# Qualité de l'information dans les réseaux sociaux : une méthode collaborative pour détecter les spams dans les tweets

**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. *Détecter les actions des utilisateurs mal intentionnés dits "spammeurs" est un réel défi pour maintenir un haut niveau de performance dans les applications mises en œuvre dans les réseaux sociaux. Les méthodes conventionnelles de détection de spams imposent des délais de traitement importants et inévitables, allant jusqu'à des mois pour traiter de grandes collections de tweets. Ces méthodes entièrement dépendantes de l'approche d'apprentissage supervisé pour la 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 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. *Detecting 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 time, for example up to months for processing large collections of tweets. These methods entirely depend on the supervised learning approach for classification, and 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 from used hashtags. Our method has been experimented on more than 6 million tweets posted on 100 trending topics. Facebook is used in parallel as the ground truth allowing the collaboration of two different social networks. Our experiments show an efficiency with regard to processing time and classification effectiveness, compared to the conventional methods for detecting spams in tweets.*

MOTS-CLÉS : *spam social, réseaux sociaux, collaboration, thématiques tendances.*

KEYWORDS: *social spam, social networks, collaboration, trending topics.*

DOI:10.3166/DN.21.3.33-53

## 1. Introduction

L'une des principales caractéristiques des réseaux sociaux en ligne (OSN) est leur dépendance envers les utilisateurs en tant que contributeurs principaux dans la génération et la publication de contenu. La dépendance à l'égard des contributions des utilisateurs pourrait être exploitée de manière positive, y compris la compréhension des besoins des utilisateurs à des fins de marketing, l'étude des opinions des utilisateurs et l'amélioration des algorithmes de récupération de l'information.

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 définie par l'aptitude de l'information à être utilisée dans un contexte particulier (Agarwal, Yiliyasi, 2010b).

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, Yiliyasi, 2010b) : (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".

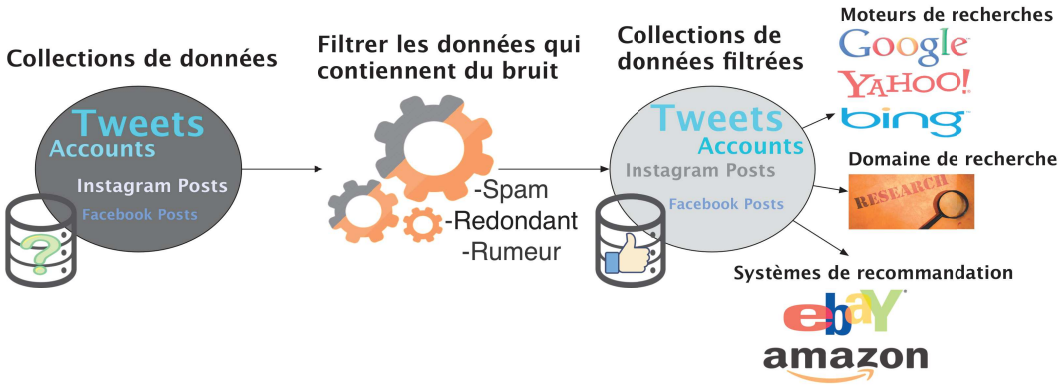


*Figure 1. Un aperçu du processus de filtrage pour maintenir un niveau de qualité de l'information dans les réseaux sociaux*

Divers types de bruits 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 réalisons nos expérimentations.

Dans 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, Widjaja, Wang, 2012 ; 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 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 à cause des facteurs suivants : 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 (spam ou non spam) 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. Dans cet article nous adoptons donc l'hypothèse que les postes

1. https://dev.twitter.com/rest/public
2. hhttps://www.facebook.com/policies

sur Facebook ne sont pas des spams. Ainsi, une correspondance entre le tweet et les postes Facebook peut être interprétée comme une indication de non-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 définit les différentes terminologies utilisées dans ce papier. La section 3 donne un aperçu sur les méthodes de détection de spam basées sur Twitter. La section 4 présente les notations, la formalisation du problème et la conception de la méthode collaborative que nous proposons pour détecter les tweets spam. La section 5 décrit l'ensemble de données utilisé pour valider notre approche. La section 6 présente les résultats. La section 7 conclut ce travail en fournissant un bilan et en dressant des perspectives.

## 2. Contexte

Cette section présente des informations générales et des terminologies relatives aux sites de micro-blogging, à la qualité de l'information, et les spammeurs sociaux pour avoir plus d'informations sur le problème du spam social sur Twitter.

