Une méthodologie pour la détection automatique de comptes multiples dans les réseaux sociaux

Application à Wikipédia

Zaher Yamak, Julien Saunier, Laurent Vercouter

INSA Rouen, Laboratoire LITIS
685 Avenue de l'Université
76801 Saint Etienne du Rouvray cedex, France
zaher.yamak@insa-rouen.fr
julien.saunier@insa-rouen.fr
laurent.vercouter@insa-rouen.fr

RÉSUMÉ. *Avec la croissance des médias sociaux (MS) comme l'élément le plus important d'Internet en termes de visiteurs, la détection des faux comptes est devenue l'un des problèmes de sécurité les plus difficiles de ces médias. Durant les dernières années, les MS ont évolué de façon importante, convertissant de fait une partie de nos vies personnelles vers le virtuel. Mais cette évolution a aussi des effets négatifs. En 2012, 16,6 millions d'américains ont été victimes de vol d'identité, selon une estimation de la 'U.S. Bureau of Justice Statistics', avec des pertes financières évaluées jusqu'à $24,7 milliards pour l'ensemble de ces victimes.*
Diverses techniques sont utilisées pour manipuler les utilisateurs dans les environnements des MS telles que le spam social, le vol d'identité, le spear phishing et les attaques sybilles... Dans cet article, nous nous intéressons à l'analyse du comportement des comptes multiples qui essaient de contourner les régulations des MS. Dans ce contexte de détection de manipulation dans les MS, nous nous concentrons sur la détection des identités multiples (Faux-nez, ou sock-puppets) créées sur la plateforme Wikipédia anglophone (EnWiki). Nous mettons en place une méthodologie complète allant de l'extraction des données nécessaires de l'EnWiki, jusqu'à l'apprentissage et les tests de nos données sélectionnées à l'aide de plusieurs algorithmes d'apprentissage automatique. Nous proposons un ensemble des caractéristiques étendant celles trouvées dans la littérature existante, de façon à les utiliser dans l'analyse automatique des données afin de détecter les comptes Faux-nez créés sur EnWiki. Nous les appliquons sur une base de données de 10 000 *comptes d'utilisateurs. Les résultats comparent plusieurs algorithmes d'apprentissage automatique afin de montrer que ces nouvelles caractéristiques et données d'apprentissage permettent de détecter 99 % de faux comptes, améliorant ainsi les résultats de la littérature.*

ABSTRACT. *With the growth of social media as the most important element of internet in term of visitors, fake accounts detection has become one of the hardest social media security challenges. Over the years, online social media (OSN) have evolved widely, converting part of our personal lives to virtual ones. But this evolution also has negative effects. In 2012, 16.6 million*

of Americans were victims of identity theft according to an estimate from the U.S. Bureau of Justice Statistics, with up to $24.7 billion of financial losses for these victims.
Various techniques are used to manipulate users in OSN environments such as social spam, identity theft, spear phishing and Sybil attacks... In this article, we are interested in analyzing the behavior of multiple fake accounts that try to bypass the OSN regulation. In the context of social media manipulation detection, we focus on the special case of multiple Identity accounts (Sockpuppet) created on English Wikipedia (EnWiki). We set up a complete methodology spanning from the data extraction from EnWiki to the training and testing of our selected data using several machine learning algorithms. In our methodology we propose a set of features that grows on previous literature to use in automatic data analysis in order to detect the Sockpuppets accounts created on EnWiki. We apply them on a database of 10 000 *user accounts. The results compare several machine learning algorithms to show that our new features and training data enable to detect 99 % of fake accounts, improving previous results from the literature.*

MOTS-CLÉS : *faux-nez, application d'apprentissage automatique, manipulation, identité, wikipédia, projets collaboratifs, média social.*

KEYWORDS: *sockpuppet, machine learning application, manipulation, deception, identity, wikipedia, collaborative project, social media.*

DOI:10.3166/RIA.30.419-439

1. Introduction

Depuis 2004, le web social est devenu une force dominante sur l'Internet. Dès 2014, 52 % des adultes utilisateurs réguliers d'Internet utilisent deux ou plusieurs sites de médias sociaux, ce qui présente une augmentation significative par rapport à 2013, où ils avaient déjà atteint 42 % des utilisateurs (Maeve *et al*., 2015). Par exemple, en 2015, il y a environ 1,4 milliard d'utilisateurs de Facebook (Jeff, 2015), ce qui représente 47 % du total des utilisateurs d'Internet.

Au fil des années, les médias sociaux se sont imposés, mettant une partie de nos vies personnelles dans des espaces virtuels. Mais cette évolution a aussi des effets négatifs. Selon les statistiques de l'*American Bureau of Justice*, environ 16,6 millions d'Américains ont été victimes de vol d'identité en 2012. Les pertes financières pour ces victimes se sont élevées à $24,7 milliards. Pour les voleurs d'identité, les médias sociaux sont un terrain de chasse fertile pour acquérir des informations personnelles (Ambika, 2014). Étant désormais une force majeure de l'Internet, il est naturel que les médias sociaux soient maintenant visés par une série d'attaques, allant de l'ingénierie sociale (David, 2015) aux cyberattaques (Goolsby *et al*., 2013).

Aujourd'hui, les sources traditionnelles d'informations (journaux, télévision, radio...) ne sont plus la seule source de nouvelles, parce que les informations sont régulièrement diffusées sur Twitter avant d'être publiées par ces médias traditionnels (Mathew, 2012). Ainsi, les médias sociaux sont devenus un moyen privilégié pour la diffusion d'informations (par exemple une célébration collective d'anniversaire) ou

des avis (par exemple sur un produit ou sur un personnage public) et la promotion d'opinions et idées (c.-à-d. activisme, invitation au vandalisme ou aux émeutes). Dans le cas d'événements qui peuvent conduire à des problèmes de sécurité civile ou de sécurité intérieure, la rapidité pour détecter ces messages, pour identifier leurs auteurs, les « aboyeurs », et les suiveurs ainsi que les processus de diffusion de cette information est nécessaire pour prévenir les actes ou les événements dangereux avant qu'ils se produisent. De même, pour les gouvernements, les personnes publiques et les entreprises, le contrôle de la diffusion de l'information est devenu une question majeure.

Grâce à des API comme *Sign in with Twitter*, les utilisateurs peuvent s'enregistrer et se connecter à des sites tiers en utilisant leur compte de réseau social. La simplicité de ce processus a permis à ces sites de bénéficier de l'accès aux renseignements personnels d'un grand nombre de comptes. Cependant, comme nous l'avons écrit précédemment, comme l'importance des médias sociaux se développe, le nombre de manipulateurs augmente.

Un manipulateur est une personne qui utilise les règles de la vie en société, dans ce cas ceux des médias sociaux, pour obtenir des avantages et des bénéfices personnels ou exercer un contrôle sur un ou plusieurs utilisateurs. La manipulation sur les réseaux sociaux peut être faite par une communication verbale (par exemple, vidéo, audio, ou texte) et/ou un comportement non verbal (par exemple : nombre de demandes d'amis par heure, ou la taille/qualité du texte ajouté. . .) en utilisant de nombreuses techniques comme des spams ou des attaques Sybilles (Douceur, 2002).

