Une méthode collaborative pour identifier les spams: contribution à la qualité de l'information dans les réseaux sociaux

Mahdi Washha, Manel Mezghani, Florence Sèdes

Une méthode collaborative pour identifier les spams: contribution à la qualité de l'information dans les réseaux sociaux

Mahdi Washha — Manel Mezghani — Florence Sèdes

Institut de Recherche en Informatique de Toulouse (IRIT), Université de Toulouse, CNRS, INPT, UPS, UT1, UT2J, 31062 TOULOUSE Cedex 9, France
mahdi.washha,manel.mezghanni,florence.sedes@irit.fr

RÉSUMÉ. *Contrer les actions des utilisateurs mal intentionnés dits "spammeurs" est un réel défi pour maintenir un haut niveau de performance dans les applications mises en oeuvre dans les réseaux sociaux. Les méthodes conventionnelles de détection de spams imposent des délais de traitement importants et inévitables, allant par exemple jusqu'à des mois pour traiter de grandes collections de tweets. Ces méthodes entièrement dépendantes de l'approche d'apprentissage supervisé choisie pour produire des modèles de classification, requièrent un ensemble de données vérité terrain qui n'est pas disponible pour ce type d'applications. Nous proposons donc une méthode basée sur un modèle linguistique non supervisé qui effectue une collaboration avec d'autres réseaux sociaux pour détecter les tweets spam dans des sujets qui génèrent de gros volumes d'échanges, par exemple à partir des hashtags utilisés. Notre méthode a été expérimentée sur plus de 6 millions de tweets postés dans 100 thématiques "tendances". Facebook est utilisé en parallèle comme vérité terrain permettant ainsi la collaboration de deux réseaux sociaux différents. Nos expérimentations démontrent une efficacité en ce qui concerne le temps de traitement et la performance de classification, par rapport aux méthodes classiques de détection de spam dans les tweets.*

ABSTRACT. *Prevent the actions of malicious users called "spammers" is a real challenge to maintain a high level of performance in applications implemented in social networks. Conventional spam detection methods impose large and unavoidable processing times, for example up to months for processing large collections of tweets. These methods entirely dependent on the supervised learning approach chosen to produce classification models, require a set of ground truth data that is not available for this type of applications. We propose a method based on an unsupervised linguistic model that collaborates with other social networks to detect spam tweets in subjects that generate large volumes of exchanges, for example from used hashtags. Our method has been experimented on more than 6 million tweets posted in 100 trending topics. Facebook is used in parallel as a ground truth allowing the collaboration of two different social networks. Our experiments show an efficiency with regard to processing time and classification performance, compared to the conventional methods of detecting spam in tweets.*

MOTS-CLÉS : *Spam social*$_1$, *Réseaux sociaux*$_2$, *Collaboration*$_3$, *Thématiques tendances*$_4$.

KEYWORDS: *Social spam*$_1$, *Social networks*$_2$, *Collaboration*$_3$, *Trending topics*$_4$.

1. Introduction

Avec l'énorme popularité des réseaux sociaux en ligne (OSN), les utilisateurs "indésirables" dits spammeurs se multiplient pour diffuser du contenu considéré comme spam (par exemple : publicité, matériel pornographique et sites web d'hameçonnage) (Benevenuto *et al.*, 2010). Cette diffusion peut causer des problèmes majeurs tels que : (i) polluer les résultats de recherche ; (ii) dégrader l'exactitude des statistiques obtenues à travers des outils d'extraction d'information ; (iii) consommer des ressources de stockage ; (iv) violer la vie privée des utilisateurs. Les mécanismes anti-spam s'avèrent insuffisants pour mettre fin au problème de spam, ce qui suscite de réelles inquiétudes quant à la qualité des collections de données "aspirées". La qualité de l'information est défini par l'aptitude de l'information à être utiliser dans un contexte particulier (Agarwal et Yiliyasi, 2010).
Le filtrage des données "bruitées" pour avoir des informations de meilleure qualité est d'évidence la solution efficace pour améliorer les résultats des moteurs de recherche et les systèmes de recherche d'information. Le processus de qualité de l'information dans les réseaux sociaux, décrit dans la figure 1, est synthétisé génériquement en trois étapes dépendantes (Agarwal et Yiliyasi, 2010) : (i) sélectionner des collections de données (par exemple : Comptes Facebook, Tweets, messages Facebook) qui nécessitent des améliorations ; (ii) déterminer le type de bruit (par exemple spam, rumeur) à filtrer ; enfin (iii) appliquer des algorithmes pré-conçus en fonction du type de bruit choisi pour produire des collections de données non "bruitées".

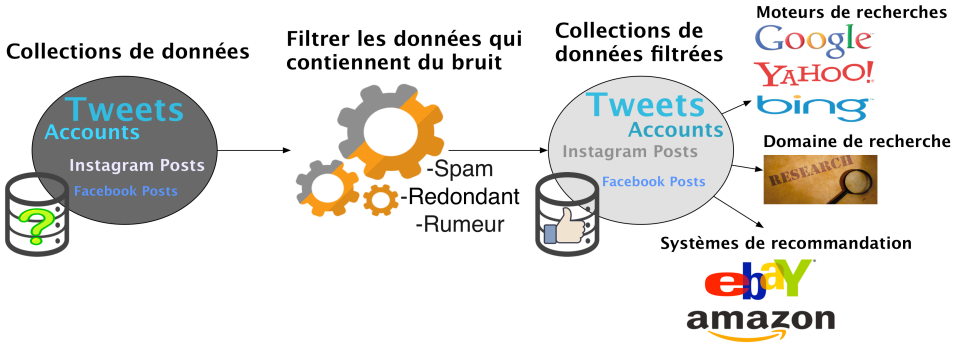

Figure 1. *Un aperçu du processus de qualité de l'information dans les réseaux sociaux.*

Divers types de bruit existent dans les réseaux sociaux. Notre contribution se focalise sur la question liée au problème du spam social. Notre équipe mène des recherches (Mezghani *et al.*, 2014 ; Abascal-Mena *et al.*, 2015 ; Canut *et al.*, 2015) abordant un large éventail de problèmes dans les réseaux sociaux comme le profilage social, l'enrichissement des profils, la détection des intérêts sociaux et la détection de communautés dites socio-sémantiques. La plate-forme Twitter a été adoptée afin d'effectuer les expérimentations nécessaires à la validation de nos contributions. Un des facteurs centraux de réussite de nos évaluations est donc la qualité des données constituant les collections sur lesquelles nous expérimentons.