**Sites de microblogging.** Le micro-blogging est défini comme un court média de diffusion qui aide les utilisateurs à se tenir au courant de l'action des blogs (Kaplan, Haenlein, 2011). Un grand nombre de sites de microblogging existe sur Internet permettant aux utilisateurs d'échanger des images, du contenu, des liens vidéo et autres. Ces sites permettent aux utilisateurs de poster leurs idées et leurs opinions. Le contenu partagé peut être lié à de nombreux sites fonctionnant sur Internet. Comparé aux sites de blogs généraux, chaque microblogging a des propriétés ou des services distinctifs ayant un contenu d'affichage faible. Parmi ces sites, Twitter est le site de microblogging social le plus populaire dédié aux nouvelles en ligne et au service de réseau social où les utilisateurs peuvent poster et interagir avec des messages, appelés tweets, limités à 140 caractères.

**Définition de thématique.** Le concept de "thématique" [5] peut être défini comme une représentation de structures sémantiques cachées dans une collection de textes (par exemple, des documents textuels, ou des tweets). D'un autre côté, "thématique tendance" ("*Trending Topic*") est un mot clé ou une phrase (par exemple #Trump, #KCA et #TopChef) qui est mentionné à un taux plus élevé que les autres. Les "thématiques tendances" deviennent populaires en raison des efforts concentrés par les utilisateurs, ou en raison d'un événement qui encourage les utilisateurs à parler d'un sujet particulier. L'objectif principal des "thématiques tendances" est d'aider les utilisateurs à comprendre ce qui se passe dans le monde et les opinions des utilisateurs à ce sujet en temps réel. Les "thématiques tendances" sont automatiquement identifiées par Twitter

3. elles sont affichées dans Twitter en haut à gauche d'un compte en tant que *Trending Topics*
4. http://osirim.irit.fr/site/
5. https://support.twitter.com/articles/101125#

*Tableau 1. Définition des quatres catégories de la qualité d'information (QI) et leurs dimensions*

| Mesure | Description |
|---|---|
| QI intrinsèque: conformité entre l'information et la vision du monde réel. | |
| Exactitude, Validité | Mesure dans laquelle les informations sont valides en fonction de références stables [a] . |
| Crédibilité | Degré de crédibilité de l'information. |
| Objectivité | Mesure dans laquelle l'information est sans préjugés et impartiale. |
| Réputation | Mesure dans laquelle l'information est de haut niveau. |
| Vérifiabilité | Mesure dans laquelle l'information peut être vérifiée pour l'exactitude. |
| QI contextuelle: degré d'adéquation de l'information dans un contexte donné ou une tâche en cours. | |
| Quantité d'Information | Mesure dans laquelle la quantité d'information est appropriée à l'utilisation. |
| Pertinence | Degré auquel l'information est applicable pour une tâche donnée. |
| A jour | Mesure dans laquelle les informations sont à jour pour une tâche donnée. |
| Complète | Mesure dans laquelle les informations correspondent à l'exhaustivité et à la précision requises dans un contexte donné. |
| Valeur ajoutée | Mesure dans laquelle l'utilisation de l'information profite aux consommateurs de l'information. |
| Compréhension contextuelle | Mesure dans laquelle l'information est facilement comprise sans ambiguïté. |
| Feedback des utilisateurs | Capacité des utilisateurs à fournir une évaluation de qualité implicite ou explicite du contenu. |
| QI Représentative : la mesure dans laquelle l'information est bien présentée et utilisable pour tous les utilisateurs en tenant compte de l'aspect technique. | |
| Concision de la représentation | Mesure dans laquelle la structure et la présentation de l'information est compacte sans être écrasante. |
| Cohérence représentative | Mesure dans laquelle la définition, le format et la valeur de l'information sont cohérents entre les applications et les systèmes. |
| Facilité de compréhension | Mesure dans laquelle l'information est compréhensible et claire. |
| Manipulabilité | La mesure dans laquelle l'information peut être mise à jour, modifiée, transférée, reproduite, intégrée et personnalisée. |
| QI Accessible : mesure dans laquelle l'information est accessible et sécurisée. | |
| Accès | Mesure dans laquelle l'information est récupérable et disponible. |
| Sécurité | Degré de protection des informations contre un accès non autorisé. |

*a*. Telles qu'un dictionnaire ou un ensemble de normes et de contraintes de domain.

grâce à un algorithme qui identifie les thématiques qui circulent massivement plus que d'autres thématiques.

