Prendre en compte le capitalisme social dans la mesure de l'influence sur Twitter

Maximilien Danisch, Nicolas Dugué, Anthony Perez

Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, F-75005 Paris, France.
CNRS, UMR 7606, LIP6, F-75005 Paris, France.
Univ. Orléans, INSA Centre Val de Loire, LIFO EA 4022 45067, Orléans, France.

ABSTRACT. *L'influence sur Twitter est un sujet particulièrement discuté avec l'explosion de l'utilisation de ce service de micro-blogging. En effet, afin de fouiller efficacement dans la masse de tweets produite par les millions d'utilisateurs de Twitter, de déterminer les tendances et les informations pertinentes, il est important de pouvoir détecter les utilisateurs influents. Ainsi, plusieurs outils fournissant un* score d'influence *ont été proposés et font référence. Cependant, les algorithmes utilisés par les sociétés qui les développent restent secrets. Dans des travaux récents, il a été montré que des comptes automatiques peuvent obtenir des scores élevés sans raison. De façon à étendre et compléter ces travaux, nous montrons que ces outils sont incapables de distinguer les utilisateurs réels de ceux appelés* capitalistes sociaux*, qui obtiennent à tort des scores d'influence élevés. Afin de résoudre ce problème, nous définissons un classifieur qui réalise cet objectif et rétablit ainsi des scores réalistes pour les capitalistes sociaux. Pour réaliser ce classifieur, nous avons réuni un jeu de données contenant des exemples de capitalistes sociaux et d'utilisateurs réguliers du réseau ainsi que leurs informations de profils et d'utilisation. Pour finir, nous avons développé une application en ligne qui utilise ce classifieur.*

KEYWORDS: *capitalisme social, mesure d'influence, Twitter*

DOI:10.3166/RMPD..1–12

1. Introduction

Contexte. Twitter est un service de micro-blogging largement utilisé pour partager, rechercher et débattre d'informations ou d'évènement du quotidien (Java *et al.*, 2007). Le nombre de ses utilisateurs est passé de 200 millions en avril 2011 (*Twitter: We Now Have Over 200 Million Accounts*, 2011) à 500 millions en octobre 2012 (*The Telegraph : Twitter in numbers*, 2013) et 1 milliard de *tweets* -courts messages de moins de 140 caractères- sont postés tous les deux jours et demi (Rodgers, 2013). Twitter est donc à la fois un service de micro-blogging et un outil média. Mais c'est également un réseau social en ligne qui inclut de nombreux outils d'échange entre utilisateurs.

Par exemple, pour voir les *tweets* d'autres utilisateurs s'afficher sur son fil d'actualité (*timeline* en anglais), il est nécessaire de s'abonner à ces utilisateurs. Si u s'abonne à v, on dit que u est un *abonné* de v. Réciproquement, v est un abonnement de u. De plus, un utilisateur peut *retweeter* (Suh *et al.*, 2010) les *tweets* d'autres utilisateurs, par exemple pour partager avec ses abonnées une information pertinente. Par ailleurs, les utilisateurs peuvent *mentionner* d'autres utilisateurs pour attirer leur attention en ajoutant `@NomUtilisateur` dans leur message.

L'augmentation du nombre d'abonnés, l'explosion du nombre de tweets par jour, l'importance de Twitter dans l'actualité et ses fonctionnalités sociales ont amené les entreprises et les académiques à s'intéresser à la notion d'influence sur ce réseau (Anger, Kittl, 2011 ; Cha *et al.*, 2010 ; Sameh, 2013 ; Tinati *et al.*, 2012 ; Waugh *et al.*, 2013). La plupart des travaux considèrent le nombre d'abonnés et d'interactions - les retweets et les mentions. Intuitivement, plus un utilisateur a d'abonnés ou plus il est retweeté et mentionné, plus son influence sur le réseau est considérée comme importante (Cha *et al.*, 2010). Différents outils ont ainsi été proposés par l'industrie dans le but d'associer à chaque utilisateur un *score* d'influence sur le réseau. Parmi les plus utilisés, on retrouve Klout (*Klout, the standard for influence*, s. d.), Kred (*Kred story*, s. d.), Twitalyzer (*Twitalyzer, serious analytics for social business*, s. d.). Dans tous les cas, l'algorithme utilisé pour calculer le score d'un utilisateur est gardé secret, même si Klout et Kred fournissent quelques informations non détaillés (voir *e.g.* (Anger, Kittl, 2011)). Ce que l'on retient de ces informations, c'est que le nombre d'abonnés n'est pas un paramètre clé de l'algorithme : ces outils se concentrent sur les interactions.