Les techniques utilisées pour la manipulation varient selon le type de médias sociaux. Par exemple, les utilisateurs de Facebook sont soumis à des techniques de spam social, profitant de mécanismes de partage et d'étiquetage d'un grand nombre d'utilisateurs sur leurs postes, tandis que les utilisateurs de LinkedIn sont plus soumis à des attaques de type *spear phishing* en raison de la nature professionnelle/personnelle mixte de l'information partagée (Norton, s. d.).

Dans cet article, nous proposons de détecter un type particulier d'attaque, les comptes multiples, qui est à l'origine de plusieurs types de manipulation différentes (Sybille, manipulation d'information, spams sociaux ...). Pour ce faire, nous avons sélectionné le médium social Wikipedia qui permet l'extraction publique d'une grande partie de ses données.

Dans le contexte de Wikipedia, l'utilisation inappropriée de plusieurs comptes est appelée « faux-nez » (ou « sockpuppet »). L'objectif de cet article est de détecter automatiquement les comptes « faux-nez » sur Wikipédia en utilisant des indicateurs non verbaux. Les principales contributions de cet article sont (1) une méthodologie pour extraire et analyser de grandes quantités de données afin de détecter automatiquement les techniques de manipulation dans les médias sociaux et (2) son application dans le cas de la détection de « faux-nez » dans Wikipedia. Les étapes principales sont les suivantes :

– Nous recueillons les données de Wikipedia des comptes (bloqués et non bloqués), ainsi que toutes leurs activités verbales et non verbales.

– Nous filtrons ces données pour sélectionner 10 000 comptes actifs et « faux-nez ».

– Nous créons un ensemble de caractéristiques pour les comportements non verbaux, afin de détecter les comptes « faux-nez » sur Wikipedia.

– Nous calculons les valeurs des caractéristiques proposées.

– Nous évaluons plusieurs algorithmes d'apprentissage supervisés qui utilisent ces caractéristiques et nous comparons nos résultats à d'autres chercheurs qui ont utilisé des caractéristiques de communication verbale et non verbale.

Le reste de l'article est organisé comme suit : dans la section 2, nous présentons les différents types de média sociaux, motivons le choix de Wikipédia, et exposons la méthode actuelle pour détecter les « faux-nez » manuellement. Dans la section 3, nous présentons l'état de l'art dans le domaine de la détection de la tromperie et de la manipulation dans les médias sociaux. Puis, dans la section 4, nous décrivons notre méthodologie proposée à travers l'extraction de données réelles, la sélection des caractéristiques, et nos mesures expérimentales. Dans la section 5, nous présentons les résultats de notre méthode proposée et discutons ces résultats dans la section 6. Enfin, dans la section 7, nous concluons et montrons nos futures orientations.

2. Les projets collaboratifs et Wikipédia

(Kaplan, Haenlein, 2010) ont proposé la classification suivante pour tous les types de médias sociaux. Ils ont identifié six catégories :

1. **Projets collaboratifs** : l'utilisateur peut ajouter, modifier et supprimer les contenus multimédias dans ces sites (par exemple, Wikipedia).

2. **Blogs** : l'utilisateur peut partager des informations, du contenu (texte, audio, vidéo), ou écrire des commentaires sur ces sites (par exemple, Twitter, TripAdvisor).

3. **Communautés de contenu** : l'utilisateur peut partager du contenu comme du texte, audio, photos ou vidéo (par exemple, YouTube, SoundCloud).

4. **Réseaux sociaux** : l'utilisateur peut créer un profil avec des informations personnelles et partager le contenu avec des amis (par exemple, Facebook).

5. **Mondes de jeux virtuels** : l'utilisateur peut créer un profil virtuel et jouer avec d'autres joueurs (par exemple, World of Warcraft).

6. **Mondes sociaux virtuels** : l'utilisateur peut créer un profil virtuel et interagir avec d'autres utilisateurs comme une vie réelle dans un monde virtuel (par exemple, Second Life).

Si les contenus et les modes d'interaction diffèrent entre les médias sociaux, ils ont néanmoins tous une caractéristique commune : ils permettent aux utilisateurs de partager des informations et d'interagir, d'une manière directe (un-à-un, un-à-plusieurs) ou indirecte. Pour maintenir les relations sociales, ils sont généralement basés sur des comptes pour identifier les utilisateurs.

Les projets collaboratifs sont l'un des types de médias sociaux les plus importants au niveau de l'audience et des contributions. Par exemple, Wikipedia a été visité en Avril 2015 par 104,40 millions de visiteurs seulement aux États-Unis (Statista, 2015).

Les projets collaboratifs abritent également des caractéristiques uniques dans les médias sociaux en ligne : ils peuvent être édités par plusieurs personnes approuvées, toutes les modifications sont suivies dans l'historique de la page, toute mauvaise édition peut être facilement annulée... Des exemples de ces sites sont Wikipedia, WikiHow et Wikiversity... À ce titre, ils nous intéressent parce qu'ils ont à la fois un grand nombre de collaborateurs et une histoire publique de toutes les modifications et tous les comptes bloqués, ce qui permet de faire une analyse automatique des données sur une base de données importante. Une autre caractéristique importante est qu'ils peuvent être généralisés aux médias sociaux en utilisant des caractéristiques communes comme la possibilité de faire des commentaires sur la page d'un autre utilisateur et l'écriture dans un style personnel.

Wikipedia est un projet collaboratif où chacun peut modifier la plupart des articles avec ou sans la création d'un compte. Nous avons choisi la version anglophone (EnWiki) pour notre étude, car il a à notre connaissance le plus grand nombre de collaborateurs et de manipulateurs parmi les projets de collaboration.

L'objectif de Wikipedia est d'être une encyclopédie participative utilisant les connaissances de ses utilisateurs grâce à une communauté de bénévoles qui l'entretiennent et ajoutent des contenus. L'édition de son contenu est ouverte au public, permettant ainsi des tentatives de manipulation par des utilisateurs malveillants.

Wikipédia est composé d'un ensemble de pages qui sont regroupées dans des espaces de noms (*namespaces*) en fonction du type d'informations qu'elles contiennent. Actuellement, Wikipedia a 26 espaces de noms [tableau 1] : 12 espaces de noms d'objets, 12 espaces de noms de discussion correspondante, et 2 espaces de noms virtuels. Chaque espace de noms d'un objet dispose d'un espace de discussion correspondant, par exemple l'espace de noms principal qui contient les articles a un espace de noms appelé page de discussion, et l'espace de noms des utilisateurs contient à la fois les pages de l'utilisateur et d'autres pages créées pour un usage personnel dispose également d'un espace de noms correspondant appelé page de discussion d'utilisateur.

L'utilisation incorrecte de plusieurs comptes utilisateurs par une même personne n'est pas autorisée selon les règles de Wikipédia, parce qu'il est attendu que les éditeurs de Wikipédia ne contribuent qu'à l'aide d'un seul compte. Dans ce cadre, on appelle « faux-nez » le ou les comptes supplémentaires grâce auquel un utilisateur va tenter de manipuler les informations contenues dans les pages Wikipédia. Les objectifs usuels sont de soutenir un côté d'un différend dans des informations d'un article ou dans une discussion avec d'autres utilisateurs, participer à un vote ou s'opposer à des changements de règles. Les faux-Nez sont également utilisés pour la manipulation d'autres utilisateurs, afin d'éviter la détection ou pour vandaliser des pages.