**Qualité de l'information.** Dans la littérature, différents efforts ont été faits pour introduire une définition évidente de la qualité de l'information (QI). (Juran, Godfrey, 1999) ont défini la QI comme le degré d'utilité ou «d'aptitude à l'emploi» de l'information dans une tâche ou un contexte particulier. (Wand, Wang, 1996) ont introduit une définition de la qualité des données et de la qualité de la cartographie entre l'état du système d'information virtuel et un état du monde réel. Cependant, comme ces définitions sont conceptuellement qualitatives, les chercheurs ont introduit diverses dimensions pour décrire ou mesurer la qualité de l'information telle que définie dans le tableau 1. Ces dimensions peuvent être classées en quatre catégories principales : intrinsèque, contextuelle, représentative et accessibilité. Cependant, pour choisir les dimensions ou catégories appropriées, il est nécessaire de définir l'entité souhaitée (par exemple, tweet, compte Twitter) et le problème de la QI (par exemple, rumeur sociale et spam social). Une fois le problème de la QI et les dimensions correspondantes déterminés, l'étape suivante consiste à quantifier les dimensions (ou caractéristiques). Enfin, toutes les métriques peuvent être combinées avec un indicateur unique, qui fournit des informations sur la qualité de l'entité. Par exemple, l'information sur les rumeurs est un problème sur la QI bien connu apparaissant dans les réseaux sociaux en ligne et les sites de micro-blogging (Zubiaga *et al.*, 2015). Ainsi, pour mesurer le degré d'un tweet d'être une rumeur, la précision dimensionnelle de la catégorie intrinsèque est l'une des nombreuses possibilités qui peuvent être utilisées. Pour quantifier une telle dimension, vérifier l'existence de l'information dans un tweet donné sur différents réseaux sociaux en ligne est une métrique possible où l'existence positive est une indication pour ne pas être une information de rumeur.

**Spam social et qualité de l'information.** Le spam social est un contenu textuel absurde ou charabia apparaissant sur les réseaux sociaux en ligne et tout site Web traitant du contenu généré par l'utilisateur tel que les chats et les commentaires charabia (Agarwal, Yiliyasi, 2010b). Le spam social peut prendre une grande variété de formes, y compris le blasphème, les insultes, les discours de haine, les critiques frauduleuses, faux amis, messages en vrac, phishing et liens malveillants, et matériel pornographique. On pourrait considérer le spam social comme une information non pertinente. Cependant, cette interpré tation est tout à fait inexacte. Nous justifions cette interprétation erronée par la définition de systèmes de recherche d'informations (RI) (Manning *et al.*, 2008) dans lesquels la disponibilité des documents dans les systèmes de RI dépend de la requête de recherche d'entrée. Ainsi, les documents non pertinents par rapport à une requête d'entré e ne sont pas né cessairement des spams. Par consé quent, en tant que définition supplémentaire, le spam social peut être défini comme une information non pertinente qui n'a pas d'interprétation dans aucun contexte tant que l'entrée n'est pas un contenu de spam. Puisque le spam social est un problème de QI pur, nous projetons le problème sur cinq dimensions de QI, y compris: la précision, la crédibilité, la réputation, la valeur ajoutée et la pertinence. En effet, le contenu de spam ne représente pas une donnée réelle du monde et il a donc un faible degré de précision et de crédibilité. En outre, la réputation du spam est également faible, car les utilisateurs normaux ont tendance à diffuser des informations précises en général. Enfin, le contenu de spam ne présente aucun avantage pour les utilisateurs d'OSN, ce qui conduit à un faible degré en termes de dimensions à valeur ajoutée et de pertinence. Bien que de projeter le problème de spam social sur le monde de la QI fournit plus d'idées concernant la gestion efficace du problème, les spammeurs sociaux font de grands efforts pour augmenter le degré de la QI. Par conséquent, comprendre et connaître les faits sur les spammeurs sociaux et leurs comportements peuvent contribuer à fournir des solutions efficaces pour le problème du spam social.

**Perspective collective.** En manipulant les comptes spam sur les OSNs, les spammeurs sociaux lancent leurs campagnes de spam (*bots*) en créant des milliers de comptes spam de manière automatisé e. Les spammeurs utilisent les API REST fournies par Twitter pour coordonner automatiquement leurs comptes spam. Par exemple, les spammeurs peuvent automatiser le comportement d'affichage de chaque compte en tweetant à une fréquence fixe. En outre, les API REST fournissent des "thématiques tendances" en cours qui sont diffusé es parmi les utilisateurs, facilitant les campagnes des spammeurs dans l'attaque des thématiques tendances par du contenu spam. Par conséquent, é tant donné le fait que chaque spammeur peut utiliser des milliers de comptes spam pour diffuser un contenu spam particulier, la probabilité est é levé e pour trouver une corré lation entre les tweets spam ainsi que leurs comptes. Une similarité dans le style d'écriture des tweets spam, y compris le nom des comptes, donne une forte indication qu'un spammer social individuel contrôle ces comptes spam afin d'agir comme étant une campagne de spam. Ainsi, le passage de la perspective individuelle de l'inspection des tweets ou des comptes pour l'existence de spam à la perspective collective augmente l'efficacité de la détection des comptes spam d'une manière rapide, en particulier lors du ciblage des collections à grande échelle.