Capitalistes sociaux. Un grand nombre d'utilisateurs du réseau Twitter tentent d'accroître leur nombre d'abonnés de façon artificielle (Dugué, Perez, 2014 ; Ghosh *et al.*, 2012). Ces utilisateurs appelés capitalistes sociaux détournent le principe d'abonnement sur Twitter en promettant des abonnements réciproques. Ils s'abonnent ainsi uniquement à leurs -futurs- abonnés, sans aucune considération pour le contenu tweeté par ces utilisateurs. Leur objectif est d'augmenter leur visibilité sur le réseau et sur les moteurs du recherche du réseau (Ghosh *et al.*, 2012). Ces comptes qui sont des utilisateurs réels et actifs, non pas des robots ou des comptes automatisés, parviennent par ces moyens à obtenir des scores d'influence élevés.

Une question survient donc naturellement : *Étant donné que les nombres d'abonnés et les retweets de ces utilisateurs sont obtenus de façon artificielle et indépendante de la pertinence du contenu qu'ils tweetent, doivent ils être considérés comme influents ?* Évidemment non, mais nous verrons que les outils actuels de mesure d'influence semblent en désaccord avec cette réponse.

Travaux liés. Récemment, Messias et al. (Messias *et al.*, 2013) ont créé deux comptes agissant de façon automatique (des *bots*) selon des stratégies très simples. L'un des bots, qui tweetait automatiquement à propos de sujets populaires obtint 500 abonnés et des scores Twitalyzer et Klout élevés. Il est intéressant de constater que ces outils considèrent de tels utilisateurs influents : ces comptes peuvent efficacement être détectés comme automatiques en utilisant les sources utilisées pour poster les tweets ou le rythme régulier des posts (Chu *et al.*, 2010).

Dans ces travaux, l'influence des capitalistes sociaux n'est pas traitée. Dugué and Perez (Dugué, Perez, 2014) observent que ces utilisateurs sont des comptes réels, administrés manuellement dans un peu moins de 90% des cas et qui parviennent néanmoins à gagner un grand nombre d'abonnés et à être largement retweetés.

Contributions. Nous nous intéressons donc à la problématique de l'influence des capitalistes sociaux. Nous montrons que les outils actuels ne sont pas capables de distinguer les capitalistes sociaux des utilisateurs réguliers et qu'ils leurs accordent ainsi à tort des scores d'influence potentiellement élevés. Pour ce faire, nous décrivons tout d'abord un premier jeu de données Twitter obtenu sur des hashtags dédiés au capitalisme social tels que *#TeamFollowBack* (Section 3). En étudiant ces utilisateurs certifiés capitalistes sociaux selon Dugué et Perez (Dugué, Perez, 2014), nous observons que certains d'entre eux sont considérés comme très influents par les outils de mesure tels que Klout et Kred (Section 4.1). Afin de remédier à cela, nous implémentons un classifieur discriminant les capitalistes sociaux des utilisateurs réguliers. Ce classifieur, capable de retourner la probabilité pour un utilisateur d'être un capital social, nous permet de pondérer le score d'influence produit par Klout (Section 5.2). Enfin, nous terminons en présentant l'application en ligne que nous avons développé qui estime la probabilité pour un utilisateur d'être un capitaliste social (Section 5.3).

2. Capitalistes sociaux

Les capitalistes sociaux ont été mis en lumière par Ghosh et al. (Ghosh *et al.*, 2012) lors d'une étude axée sur les spammeurs. Les auteurs observent avec surprise que les utilisateurs qui s'abonnent le plus aux spammeurs sont des utilisateurs réels et non des robots ou des spammeurs. Ces utilisateurs qui cherchent à augmenter à tout prix leur nombre d'abonnés mettent en place deux techniques très simples et basées sur des abonnements réciproques :

- Follow Me and I Follow You : l'utilisateur assure à ses abonnés qu'il s'y abonnera en retour.
- I Follow You, Follow Me : au contraire, ces utilisateurs s'abonnent massivement à d'autres utilisateurs en espérant que ceux-ci s'abonnent à eux en retour.

Ils s'abonnent à d'autres utilisateurs sans considération pour le contenu de leurs tweets et sont donc nocifs pour le réseau. Ils rendent en effet plus difficile la détection d'utilisateurs et de contenus pertinents.

