrases complètes mais aligne des noms et verbes (LD fort) ou se limite à quelques liens hypertextes (LD faible).

Les êtres humains se distinguent par l'emploi fréquent de retweets (rt) ou des mentions (symbole @) (voir tableau 3). Ils utilisent aussi fréquemment le génitif ('s) et les pronoms (e.g., *I, me*). Au niveau de la ponctuation, les différences sont également très visibles entre l'usage de tirets (e.g., - – /) par la machine et la présence de points chez les humains. Cela indique la nécessité pour les humains de faire des phrases d'une part, et d'autre part, de recourir à l'emploi des points de suspension.

Entre hommes et femmes, on observe clairement des distinctions. Ainsi les femmes recourent plus facilement aux hashtags (#) et retweets (rt) que les hommes. Des études précédentes soulignaient que les femmes sur-employaient les pronoms personnels. Notre étude confirme

ce phénomène mais en partie seulement. En effet, certains pronoms de la troisième personne sont significativement plus fréquents chez les hommes (e.g., *he*, *they*, *it*).

L'emploi des emojis permet également de distinguer les bots des êtres humains (voir tableau 4). Ainsi les premiers possèdent de plus haute fréquence pour les emojis d'animaux, d'indications de direction (e.g.,), d'objets ou de drapeaux (e.g., ,). Pour les êtres humains, les visages sont plus fréquents (e.g., ,). Par contre, la distinction entre ceux plus récurrents chez les hommes ou les femmes s'avère plus difficile à déterminer. Toutefois, les femmes tendent à inclure un peu plus d'emojis dans leurs tweets.

De plus, la technologie n'est pas l'apanage exclusif des hommes (e.g., *iot*, *playlist*, *macintosh*) car les termes *mailonline*, *video*, ou *apps* sont surreprésentés chez les femmes. Au niveau des autres thèmes abordés par les tweets, chaque catégorie expose des sujets qui lui sont propres comme les termes *job*, *technology*, *business*, *application* pour les bots, *girlstreamer*, *travel*, *abortion*, *social*, *homebusiness* pour les femmes ou *golfbreak*, *theatre*, *marijuama*, *brexit*, *beer*, ou *celtic* pour les hommes.

Lors de la classification automatique, les systèmes obtiennent des taux de succès élevés (proche des 93 à 95 %) dans la distinction entre tweets émis par une machine ou ceux rédigés par un être humain. Les deux catégories possèdent des caractéristiques très distinctes. Pour distinguer entre hommes et femmes, les taux de réussite s'avèrent plus faibles (70 % dans cette étude, ou 80 % lors de la campagne CLEF PAN 2019). Pour s'en convaincre, on peut observer les emojis les plus usités par les deux genres (voir tableau 4) et constater un large recouvrement.

Références

Argamon, S., Koppel, M., Pennebaker, J.W. & Schler, J. (2009). Automatically profiling of the author of an anonymous text. *Communications of the ACM*, 52(3), 119–123.

Baayen H.R. (2008). Analysis linguistic data: A practical introduction to statistics using R. Cambridge University Press, Cambridge.

Biber, D. & Conrad, S. (2009). *Register, Genre, and Style*. Cambridge University Press, Cambridge.

Boyd, R.L., & Pennebaker, J.W. (2017). Language-based Personality: A new approach to Personality in Digital World. *Current Opinion in Behavioral Sciences*, 18, 63–68.

Burrows, J.F. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. *Literary & Linguistic Computing*, 17(3), 267–287.

Coppersmith, G., Dredze, M. & Harman, C. (2014). Quantifying Mental Health Signals in Twitter. In *Workshop on Computational Linguistics and Clinical Psychology* (ACL), 51–60.

Crystal, D. (2006). *Language and the Internet*. Cambridge University Press, Cambridge.

Eckert, P. & McConnell-Ginet, S. (2013). *Language and Gender*. Cambridge University Press, Cambridge.

Efron, B., & Hastie, T. (2016). *Computer Age, Statistical Inference*. Cambridge University Press, Cambridge.

Eisenstein, F.J. (2019). *Introduction to Natural Language Processing*. The MIT Press, Cambridge.

Guntuku, S.C., Yaden, D.B., Kern, M.L., Ungar, L.H., & Eichstaedt, J.C. (2017). Detecting depression and mental illness on social media: An integrative review. *Current Opinion in Behavioral Sciences*, 18, 43–49.

Harman, D. (1991). How effective is suffixing? *Journal of the American Society for Information Science*, 42(1), 7–15.

Kestemont, M., Stamatatos, E., Manjavacas, E., Daelemans, W., Potthast, M. & Stein, B. (2019). Overview of the cross-domain authorship attribution task at PAN 2019. In *Working Notes of the CLEF 2019 Evaluation Labs*, vol. 2380, CEUR, Aachen.