## 3. Aperçu des méthodes de détection de spam basées sur Twitter

Les spammeurs sociaux exploitent la flexibilité d'utiliser les OSN pour abuser de tous les services légaux et possibles supportés par les OSNs pour diffuser leur contenu de spam. Quel que soit le type du OSN ciblé, les spammeurs sociaux adoptent les mêmes faits ou principes dans leurs buts et comportements, résumés comme suit:

– Les spammeurs sociaux sont des personnes qui visent à atteindre des objectifs contraires à l'éthique (par exemple, promouvoir des produits), et ainsi, ils utilisent leur intelligence pour accomplir leurs tâches de spammeur de manière efficace et rapide.

– Les spammeurs sociaux exploitent les "thématiques tendances" pour lancer leur contenu spam.

– Les spammeurs sociaux créent et lancent souvent une campagne de comptes spam sur une courte période (par exemple, un jour) pour maximiser leurs profits monétaires et accélérer leur comportement de spam.

– Comme un ensemble d'API est fourni par les réseaux sociaux, les spammeurs sociaux les utilisent pour automatiser leurs tâches de lancer des spams de manière systématique (par exemple, tweeter toutes les 10 minutes). Le comportement de publication aléatoire est évité et n'est pas une solution préférable pour les spammeurs sociaux, car elle peut diminuer le profit cible et ralentir leur comportement de "spamming".

Sur Twitter, les spammeurs sociaux tirent parti de différents services fournis pour lancer leurs attaques de spam via: (i) URL ; (ii) Hashtag ; et (iii) service de *Mention*.
La taille du tweet étant limitée à 140 caractères, les spammeurs utilisent des services d'URL (par exemple, Bitly et TinyURL) sur Twitter pour convertir de longues URL en petites URL. Cette option permet aux spammeurs sociaux d'abuser de ce service en publiant des sites Web de spams avec un raccourci d'URL pour masquer le domaine.
Les hashtags sont largement utilisés dans les OSN en tant que service pour grouper les tweets par thème, facilitant ainsi le processus de recherche. Ce service est mal utilisé par les spammeurs sociaux en taggeant des hashtags dans leurs tweets de spam qui peuvent aussi contenir des URL. Ceci augmente les chances d'être recherché par les utilisateurs.

Le service *Mention* fournit un mécanisme d'envoi direct de message à un utilisateur particulier en utilisant le symbole @ suivi du nom d'écran de l'utilisateur souhaité. Différemment des URL et de l'hashtag, les spammeurs sociaux utilisent abusivement ce service pour envoyer leurs tweets à une liste définie d'utilisateurs.

Outre ces services, Twitter fournit des API pour les développeurs à utiliser dans leurs applications tierces. Les spammeurs sociaux exploitent ce service distinctif
comme une opportunité d'automatiser leur comportement de spam. Twitter donne la possibilité aux utilisateurs de signaler les comptes de spam en cliquant sur "Signaler: ils affichent spam" option disponible dans tous les comptes. Lorsqu'un compte est signalé, les administrateurs de Twitter examinent manuellement et analysent en profondeur ce compte pour prendre ultérieurement une décision de suspension. Cependant, un tel mé-

*Tableau 2. Description des différentes terminologies utilisées dans la taxonomnie de la détection du spam social*