Dugué et Perez (Dugué, Perez, 2014) ont récemment fourni une méthode capable de détecter les capitalistes sociaux en utilisant la mesure de similarité appelée indice de chevauchement (Simpson, 1943) qui calcule la relation qui existe entre l'ensemble A des abonnements et l'ensemble B : $I(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)}$ Un utilisateur avec un indice de chevauchement supérieur à 1 est abonné à un grand nombre de ses abonnés et a ainsi appliqué les principes mentionnés ci-dessus : **FMIFY** et **IFYFM**.

De plus, Dugué et Perez (Dugué, Perez, 2014) ont montré que les capitalistes sociaux appliquent leurs méthodes de façon très efficace en insérant dans leur tweets des

mots-clés appelés *hashtags* et dédiés au processus de capitalisme social. Ces utilisateurs utilisent des hashtags tels que *#IFollowBack* ou *#TeamFollowBack* pour interagir entre capitalistes sociaux (voir Figure 1). Avec ces hashtags, les capitalistes sociaux demandent explicitement des retweets ou des mentions en échange d'abonnements. Nous utilisons ces moyens d'identifier les capitalistes sociaux pour créer le jeu de données que nous détaillons dans la prochaine section.

FIGURE 1 – Timeline du capitaliste social "followback_707"

3. Jeu de données

Exemples positifs. Dans le but de constituer un jeu de données de capitalistes sociaux certifiés tels que ceux décrits par Dugué and Perez (Dugué, Perez, 2014), nous avons récolté des tweets postés sur les hashtags *#TeamFollowBack*, *#instantfollowback* et *#teamautofollow* dédiés au capitalisme social. Nous avons identifié les utilisateurs ayant posté au moins trois tweets avec ces hashtags en quelques jours et ainsi obtenu un échantillon d'à peu près 25 000 capitalistes sociaux.

Exemples négatifs. La première étape pour obtenir des exemples négatifs consiste à échantillonner aléatoirement Twitter. Nous avons ainsi choisi aléatoirement 15 000 utilisateurs de ce réseau qui en contient plus de 550 millions d'après Twitter. Cependant, puisque la grande majorité de ces utilisateurs a peu de connexion avec le reste du réseau, ils ne constituent pas un échantillon suffisamment pertinent. Nous avons ainsi choisi aléatoirement 55 000 utilisateurs parmi les abonnements de ceux-ci. Ces utilisateurs plus connectés et plus actifs sont très probablement des utilisateurs réguliers et nous fournissent donc nos exemples négatifs. Par ailleurs, d'après Dugué and Perez (Dugué, Perez, 2014), les capitalistes sociaux représentent 0.2% du réseau. Ainsi, même en choisissant parmi les abonnements d'utilisateurs aléatoires ce qui favorise le choix d'utilisateurs bien connectés comme les capitalistes sociaux, notre échantillon contient assurément un nombre négligeable de capitalistes sociaux (0.2% de notre échantillon, soit 110 utilisateurs).

Ensuite, nous avons obtenu via l'API REST fournie par Twitter, pour chaque compte, un ensemble d'informations pertinentes pour les caractériser. Ces informations sont classées en différentes catégories (voir Tableau 1). Remarquons que les restrictions imposées par Twitter quant à l'utilisation de leur API nous poussent à

considérer uniquement les 200 derniers tweets des utilisateurs.

Tableau 1 – Description des différentes informations de compte étudiées.

CATÉGORIE	CARACTÉRISTIQUES
Activité	Nombre de : 1 tweets 2 Listes Twitter qui contiennent l'utilisateur 3 tweets favoris
Topologie locale	Nombre de : 4 Abonnements 5 Abonnés 6 Utilisateurs qui sont à la fois des abonnés et des abonnements
Contenu des tweets	Nombre moyen de : 7 caractères par tweet 8 hashtags par tweets 9 url par tweets 10 mentions par tweets
Caractéristiques des tweets	11 Nombre moyen de retweets pour un tweet 12 nombre moyen de retweets pour un retweet 13 pourcentage de retweets parmi les tweets
Sources	Proportion d'utilisation de : 14 Application Twitter officielle 15 Outil de gestion de compte 16 Outil d'abonnement ou de désabonnement automatique 17 Outil de post de tweet automatique 18 Autres applications (Vine, Wiki, Soundcloud...) 19 Applications smartphones ou tablettes

Les sources utilisées pour poster les tweets et le nombre d'url moyen par tweet ont prouvé être efficaces pour séparer les comptes humains des comptes automatisés (Chu *et al.*, 2010). Cependant, Dugué et Perez (Dugué, Perez, 2014) ont montré qu'un peu moins de 90% des capitalistes sociaux détectés sur ces hashtags n'automatisent pas leurs comptes. Il est ainsi nécessaire d'étudier d'autres types d'informations pour construire un classifieur robuste.

