UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

Département d'ingénierie et d'informatique

Analyse et détection de pourriels textuels dans les réseaux sociaux par apprentissage

ESSAI PRÉSENTÉ

COMME EXIGENCE PARTIELLE POUR L'OBTENTION

DU GRADE DE MAÎTRISE EN SCIENCES ET TECHNOLOGIES DE

L'INFORMATION

Par

Ekpao Anani, PASSIGUE

18 Août 2015

Jury d’évaluation

Président du Jury **:**	Pr. Kamel Adi
Directeur de recherche :	Pr. Mohand Saïd Allili

Dédicace

À mes parents,

Christine Kouloum, Alfa (1938 - 2008)

&

Antoine Simfeikiayakou, Passigue

Remerciements

L’auteur de ce document tient à remercier monsieur ***Mohand Saïd Allili*** pour sa disponibilité, et sa franche collaboration tout au long de la réalisation du présent projet, et à monsieur ***Kamel Adi*** pour ses précieuses remarques et suggestions. Ses remerciements vont également, à madame ***Nadia Baaziz*** pour son soutien et ses bonnes idées d’orientation, et à tout le ***corps professoral*** du département de l’ingénierie et de l’informatique pour leur enseignement. Toutes ses reconnaissances vont à la famille ***PASSIGUE*** pour leur affection.

De près ou de loin, merci à tous ceux qui ont une forte pensée vers lui!

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

***SVM** : Support Vector Machine*

***CNIL** : Commission Nationale de l'Informatique et des Libertés (Suisse)*

***SMS**: Short Messaging Service*

***SPIM:** Spam over Instant Messaging*

***UIT** : Union Internationale de Télécommunication*

***URL**: Uniform Resource Locator*

***FXL**: Feature Extraction Language*

***HTML**: HyperText Mark-Up Language*

***SCR** : Somme des Carrés résiduels*

***TREC** : Text REtrieval Conference*

***NUS**: National University of Singapore*

***SLIQ**: Supervised Learning in Quest*

***CART**: Classification and Regression Trees*

***SMTP :** Simple Mail Transfer Protocol*

Résumé

Le développement des technologies multimédias a donné un nouvel élan aux réseaux sociaux, qui deviennent de plus en plus une ressource importante et diversifiée pour l'individu. Cependant, le potentiel de croissance et la portée de ces réseaux les rend vulnérables aux pourriels et à la cybercriminalité. Ceci a encouragé l'introduction de techniques de filtrage pour s'attaquer au problème. Récemment, les techniques d'apprentissage automatique ont suscité beaucoup d'intérêt et démontré un succès pour le filtrage de pourriels multimédias. La combinaison de ces techniques avec les méthodes statistiques permet une meilleure compréhension de la structure des données manipulées par les polluposteurs.

Dans cet essai, nous exposons les différentes formes de pourriels présentes dans les réseaux sociaux et leurs modes d'intrusion. Ensuite, nous nous focaliserons spécifiquement sur les pourriels textuels et les techniques d'apprentissage adaptées à leur classification. Ainsi, les classificateurs probabilistes (ex. bayésien naïf et régression logistique) et discriminants (ex. arbres de décision et machines à vecteurs de support (SVM)) sont explorés pour l'analyse des activités des polluposteurs dans les réseaux sociaux. Enfin, ces méthodes seront utilisées pour classifier différents corpus de données textuelles collectés de différents réseaux sociaux pour détecter des polluposteurs potentiels. Nous avons testé notre approche sur des messages réels et obtenu des résultats très probants. Entre autres, ces résultats démontrent la fiabilité des méthodes d'apprentissage pour le filtrage des pourriels sociaux.

Chapitre 1- Introduction

1.1 Présentation du sujet

Au cours de cette dernière décennie, les cybercriminels ont manifesté un intérêt grandissant pour les réseaux sociaux tels que Facebook, YouTube, Twitter, etc. adaptant avec succès leurs méthodes aux évolutions des technologies de l'Internet. Selon un rapport publié de Nexgate au cours de la première moitié de 2013 [1], le taux de croissance du pourriel sociale a été de 355%. Cette croissance est nettement plus rapide que celle des comptes et des messages sur les réseaux sociaux les plus marqués. Les réseaux sociaux font aujourd'hui partie du quotidien de nombreux usagers de l'Internet, et se présentent comme le premier lieu de rencontre avec ses amis. C'est aussi le lieu par excellence pour la plupart des usagers pour lire des commentaires ou voir des propositions d'autres personnes et se faire une opinion avant de choisir un produit spécifique. On peut également consulter les avis d'autres consommateurs sur des produits qui nous intéressent. Ces réseaux génèrent alors des volumes considérables de données personnelles à travers les différents profils qui se créent. Les informations personnelles des usagers sont précieuses pour une publicité ciblée et adaptée. Elles attirent les polluposteurs et les auteurs de virus, qui tentent régulièrement d'arnaquer des membres innocents afin d'exploiter à leur profit la confiance accordée à ces réseaux. Ainsi, les réseaux sociaux deviennent la cible privilégiée des malveillants.