| Terminologie | Description |
|---|---|
| Tweet-Level Detection | Prédiction de l'étiquette de classe de tweet s'il s'agit d'un spam ou non-spam. |
| Account-Level Detection | Analyse du profil ou du compte de l'utilisateur afin de prédire s'il s'agit d'un spammeur ou d'un utilisateur légitime. |
| Campaign-Level Detection | Prendre la perspective collective de sorte qu'un groupe de comptes est profondément examiné pour déterminer s'il s'agit d'une campagne de spam ou pas. |
| User Features | Fonctions extraites d'attributs (par exemple, nom d'utilisateur, pseudonyme) existant dans l'objet utilisateur, tels que l'âge du compte et le nombre de tweets affichés par l'utilisateur. |
| Content Features | Fonctionnalités extraites du contenu d'un ou de plusieurs tweets, tels que le nombre de hashtags, et le nombre d'URL. |
| Graph Features | Fonctionnalités qui nécessitent de construire d'abord un graphe bidirectionnel contenant les voisins de l'utilisateur pour extraire des fonctionnalités telles que cluster local, noeud betweenness. |
| Timing Features | Fonctionnalités extraites en analysant l'heure de publication d'un groupe de tweets affichés par un utilisateur[a]. |
| Automation Features | Fonctionnalités d'automatisation liées à l'utilisation d'API externes prises en charge par des sites Web externes. |
| Link Features | Fonctionnalités extraites de l'analyse des URL publiées dans les tweets (tweets et la similarité du contenu des URL). |

*a*. tels que la fréquence de "tweeting"

canisme de rapport est inefficace pour combattre et supprimer les spammeurs sociaux car il a besoin d'efforts significatifs et importants de la part des administrateurs. De plus, de nombreux utilisateurs peuvent fournir de faux rapports et tous les rapports ne sont donc pas forcément fiables.

Comme tentative supplémentaire de réduire le problème de spam, Twitter a défini des règles générales (par exemple, publier du matériel pornographique est interdit) avec la suspension permanente des comptes qui violent ces règles (Twitter, 2016). Malheureusement, les spammeurs sociaux sont assez intelligents pour contourner les règles de Twitter. Par exemple, les spammeurs sociaux peuvent coordonner plusieurs comptes avec la répartition de la charge de travail souhaitée parmi ces comptes pour induire en erreur la détection. Par conséquent, les approches de Twitter sont inefficaces pour le filtrage anti-spam en temps réel. Les lacunes du mécanisme anti-spam de Twitter ont incité les chercheurs à introduire des méthodes plus robustes afin d'augmenter la qualité des données entrantes pour les applications qui utilisent Twitter comme source d'information principale.

Après avoir un aperçu approfondi d'un large éventail de travaux de recherche scientifique liés aux méthodes de détection de spam sur Twitter, nous construisons une taxonomie détaillée pour ces méthodes, illustrée à travers la figure 2, en fonction de différents critères, notamment : (i) le type d'approche de détection (Machine Learning ou Honeypot) ; (ii) le niveau de détection (Tweet, compte et campagne) ; et (iii) type de fonctionnalités (utilisateur, contenu, lien, automatisation, graphique et minutage) exploitées dans les méthodes de détection. Le tableau 2 fournit une description de ces terminologies. L'axe d'apprentissage automatique se concentre sur la détection du spam social de manière automatisée, tandis que l'approche de l'honeypot social nécessite une intervention des administrateurs des systèmes.

**Approche d'apprentissage automatique.** La plupart des méthodes de détection de spam ont utilisé des algorithmes d'apprentissage automatique supervisé à trois niveaux de détection, répartis entre la détection au niveau du tweet, la détection au niveau du compte et détection au niveau de la campagne, décrite comme suit :