Dans la bataille de la lutte contre le spam sur Twitter, de nombreuses méthodes (Wang, 2010 ; Benevenuto *et al.*, 2010 ; Yardi *et al.*, 2009 ; Stringhini *et al.*, 2010 ; Yang *et al.*, 2011 ; Chu *et al.*, 2012b ; Amleshwaram *et al.*, 2013) ont été proposées pour détecter

les comptes spam et les campagnes de spam, mais peu d'attention dédiée à la détection des tweets "spam" (tweets indésirables, parasites, malveillants, "malicieux", ...). Les méthodes de détection au niveau des comptes et des campagnes prennent du temps, nécessitant des mois pour traiter de grandes collections de millions d'utilisateurs de Twitter. La principale source de consommation de temps est l'utilisation restreinte de l'API REST[1] pour récupérer une information requise (par exemple : *followers*, *followees*, activité de l'utilisateur). Les méthodes de détection de spam existantes au niveau des tweets sont fondées sur l'exploitation du concept d'extraction des algorithmes d'apprentissage automatique supervisés pour construire un modèle prédictif à l'aide d'un ensemble de données. La principale force des méthodes basées sur les tweets est la détection rapide puisque le processus de détection est effectué sur les informations disponibles dans le tweet lui-même. Toutefois, compte tenu de la dynamique des comportements de spammeurs, les méthodes de détection de spam au niveau du tweet s'avèrent limitées : i) utilisation de caractéristiques non discriminatoires et inefficaces telles que le nombre de mots dans le tweet ; (ii) nécessité pour un ensemble de données annotées de construire un modèle de classification ; et (iii) utilisation d'algorithmes d'apprentissage supervisés produisant des modèles biaisés par les données d'apprentissage choisies (non-généricité, non-exhaustivité, etc.).

Dans cet article, nous présentons une méthode non supervisée pour filtrer les spams dans les tweets dans des collections à grande échelle. Notre méthode réalise une collaboration avec d'autres OSN par la recherche et la collecte d'informations pertinentes concernant les thématiques choisies. La correspondance de contenus est effectuée entre un tweet donné et des informations pertinentes extraites, comme par exemple à partir des publications Facebook, afin de décider plus tard la classe qui correspond au mieux à ce tweet. Dans ce travail, nous supposons que le volume et le contenu du spam sur les OSN varient en fonction des règles de confidentialité suivies par les OSN. Par exemple, le réseau social Facebook[2] adopte des règles plus restrictives que celles de Twitter dans l'ouverture de nouveaux comptes (par exemple : la vérification du numéro de téléphone), ce qui rend plus difficile le lancement de grandes campagnes de spam. Nos expérimentations ont été menées en choisissant des thématiques "tendances"[3] et testées sur la plateforme OSIRIM[4].

Le reste de l'article est organisé comme suit. La section 2 donne un aperçu sur les méthodes de détection de spam basées sur Twitter. La section 3 présente les notations, la formalisation du problème et la conception de la méthode collaborative que nous proposons pour détecter les tweets spam. La section 4 décrit l'ensemble de données utilisé pour valider notre approche. La section 5 présente les résultats. La section 6 conclut ce travail en fournissant un bilan et en dressant des perspectives.

1. https ://dev.twitter.com/rest/public
2. hhtps ://www.facebook.com/policies
3. elles sont affichées dans Twitter en haut à gauche d'un compte en tant que *Trending Topics*
4. http ://osirim.irit.fr/site/

2. Aperçu des méthodes de détection de spam basées sur Twitter

La plupart des travaux existants pour lutter contre le spam sur Twitter se sont concentrés sur le compte Twitter et la campagne Twitter (ce terme signifie un ensemble de comptes généré automatiquement à travers un bot (robot) informatique) comme niveaux de détection avec peu d'efforts consacrés à la détection de spams.

Niveau tweet : À ce niveau, les tweets individuels sont vérifiés afin d'éliminer d'éventuels contenus indésirables.. Benevenuto (Benevenuto *et al.*, 2010) a extrait un ensemble de caractéristiques statistiques simples du tweet telles que le nombre de mots, le nombre de hashtags et le nombre de caractères. Ensuite, un classifieur binaire est construit sur un petit ensemble de données annotées. Martinez-Romo et Araujo (Martinez-Romo et Araujo, 2013) ont détecté des tweets spam dans les thématiques tendances à travers l'utilisation de modèles de langage pour extraire plus de fonctionnalités telles que la divergence de distribution de probabilité entre un tweet donné et d'autres tweets. Le problème majeur à ce niveau de détection provient du manque d'informations qui peuvent être extraites du tweet lui-même. En outre, la construction de modèles de langues à l'aide de tweets dans les thématiques tendances échoue définitivement quand il y a d'énormes attaques de spam. Notre travail permet de surmonter ces lacunes en exploitant les informations pertinentes dans d'autres OSN.