Actuellement, chaque utilisateur qui estime qu'un autre compte est un « faux-nez » peut ouvrir une page d'enquête sur ce compte en fournissant une preuve claire de

Tableau 1. Espaces de noms de Wikipédia

Numéro	**Espace de noms d'objet**	**Espace de noms de discussion**	**Numéro**
0	(Principale/Article)	Discussion	1
2	Utilisateur	Discussion utilisateur	3
4	Wikipédia	Discussion Wikipédia	5
6	Fichier	Discussion fichier	7
8	MediaWiki	Discussion MediaWiki	9
10	Modèle	Discussion modèle	11
12	Aide	Discussion aide	13
14	Catégorie	Discussion catégorie	15
100	Portail	Discussion Portail	101
102	Projet	Discussion Projet	103
104	Référence	Discussion Référence	105
828	Module	Discussion module	829
Espace de noms virtuel			
-1	Spécial		
-2	Media		

sa malveillance aux administrateurs expérimentés de Wikipedia. Ces administrateurs tentent alors de détecter le « faux-nez » manuellement, en étudiant le comportement du « faux-nez » sur Wikipedia ou par détection de la similitude dans le style d'écriture.

Dans de nombreux cas, l'administrateur demande l'intervention d'un *'checkuser'*, un ensemble d'utilisateurs plus privilégiés qui ont accès à l'adresse IP de tous les comptes, d'intervenir en vérifiant et en comparant l'adresse IP avec d'autres comptes. Ensuite, le choix se fait selon la similarité de l'emplacement avec d'autres comptes et selon les styles d'interventions similaires afin de bloquer ou non le compte.

Toutes ces données, avec les raisons de l'interdiction sont accessibles au public [1], ce qui fait un cas d'utilisation vérifiable pour tester des algorithmes/caractéristiques pour détecter ce genre de manipulation dans les médias sociaux.

3. Travaux connexes

Les faux comptes sont utilisés pour augmenter la visibilité de contenus de niche, de messages de forums ou encore de pages de fans par la manipulation des votes ou du nombre de vues (Sture, 2010 ; Norajong, 2010). D'autres sont créés pour les comportements malveillants, tels que la distribution de spams, la distribution de logiciels malveillants et la fraude d'identité (Riva, 2010).

1. par https://en.wikipedia.org/w/index.php?title=Wikipedia:Sockpuppet_investigations/Cases/Overview&offset=&limit=500&action=history

De nombreuses solutions ont été proposées afin de détecter les différents types de comportements malveillants dans les médias sociaux. Deux principales approches sont utilisées : l'analyse des communications verbales et l'analyse du comportement non verbal. Dans cette section, nous passons en revue les travaux qui sont proches de notre objet d'étude.

3.1. Communication verbale

(Gao *et al*., 2010) proposent une solution pour détecter les spammeurs sur Facebook. Ils commencent en filtrant les données analysées à partir de Facebook pour garder uniquement les messages contenant des URL, leur hypothèse étant que chaque spammeur va essayer de rediriger l'utilisateur de médias sociaux vers un faux site à l'extérieur de FaceBook. Puis, ils relient les messages similaires qui partagent la même destination ou le même contenu textuel. Enfin, ils regroupent 1,402,028 messages liés en vérifiant si le lien guide vers un faux site ou non, afin de détecter les utilisateurs malveillants et les messages. Les auteurs obtiennent de bons résultats avec cette méthode en détectant 93,9 % des messages de mur malveillants. Cependant, cette méthode ne peut s'appliquer qu'aux tentatives de redirection frauduleuse et non aux manipulations de contenu.

(Solorio *et al*., 2013a) utilisent des techniques de traitement de la langue naturelle pour détecter sur Wikipedia les utilisateurs qui maintiennent plusieurs comptes en fonction de leur comportement verbal. Des caractéristiques textuelles sont utilisées tels que le nombre de lettres de l'alphabet utilisées, le nombre de zones de textes, le nombre d'émoticônes ou l'utilisation de mots sans voyelles. Ces caractéristiques sont testées sur toutes les révisions faites par les utilisateurs sur des pages à travers Wikipedia. L'algorithme de machine à vecteurs de support (SVM) a montré une précision globale de 68,83 % en utilisant un ensemble de données expérimentale de 77 cas d'utilisateurs légitime et « faux-nez ».

En fait, la détection automatique du comportement verbal n'est pas toujours efficace parce que les manipulateurs peuvent comprendre la méthode de détection et ainsi changer leur méthode d'écriture, ce qui rend la détection automatique plus difficile. Ce changement de comportement volontaire est plus difficile lorsque la détection automatique utilise le comportement non verbal.

3.2. Comportement non verbal

(Sarita *et al*., 2009) ont réalisé une expérience relative au spam sur Twitter. Suite à la création d'un mot dièse (*hashtag*), ils observent l'interaction des spammeurs avec celui-ci afin de découvrir les modèles de comportement entre les comptes des spammers. Ils étudient également certaines caractéristiques qui pourraient permettre de distinguer un spammeur d'un utilisateur légitime, tels que le nombre d'amis et le nombre de suiveurs.

(Yang *et al*., 2014) intègrent une détection en temps réel dans le média social chinois RenRen pour la détection des comptes sibylles à l'aide d'une analyse du comportement non verbal. La distinction entre les comptes légitimes et les comptes sybilles est réalisée à l'aide des caractéristiques suivantes : la fréquence d'invitation, les demandes sortantes acceptées et les demandes entrantes acceptées. Ils appliquent une machine à vecteurs de support (SVM) à un ensemble de données de 1000 utilisateurs normaux et 1000 utilisateurs sibylles. Ils distribuent aléatoirement l'échantillon initial en 5 sous-échantillons, dont 4 sont utilisés pour l'apprentissage du classificateur, et le dernier est utilisé pour le tester. Les résultats montrent que le classifieur est très précis, car il a correctement identifié 99 % des deux types de comptes sybilles et non sybilles.

(Cao *et al*., 2012) ont introduit un nouvel outil appelé SybilRank. Il permet aux médias sociaux de classer les utilisateurs en fonction de leur probabilité perçue d'être faux (sibylles). SybilRank peut s'adapter à des graphes relationnels possédant des centaines de millions de nœuds. SybilRank a été déployé dans le centre d'opération de Tuenti et permet de grouper des grands nombres d'utilisateurs sibylles au sein de topologies régulières (étoiles, maille, arbres,. . .) qui sont connectées aux communautés "honnêtes" à travers un nombre limité de bords d'attaque.