Compte tenu de l'ampleur mondiale du préjudice causé par le pourriel social, nombreux sont des chercheurs qui combattent ce fléau, par l'usage des méthodes d'apprentissage automatique « Machine Learning », une nouvelle discipline émergente [2]. Celle-ci permet de développer une intelligence artificielle capable d'analyser automatiquement des données, de détecter des motifs et associations de données. Elle permet également d'utiliser des motifs pour la prédiction de données ou encore de prendre des décisions de façon automatique. C'est une discipline en plein essor en raison de ses nombreuses applications. Mise à part le filtrage du pourriel, elle est exploitée pour le forage de données, la reconnaissance des formes, la vision artificielle, la sécurité, la génétique, le web, l'économie, etc. Tan et al. [3] présentent une revue complète des approches récentes dans l'application des algorithmes d'apprentissage automatique pour filtrer le pourriel social.

Le filtrage de pourriels basé sur l'analyse sémantique des messages est un exemple de catégorisation de textes. Il est donc nécessaire de construire un vaste corpus de données (pourriel et messages légitimes) et extraire des attributs les plus représentatifs des messages qui serviront à entrainer les algorithmes de filtrage. Cormack précise que cela permet d'attribuer à un document textuel, un ensemble de classes prédéfinies [4, 5]. La phase conjointe d'analyse et de sélection de caractéristiques est essentielle dans les systèmes de classification de messages. En effet, il s'agit de constituer un sous-ensemble minimum de $\mathcal{N}$ caractéristiques à partir d'un ensemble original, contenant $\mathcal{D}$ ($\mathcal{N} \leq \mathcal{D}$), de sorte que l'espace de caractéristiques soit réduit de façon optimale selon certains critères de sélection [5]. L'objectif est de maintenir ou d'améliorer les performances des classificateurs, y compris la complexité des algorithmes dans le système, tout en réduisant le nombre de caractéristiques à utiliser.

La section suivante dans le présent chapitre portera sur les modalités du pourriel social et des réseaux sociaux. Plus précisément, il s'agit de passer en revue les différents réseaux sociaux et leurs caractéristiques, les catégories de pourriels qui règnent dans ces réseaux et leurs modes opératoires.

1.2 Le pourriel et les réseaux sociaux

En Janvier 1994, le réseau Usenet reçoit le premier pourriel à l'échelle mondiale. Dans ce pourriel, un administrateur système de l'Université Andrews annonçait la venue prochaine d'un soi-disant "messie" [6]**.** Le pourriel n'est donc pas un phénomène nouveau. Il date de très longtemps, et a su évoluer également avec le temps. Ces évolutions d'une part, sont purement l'initiation des polluposteurs eux-mêmes qui ne sont jamais à court d'idées et, d'autre part, sont induites par la ténacité des acteurs de la lutte anti-spam.

1.2.1 Définition du concept de pourriel

Selon la CNIL (La Commission nationale française de l'informatique et des libertés), le pourriel est défini de la manière suivante : « Le "spamming" ou "spam" est l'envoi massif, et parfois répété, de messages électroniques non sollicités, à des personnes avec lesquelles l'expéditeur n'a jamais eu de contact et dont il a capté l'adresse électronique de façon irrégulière » [6]. Cette définition illustre bien le phénomène, mais l'on peut toutefois la nuancer, puisqu'auparavant le regard était seulement rivé sur les courriers électroniques. À l'heure actuelle, le pourriel s'est

répandu à d'autres secteurs d'activités, tels que la messagerie instantanée (Chat), les services de texte court (SMS), les appels téléphoniques, les forums, les blogues et aussi vers les réseaux sociaux. Ainsi, la définition du concept de pourriel peut être moins restrictive en termes de fréquence ou de la quantité de moyens de collecte des renseignements personnels, ou encore en termes du mode de diffusion du contenu.