Niveau compte : Les méthodes conçues dans (Wang, 2010 ; Benevenuto *et al.*, 2010 ; Stringhini *et al.*, 2010 ; Mccord et Chuah, 2011 ; Cao et Caverlee, 2015) construisent d'abord des vecteurs en extrayant des caractéristiques extraites "à la main" telles que le nombre de *followers*, et l'intermédiation de noeuds. Ensuite, des algorithmes d'apprentissage automatique supervisés sont appliqués pour construire un modèle de classification sur un ensemble de données annotées. Malgré un taux de détection élevé en exploitant ces fonctionnalités, les extraire est chronophage en raison du temps nécessaire au recueil des informations du serveur de Twitter via l'utilisation de l'API REST. En effet, ces API sont limitées à un certain nombre prédéfini d'appels, ce qui rend l'extraction de la plupart des fonctionnalités impossible, en particulier dans le traitement de données à grande échelle.

Niveau campagne : Chu et al. (Chu *et al.*, 2012b) ont proposé une méthode de détection de campagne à travers le regroupement des comptes spam selon les URL disponibles dans les tweets. Un vecteur est ensuite représenté, via des caractéristiques similaires aux méthodes de détection au niveau du compte. Dans (Chu *et al.*, 2012a), un modèle de classification a été conçu pour capturer les différences entre bot, humain et *cyborg*. Malheureusement, ce niveau de détection présente des inconvénients similaires à ceux mentionnés pour le niveau "compte", ce qui rend ces solutions non évolutives pour de grandes collections d'utilisateurs ou de tweets.

3. Conception du modèle collaboratif

Notre approche se concentre sur la recherche d'une information appariée à partir d'autres OSN pour un tweet donné lié à une thématique donnée. Comme le but évident de l'utilisation de la modélisation des thématiques dans les OSN est de regrouper des informations similaires, la probabilité de trouver la même information parlant de la même thématique sur différents OSN est relativement élevée. Inversement, la probabilité de trouver le même contenu de spam

affiché sous la même thématique est relativement faible en raison de sa dépendance à l'égard des objectifs des spammeurs et de l'ouverture des OSN eux-mêmes. Par conséquent, au lieu d'extraire des fonctionnalités non informatives (par exemple le nombre de mots du tweet) pour apprendre le modèle à l'aide d'algorithmes d'apprentissage automatique, nous utilisons le concept de modèle de langage statistique (*statistical language model concept*) pour détecter les tweets considérés comme spam.

3.1. *Notations et définitions*

Notons $C_H = \{T_1, T_2, ...\}$ une collection de tweets pour une thématique donnée H où $T_\bullet$ représente le tweet modélisé comme étant 2-tuple $T_\bullet =< Texte, Actions >$. De plus, nous modélisons les informations récupérées sur la thématique, H, à partir de réseaux sociaux définis (par exemple : Facebook, Instagram), $SN_\bullet$, comme étant un ensemble fini $S_H = \{SN_{Facebook}, SN_{Instagram}, ...\}$. Chaque $SN_\bullet$ est modélisé comme étant un ensemble fini de *posts* $SN_\bullet = \{O_1, O_2, ...\}$, où l'élèment $O_\bullet$ est défini par 2-tuple $O_\bullet =< Texte, Actions >$. Chaque élément à l'intérieur de l'article O et tweet T tuple est défini comme suit :

Texte : Comme chaque message peut être constitué de texte, nous représentons le contenu du message comme un ensemble fini de mots, $Texte = \{w_1, w_2, ...\}$.

Actions : Les utilisateurs des réseaux sociaux peuvent effectuer des actions sur les *posts*[5] en réaction au contenu des *posts* ou des tweets. Nous définissons les actions comme un ensemble fini de 2-tuple, $Actions = \{< a_{name_1}, a_{val_1} >, < a_{name_2}, a_{val_2} >, ...\}$, où a_{nom} représente le nom de l'action (par exemple : aimer, partage et commenter sur Facebook) en fonction du réseau social considéré, et $a_{val} \in N_{ge0}$ est le nombre de fois que l'action correspondante effectuée par les utilisateurs du réseau social sur le *post* considéré ou le tweet.

3.2. *Formalisation du problème*

Étant donné un ensemble de tweets C_H associés à une thématique tendance H, et affichés par un ensemble d'utilisateurs distincts U_H tels que $U_H \leq |C_H|$, notre problème principal est de filtrer les tweets spam dans la collection donnée C_H sans impliquer l'information nécessitant des appels à l'API REST. Plus formellement, nous cherchons à concevoir une fonction f telle qu'elle prédise l'étiquette de classe de chaque tweet dans une collection donnée, définie comme $f(T) : T \rightarrow \{spam, non - spam\}, T \in C_H$.

3.3. *Probabilité, Priorité et Classification de tweet*

Probabilité de tweet : Nous utilisons les modèles de langage statistique (Ponte et Croft, 1998) pour estimer le degré de pertinence des *posts* dans d'autres OSN par rapport à un tweet donné. Ceci nous permettra de prendre une décision plus tard sur le tweet (il est un

5. Un *post* est une information publiée par un utilisateur

spam ou non-spam). La méthode de modélisation de langage calcule la probabilité $P(D|Q)$ d'un document D généré par une requête Q pour classer un ensemble de documents. Nous transformons le même concept pour obtenir le *post* le plus pertinent dans d'autres réseaux sociaux pour un tweet donné. Ainsi, nous traitons les tweets comme des requêtes et les *posts* comme des documents, en calculant la probabilité d'un *post* O à être générée par un tweet T comme :

$$P^{SN_i}(O|T) \stackrel{\text{rank}}{=} P^{SN_i}(O).P^{SN_i}(T|O) = P^{SN_i}(O). \prod_{w \in T.Text} P^{SN_i}(w|O) \quad [1]$$