Plus proche de notre sujet d'étude, (Tsikerdekis, Zeadally, 2014) proposent une méthode pour détecter les identités multiples à travers l'analyse de l'activité non verbale de l'utilisateur dans l'environnement des médias sociaux. Ils utilisent les données publiquement disponibles de Wikipedia et des algorithmes d'apprentissage automatique. Certaines variables représentent le comportement non verbal des utilisateurs, comme le nombre de révisions effectuées par un utilisateur dans une fenêtre de temps spécifique depuis l'inscription initiale sur le site, le nombre total d'octets ajoutés et le nombre total d'octets supprimés,. . . Pour les tests ils divisent les données en 7 500 cas de « faux-nez » et ils utilisent les algorithmes SVM (Cortes, Vapnik, 1995), Forêt d'arbres décisionnels (RF) (Breiman, 2001), et Adaptive Boosting (ADA) , qui montrent une précision globale de 71,3 %.

Dans ces études, nous avons vu que les communications verbales et non verbales sont des axes intéressants, et que, globalement, la détection des « faux-nez » n'est pas à ce jour suffisante pour la découverte automatique d'un tel comportement. Dans la section suivante, nous allons décrire notre méthodologie pour détecter les « faux-nez » en sélectionnant de nouvelles caractéristiques de comportement non verbal à l'aide des données réelles récupérées de EnWiki.

4. Méthodologie et indicateurs

Dans cette étude, nous travaillons initialement sur le comportement non verbal de l'utilisateur. La méthodologie globale est illustrée dans la *figure* 1. Les quatre étapes principales de notre méthodologie sont :

1. Extraction des données
2. Sélection des comptes

3. Calcul des caractéristiques
4. Apprentissage et test des algorithmes

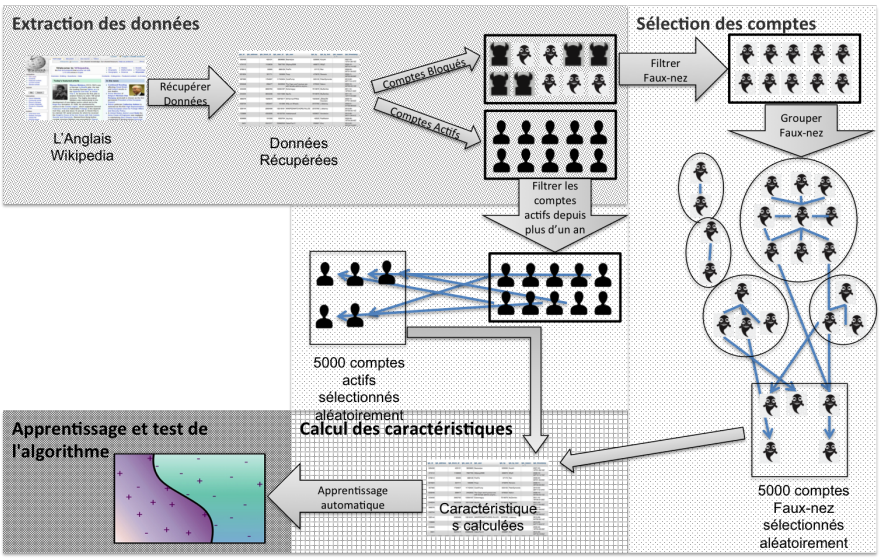

Figure 1. Les quatre étapes principales de notre méthodologie

Extraction des données

Afin d'extraire les informations pertinentes pour les algorithmes d'apprentissage automatique, la première étape consiste à extraire les données disponibles dans EnWiki. La taille totale des données contenues dans EnWiki est d'environ 10 TB non compressés [2], donc ce n'est pas efficace de récupérer de façon aveugle l'ensemble des données pour les traiter. Par conséquent, notre point de départ est de récupérer en premier lieu tous les utilisateurs bloqués, ainsi que leurs contributions sur Wikipédia, entre Février 2004 et Avril 2015.

Plusieurs raisons peuvent mener à bloquer un compte, telles que le spam, le vandalisme, ou les menaces. Dans notre cas, nous sommes intéressés à étudier les comptes « faux-nez », donc nous avons filtré uniquement les comptes bloqués pour cette raison pour une durée illimitée. Le nombre total de comptes bloqués pour cette raison est de 118 414.

Quand un compte est bloqué parce qu'il est détecté comme étant un « faux-nez », les administrateurs déclarent a quel compte d'origine il est lié, ce qui nous permet de récupérer l'ensemble des comptes liés à un même utilisateur initial.

2. 100 Go compressés en utilisant 7-Zip

Une fois que ces comptes ont été trouvés, nous avons extrait toutes les données liées dans les trente espaces de noms mentionnés dans le tableau 1. Nous avons en sus récupéré toutes les informations sur ces pages, y compris le comportement des autres utilisateurs sur les modifications des utilisateurs bloqués. Cela permet par exemple de détecter automatiquement si les modifications effectuées par un utilisateur spécifique sont conservées par d'autres utilisateurs ou annulées.

La base de données finale utilisée dans cette étude contient un total de 71 Go d'informations, stockées dans une base de données locale.

Sélection des comptes

La deuxième étape est la sélection des comptes actifs et « faux-nez » qui nous serviront de base d'entraînement et de test pour les algorithmes de détection.

En ce qui concerne les « faux-nez », nous avons d'abord regroupé les 118 414 comptes et les avons référé à leur compte utilisateur d'origine, *c-à-d* le premier compte créé par l'utilisateur. Nous avons obtenu 12 088 groupes, allant de 2 à 557 membres. Parmi ces groupes, on a choisi au hasard 5,000 comptes « faux-nez » qui sont dans un groupe contenant au moins 3 comptes, afin d'être sûrs que nous sélectionnons des comptes d'utilisateurs qui ont essayé de nombreuses fois de manipuler.

Un exemple de ces groupes est celui du compte original *Web Crawler*, qui a créé 4 comptes supplémentaires (*Fixthemistakes*, *Wordslayer*, *Wikidirt* et *PageOneEditor*) pour manipuler sur Wikipedia. Tous ces comptes sont bloqués comme « faux-nez » dans EnWiki parce qu'ils ont été créés uniquement pour la suppression des informations de la page *Dennis L. Montgomery*. Tous ces comptes utilisent une rhétorique similaire sur la protection des membres de la famille.

Une fois que les comptes « faux-nez » sont sélectionnés, nous devons aussi sélectionner les comptes actifs pour que l'algorithme ait en référence des comportements acceptables. Ceux-ci sont appelés comptes « actifs » parce qu'ils ne sont pas bloqués et ils peuvent utiliser Wikipedia sans aucune restriction. Nous sélectionnons au hasard 5000 comptes actifs, qui remplissent deux critères : ils sont actifs depuis plus d'un an et ils ont fait au moins une contribution sur Wikipedia. Ces critères ont été choisis afin d'avoir une forte probabilité que ces comptes ne sont pas des comptes « faux-nez » encore inconnus, et d'avoir un comportement non verbal suffisant.

Calcul des caractéristiques

Nous avons sélectionné et calculé un ensemble de caractéristiques, détaillées dans la section 4.1. Aucune de ces caractéristiques n'est directement disponible dans les données extraites.

Par conséquent, la troisième étape de notre méthode consiste à pré-traiter les données brutes pour la constitution des vecteurs de données d'apprentissage et de test. Cela se fait via des scripts PHP qui fonctionnent sur les données récupérées pour calculer toutes les valeurs des caractéristiques de chaque compte.