Le contenu du pourriel n'est ni pertinent ni sollicité et de nature souvent nuisible. Son but est de détourner l'attention des internautes des informations jugées dignes d'intérêts, et de proposer ou de promouvoir certains produits et services non accessibles à ceux-ci dans le monde réel. Pour lutter efficacement contre ces formes de pollution dans les réseaux sociaux et adopter des mesures de filtrage adéquates, il est indispensable de bien saisir le concept de pourriel, ses caractéristiques, ses modes opératoires et le vocabulaire qu'il est susceptible de véhiculer.

1.2.2 Vocabulaire associé au pourriel social

Selon une étude menée par Sophos en 2013 [7], certains termes sont fréquents dans les thèmes véhiculés par les pourriels. Il s'agit des termes comme «"viagra", "argent", "loterie", "régime amaigrissant", "chèque", "casino en ligne", "poker", "revenu complémentaire"», etc. Ces termes pourront être associés aux algorithmes de filtrage antispam. Alors, la probabilité d'appartenance à un pourriel sera plus élevée si ces termes sont fréquents dans le contenu du message électronique. Cette même étude de Sophos a montré que plus de 50% de pourriels appartiennent aux catégories de messages classés selon les thèmes suivants:

- Adulte et pharmaceutique (pornographie et érotisme)
- Escroquerie financière (hameçonnage, et casino en ligne)
- Informatique et multimédia (optimisation de site Internet, antispam, antivirus)
- Éducation et formation (formation accessible à un tarif intéressant, parfois de faux diplômes)
- Messages politiques (diffusion des menaces politiques ou terroristes réalisées par des groupes extrémistes)
- Appel à la charité et à la spiritualité (évangélisation spirituelle ou religieuse, astrologie, lettre de chaîne, etc.)
- Pourriel destiné à distribuer des logiciels malveillants, etc.

1.2.3 Subjectivité du pourriel

Pour écarter toute ambiguïté autour du concept de pourriel, il est indispensable de clarifier sur le contenu de celui-ci et sur celui du message légitime. Le pourriel peut se présenter sous forme d'une campagne publicitaire (image, vidéo), avec un contenu qui n'est pas souvent recherché par les destinataires, et donc non pertinent. Il peut également se présenter sous forme de messages textuels contenant des liens vers des ressources potentiellement dangereuses. En revanche, une proposition commerciale légitime, une invitation adressée personnellement à un destinataire ou un bulletin d'information peuvent aussi paraître comme un message non sollicité, mais loin d'être du pourriel. C'est l'exemple d'un message dont l'expéditeur s'est trompé d'adresse, ou des messages en provenance des administrateurs systèmes ou d'anciens amis, qui n'avaient pas encore l'occasion de correspondre par courriels. Effectivement, ces messages ne sont pas sollicités, mais pas nécessairement indésirables. Toutefois, il arrive aussi que des messages en provenance des bulletins d'information et autres maquettes publicitaires intéressent certains, pendant que d'autres les classent dans les pourriels. Le concept de pourriel est dans une grande mesure subjectif [8].

Dans un contexte où le concept de pourriel (spam) peut porter à confusion, seule la pertinence du contenu permettra de distinguer le pourriel d'un message légitime. Mais comment évaluer la pertinence du contenu d'un message sachant que l'appréciation ou l'intérêt porté à ce contenu peut varier d'un utilisateur à un autre? Pour mieux répondre à cette inquiétude, il convient alors de classifier des messages sur des critères stricts reposant sur des caractéristiques simples et exemptes de toute ambiguïté.

À défaut de prédire le thème que véhicule un message, et surtout, d'être en mesure d'émettre un jugement en lieu et place de l'utilisateur, il est souhaitable, dans la plupart des situations, de se limiter à l'analyser de la structure, l'origine, la destination et les effets de bord d'un message. Cette analyse permet de détecter d'éventuelles déviations de comportements et de contextes des utilisateurs. Pour détecter les déviations et les caractéristiques suspectes, les techniques actuelles vont s'appuyer essentiellement sur l'analyse du comportement de l'utilisateur, et les thématiques abordées. Enfin la combinaison des données statistiques permet de réaliser une classification basée sur la sémantique.

1.2.4 Évolution du pourriel

Selon le rapport ***Statista*** [9], le nombre d'utilisateurs de réseaux sociaux atteindra les 1,96 milliard à la fin 2015, et estimé à environ 2,44 milliards d'utilisateurs sur le globe, jusqu'en fin 2018. En revanche, le volume des pourriels enregistré sur les réseaux sociaux a atteint un taux de croissance de 66,41% du trafic mondial en décembre 2014 [1].