$P^{SN_i}(O)$ est la probabilité a priori d'un *post* O telle que $O \in SN_i$. Cette probabilité peut être considérée comme une fonction indépendante du tweet (c'est-à-dire les caractéristiques **non** extraites du tweet) représentant la probabilité d'être du contenu non-spam dans le réseau social SN_i. On peut calculer l'autre composante de probabilité $P^{SN_i}(T|O)$ en utilisant différents modèles (Jelineck Mercer, Dirichlet) (Ponte et Croft, 1998) pour calculer $P^{SN_i}(O)$ ou (divergence de Kullback-Leibler) (Kullback, 1987) pour calculer le degré de dissimilarité entre les modèles de tweet et du langage du *post*. Dans cet article, nous utilisons le modèle de langage *uni-gram* pour représenter les tweets et les *posts* en raison de sa remarquable performance dans le domaine de la recherche d'information. En outre, nous adoptons la méthode de divergence de Kullback-Leibler (KL) en raison de son temps de calcul rapide par rapport aux autres méthodes. Cependant, la version classique de la méthode KL ne peut pas être exploitée directement dans le calcul de la probabilité $P^{SN_i}(T|O)$ puisque la valeur nulle de KL signifie que les modèles de langue de tweet et de *post* sont complètement similaires. De plus, l'intervalle de la méthode KL est non borné, ce qui signifie que la valeur ∞ apparaît lorsque deux modèles de langage sont différents. Par conséquent, nous personnalisons la version courante de la méthode KL pour inverser la sémantique des valeurs de KL (c-à-d $0 \implies$ non silimaire et $1 \implies$ similaire) en limitant ses valeurs, où la composante de probabilité $P^{SN_i}(T|O)$ est définie comme :

$$P^{SN_i}(T|O) = \frac{\log|T.Text| - \sum_{w \in T.Text} P(w|M_T) * min(|\log \frac{P(w|M_T)}{P(w|M_O)}|, \log|T.Text|)}{\log|T.Text|} \quad [2]$$

Où $P(w|M_T)$ et $P(w|M_O)$ sont les probabilités que le mot w soit généré par les modèles de langue du tweet et du *post* (M_T, M_O), respectivement.

Priorité de tweet : Comme les *posts* récupérés du réseau social SN_i peuvent être des contenus spams, nous estimons la probabilité d'être non-spam en exploitant les actions effectuées par les utilisateurs sur les *posts* récupérés (c'est-à-dire plus d'actions $\implies$ faible probabilité d'être un spam). Nous supposons que les actions (par exemple aimer, faire un commentaire et partager) sont des fonctionnalités indépendantes, et donc la formule générale pour calculer la priorité du *post* est calculée comme suit :

$$P^{SN_i}(O) = \prod_{A \in O.Actions} P(A) \quad [3]$$

Tableau 1. *Statistiques des base de données récupérées de Twitter et Facebook.*

Twitter		**Facebook**	
Propriété	**Valeur**	**Propriété**	**Valeur**
# de comptes	2,088,131 (4.9% spammeurs)	# d'utilisateurs	3,122
# de tweets	6,470,809(11.8% spam)	# de *posts*	6,880
# de réponse aux tweets	76,393	# de commentaires	2,398,611
# de re-tweeted tweets	3,129,237	# de réactions	64,083,457

Où $P(A)$ est estimé à l'aide de la probabilité maximale d'exécuter l'action A sur le *post* O, calculée comme $P(A) = \frac{Count(A,O)}{Count(A,SN_i)}$. $Count(A,O) = A.val$. Cela signifie que le nombre de fois que l'action A a été effectuée sur le *post* O. $Count(A, SN_i)$ représente la somme de l'action A sur les *posts* disponibles dans SN_i.

Classification de tweet : Lors de l'inférence pour un tweet donné sur un ensemble de *posts* dans SN_i, on obtient un vecteur de valeurs de probabilité où chacune représente le degré d'appariement entre un *post* et un tweet donné. Nous exploitons ces valeurs pour prendre une décision sur la classe d'un tweet donné. Pour ce faire, nous définissons une fonction de seuil qui décide d'étiqueter des tweets en tant que non-spam dans le cas de trouver au moins un *post* sur un réseau social ayant une probabilité supérieure à un seuil fixe. Formellement, nous définissons la fonction de seuil comme suit :

$$F(T, S_H) = \begin{cases} non-spam & max\{\frac{P^{SN_i}(O|T)}{Sum(SN_i,T)} | SN_i \in S_H, O \in SN_i\} \geq \Delta \\ spam & \text{otherwise} \end{cases} \quad [4]$$

Où la fonction $Sum(SN_i, T) = \sum_{O \in SN_i} P^{SN_i}(O|T)$ normalise la probabilité de chaque *post* extrait d'un certain réseau SN_i, rendant leur somme égale à un. Δ est un seuil interprété comme la probabilité minimale (c'est-à-dire le degré correspondant) requise pour classer le tweet T considéré comme non-spam.

4. Description de la base de données et la vérité terrain

Comme il existe différents réseaux sociaux disponibles sur le web, dans ce papier, nous expérimentons notre méthode à travers la collaboration avec le réseau social Facebook seulement. Parmi les différents réseaux sociaux disponibles, nous avons choisi Facebook pour mener l'expérimentation décrite dans ce papier. Ci-après sont décrits les ensembles de données exploités dans ladite expérimentation. Par conséquent, dans cette section, nous décrivons les ensembles de données Twitter et Facebook qui ont été exploités pour valider notre méthode.