Les données traitées sont alors composées d'un ensemble de 11 caractéristiques pour chaque compte.

Apprentissage et test de l'algorithme

La dernière étape de notre méthode est l'apprentissage supervisé et les tests. Nous avons décidé de comparer nos caractéristiques de données extraites sur plusieurs algorithmes d'apprentissage automatique pour évaluer l'exactitude et la robustesse de notre méthode.

Afin d'évaluer l'efficacité de notre méthode, nous avons utilisé les algorithmes d'apprentissage automatique suivants : Machine à vecteurs de support (SVM) (Cortes, Vapnik, 1995), Forêt d'arbres décisionnels (RF) (Breiman, 2001), Bayésien Naïf (NB) (Russell *et al*., 1995), k plus proches voisins (KNN) (Altman, 1992), Réseaux Bayésiens (BN) (Heckerman, 2008) et Adaptive Boosting (ADA) (Freund, Schapire, 1995).

Ces algorithmes ont été testés à l'aide d'une validation croisée (*10-fold*) : on divise les données de 10,000 comptes en sous-ensembles complémentaires, on analyse le sous-ensemble de l'apprentissage, et on valide l'analyse sur l'ensemble de validation. La partition aléatoire, l'apprentissage et la validation sont répétés dix fois.

Pour quantifier la qualité de la prévision, nous avons sélectionné un ensemble de mesures décrites dans la section 4.2.

4.1. Sélection des caractéristiques

La qualité de l'apprentissage est déterminée par les caractéristiques formant le vecteur représentatif des données d'un compte. Nous proposons plusieurs caractéristiques pour différencier entre un compte « faux-nez » et un compte actif. Celles-ci peuvent être divisées en trois ensembles : comportement de contribution de l'utilisateur, comportement des autres utilisateurs envers ces contributions, et comportement du compte.

Puisque notre objectif est de détecter les « faux-nez », dont le but est de manipuler les contributions de l'encyclopédie, nous sommes intéressés par le comportement de contribution des utilisateurs. En particulier, nous sommes intéressés par le nombre et les zones -en termes d'espaces de noms- des contributions de chaque utilisateur dans Wikipedia. Nous sélectionnons aussi la moyenne d'octets ajoutés et supprimés dans chaque contribution, et la moyenne des contributions de chaque utilisateur dans le même article.

Deuxièmement, nous étudions la fréquence d'annulation directe d'une contribution d'un compte dans un article par un autre utilisateur (aussi appelé *revert*). Ceci est une utilisation indirecte des expériences d'autres utilisateurs qui peut être calculée automatiquement à partir des données.

Enfin, nous comparons l'intervalle entre le moment d'enregistrement de chaque utilisateur et leur première contribution.

Dans la deuxième étape de notre méthode, nous calculons pour chaque compte les valeurs des caractéristiques sélectionnées qui captent le comportement des utilisateurs sur Wikipedia. Ces caractéristiques sont détaillées ci-dessous.

Nombre de contributions de l'utilisateur par espace de noms

Nous composons les contributions de l'utilisateur selon les espaces de noms en six catégories : articles, discussions des articles, pages des utilisateurs, discussions des pages des utilisateurs, projets et *autres* (tous les autres espaces de noms sont réunis sous une caractéristique). Le choix de ces espaces de noms est motivé par leur importance pour détecter le comportement de contribution et les intérêts des contributeurs de Wikipedia.

Fréquence de revert après chaque contribution dans les articles

Nous faisons l'hypothèse que la plupart du temps, la manipulation d'un « fauxnez » est détectée et annulée par un autre utilisateur, car chaque page est gérée par plusieurs contributeurs et quand ils trouvent une contribution malveillante ils l'annulent directement. Dans cette caractéristique nous calculons donc la fréquence des *reverts* immédiats après chaque contribution dans un article.

On considère que Na est le nombre des contributions dans les pages d'article et que Nr est le nombre de reverts, la caractéristique Fr est donc :

$$Fr \quad = \frac{Nr*100}{Na} \tag{1}$$

Par exemple, un utilisateur qui a fait 9 contributions dans deux pages, mais 3 sur 9 de ses contributions ont été annulés, aura un score de 33 %.

Notons que cette caractéristique n'est pas directement disponible dans les données, et le prétraitement doit comparer tous les changements dans l'historique des pages d'articles pour détecter ces *reverts*.

Moyenne des octets ajoutés et supprimés dans chaque révision

Notre objectif pour ces deux caractéristiques est de trouver les déterminants du comportement de l'utilisateur durant l'écriture dans les articles. La première caractéristique calcule la moyenne du nombre d'octets d'informations ajoutées dans les articles pour toutes les contributions (révisions) de chaque compte, et la seconde calcule la moyenne des nombres d'octets d'informations supprimées dans les articles pour toutes les contributions de chaque compte.

$$Og+ \quad = \frac{\sum_{i=1}^{N} Op}{N} \tag{2}$$

On considère que Op est le nombre d'octets positifs et que N est le nombre de révisions.

$$Og- \quad = \frac{\sum_{i=1}^{N} On}{N} \tag{3}$$

On considère que On est le nombre d'octets négatifs et que N est le nombre de révisions.

L'hypothèse sous-jacente est qu'il est possible d'extraire des indicateurs du but de la manipulation du « faux-nez » d'après son comportement d'ajout et/ou de retrait. Le but de la manipulation d'un « faux-nez » peut être soit d'ajouter et de publier une ou des information(s) particulières(s), soit de supprimer une contribution précédente ou une partie des contributions précédentes.

Moyenne des contributions dans le même article

On calcule la moyenne des contributions dans le même article, parce que nous considérons qu'un manipulateur est plus enclin à essayer plusieurs fois de manipuler le(s) même(s) article(s). En effet, si un ensemble d'informations, à ajouter ou retirer, est visé par le manipulateur, ceux-ci seront concentrés dans un nombre de pages restreint par la structure même de l'encyclopédie.

Dans cette caractéristique, nous calculons donc le nombre de contributions par article pour chaque compte, puis nous calculons la moyenne de ces nombres.

$$Ci \quad = \sum A_i \tag{4}$$

On considère que A_i est le nombre de contributions dans le même article i

$$Cg \quad = \frac{\sum_{i=1}^{N} Ci}{N} \tag{5}$$

où N est le nombre d'articles.

Intervalle temporel entre l'enregistrement de l'utilisateur et sa première contribution

Dans cette caractéristique, on calcule la différence (en secondes) entre le temps de l'enregistrement et le temps de la première contribution dans EnWiki pour chaque compte. L'hypothèse sous-jacente est qu'un utilisateur dont l'intention est de mener à bien une manipulation crée au début de sa manipulation plusieurs comptes, puis les laisse dormir pour les utiliser séparément quand un compte actif sera bloqué.

Notons que cette dernière caractéristique et celle de la moyenne d'octets ajoutées et supprimées ont déjà été utilisées dans (Tsikerdekis, Zeadally, 2014), tandis que les caractéristiques des contributions dans les espaces de noms ont été modifiées pour qu'elles soient plus précises en les composants sur un plus grand nombre d'espaces de noms.