On peut voir que la distribution de ces attributs est différente pour les capitalistes sociaux et les utilisateurs réguliers (Figure 2). Les capitalistes sociaux sont notamment plus retweetés : le nombre de retweets est au coeur des scores Klout et Kred.

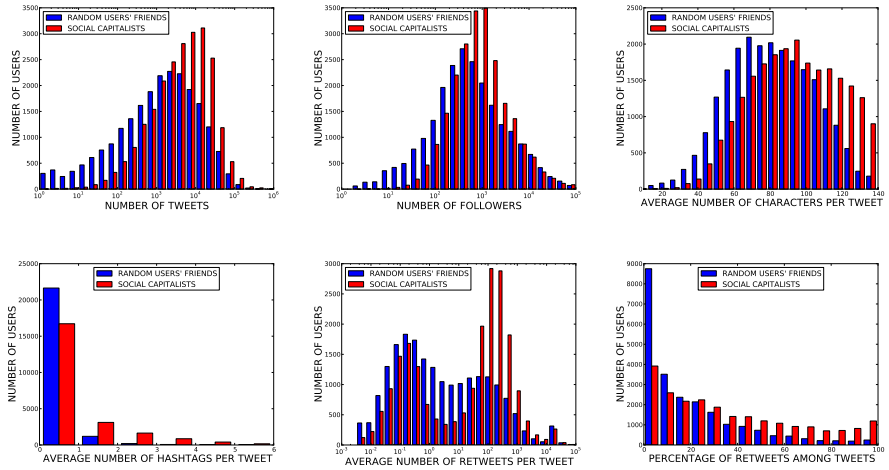


FIGURE 2 – Histogrammes montrant les statistiques pour quelques attributs pour les capitalistes sociaux d'une part et les amis d'utilisateur aléatoire.

4. Mesurer l'influence sur Twitter

L'*influence* sur Twitter a soulevé beaucoup d'intérêt ces dernières années (Anger, Kittl, 2011 ; Cha *et al.*, 2010 ; Sameh, 2013 ; Tinati *et al.*, 2012 ; Waugh *et al.*, 2013). De nombreux paramètres sont utilisés pour la mesurer : nombre d'abonnés, retweets, mentions, favoris, etc. Afin de réaliser des scores plus élaborés, il est possible de combiner certains ratios. Le plus simple est le ratio du *nombre d'abonnements sur le nombre d'abonnés*. Plus le résultat est proche de 0, plus les utilisateurs du réseau sont intéressés par le contenu fourni par l'utilisateur. Au contraire, si le résultat est bien supérieur à 1, l'utilisateur peut être considéré comme s'abonnant en masse. Ce ratio peut néanmoins conduire à de mauvaises interprétations et est ainsi associé à des paramètres liés au degré d'*interaction* de l'utilisateur tels que *Le ratio Retweet et Mention* et *le ratio d'interaction* (Anger, Kittl, 2011). Le premier compte le nombre de tweets postés par un utilisateur qui sont retweetés ou mènent à une conversation et le divise par le nombre de tweets quand le second considère le nombre d'utilisateurs distincts qui retweetent ou mentionnent l'utilisateur divisé par son nombre d'abonnés.

Dans cet article, nous nous concentrons sur les outils disponibles en ligne les plus largement utilisés comme mesures d'influence. Ces outils restent populaires [1] malgré leurs défauts (Nathanson, 2014). C'est en particulier le cas de Klout (*Klout, the standard for influence*, s. d.), Kred (*Kred story*, s. d.) et Twittalyzer (*Twitalyzer, serious analytics for social business*, s. d.) (malgré la fin de son activité commerciale en sep-

1. Klout a récemment été acheté par Lithium Technologies pour 200 millions de dollars (Shu, 2014).

tembre 2013). Klout mesure l'influence d'un utilisateur en se basant sur les principaux réseaux sociaux (par exemple Facebook, Linkedin et Instagram), pas uniquement sur Twitter. Il est cependant possible de savoir quels réseaux sont utilisés pour obtenir le Klout score d'un utilisateur. Nous préciserons cela lorsque nous utiliserons le Klout score dans cet article. Dans la prochaine section, nous montrons les limites des outils de mesure de l'influence populaires en montrant que ces outils considèrent certains capitalistes sociaux comme très influents.