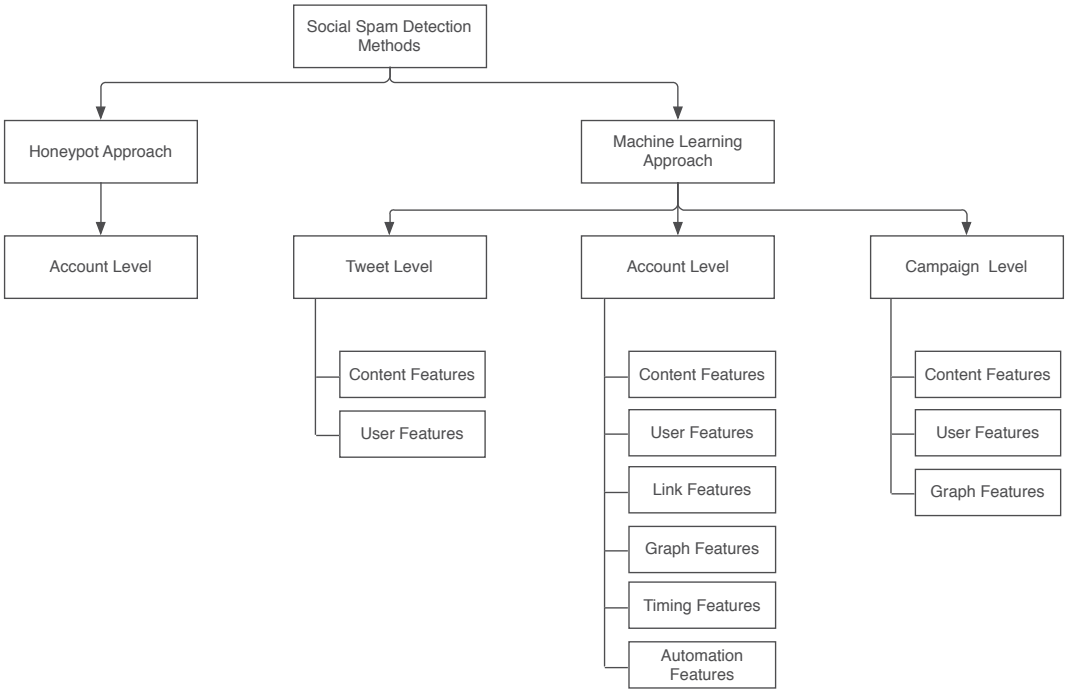


*Figure 2. Taxonomie des méthodes de détection de spam social dans Twitter. La description de quelques terminologies est fournie dans le tableau 2*

– **Niveau tweet :** à ce niveau, les tweets individuels sont vérifiés afin d'éliminer d'éventuels contenus indésirables. (Benevenuto *et al.*, 2010) ont extraits 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, 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 langages 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, Chuah, 2011 ; Cao, 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 nœuds. 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'ex-

traction de la plupart des fonctionnalités impossible, en particulier dans le traitement de données à grande échelle.

– **Niveau campagne :** (Chu, Widjaja, Wang, 2012) 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, Gianvecchio *et al.*, 2012), 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.

**Approche de *honeypot.*** Le *honeypot* (pot de miel) social est considéré comme une ressource de système d'information capable de surveiller les comportements des spammeurs sociaux à travers la journalisation de leurs informations telles que l'information des comptes et de tout contenu disponible (Lee *et al.*, 2010). Cette approche analyse manuellement les comptes afin de prendre une décision s'il s'agit de spams ou pas. En fait, il n'y a pas de différence significative entre le mécanisme anti-spam de Twitter et l'approche du *honeypot* social. Tous les deux ont besoin d'un contrôle administratif pour prendre une décision sur les comptes qui sont tombés dans "le pot de miel" piège. Le contrôle administratif vise à réduire le taux de faux positifs. Ceci est considéré comme une solution alternative au classement aveugle de tous les utilisateurs qui ont été supprimés dans ce piège en tant qu'utilisateurs spam.

## 4. 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 tendance" donnée. Comme le but de l'utilisation 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. 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 estimer la correspondance entre un tweet donné et les postes dans d'autres OSN. Si la correspondance est faible, le tweet sera considéré comme spam.

### *4.1. 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-tuples $T_\bullet =< Texte, Actions >$.

**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*[6] en réaction au contenu des *posts* ou des tweets. Nous définissons les actions comme un ensemble fini de 2-tuples, $Actions = \{< a_{name_1}, a_{val_1} >, < a_{name_2}, a_{val_2} >, ....\}$, 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.

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'élément $O_\bullet$ est défini par 2-tuples $O_\bullet =< Texte, Actions >$.

### *4.2. 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 $f(T) : T \rightarrow \{spam, non - spam\}, T \in C_H$.

### *4.3. Probabilité, priorité et classification de tweet*

**Probabilité de tweet :** nous utilisons les modèles de langage statistique (Ponte, 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 (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 :

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

Le symbole $\stackrel{\text{rank}}{=}$ est utilisé dans le domaine de la recherche d'information pour désigner le classement des documents selon la probabilité de chaque document. Dans notre

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

cas, nous classons le post Facebook pour un tweet donné.
$P^{SN_i}(O)$ 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 comme Jelineck Mercer, Dirichlet (Ponte, Croft, 1998) pour calculer $P^{SN_i}(O)$ ou la divergence de Kullback-Leibler (Kullback, 1987) pour calculer le degré de dissimilarité entre les modèles de tweet et du langage du *post*.
$T.Texte$ est le texte associé au tweet T. $Texte$ est déjà défini dans dans la section 4.1.