Les caractéristiques de la fréquence de *revert* après chaque contribution dans les articles et de la moyenne de contribution dans le même article sont, quant à elles, de nouvelles propositions.

4.2. Algorithmes et mesures

Afin d'évaluer l'efficacité de notre modèle, nous montrons d'abord la précision n de notre modèle et puis nous utilisons la matrice de confusion indiquée dans le tableau 2. Cette matrice est utilisée pour visualiser les performances des différents algorithmes en utilisant les mesures suivantes :

$$TauxdeVraisPositifs(TVP) \quad = \frac{VP}{VP+FN} \tag{6}$$

Cette mesure indique le taux de « faux-nez » vraiment détecté.

$$TauxdeFauxPositifs(TFP) \quad = \frac{FP}{FP+VN} \tag{7}$$

Cette mesure indique le taux de « faux-nez » faussement détecté.

$$Precision \quad = \frac{VP}{VP+FP} \tag{8}$$

Cette mesure indique la fraction des cas retournés qui sont des « faux-nez » valides.

$$F - mesure \quad = \frac{2*Precision*TVP}{Precision+TVP} \tag{9}$$

Cette mesure indique la fraction de la combinaison de TVP et de précision.

$$MCC = \frac{VP * VN - FP * FN}{\sqrt{(VP+FP)(VP+FN)(VN+FP)(VN+FN)}} \tag{10}$$

Le coefficient de corrélation de Matthews (MCC) indique une mesure équilibrée qui peut être utilisée même si les classes sont de tailles très différentes.

Tableau 2. Matrice de confusion utilisée pour évaluer l'efficacité de notre méthode

	Faux-nez	Comptes Actifs
Faux-nez Prédits	Vrai Positif (VP)	Faux Positif (FP)
Comptes Actifs Prédits	Vrai Negatif (VN)	Faux Negatif (FN)

Dans les travaux précédents, (Solorio *et al.*, 2013a) a évalué sa méthode proposée en utilisant Machine à vecteurs de support (SVM) seulement, tandis que (Tsikerdekis, Zeadally, 2014) a utilisé Machine à vecteurs de support (Cortes, Vapnik, 1995), Forêt d'arbres décisionnels (RF) (Breiman, 2001), et Adaptive Boosting (ADA) (Freund, Schapire, 1995) pour évaluer sa méthode.

Nous résumons nos résultats obtenus en utilisant les indicateurs de performance mentionnés et la matrice de confusion dans le tableau 2 de la section suivante.

5. Expérimentation et résultats

Nous avons utilisé l'ensemble de données présentées dans la section 4 qui comprend 10 000 comptes, la moitié comporte les données de comptes bloqués et l'autre moitié celles de comptes actifs.

Nous avons utilisé la boîte à outils Weka [3] pour expérimenter les résultats d'algorithmes de classification sur les caractéristiques proposées, selon une validation croisée *ten-fold*. Les données ont été séparées de façon aléatoire à 2/3 pour l'apprentissage et 1/3 pour les tests. Pour les données d'apprentissages nous avons utilisé la validation croisée *ten-fold* pour trouver les meilleurs paramètres de chaque algorithme d'apprentissage automatique, puis on a choisi les meilleurs paramètres trouvés pour la validation sur les données de test. Les résultats obtenus sont présentés dans les tableaux 3 et 4, ainsi que la *figure* 2. Nous discutons ensuite ces résultats dans la section suivante.

5.1. Comparaison des algorithmes

Le tableau 3 compare le pourcentage de précision de notre modèle entre différents algorithmes d'apprentissage automatique, Il montre que nous avons obtenu la meilleure précision en utilisant Adaptive Boosting (99,9 %), Forêt d'arbres décisionnels (99,8 %) et les Réseaux Bayésiens (99,6 %).

Tableau 3. Pourcentage de précision de notre modèle proposé entre différents algorithmes d'apprentissage automatique

Algorithme d'apprentissage automatique	Précision du modèle
SVM	78,1 %
Forêt d'arbres décisionnels	99,8 %
Naïve Bayésienne	50,1 %
K Plus Proches Voisins	63 %
Réseau Bayésien	99,6 %
Adaptive Boosting	99,9 %

Les résultats des algorithmes restants sont nettement inférieurs, avec SVM à $78,1\,\%$, KNN qui a comme résultat 63 %, et enfin le résultat le plus faible est le Bayésien Naïf à 50,1 %. Le tableau 4 et la *figure* 2 comparent les différentes valeurs des indicateurs de notre modèle proposé avec plusieurs algorithmes d'apprentissage automatique. Ils montrent que le TVP, la précision, le rappel et la F-mesure en utilisant l'algorithme Forêt d'arbres décisionnels sont 0,998/1, et que le MCC est 0,997. La F-mesure en utilisant SVM donne 0,771 et 0,628 en utilisant KNN.

Ces résultats montrent que seuls les algorithmes *Adaptive Boosting*, forêt d'arbres décisionnels et réseaux bayésiens sont efficaces en utilisant nos caractéristiques pour

3. http://weka.wikispaces.com

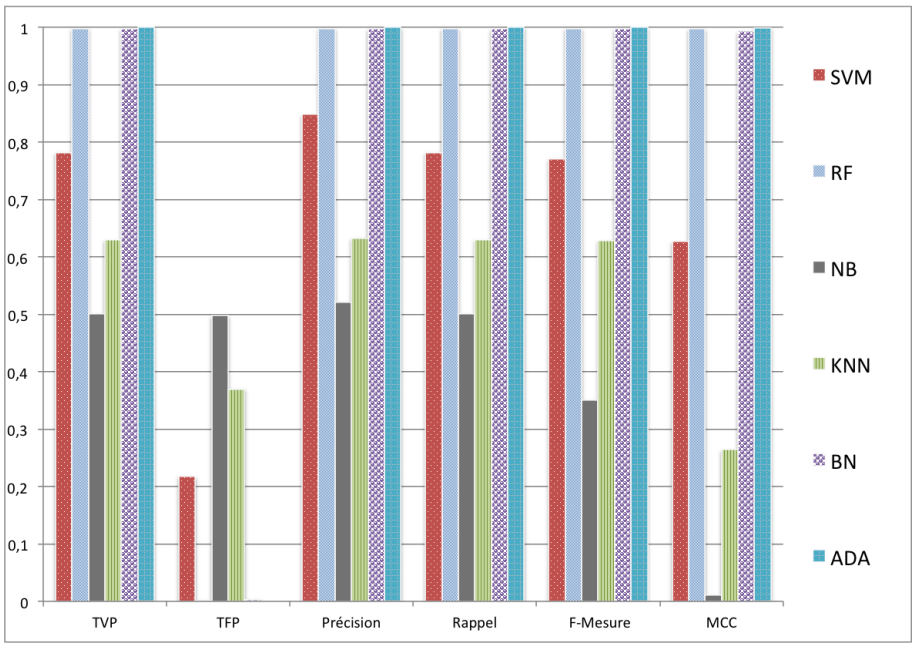

Figure 2. Performances des différents algorithmes d'apprentissages automatiques

Tableau 4. Performances des algorithmes d'apprentissage automatique

	TVP	TFP	Précision	Rappel	F-Mesure	MCC
SVM	0,782	0,218	0,848	0,782	0,771	0,627
Forêt d'arbres décisionnels	0,998	0,002	0,998	0,998	0,998	0,997
Naïve Bayésienne	0,501	0,498	0,521	0,501	0,351	0,011
K Plus Proches Voisins	0,630	0,370	0,633	0,630	0,628	0,264
Réseau Bayésien	0,997	0,003	0,997	0,997	0,997	0,993
Adaptive Boosting	1,000	0,000	1,000	1,000	1,000	0,999

la suppression automatique des comptes « faux-nez », et que l'algorithme SVM montre également des résultats acceptables.