4.1. L'impact du capitalisme social

Avec la promesse de suivre en retour les utilisateurs qui les suivent, les capitalistes sociaux parviennent à accroître leur nombre d'abonnés efficacement. Cependant, comme mentionné précédemment, ceci ne conduit pas nécessairement à une augmentation de leur score d'influence puisque que ce paramètre n'est pas considéré comme important par les principaux outils de mesure. Néanmoins, avoir un grand nombre d'abonnés facilite les interactions telles que les retweets et les mentions, qui sont considérés comme des indicateurs d'influence. C'est surtout le cas pour les capitalistes sociaux qui postent des tweets dont le contenu demande des retweets et des mentions en échange d'un abonnement (voir Figure 1).

Nous illustrons cela en calculant les scores Klout, Kred et Twitalyzer de capitalistes sociaux certifiés extraits du jeu de données collecté en utilisant les hashtags décrits par Dugué and Perez (Dugué, Perez, 2014).

Tableau 2 – Scores d'influence des capitalistes sociaux extraits de nos jeux de données : Klout, Kred et Twitalyzer.

Name	**Abonnements**	**Abonnés**	**Klout**	**Kred**	**Twitalyzer**
teamukfollowbac	120,065	134,669	79	98.9	25.8
berge31	2,522	2,434	76	77,8	1
TheDrugTribe	26,266	28,832	69	98.2	27.2
globalsocialm2	5,603	5,624	69	95.1	3.3
repentedhipster	3,148	2,940	66	78,2	1
LIGHTWorkersi	112,963	103,475	66	96,2	22.4
ilovepurple__	49,666	52,448	65	97.5	22.9
TEAMF0LL0W	13,246	78,615	65	99.2	21.2
TEEMF0LL0W	10,977	92,412	64	99.3	21.1

Les capitalistes sociaux du Tableau 2 sont considérés comme influent par les trois mesures, malgré des comportements, des biographies ou même des noms parfois très explicites.

Pour compléter les observations précédentes, nous comparons dans la Figure 3 les différents scores de deux comptes populaires et actifs, ceux de **Barack Obama** et

Oprah Winfrey, à ceux de capitalistes sociaux avérés ou de comptes automatiques (*Carina Santos*, le compte créé par Messias et al. (Messias *et al.*, 2013)).

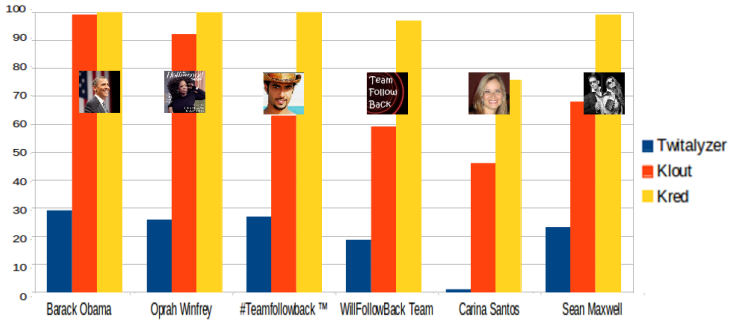


FIGURE 3 – Comparaison des trois mesures pour les comptes de **Barack Obama** et **Oprah Winfrey** et ceux de capitalistes sociaux avérés.

Les capitalistes sociaux ont des scores Kred et Twitalyzer similaires à ceux des comptes de références quand leur Klout est par contre plus faible. Cette différence peut s'expliquer par le fait que Klout utilise l'activité de plusieurs réseaux sociaux, et celle de **Wikipédia**. **Barack Obama** et **Oprah Winfrey** avec leur page Wikipédia bien documentée et consultée accèdent donc à un score plus élevé. Excepté Carina Santos, les comptes que nous considérons sont des capitalistes sociaux avérés et très faciles à identifier : leur noms et biographies sont explicites. Par ailleurs, ces comptes tweetent *exclusivement* du contenu lié au capitalisme social et ne produisent aucun contenu pertinent. Pourtant, aucun des outils en ligne n'est capable de détecter ces comportements évidents.