Ensemble de données Twitter : Les ensembles de données utilisés lors de la détection au niveau tweet (Benevenuto *et al.*, 2010 ; Martinez-Romo et Araujo, 2013) ne sont pas publiquement disponibles pour la recherche. De plus, les politiques de Twitter permettent de publier uniquement les ID des comptes et des tweets d'une partie de l'ensemble de données Twitter. En effet, dans le contexte du problème de spam social, l'utilisation de l'ID n'est pas une solution puisque Twitter peut déjà avoir supprimé l'objet correspondant (compte

ou tweet) et donc aucune information n'est disponible pour le récupérer. Par conséquent, nous avons développé un robot d'exploration pour collecter des tweets en utilisant la méthode de diffusion en temps réel fournie par Twitter. Ensuite, nous avons lancé notre robot d'exploration pendant cinq mois, du 1er janvier 2016 au 31 mai 2016, avec le stockage des thématiques qui constituaient des tendances pour la période spécifiée. Ensuite, nous avons regroupé les tweets analysés sur la base des thématiques disponibles dans le texte des tweets, en supprimant les tweets qui ne relèvent pas pas de thématique tendance. Comme des milliers de thématiques sont disponibles dans notre collection de tweets, nous avons sélectionné les tweets de 100 thématiques tendances échantillonnées au hasard pour valider notre approche. Pour créer un ensemble de données annotées composé de tweets spam et non-spam, nous utilisons un processus d'annotation très répandu dans les recherches de détection de spam social, nommé «Twitter Spammers suspendus (TSS)» (Martinez-Romo et Araujo, 2013). Le processus vérifie si l'utilisateur de chaque tweet a été suspendu par Twitter. En cas de suspension, l'utilisateur est considéré comme un spammeur ainsi que le tweet correspondant est étiqueté comme un spam ; sinon le tweet est assigné comme non-spam et l'utilisateur comme "légitime". Nous avons effectué ce processus le 1 novembre 2016 afin d'avoir un grand ensemble de tweets spam annotés (763 555 exactement) et environ 102 318 spammeurs (comptes "spam"), comme indiqué dans le Tableau 1.

Ensemble de données Facebook : Pour les 100 thématiques sélectionnées, nous avons analysé les messages Facebook correspondants qui contiennent ces thématiques et qui sont postés pendant la période du 1 janvier 2016 au 31 mai 2016. Il est important de mentionner que la communauté Facebook s'est arrêtée récemment après la recherche des API à la dernière version, v2.8, de Graph API [6] publiée en août 2016. Ainsi, nous surmontons cet obstacle en développant un robot d'exploration Facebook qui recherche une thématique particulière en utilisant un compte Facebook normal, puis analyse les balises HTML des articles récupérés. Nous automatisons ce processus en utilisant l'outil d'automatisation de navigateur Web Selenium open source[7]. Au total, comme indiqué dans le Tableau 1, nous avons analysé plus de 6 880 messages Facebook générés par environ 1 212 utilisateurs différents en moins d'une heure.

5. Résultats et évaluation

5.1. *Configuration expérimentale*

Métriques de précision : Comme la vérité terrain de chaque classe d'étiquette de chaque tweet est donnée, nous utilisons l'exactitude, la précision, le rappel, la F-mesure, la précision moyenne, le rappel moyen et la F-mesure moyenne ; calculée en fonction de la matrice de confusion de l'outil Weka (Hall *et al.*, 2009) ; comme métriques couramment utilisées dans les problèmes de classification. Comme notre problème est la classification en deux classes (binaires), nous calculons la précision, le rappel et la F-mesure pour la classe «spam», alors que les métriques de moyennes combinent les deux classes en fonction de la fraction de

6. https ://developers.facebook.com/docs/graph-api/using-graph-api
7. http ://docs.seleniumhq.org/

Tableau 2. *Description de l'état de l'art associé au fonctionnalités de "tweet" utilisées dans la construction de modèles de classification supervisée.*

Caractéristique	Description
Nombre de hashtags	Nombre de hashtags disponibles dans le texte du tweet.
Nombre de mots spam	Nombre de mots répertoriés comme spam dans le texte du tweet.
Ratio de Hashtags	Ratio du nombre de hashtags par rapport au nombre de mots dans le tweet.
Ratio d'URLs	Ratio du nombre d'URL affiché dans le tweet par rapport au nombre de mots du tweet.
Nombre de mots	Nombre de mots dans le tweet.
Nombre de caractères numériques	le nombre de caractères numériques dans le texte tweet.
Nombre d'URLs	Nombre d'URL affichées dans le tweet.
Nombre de mentions	Nombre de comptes (utilisateurs) mentionnés dans le tweet.
Tweet répondu	Vérifie si le tweet est un tweet répondu ou non.
Similarité du Tweet et du contenu URL	Mesure la similarité entre le texte du tweet et le texte de l'URL posté dans Tweet.

chaque classe (par exemple 11,8 % * "Précision de spam" + 88,2 % * "précision de non-spam").

***Baselines* ou données de réfèrence :** Nous définissons deux baselines pour comparer notre méthode avec eux, à savoir : (i) baseline "A" qui représente les résultats lors de la classification de tous les tweets comme non-spam directement sans classement ; (ii) baseline "B" qui montre les résultats obtenus lors de l'application d'algorithmes d'apprentissage supervisés selon les fonctionnalités associées au "tweet" décrites dans le tableau 2. Comme de nombreux algorithmes d'apprentissage fournis par l'outil Weka, nous exploitons *Naive Bayes*, *Random Forest*, *J48,* et *Support Vector Machine* (SVM) comme méthodes d'apprentissage supervisées connues pour évaluer la performance des caractéristiques mentionnées.