Kocher, M. & Savoy J. (2017). Distance Measures in Author Profiling. *Information Processing & Management*, 53(5), 1103-1119.

Manning, C.D. & Schütze, H. (2000). *Foundations of Statistical Natural Language Processing*. Cambridge University Press, Cambridge.

Muller, C. (1992). *Principes et Méthodes de Statistique Lexicale*. Honoré Champion, Paris.

Neuman, Y. (2016). Computational Personality Analysis: Introduction, Practical Applications and Novel Directions. Springer-Verlag, Cham.

Pennebaker, J.W. (2011). *The Secret Life of Pronouns*. Bloomsbury Press, New York.

Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P. & Stein, B. (2017). Overview of PAN'17. In Gareth J. F. Jones et al. (Eds), *Experimental IR meets multilin-guality, multimodality, and interaction*. 7th International Conference of the CLEF Initiative (CLEF 2017), Berlin, Springer.

Rangel, F. & Rosso, P. (2019). Overview of the 7th Author Profiling Task at PAN 2019. In *Working Notes of the CLEF 2019 Evaluation Labs*, vol. 2380, CEUR, Aachen.

Rangel, F., Celli, R., Rosso, P., Potthast, M., Stein, B. & Daelemans, W. (2015). Overview of the 3rd author profiling task at PAN 2015. In: Cappanello L., Ferro, N., SanJuan, E. (eds.) *Notebook Papers of CLEF 2015 Labs and Workshop*, vol. 1391, CEUR, Aachen.

Rangel, F. & Rosso, P. (2016). On the impact of emotions on author profiling. *Information Processing & Management*, 52(1), 73–92.

Rangel, F., Rosso, P., Chugur, I., Potthast, M. Trenkmann, M., Stein, B., Verhoeven, B. & Daelemans, W. (2014). Overview of the 2nd author profiling task at PAN 2014. In: Cappellato L., Ferro N.,

Halvey M., Kraaij W. (eds.) *Notebook Papers of CLEF 2014 Labs and Workshop*, vol. 1180, CEUR, Aachen.

Rangel, F., Rosso, P., Montes-y-Gòmez, M., Potthast, M. & Stein, B. (2018). Overview of the 6th Author Profiling Task at PAN 2018: Multimodal gender identification in Twitter. In *Working Notes of the CLEF 2018 Evaluation Labs*, vol. 2125, CEUR, Aachen.

Rosso, P., Potthast, M., Stein, B., Stamatatos, E., Rangel, F. & Daelemans, W. (2019). Evolution of the PAN lab on digital text forensics. In N. Ferro & C. Peters, *Information retrieval evaluation in a changing world*, Springer, Cham, 461–486.

Rosso, P., Rangel, F., Potthast, M., Stein, B., Stamatatos, E., Tschuggnall, M. & Stein, B. (2016). Overview of PAN 2016. In: Fuhr, N., Quaresma, P., Gonçalves, T., Larsen, B., Balog, K., Macdonald, C., Ferro, N. (Eds) *Experimental IR meets multilinguality, mul-timodality, and interaction*, 332–350. Springer, Heidelberg.

Savoy, J. (2015). Comparative evaluation of term selection functions for authorship attribution. *Digital Scholarship in the Humanities*, 30(2), 246–261.

Schler, J., Koppel, A., Argamon, S. & Pennebaker, J. (2006). Effects of age and gender on blogging. *Proceedings AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, 191–197.

Schwartz, H.A, Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seliman, M.E.P. & Ungar, L.H. (2013). Personality, gender, and age in the language of social media. *PLOS One*, 8(9).

Tausczik, Y.R. & Pennebaker, J.W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1), 24–54.

Young, L. & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29, 205–231.

Yule, G. (2010). *The Study of Language*. Cambridge University Press, Cambridge.

Rang	**Bots**	**Humains**	**Hommes**	**Femmes**
1	18.6 urllink	42.7 @	6.6 **.**	4.3 #
2	12.1 **,**	13.6 rt	3.9 the	3.4 rt
3	11.0 **-**	6.6 I	1.3 **'**	2.5 **...**
4	6.6 _	4.9 _	1.2 a	2.2 I
5	4.8 job	3.5 **'**	1.2 that	1.5 you
6	3.7 #	3.3 !	0.9 is	1.5 my
7	3.1 developper	2.9 this	0.9 **,**	1.5 **!**
8	2.9 engineer	2.6 my	0.8 it	1.1 _
9	2.8 **/**	2.3 s	0.8 he	1.0 so
10	2.5 software	2.2 to	0.7 ?	1.0 me

***Tableau 1** : Termes avec leur probabilité de différence*