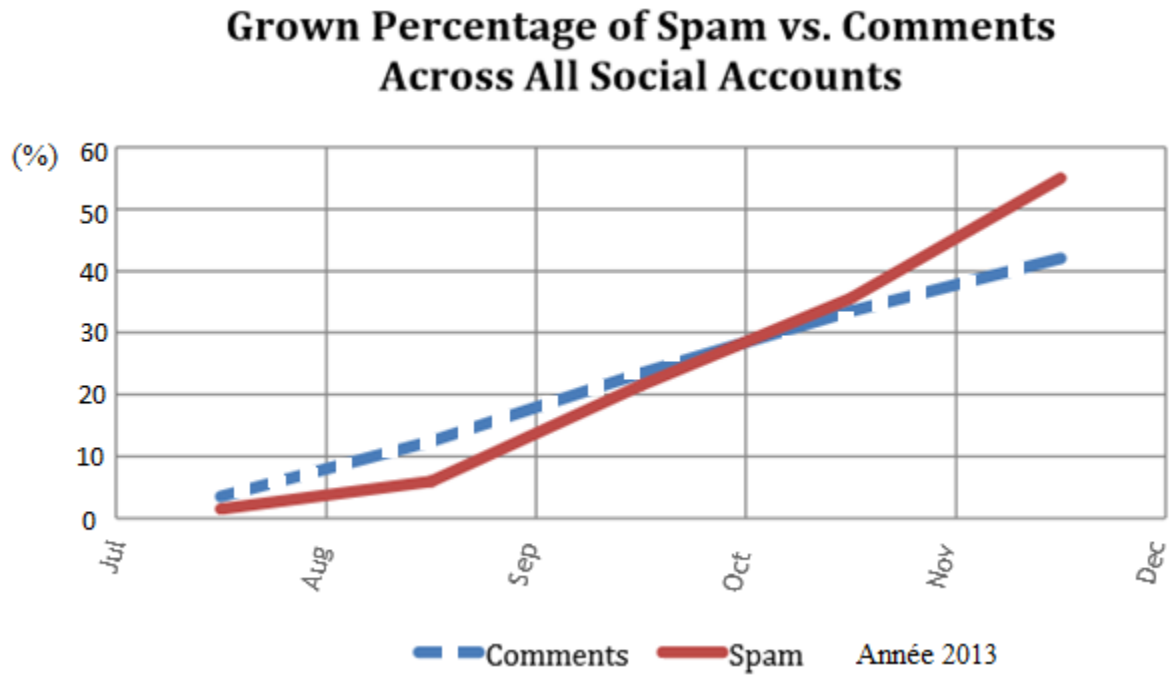

Figure 1: *Évolution du pourriel par rapport au volume de messages transitant sur les réseaux sociaux. Source image, www.nexgate.com*

Selon le classement effectué par ***Sophos*** [7] entre juillet et septembre 2014, les États-Unis ont été les principaux relayeurs de pourriel avec 11,5% de tous les pourriels distribués. Ensuite vient la Chine avec un taux de 9,1% du pourriel mondial, et de la France qui occupe la troisième place avec une moyenne de 6,4%.