Pour conclure cette section, nous mettons en relation le Klout score moyen d'exemples positifs de notre jeu de données et le nombre d'abonnés de ces utilisateurs. Souvenons que Klout considère le nombre de retweets et de mentions comme plus important que le nombre d'abonnés. Cependant, comme le montre la Figure 4, le nombre d'abonnés des utilisateurs est fortement corrélé au Klout score. Ceci s'explique par le fait que les capitalistes sociaux réussissent à obtenir plus d'interactions et de visibilité à mesure que leur nombre d'abonnés augmente.

5. Une nouvelle approche pour mesurer l'influence

Nous présentons maintenant un *c*lassifieur qui discrimine les capitalistes sociaux des utilisateurs normaux. De plus, notre classifieur estime la probabilité qu'un utilisateur soit un capitaliste social. Nous utilisons ce résultat afin d'équilibrer le score d'influence donné par Klout, cf. Équation 1, ce qui diminue efficacement le score d'influence des capitalistes sociaux tout en conservant le score d'influence des utilisateurs normaux.

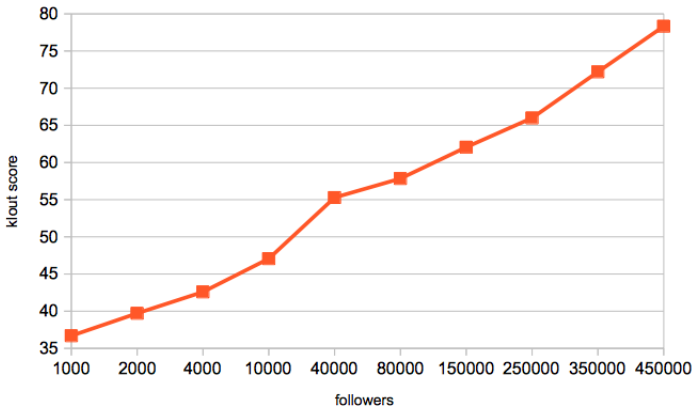


FIGURE 4 – Scores Klout moyens en fonction d'une borne inférieure sur le nombre d'abonnés.

5.1. Classification des capitalistes sociaux

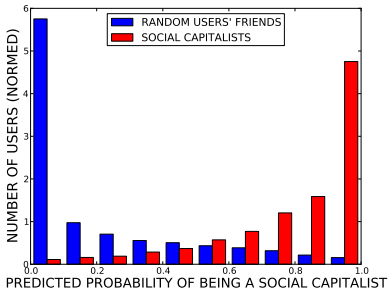
Dugué et Perez (Dugué, Perez, 2014) ont développé un algorithme performant pour discriminer les capitalistes sociaux des utilisateur normaux. Cependant, leur but était d'utiliser seulement des attributs topologiques. Les résultats de ces travaux peuvent être améliorés en utilisant d'autres attributs tels que dans le jeu de données décrit dans 3. Celui-ci contient 77 102 utilisateurs, dont 22 845 capitalistes sociaux et 54 257 utilisateur normaux et leurs attributs (Section 3).

Nous avons partagé aléatoirement le jeu de données en un ensemble d'apprentissage de 70%, utilisé pour entraîner le classifieur, et un ensemble de test de 30%, utilisé pour évaluer les performances du classifieur. Nous avons utilisé des algorithmes classiques tels que les K-plus proches voisins (KNN), la machine à vecteurs de support (SVM), les forêts aléatoires (RF) et la régression logistique (LR) (Bishop *et al.*, 2006), implémentés à l'aide de la bibliothèque python sklearn (Pedregosa *et al.*, 2011).

RF et LR donnent des résultats à la précision [2] légèrement supérieure, ce qui corrobore les résultats obtenus dans (McCord, Chuah, 2011). Nous utilisions ainsi LR qui est conçu pour estimer la probabilité qu'un exemple soit positif (résultat que nous utiliserons pour équilibrer le score Klout) et qui, une fois ajusté, nécessite seulement le stockage de quelques paramètres. Afin d'obtenir de meilleures performances, nous avons transformé les attributs de la manière suivante : (i) nous avons utilisé un attribut constant additionnel, (ii) nous avons passé certains attributs en échelle logarithmique quand cela était nécessaire et (iii) nous avons utilisé les produits des attributs deux à deux comme attributs supplémentaires.