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. Nous adoptons la méthode de divergence de Kullback-Leibler ($KL$). 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 langage 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 $\infty$ 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 signifie non similaire et 1 signifie similaire) en limitant ses valeurs, où la composante de probabilité $P^{SN_i}(T|O)$ est définie comme :

$$P^{SN_i}(T|O) = \frac{\log|T.Text| - F}{\log|T.Text|} \tag{2}$$

$$F = \sum_{w \in T.Text} P(w|M_T) * min(|\log \frac{P(w|M_T)}{P(w|M_O)}|, \log|T.Text|) \tag{3}$$

Où $P(w|M_T)$ et $P(w|M_O)$ sont les probabilités que le mot $w$ soit généré par les modèles de langage 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 signifie 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 :

$$P^{SN_i}(O) = \prod_{A \in O.Actions} P(A) \tag{4}$$

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

*Tableau 3. 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 |

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:

$$F(T, S_H) = \begin{cases} nonSpam & max\{\frac{P^{SN_i}(O|T)}{Sum(SN_i,T)} | SN_i \in S_H, O \in SN_i\} \geq \Delta \\ spam & \text{autrement} \end{cases} \quad (5)$$

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 1. Nous allons tester avec différentes valeur de Delta. $\Delta$ 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.

## 5. Description de la base de données et la vérité terrain

Dans ce papier, nous expérimentons notre méthode à travers la collaboration avec le réseau social Facebook. Ci-après sont décrits les ensembles de données exploités dans notre expérimentation pour l'ensemble de données Twitter et Facebook afin de 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, 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 de la "thématique tendance". Nous rappelons que l'information qu'un tweet appartient à une "thématique tendance" est donnée par Twitter. 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, 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". La vérification est effectué après la collecte des tweets. 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 3.

**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 a arrêté récemment l'utilisation des API pour la recherche d'information depuis la dernière version, v2.8, de Graph API [7] 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 [8]. Au total, comme indiqué dans le tableau 3, nous avons analysé plus de 6 880 messages Facebook générés par environ 1 212 utilisateurs différents en moins d'une heure.

## 6. Résultats et évaluation

Dans cette section, nous présentons d'abord la configuration expérimentale qui consiste à détailler les différentes métriques utilisées, l'ensemble des données de référence (*Baseline*), le paramétrage et la procédure d'expérimentation. Ensuite, nous détaillons les résultats obtenus.

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

*Tableau 4. Description des caractéristiques de "tweet" utilisées dans la construction de modèles de classification supervisée de l'état de l'art*

| 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 | 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. |

### *6.1. Configuration expérimentale*

**Métriques de précision :** comme la vérité terrain de chaque classe (son 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). Ce sont des métriques couramment utilisées dans les problèmes de classification. Comme notre problème est la classification en deux classes (binaire), 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 chaque classe (par exemple 11,8 % * "Précision de spam" + 88,2 % * "précision de non-spam").

***Baselines* ou données de référence :** 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 4. 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 supervisé 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, $\Delta$ 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, Lin, 2011) intégrée à l'outil Weka pour définir la fonction du noyau sur *Radial*

*Tableau 5. 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 % |

*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 (validation croisée 100 fois), résumées dans les étapes suivantes : (i) pour chaque thématique, nous construisons un espace vectoriel de caractéristique en utilisant les caractéristiques de l'état de l'art décrites dans le tableau 4 ; (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 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 $\Delta$, 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.

*Tableau 6. 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($\Delta$) | 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 % |

### *6.2. Résultats expérimentaux*

Selon les résultats des baselines rapportées dans le tableau 5 [9], les modèles de classification supervisé 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 tweets des 100 "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

9. Pour rappel : la précision, le rappel et la F-mesure sont associés à la classe «spam». Les métriques de moyennes combinent les deux classes en fonction de la fraction de chaque classe

des techniques de l'état de l'art traitant le tweet. Dans l'ensemble, les résultats obtenus par les 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 ; (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 6 [10], le comportement est complètement différent en classifiant les tweets «spam», surtout lorsque la valeur de $\Delta$ devient plus élevée. Les résultats de rappel sont tout à fait compatibles avec l'équation 6 conçue pour la classification des tweets. Pour les valeurs élevées de $\Delta$, 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 $\Delta$. 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 $\Delta$ 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 expliquons ce comportement par 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.

---

10. Pour rappel : la précision, le rappel et la F-mesure sont associés à la classe «spam». Les métriques de moyennes combinent les deux classes en fonction de la fraction de chaque classe.

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 (contre plusieurs mois pour les approches de l'état de l'art). Enfin, comme diverses expériences sont données pour différentes valeurs $\Delta$ 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 $\Delta$ élevée est recommandée pour avoir une collection de qualité élevée mais avec une forte probabilité de perdre des informations non "bruitées".

## 7. 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 traite les informations de manière rapide, 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. 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 effet, nous envisageons d'étudier différents comportements d'utilisation selon d'autres modèles de langage.