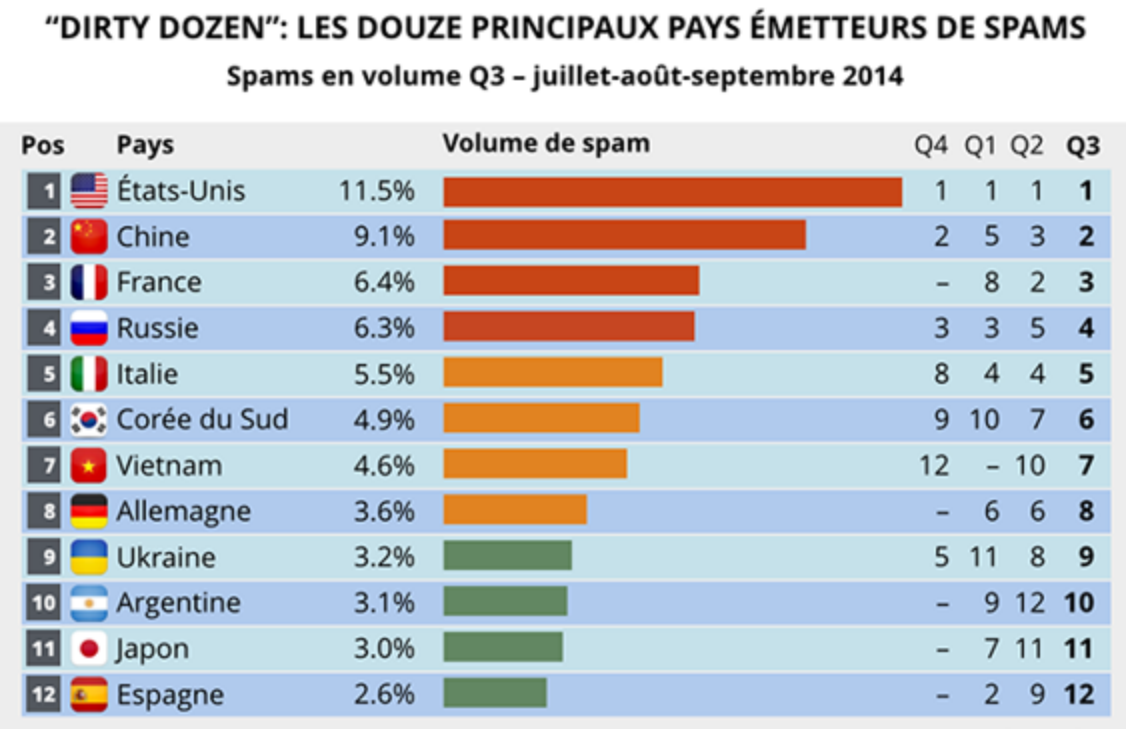

Figure 2: *Distribution du pourriel dans le monde, Source image www.sophos.com*

1.3 Modèles de réseaux sociaux

Les réseaux sociaux en ligne ont fait leur apparition aux États-Unis en 1995, mais leur extension vers le reste du globe n'a été effectif qu'à partir de 2004. Ils servent à relier les amis, les associés, les professionnels, organisés autour d'un thème fédérateur à savoir : le domaine d'étude, la classe sociale, la religion, etc. Ils permettent enfin de partager des intérêts, des connaissances, des métiers, des passions communes. En outre, c'est le premier lieu pour lire les commentaires avant de choisir par exemple son prochain téléphone intelligent, voir les propositions d'autres personnes avant de choisir son hôtel de destination, ou encore avant la réservation de son billet d'avion chez une compagnie, etc.

En plus du phénomène de la globalisation qui nous envahit, s'ajoutent les fonctionnalités multimédias considérablement enrichies des plateformes de réseaux sociaux. Les internautes ont finalement accès à une masse de nouveaux produits à découvrir. D'autres opportunités sont offertes par l'ouverture du web avec l'avènement du web 2.0 qui permet aux utilisateurs de créer leurs propres contenus. En dépit de ces énormes opportunités, la contrepartie en est que les

réseaux sociaux sont vulnérables à toute manipulation. Ces multiples opportunités expliquent l'envahissement des réseaux sociaux par les polluposteurs.

À travers la littérature, on distingue l'existence de deux catégories de réseaux sociaux : les réseaux sociaux qui exposent publiquement les informations de profil (ex. MySpace), et les réseaux basés sur la confiance (ex. Facebook, Twitter) [2]. Dans la première catégorie, les renseignements personnels d'un utilisateur sont accessibles à tous les utilisateurs ou visiteurs du réseau. Dans la deuxième, l'accès à l'information d'un profil et l'interaction avec l'utilisateur devraient se limiter aux personnes ayant été acceptées dans un cercle, bâti sur la confiance. Malheureusement, il arrive de retrouver des intrus dans des cercles de confiances. Les analyses révèlent que cette intrusion est liée d'une part à la vulnérabilité des mécanismes d'authentification érigés par les concepteurs des réseaux sociaux, et d'autre part à l'imprudence des utilisateurs eux-mêmes.

D'autres études révèlent que plus de 83% des utilisateurs acceptent sans réfléchir d'ajouter à leur liste de confiance, des ami(e)s dont ils ignorent totalement l'existence réelle. Ensuite, 45% des utilisateurs ont le clique facile sur les liens qui sont postés sur leur mur [2]. Les spécialistes de la sécurité trouvent que la confiance devrait constituer l'une des caractéristiques fondamentales pour sécuriser tout réseau social. Mais les expériences illustrent à quel point les utilisateurs se précipitent sans aucune vérification à ajouter des amis à leurs profils, ou rejoindre des groupes inconnus, voire devenir fans de pages dangereuses. Ces attitudes rendent complexes la compréhension du concept de pourriel et les méthodes d'intrusion des polluposteurs.