5.2. Comparaison avec la littérature

Nous présentons une comparaison entre les résultats de notre méthode proposée et trois méthodes précédentes dans le tableau 5. On peut observer que notre méthode donne le meilleur pourcentage de précision entre toutes les différentes méthodes en atteignant 99,9 %. La méthode de détection des violations fondée sur les indices non verbaux (Tsikerdekis, Zeadally, 2014) atteint un résultat de 71,3 %. l'algorithme des

attributs du texte (Solorio *et al*., 2013b) a une précision de 73 %, et enfin l'algorithme de traitement du langage naturel (Solorio *et al*., 2013a) atteint 68,8 % comme précision globale.

Une différence significative entre les deux dernières études est le nombre de données traitées. Alors que nous avons utilisé 10 000 cas, les deux études de Solorio ne s'appuyaient que sur respectivement 623 et 77 cas. En effet, le calcul de l'analyse du comportement verbal est plus lourd que celui du comportement non verbal, et ne peut pas être aussi facilement traité.

Par rapport à la première étude, qui a utilisé 15 000 cas, seule la sélection des caractéristiques est sensiblement différente et explique la variation des résultats.

Tableau 5. Comparaison des résultats entre notre méthode proposée et les méthodes précédentes

	Précision globale	L'ensemble de données
Algorithme des attributs de texte (Solorio *et al*., 2013b)	73 %	623 cas
Algorithme de Traitement du Langage Naturel (Solorio *et al*., 2013a)	68,8 %	77 cas
Méthode de détection des violations non verbaux (Tsikerdekis, Zeadally, 2014)	71,3 %	15 000 cas
Notre méthode proposée	99,9 %	10 000 cas

6. Discussion des résultats

Les résultats présentés dans la section précédente sont les meilleurs par comparaison entre notre méthode et les trois méthodes précédentes (Solorio *et al*., 2013b ; 2013a ; Tsikerdekis, Zeadally, 2014). Nous avons constaté que l'*Adaptive Boosting* a la meilleure précision entre tous les algorithmes, et que les algorithmes forêt d'arbres décisionnels et réseaux bayésiens sont pratiquement aussi efficaces. Le fait que trois algorithmes montrent des résultats de précision à plus de 99,5 % est une indication de la solidité de notre sélection de caractéristiques.

Sans surprise, nous avons également constaté que la précision minimale est obtenue avec l'algorithme bayésien Naïf avec 52 %, car cet algorithme travaille sur la probabilité de chaque valeur dans les paramètres indépendamment, ce qui peut mener à des mauvais résultats. Par contre, si on utilise des données discrétisées, c-à-d, on transforme les variables quantitatives en qualitatives ordinales en les découpant en classes (intervalles), les résultats augmentent à 95 %.

La meilleure méthode pour la détection de « faux-nez » avant notre contribution a été la méthode de (Tsikerdekis, Zeadally, 2014). Si nous comparons notre méthode à celle-ci, nous trouvons que les deux méthodes utilisent seulement les indicateurs non verbaux et que la précision globale est de 71 % pour la méthode de Tsikerdekis

(Tsikerdekis, Zeadally, 2014) en face de 78 % pour la nôtre en utilisant SVM avec fonction de radiale de base (RBF) comme noyau. Cependant, le SVM linéaire n'a pas la meilleure précision, ce qui est normal parce que les données ne peuvent pas être détectées facilement à l'aide d'une fonction linéaire.

Grâce à notre méthode nous pouvons dire que, au mieux de notre connaissance, nous avons sélectionné le meilleur ensemble de caractéristiques qui permette de détecter les manipulateurs d'identités, arrivant à une précision de 99,9 %.

Cette précision est due aux caractéristiques non verbales sélectionnées, nécessitant des prétraitements, et au grand nombre de données qui ont aidé l'algorithme à bien apprendre en utilisant les données qui représentent le comportement des « faux-nez ».

Nous évaluons notre liste de caractéristiques en utilisant la méthode d'évaluation de l'attribut de corrélation et nous avons constaté que les trois meilleures - plus importantes- caractéristiques sont la fréquence des *reverts* après chaque contribution dans les articles, puis la moyenne d'octets ajoutés dans chaque révision, et la troisième est la moyenne de contribution dans le même article. Ceci valide et explique nos résultats parce que les premières et troisièmes caractéristiques différencient notre méthode de celle de (Tsikerdekis, Zeadally, 2014).

Par ailleurs, il est possible d'utiliser ou adapter ces caractéristiques pour plusieurs medias sociaux. Par exemple, nous montrons dans le tableau 6 un example des caractéristiques qui pourraient être utilisées sur des données de Facebook. Notons que certaines de ces données ne sont pas accessibles publiquement, et qu'il faudrait donc pour sa mise en place accéder aux données d'administration de ces sites.

Tableau 6. Comparaison entre les caractérisques utilisés sur wikipedia et les caractéristiques qu'on peut utiliser pour des données de Facebook

Caractéristiques projets collaboratifs	Caractéristiques Facebook
Nombre de contributions de l'utilisateur par espace de noms	Nombre de contributions de l'utilisateur dans les different types de pages Facebook (groupes, profil, page d'ami,...)
Fréquence de *revert* après chaque contribution dans les articles	Fréquence des rapports et des suppressions des commentaires faits par les autres utilisateurs
Moyenne des octets ajoutés et supprimés dans chaque révision	Moyenne des octets ajoutés dans le profil personnel et celui des autres
Moyenne des contributions dans le même article	Moyenne des contributions dans le profil du même ami ou page ou groupe
Intervalle temporel entre l'enregistrement de l'utilisateur et sa première contribution	Intervalle temporel entre l'enregistrement de l'utilisateur et sa première contribution sur Facebook

Nous pouvons considérer que notre méthode est efficace dans la détection des « faux-nez » parce que le manipulateur n'est pas nécessairement au courant des indicateurs non verbaux. En outre, il ne peut pas les manipuler aussi facilement, par exemple la fréquence de *revert* après chaque révision est un très bon indicateur parce

que le manipulateur ne peut pas contrôler cette fonction pour essayer de se cacher de la détection, celle-ci étant issue des autres contributeurs de Wikipédia. Le prétraitement de données non triviales peut également créer une difficulté pour les manipulateurs d'être conscients qu'il y a une méthode qui détecte l'intervalle de temps entre l'enregistrement et la première contribution.

Enfin, les utilisateurs qui ont un objectif de manipuler dans un article précis ont une difficulté de fuir la détection et ne peuvent pas changer facilement leur comportement d'écriture non verbale s'ils veulent atteindre leur objectif.