2. La précision est la proportion d'exemples correctement étiquetés.

Après avoir ajustés les paramètres du classifieur sur l'ensemble d'apprentissage, nous avons évalué ses performances sur l'ensemble de test. La Figure 5a montre l'histogramme des probabilités d'être positif. Nous voyons qu'il existe une très forte corrélation entre la probabilité prédite et le fait qu'un exemple soit effectivement un capitaliste social ou un utilisateur normal. De plus, en coupant les probabilités à 0.5 afin d'obtenir un classifieur binaire, nous obtenons un F-score de 87%.



(a)

Name	**Klout score**	P_{Ksoc}	S_{DDP}
barackobama	99	$8.42 \cdot 10^{-4}$	99
oprah	93	$5.86 \cdot 10^{-9}$	93
followback_707	64	0.999	1
seanmaxwell	69	0.937	9
scarina91	46	0.110	46
teamukfollowbac	80	0.838	26
TEEMF0LLOW	69	0.416	69
zuandoemkta	62	0.861	18
kosma003	53	0.747	27
NicolasDugue	33	0.120	33

(b)

FIGURE 5 – (5a) Histogramme des probabilités d'être un capitaliste social. (5b) Klout scores, DDP scores et P_{Ksoc}.

5.2. Rééquilibrage du score Klout

Nous utilisons ici la probabilité prédite P_{Ksoc} qu'un utilisateur soit un capitaliste social pour rééquilibrer le score Klout.

$$S_{DDP} = \begin{cases} S_{Klout} & \text{si } P_{Ksoc} \leq 0.5 \\ 2(1 - P_{Ksoc})S_{Klout} & \text{si } P_{Ksoc} > 0.5 \end{cases} \quad (1)$$

Nous avons évalué ce nouveau score sur des capitalistes sociaux, des utilisateurs normaux tirés aléatoirement de notre ensemble de test, ainsi que sur d'autres utilisateurs influents de twitter, cf. Table 5b.

Quand nous considérons des exemples positifs pris de notre ensemble de test (comme **teamukfollowbac**), nous observons que sa probabilité prédite d'être un capitaliste social est élevée. Ainsi, le score Klout rééquilibré reflète cette observation et voit l'utilisateur moins influent qu'avec le score Klout. Ceci est encore plus frappant avec **seanmaxwell** et **followback_707**, deux autres utilisateurs positifs extraits de notre ensemble de test. Ces deux utilisateurs tweetent afin de joindre d'autres capitalistes sociaux et d'obtenir des followers, un comportement qui ne devrait pas être considéré comme influent sur Twitter. Le score Klout rééquilibré prend en considération ces remarques et décroit l'influence de ces utilisateurs. **TEEMF0LL0W** est un exemple positif sur lequel les résultats du classifieur ne sont pas particulièrement convaincant. Cet utilisateur tweete seulement des messages très courts comportant

seulement des hashtags, ce qui diffère de l'activité des autres capitalistes sociaux et pourrait expliquer les problèmes rencontrés par notre classifieur.

5.3. Application en ligne

Nous avons développé une application en ligne qui calcule la probabilité d'un utilisateur de twitter d'être un capitaliste social. Cette application est disponible à l'adresse http://www.bit.ly/DDPapp (notons qu'il faut avoir un compte Twitter pour utiliser l'application).Étant donné le nom d'utilisateur d'un compte Twitter, l'application calcule les attributs décrits et utilise ensuite le classifieur LR afin de calculer la probabilité estimée que l'utilisateur de ce compte soit un capitaliste social. En haut à gauche de la page web, les scores Klout et Kred sont affichés. Le score Klout rééquilibré est affiché sur la photo de profil de l'utilisateur. En plus des informations basiques du compte twitter (nombre d'amis, de followers, de tweets, de listes et de favoris), des informations plus complexes sont affichées, telles que la proportion de tweets originaux et de retweets ou les sources utilisées pour poster les tweets.