Paramétrage : Dans le calcul de la probabilité a priori du *post*, nous adoptons les «Likes», «Shares», «Comments», «Wow», «Love», «Sad», «Haha» et «Angry» comme des actions. Dans notre méthode, Δ est la variable principale dans la classification des tweets et nous étudions donc l'impact du changement de sa valeur à travers des expériences à différentes valeurs de $\Delta \in [0.1, 1.0]$ avec un pas d'incrémentation égal à 0.1. Pour la méthode *Naive Bayes*, nous définissons les options "*useKernelEstimator*" et "*useSupervisedDiscretization*" à false comme valeurs par défaut définies par Weka. Pour *Random Forest*, nous avons mis l'option "*max depth*" à 0 (illimité), en étudiant l'effet du changement du nombre d'arbres $\in \{100, 500\}$. Pour la méthode $J48$, nous fixons le nombre minimum d'instances par feuille à 2, le nombre de plis à 3 et le facteur de confiance à 0,2. Pour la méthode SVM, nous utilisons l'implémentation *LibSVM* (Chang et Lin, 2011) intégrée à l'outil Weka pour définir la fonction du noyau sur *Radial Basis* et examiner l'impact de gamma $\in \{0.5, 1\}$, où les paramètres restants sont par défaut.

Procédure d'expérimentations : Pour le baseline «B», nous utilisons le concept de validation croisée pour les 100 thématiques tendances dans notre ensemble de données, résumées dans les étapes suivantes : (i) pour chaque thématique, nous construisons un espace vectoriel de caractéristique en utilisant l'état de l'art des caractéristiques décrites dans le tableau 2 ; (ii) ensuite, un espace vectoriel de caractéristiques d'une thématique sélectionnée (ensemble d'apprentissage) est utilisé uniquement pour construire un modèle prédictif en utilisant un algorithme d'apprentissage choisi ; (iii) les espaces vectoriels caractéristiques des thématiques restantes (c'est-à-dire 99 thématiques à tester), sont validés sur le modèle de classification construit à l'étape précédente ; (iv) les résultats de validation en termes de vrai positif, de

Tableau 3. *Résultats de performance des baselines A et B en fonction de différentes mesures.*

Algorithme d'apprentissage	Exactitude	Précision	Rappel	F-Mesure	Précision moy.	Rappel moy.	F-Mesure moy.
Baseline (A) : Tous les tweets sont labelisés "non-spam"							
————	88.2%	0.0%	0.0%	0.0.%	88.2%	88.2%	88.2%
Baseline (B) : Approche d'apprentissage supervisé							
Naive Bayes	81.2%	13.7%	10.5%	11.9%	79.0%	81.2%	80.1%
Random Forest (#Trees=100)	86.4%	13.2%	2.8%	4.6%	79.0%	86.4%	80.1%
Random Forest (#Trees=500)	86.5%	12.6%	2.6%	4.7%	79.4%	86.5%	82.8%
J48 (Confidence Factor=0.2)	86.4%	13.8%	2.9%	4.9%	79.6%	86.4%	82.5%
SVM (Gamma=0.5)	87.2%	15.7%	0.2%	0.4%	78.3%	87.2%	82.5%
SVM (Gamma=1.0)	87.0%	15.9%	0.1%	0.3%	77.9%	87.0%	82.2%

vrai négatif, de faux positif et de faux négatif sont extraits et stockés ; (v) les étapes de ii à iv sont répétées sur chaque thématique de la collection ; (vi) enfin, en utilisant les résultats de validation obtenus pour chaque thématique, nous calculons les métriques de performance mentionnées ci-dessus. Il est important de mentionner que la procédure expérimentale pour le baseline «B» simule exactement les scénarios réels dans la détection des tweets spam.

En expérimentant notre méthode, nous réalisons pour chaque thématique les étapes suivantes : (i) pour une certaine valeur de seuil de classification Δ, le modèle de classification conçu dans la section 3 est appliqué sur les tweets des thématiques considérées en utilisant les termes correspondants à la thématique des *posts* dans Facebook afin de prédire les étiquettes des classes des tweets ; (ii) ensuite les résultats en termes de vrai positif, de vrai négatif, de faux positif et de faux négatif sont extraits et stockés pour les calculs des résultats finaux ; (iii) les deux étapes précédentes sont effectuées sur chaque thématique de l'ensemble de données ; (iv) dans la dernière étape, les résultats complets de chaque thématique sont additionnés pour calculer les résultats de performance en utilisant les métriques mentionnées.

5.2. *Résultats expérimentaux*

Selon les résultats des baselines rapportées dans le tableau 3, les modèles de classification supervisés ont une forte défaillance dans le filtrage des tweets spam existant dans les 100 thématiques tendances. Cet échec peut être facilement identifié à partir des valeurs basses de rappel de spam (4^{eme} colonne) où la valeur la plus élevée est obtenue par l'algorithme d'apprentissage $NaiveBayes$. Le 10,5 % du rappel de spam obtenu par $NaiveBayes$ signifie que moins de 80 000 tweets spam peuvent être détectés à partir de 736 500 tweets spam. Les faibles valeurs de précision du spam indiquent également qu'un nombre important de tweets «non-spam» a été classé en «spam». Par la suite, comme la F-mesure de spam dépend des mesures de rappel et de précision, les valeurs de la F-mesure de spam sont évidemment faibles. Les valeurs de précision du baseline «B» sont proches des valeurs de précision de la baseline «A». Toutefois, compte tenu des faibles valeurs de précision du spam et du rappel de spam, la métrique d'exactitude dans ce cas n'est pas une mesure indicative et utile pour juger l'apprentissage supervisé comme une approche efficace. Plus précisément, l'approche d'apprentissage supervisé n'ajoute pas une contribution significative à l'amélioration de la qualité des 100 tweets des thématiques tendances. L'idée clé de l'utilisation de différents algorithmes d'apprentissage est de varier leurs paramètres est de mettre en évidence la mauvaise qualité des techniques de l'état de l'art traitant le tweet. Dans l'ensemble, les résultats obtenus par les