*Remerciements*

*Ce travail s'intègre dans le cadre des contributions du projet ANR FILTER 2.*

## Bibliographie

Abascal-Mena R., Lema R., Sèdes F. (2015). Detecting sociosemantic communities by applying social network analysis in tweets. *Social Netw. Analys. Mining*, vol. 5, n° 1, p. 38:1–38:17. Consulté sur http://dx.doi.org/10.1007/s13278-015-0280-2

Agarwal N., Yiliyasi Y. (2010a). Information quality challenges in social media. In *International conference on information quality (iciq).*

Agarwal N., Yiliyasi Y. (2010b). Information quality challenges in social media. In *International conference on information quality (iciq)*, p. 234-248.

Amleshwaram A. A., Reddy N., Yadav S., Gu G., Yang C. (2013). Cats: Characterizing automation of twitter spammers. In *Communication systems and networks (comsnets), 2013 fifth international conference on*, p. 1–10.

Benevenuto F., Magno G., Rodrigues T., Almeida V. (2010). Detecting spammers on twitter. In *In collaboration, electronic messaging, anti-abuse and spam conference (ceas*, p. 12.

Canut C. M., On-at S., Péninou A., Sèdes F. (2015). Time-aware egocentric network-based user profiling. In *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. Consulté sur http://doi.acm.org/10.1145/2808797.2809415

Cao C., Caverlee J. (2015). Detecting spam urls in social media via behavioral analysis. In *European conference on information retrieval*, p. 703–714.

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

Chu Z., Gianvecchio S., Wang H., Jajodia S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *Dependable and Secure Computing, IEEE Transactions on*, vol. 9, n° 6, p. 811–824.

Chu Z., Widjaja I., Wang H. (2012). Detecting social spam campaigns on twitter. In *Applied cryptography and network security*, p. 455–472.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H. (2009, novembre). The weka data mining software: An update. *SIGKDD Explor. Newsl.*, vol. 11, n° 1, p. 10–18.

Juran J., Godfrey A. B. (1999). Quality handbook. *Republished McGraw-Hill*, p. 173–178.

Kaplan A. M., Haenlein M. (2011). The early bird catches the news: Nine things you should know about micro-blogging. *Business horizons*, vol. 54, n° 2, p. 105–113.

Kullback S. (1987). The Kullback-Leibler Distance. *The American Statistician*, vol. 41, n° 4, p. 340–341.

Lee K., Caverlee J., Webb S. (2010). Uncovering social spammers: Social honeypots + machine learning. In *Proceedings of the 33rd international acm sigir conference on research and development in information retrieval*, p. 435–442. New York, NY, USA, ACM. Consulté sur http://doi.acm.org/10.1145/1835449.1835522

Manning C. D., Raghavan P., Schütze H. (2008). *Introduction to information retrieval*. New York, NY, USA, Cambridge University Press.

Martinez-Romo J., Araujo L. (2013). Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications*, vol. 40, n° 8, p. 2992–3000.

Mccord M., Chuah M. (2011). Spam detection on twitter using traditional classifiers. In *International conference on autonomic and trusted computing*, p. 175–186.

Mezghani M., Zayani C. A., Amous I., Péninou A., Sèdes F. (2014). Dynamic enrichment of social users' interests. In *IEEE 8th international conference on research challenges in information science, RCIS 2014, marrakech, morocco, may 28-30, 2014*, p. 1–11. Consulté sur http://dx.doi.org/10.1109/RCIS.2014.6861066

Ponte J. M., Croft W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international acm sigir conference on research and development in information retrieval*, p. 275–281.

Stringhini G., Kruegel C., Vigna G. (2010). Detecting spammers on social networks. In *Proceedings of the 26th annual computer security applications conference*, p. 1–9. New York, NY, USA, ACM.

Twitter. (2016). *The twitter rules.* https://support.twitter.com/articles/18311#. ([Online; accessed 1-March-2016])

Wand Y., Wang R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, vol. 39, n° 11, p. 86–95.

Wang A. H. (2010, July). Don't follow me: Spam detection in twitter. In *Security and cryptography (secrypt), proceedings of the 2010 international conference on*, p. 1-10.

Yang C., Harkreader R. C., Gu G. (2011). Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers. In *Proceedings of the 14th international conference on recent advances in intrusion detection*, p. 318–337. Berlin, Heidelberg, Springer-Verlag.

Yardi S., Romero D., Schoenebeck G., boyd danah. (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

Zubiaga A., Liakata M., Procter R., Bontcheva K., Tolmie P. (2015). Towards detecting rumours in social media. In *Artificial intelligence for cities, papers from the 2015 AAAI workshop, austin, texas, usa, january 25, 2015*. Consulté sur http://aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10160