7. Conclusion et perspectives

Dans cet article, nous présentons une méthode pour détecter les détenteurs de comptes multiples dans les projets collaboratifs en ligne. Afin de tester et valider notre modèle, nous proposons des caractéristiques pour les comportements non verbaux en utilisant des données réelles issues de comptes Wikipédia anglophones. Toutes les caractéristiques peuvent être également appliquées sur les données issues d'autres sites de projets collaboratifs.

Le succès d'un algorithme d'apprentissage automatique dépend de la sélection des caractéristiques qui sont les entrées de l'algorithme. De cette façon, nous avons atteint une précision de 99,7 % avec trois algorithmes, *Adaptive Boosting*, forêt d'arbres décisionnels et réseau bayésien, et 78 % avec SVM. Les algorithmes d'apprentissage automatique ont été formés et testés en utilisant la méthode de validation croisée *ten-fold* pour 10 000 cas comme données.

Ces résultats favorables sont dus à un ensemble de caractéristiques bien choisies qui aident à identifier un « faux-nez ». Cet ensemble de caractéristiques est automatiquement calculé à partir des données publiquement disponibles.

Si le SVM devait être utilisé dans un algorithme de blocage automatique, l'administrateur doit être conscient qu'il y a 12 % de faux positifs, qui nécessitent donc des appels examinés par des administrateurs humains. Même avec les deux meilleurs méthodes, la révocation doit être l'objet d'appels car les faux positifs existent (même si ceux-ci sont faibles, représentant moins de 0,2 % des traitements).

Notre méthode a donné de très bons résultats. Dans l'avenir, nous prévoyons d'explorer d'autres médias sociaux pour vérifier notre ensemble de caractéristiques dans d'autres contextes, tels que les forums ou twitter. Cela peut conduire à des difficultés, d'une part parce que les comportements varient selon le type de média social, et d'autre part parce que les caractéristiques doivent être adaptées aux données produites. Par exemple, si la notion de *revert* peut être assimilée à une suppression de message dans les services de blogging, forums et réseaux sociaux, cette donnée n'est accessible que par l'utilisateur du compte et par l'administrateur du site. Dans ce cas, seule la mise en place de l'algorithme au sein même de la plateforme peut permettre la détection de faux comptes.

De plus, nous souhaitons améliorer dans ce cas la détection de comptes multiples en se fondant sur l'analyse de la communication verbale, en plus du comportement non verbal, dans les médias sociaux qui utilisent beaucoup les expressions en texte libre.

Enfin, notre travail démontre que les techniques de détection automatisées peuvent être utilisées avec succès dans EnWiki, afin qu'elles puissent également être utilisées dans d'autres projets de collaboration comme Wikitionary et Wikiversity. Une étape supplémentaire serait de détecter non seulement les comptes multiples, mais également les groupes de comptes auquels ils appartiennent.

Bibliographie

Altman N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, vol. 46, n° 3, p. 175–185.

Ambika C. M. (2014, December). *The evolution of social media 2004 - 2014: The good, the bad and the ugly of it !* (http://dazeinfo.com/2014/12/12/evolution-social-media-2004-2014-good-bad-ugly/)

Breiman L. (2001). Random forests. *Machine learning*, vol. 45, n° 1, p. 5–32.

Cao Q., Sirivianos M., Yang X., Pregueiro T. (2012). Aiding the detection of fake accounts in large scale social online services. In *Proceedings of the 9th usenix conference on networked systems design and implementation*, p. 15–15.

Cortes C., Vapnik V. (1995). Support-vector networks. *Machine learning*, vol. 20, n° 3, p. 273–297.

David B. (2015, MARS). *5 social engineering attacks to watch out for.* (http://tripwire.com/state-of-security/security-awareness/5-social-engineering-attacks-to-watch-out-for/)

Douceur J. R. (2002). The sybil attack. In *Peer-to-peer systems*, p. 251–260. Springer.

Freund Y., Schapire R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, p. 23–37.

Gao H., Hu J., Wilson C., Li Z., Chen Y., Zhao B. Y. (2010). Detecting and characterizing social spam campaigns. In *Proceedings of the 10th acm sigcomm conference on internet measurement*, p. 35–47.

Goolsby R., Shanley L., Lovell A. (2013). *On cybersecurity, crowdsourcing, and social cyber-attack*. Rapport technique. DTIC Document.

Heckerman D. (2008). A tutorial on learning with bayesian networks. In *Innovations in bayesian networks*, p. 33–82. Springer.

Jeff B. (2015). *33 social media facts and statistics you should know in 2015.* (http://www.jeffbullas.com/2015/ 04/08/ 33-social-media-facts-and-statistics-you-should-know-in-2015/)

Kaplan A. M., Haenlein M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, vol. 53, n° 1, p. 59–68.

Maeve D., Nicole E., Cliff L., Amanda L., Mary M. (2015, January). *Social media update 2014.* (http://www.pewinternet.org/2015/01/09/social-media-update-2014/)

Mathew I. (2012, February). *If you think twitter doesn't break news, you're living in a dream world.* (https://gigaom.com/2012/02/29/if-you-think-twitter-doesnt-break-news-youre-living-in-a-dream-world/)

Norajong. (2010, May). *Why the number of people creating fake accounts and using second identity on facebook are increasing.* (http://networkconference.netstudies.org/2010/05/why-the-number-of-people-creating-fake-accounts-and-using-second-identity-on-facebook-are-increasing/)

Norton. (s. d.). *Spear phishing: Scam, not sport.* (http://us.norton.com/spear-phishing-scam-not-sport/article)

Riva R. (2010, May). *Stolen facebook accounts for sale.* (http://www.nytimes.com/2010/05/03/technology/internet/ 03facebook.html)

Russell S., Norvig P., Intelligence A. (1995). A modern approach. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, vol. 25, p. 27.

Sarita Y., Daniel R., Schoenebeck G., danah b. (2009). Detecting spam in a twitter network. *First Monday*, vol. 15, n° 1. Consulté sur http://firstmonday.org/ojs/index.php/fm/article/view/2793

Solorio T., Hasan R., Mizan M. (2013a). A case study of sockpuppet detection in wikipedia. In *Workshop on language analysis in social media (lasm) at naacl hlt*, p. 59–68.

Solorio T., Hasan R., Mizan M. (2013b). Sockpuppet detection in wikipedia: A corpus of real-world deceptive writing for linking identities. *arXiv preprint arXiv:1310.6772*.

Statista. (2015). *Number of unique u.s. visitors to wikipedia.org from may 2011 to april 2015 (in millions).* (http://www.statista.com/statistics/265119/number-of-unique-us-visitors-to-wikipediaorg/)

Sture N. (2010, February). *Fake accounts in facebook - how to counter it.* (http://ezinearticles.com/?id=3703889)

Tsikerdekis M., Zeadally S. (2014). Multiple account identity deception detection in social media using nonverbal behavior. *Information Forensics and Security, IEEE Transactions on*, vol. 9, n° 8, p. 1311–1321.

Yang Z., Wilson C., Wang X., Gao T., Zhao B. Y., Dai Y. (2014). Uncovering social network sybils in the wild. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 8, n° 1, p. 2.