Tableau 4. *Nos résultats de performance de la méthode collaborative selon différentes métriques, montrant l'impact de la composante de probabilité a priori du post lors de l'exécution de la collaboration avec Facebook.*

Model(Δ)	Exactitude	Précision	Rappel	F-Mesure	Précision moy.	Rappel moy.	F-Mesure moy.
Uniforme* - Probabilité à posteriori du *post							
$\Delta = 0.1$	49.8%	10.8%	48.3%	17.7%	79.7%	49.8%	61.3%
$\Delta = 0.2$	32.3%	10.8%	69.4%	18.7%	79.1%	32.3%	45.9%
$\Delta = 0.3$	26.2%	10.8%	77.0%	18.9%	78.6%	26.2%	39.3%
$\Delta = 0.4$	22.8%	10.9%	82.3%	19.2%	78.5%	22.8%	35.3%
$\Delta = 0.5$	21.0%	11.0%	85.3%	19.4%	78.7%	21.0%	33.2%
$\Delta = 0.6$	19.4%	11.0%	87.9%	19.6%	78.8%	19.4%	31.2%
$\Delta = 0.7$	18.7%	11.1%	89.3%	19.7%	79.1%	18.7%	30.3%
$\Delta = 0.8$	17.5%	11.1%	90.9%	19.8%	79.3%	17.5%	28.7%
$\Delta = 0.9$	17.2%	11.1%	91.5%	19.8%	79.2%	17.2%	28.3%
$\Delta = 1.0$	17.2%	11.1%	91.6%	19.8%	79.4%	17.2%	28.3%
Non-Uniforme* Probabilité à posteriori du *post							
$\Delta = 0.1$	80.7%	17.0%	18.8%	17.8%	81.4%	80.7%	81.0%
$\Delta = 0.2$	80.6%	17.2%	19.3%	18.2%	81.5%	80.6%	81.0%
$\Delta = 0.3$	79.3%	15.8%	19.6%	17.5%	81.2%	79.3%	80.2%
$\Delta = 0.4$	77.8%	15.0%	21.1%	17.5%	81.1%	77.8%	79.4%
$\Delta = 0.5$	73.4%	13.5%	24.9%	17.4%	80.8%	73.4%	77.1%
$\Delta = 0.6$	64.0%	12.3%	36.4%	18.5%	80.7%	64.0%	71.4%
$\Delta = 0.7$	57.7%	11.9%	43.4%	18.7%	80.6%	57.7%	67.2%
$\Delta = 0.8$	51.9%	11.5%	49.0%	18.6%	80.3%	51.9%	63.0%
$\Delta = 0.9$	42.2%	11.0%	59.0%	18.6%	79.8%	42.2%	55.2%
$\Delta = 1.0$	34.79%	10.7%	66.0%	18.5%	79.1%	34.79%	48.3%

modèles permettent de tirer diverses conclusions : (i) les techniques de l'état de l'art ne sont pas discriminatives entre les tweets non-spam et spam, assurant la dynamique du contenu de spam ; (ii) les spammeurs ont tendance à publier des tweets presque similaires aux non-spam ; (iii) l'adoption d'une approche supervisée pour effectuer l'apprentissage sur un ensemble de données annotées de thématiques tendances et l'application du modèle de classification sur des thématiques tendances futures ou non annotées n'est *pas* la solution du tout.

En examinant les performances de notre méthode dans le tableau 4, le comportement est complètement différent en classifiant les tweets «spam», surtout lorsque la valeur de Δ devient plus élevée. Les résultats de rappel sont tout à fait compatibles avec l'équation 4 conçue pour la classification des tweets. Pour les valeurs élevées de Δ, la difficulté majeure est de trouver un "bon" *post* Facebook correspondant pour classer le tweet considéré comme «non-spam». Ainsi, cela explique la dégradation dramatique de la précision lors de l'augmentation de la valeur de Δ. Bien qu'on ait obtenu les valeurs de rappel élevées, les valeurs de précision de spam de notre méthode sont presque semblables à celles de l'approche d'apprentissage supervisé.

Uniformité contre non-Uniformité de la probabilité a priori du *post* :Le rôle de la composante de probabilité a priori du *post* est évident dans la détection des tweets spam. Travaillant sur l'hypothèse que chaque *post* de Facebook a la même probabilité (uniforme) pour être non-spam augmente les valeurs de rappel de spam lorsque la valeur de Δ devient plus élevée, conduisant à détecter la plupart des tweets spam. Au contraire, un nombre important de tweets «non-spam» a été classé comme «spam». Nous interprétons ce comportement en

raison de la faible valeur de la probabilité a priori du *post* lorsqu'on travaille sur l'hypothèse de probabilité uniforme. En effet, ce problème est réduit lorsque l'on considère les actions effectuées sur le *post* de Facebook pour calculer la composante de probabilité a priori du *post*. Ainsi, le rappel de spam a augmenté sans dégradation élevée dans les valeurs de précision. Malgré les faibles valeurs de précision de spam, les valeurs élevées de précision moyenne signifient que peu de tweets ont été classés comme «non-spam» alors qu'ils sont vraiment des «spam».