6. Conclusion

Notre étude montre qu'un grand nombre de capitalistes sociaux obtiennent à tort des scores d'influence élevés. Afin de pallier ces limitations, nous avons développé un classifieur qui utilise des attributs topologiques ainsi que d'autres attributs extraits des tweets de l'utilisateur et de son l'activité. Notre classifieur détecte efficacement les capitalistes sociaux et étend des travaux précédant basés seulement sur des attributs topologiques (Dugué, Perez, 2014). Nous utilisons ensuite les prédictions du classifieur afin de rééquilibrer le score Klout, qui est l'outil pour mesurer l'influence sur twitter le plus utilisé. Ce travail peut être étendu de plusieurs manières. Tout d'abord, nous pourrions étendre notre jeu de données à d'autres attributs et en accroître la taille pour améliorer l'efficacité du classifieur. De plus, nous nous sommes concentrés sur le développement d'un outil pour discriminer les capitalistes sociaux des utilisateurs normaux. Nous avons ensuite utilisé cet outil pour rééquilibrer le score Klout. Il serait intéressant de construire une nouvelle mesure d'influence en partant de zéro. Pour finir, ce travail est ciblé sur le réseau Twitter sur lequel les capitalistes sociaux ont été mis en évidence par Ghosh et al. (Ghosh *et al.*, 2012). D'autres applications avec une composante sociale telles qu'Instagram ou Youtube pourraient avoir les mêmes propriétés et nécessiter une amélioration des outils de mesure d'influence.

Références

Anger I., Kittl C. (2011). Measuring influence on Twitter. In *Proceedings of the 11th international conference on knowledge management and knowledge technologies*, p. 1–4. ACM.

Bishop C. M. *et al. (2006).* Pattern recognition and machine learning *(vol. 1). Springer.*

Cha M., Haddadi H., Benevenuto F., Gummadi K. (2010). Measuring user influence in Twitter: The million follower fallacy. In Proceedings of ICWSM. *AAAI.*

Chu Z., Gianvecchio S., Wang H., Jajodia S. (2010). Who is tweeting on Twitter: Human, Bot, or Cyborg? In Acsac*, p. 21–30. ACM.*

Dugué N., Perez A. (2014). Social capitalists on Twitter: detection, evolution and behavioral analysis. Social Network Analysis and Mining*, vol. 4, n° 1, p. 1-15. (Springer)*

Ghosh S., Viswanath B., Kooti F., Sharma N., Korlam G., Benevenuto F. et al. (2012). Understanding and combating link farming in the Twitter social network. In *WWW*, p. 61–70.

Java A., Song X., Finin T., Tseng B. (2007). Why we twitter: understanding microblogging usage and communities. In *Workshop on web mining and social network analysis*, p. 56–65.

Klout, the standard for influence. (s. d.). (http://www.klout.com)

Kred story. (s. d.). (http://www.kred.com)

McCord M., Chuah M. (2011). Spam detection on twitter using traditional classifiers. In *Autonomic and trusted computing*, p. 175–186. Springer.

Messias J., Schmidt L., Oliveira R., Benevenuto F. (2013). You followed my bot! transforming robots into influential users in Twitter. *First Monday*, vol. 18, n° 7.

Nathanson J. (2014). How Klout Finally Matters. (http://www.slate.com/articles/business/the_bet/2014/05/klout_is_basically_dead_but_it_finally_matters.html)

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O. *et al. (2011). Scikit-learn: Machine learning in Python.* JMLR*, vol. 12, p. 2825–2830.*

Rodgers S. (2013, august). Twitter blog. (https://blog.twitter.com/2013/behind-the-numbers-how-to-understand-big-moments-on-twitter)

Sameh A. (2013). A Twitter analytic tool to measure opinion, influence and trust. Journal of Industrial and Intelligent Information*, vol. 1, n° 1, p. 37–45.*

Shu C. (2014, march). Tech Crunch.

Simpson G. G. (1943). Mammals and the nature of continents. Am. J. of Science*, n° 241, p. 1-41.*

Suh B., H. L., Pirolli P., Chi E. H. (2010). Want to be retweeted? large scale analytics on factors impacting retweet in Twitter network. In Socialcom*, p. 177-184.*

The telegraph : Twitter in numbers. *(2013). (http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html)*

Tinati R., Carr L., Hall W., Bentwood J. (2012). Identifying communicator roles in Twitter. In International conference companion on www*, p. 1161–1168. ACM.*

Twitalyzer, serious analytics for social business. *(s. d.). (http://www.twitalyzer.com - As of September 28, 2013 Twitalyzer has decided to no longer sell new subscriptions.)*

Twitter: We now have over 200 million accounts. *(2011). (http://www.huffingtonpost.com/2011/04/28/twitter-number-of-users_n_855177.html)*

Waugh B., Abdipanah M., Hashemi O., Rahman S. A., Cook" D. M. (2013). The influence and deception of Twitter: The authenticity of the narrative and slacktivism in the australian electoral process. In Proceedings of the 14th australian information warfare conference.