Haute qualité contre faux positif : Dans le filtrage du courrier indésirable, les efforts sont focalisés sur le problème de faux positif, qui se produit ici lorsqu'un véritable «non-spam» est classé comme «spam». Cependant, dans le contexte du spam social, le faux positif est moins important en raison de la disponibilité de collections de données à grande échelle, ce qui signifie que classer le tweet «non-spam» en tant que «spam» n'est pas un réel problème. Dans le contexte des réseaux sociaux l'objectif est d'augmenter la qualité des données où un large éventail d'applications basées sur Twitter (par exemple, le résumé de tweet) a une priorité élevée pour travailler sur les collections non "bruitées". L'aspect temps de calcul est important lorsqu'il s'agit de cibler des collections à grande échelle. Par conséquent, notre méthode est parfaitement adaptée pour traiter les collections à grande échelle avec des informations de haute qualité. Par exemple, le temps nécessaire au traitement de notre ensemble de données Twitter ne dépasse pas quelques heures, réparti entre les données d'exploration de Facebook et l'application de notre modèle. Enfin, comme diverses expériences sont données pour différentes valeurs Δ où aucune valeur optimale ne peut satisfaire toutes les métriques de performance, la sélection dépend principalement des exigences souhaitées de la collection finale. Par exemple, une valeur Δ élevée est recommandée pour avoir une collection de qualité élevée mais avec une forte probabilité de perdre des informations non "bruitées".

6. Conclusion

Nous avons presenté une approche basée sur la collaboration de réseaux sociaux pour filtrer les tweets spam dans des collections à grande échelle. Nous proposons une méthode d'apprentissage non supervisée basée sur le concept de modèle de langage pour identifier des informations similaires dans d'autres réseaux sociaux pris en compte dans cette collaboration. Notre méthode surpasse les méthodes classiques de détection de spam en ce qui concerne la consommation de temps, nécessitant quelques heures pour traiter environ 6 millions de tweets publiés dans 100 thématiques tendances. À partir de cette contribution dans le cadre de la lutte contre le spam, nous prévoyons d'étudier l'effet de la collaboration avec d'autres réseaux sociaux tels que Instagram. En outre, nous avons l'intention d'améliorer la performance de classification en extrayant plus de fonctionnalités des commentaires des utilisateurs tels que les caractéristiques liées aux sentiments. En outre, nous envisageons d'étudier différents comportements d'utilisation selon d'autres modèles de langage.

Remerciements

Ce travail s'intgère dans le cadre des contirubtions du projet ANR FILTER 2.

7. Bibliographie

Abascal-Mena R., Lema R., Sèdes F., « Detecting sociosemantic communities by applying social network analysis in tweets », *Social Netw. Analys. Mining*, vol. 5, nº 1, p. 38 :1-38 :17, 2015.

Agarwal N., Yiliyasi Y., « Information quality challenges in social media », *International Conference on Information Quality (ICIQ)*, 2010.

Amleshwaram A. A., Reddy N., Yadav S., Gu G., Yang C., « CATS : Characterizing automation of Twitter spammers », *Communication Systems and Networks (COMSNETS), 2013 Fifth International Conference on*, IEEE, p. 1-10, 2013.

Benevenuto F., Magno G., Rodrigues T., Almeida V., « Detecting spammers on twitter », *In Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS*, p. 12, 2010.

Canut C. M., On-at S., Péninou A., Sèdes F., « Time-aware Egocentric network-based User Profiling », *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015, Paris, France, August 25 - 28, 2015*, p. 569-572, 2015.

Cao C., Caverlee J., « Detecting spam urls in social media via behavioral analysis », *European Conference on Information Retrieval*, Springer, p. 703-714, 2015.

Chang C.-C., Lin C.-J., « LIBSVM : A library for support vector machines », *ACM Transactions on Intelligent Systems and Technology*, vol. 2, p. 27 :1-27 :27, 2011. Software available at.

Chu Z., Gianvecchio S., Wang H., Jajodia S., « Detecting automation of twitter accounts : Are you a human, bot, or cyborg ? », *Dependable and Secure Computing, IEEE Transactions on*, vol. 9, nº 6, p. 811-824, 2012a.

Chu Z., Widjaja I., Wang H., « Detecting social spam campaigns on twitter », *Applied Cryptography and Network Security*, Springer, p. 455-472, 2012b.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H., « The WEKA Data Mining Software : An Update », *SIGKDD Explor. Newsl.*, vol. 11, nº 1, p. 10-18, November, 2009.

Kullback S., « The Kullback-Leibler Distance », *The American Statistician*, vol. 41, nº 4, p. 340-341, 1987.

Martinez-Romo J., Araujo L., « Detecting malicious tweets in trending topics using a statistical analysis of language », *Expert Systems with Applications*, vol. 40, nº 8, p. 2992-3000, 2013.

Mccord M., Chuah M., « Spam detection on twitter using traditional classifiers », *International Conference on Autonomic and Trusted Computing*, Springer, p. 175-186, 2011.

Mezghani M., Zayani C. A., Amous I., Péninou A., Sèdes F., « Dynamic enrichment of social users' interests », *IEEE 8th International Conference on Research Challenges in Information Science, RCIS 2014, Marrakech, Morocco, May 28-30, 2014*, p. 1-11, 2014.

Ponte J. M., Croft W. B., « A language modeling approach to information retrieval », *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, p. 275-281, 1998.

Stringhini G., Kruegel C., Vigna G., « Detecting Spammers on Social Networks », *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, ACM, New York, NY, USA, p. 1-9, 2010.

Wang A. H., « Don't follow me : Spam detection in Twitter », *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*, p. 1-10, July, 2010.

Yang C., Harkreader R. C., Gu G., « Die Free or Live Hard ? Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers », *Proceedings of the 14th International Conference on Recent Advances in Intrusion Detection*, RAID'11, Springer-Verlag, Berlin, Heidelberg, p. 318-337, 2011.

Yardi S., Romero D., Schoenebeck G., danah boyd, « Detecting spam in a Twitter network », *First Monday*, 2009.