1.3.1 Caractéristiques des réseaux sociaux

Dans cette section, nous allons décrire les caractéristiques de 4 réseaux sociaux les plus menacés par les activités des polluposteurs. Il s'agit de Facebook, MySpace, Twitter et Sino Weibo.

1.3.1.1 Le réseau social Facebook

D'après les statistiques publiées sur son site web [10], Facebook comptait en mars 2015, 1.41 milliard de membres actifs dans le monde, avec plus de 2 milliards d'objets (vidéo, images) en partages chaque semaine. C'est un réseau basé sur la confiance, c'est-à-dire, habituellement, les profils ne sont pas public et peuvent être vu uniquement par des « amis », bien qu'ami n'ait pas

la même définition que dans la vraie vie. Lorsqu'un membre « A » du réseau souhaite établir une relation avec un membre « B », la plateforme envoie d'abord une requête au membre « B », qui doit confirmer qu'il connait « A ». Une fois que « B » confirme reconnaitre « A », un rapport de lien est établi entre les deux. Dans la majorité des cas, les utilisateurs de Facebook, sans réfléchir sont enclins à se précipiter d'ajouter des personnes inconnues à leur profil. Dans la vie réelle, cela aurait été possible qu'après un contrôle ou une vérification de la personnalité de celles-ci.

Dans le passé, la plupart des membres de Facebook était groupée en réseaux, où les ressortissants d'un pays ou d'une ville, les camarades des écoles pouvaient se retrouver en communautés ou entre pairs. Or, les paramètres par défaut de Facebook permettent à chaque membre, la visibilité de l'ensemble des profils inscrits dans un même réseau. Ainsi, un utilisateur malicieux aurait la main mise sur un vaste réseau de données, pouvant servir à des attaques ciblées et de grande envergure. Par exemple, il est plus facile de mener une campagne ciblée, connaissant le genre, l'âge, les centres d'intérêt de certains profils. Pour cette raison, Facebook a déprécié les réseaux géographiques en octobre 2009. En revanche, les réseaux d'écoles et des fondations continuent d'exister sous cette forme, mais avec une sécurité renforcée. C'est la raison pour laquelle il faut fournir une adresse courriel valide d'une organisation membre avant tout accès.

1.3.1.2 Le réseau social MySpace

MySpace a été l'un des premiers réseaux sociaux qui a gagné une forte popularité chez les internautes. Son principe est de fournir à chaque membre, une page web que celui-ci peut personnaliser avec ses renseignements individuels et ses champs d'intérêt. À l'instar de Facebook, le concept de relation existe, sauf que les pages web des membres sont publiques par défaut. Par conséquent, il est plus facile pour un utilisateur malveillant d'obtenir des informations que sur Facebook. Une campagne de pourriel ciblée sera encore plus efficace connaissant le genre, l'âge, ou la nationalité des groupes spécifiques.

1.3.1.3 Le réseau social Twitter

Selon les statistiques, Twitter a été le réseau social dont le potentiel de croissance était le plus élevé dans Internet en 2009, soit une croissance de 660% [2]. C'est le réseau social le plus simple comparé à Facebook et MySpace. Il constitue une plateforme de micro blogue où les utilisateurs interagissent par des messages texte (tweets). Contrairement à Facebook et MySpace,

aucun renseignement personnel n'est accessible sur les pages Twitter par défaut. Les utilisateurs sont identifiés par leur nom d'utilisateur et, optionnellement, par leur nom propre.

Pour identifier un utilisateur, il faut analyser ses tweets, ce qui n'est pas une chose facile. Un utilisateur de Twitter peut commencer à suivre un autre, et ainsi il aura accès à tous ses tweets, et si l'autre personne accepte, elle peut suivre la première en retour. Les tweets peuvent être groupés par « hashtag », ce qui permet aux autres membres de vérifier avec précision, l'origine d'un tweet, posté à tel moment, et portant sur tel sujet et tel centre d'intérêt. Lorsqu'un tweet quelconque intéresse un membre, il peut réagir, et automatiquement sa réaction est diffusée à tous ses amis adeptes qui le suivent.

1.3.1.4 Le réseau social Sino Weibo

Selon Xianghan [11], Sino Weibo (réseau social le plus populaire en chine), est une plateforme de micro blogue conçu à l'instar de Twitter. La taille maximale des messages est de 140 caractères. Lorsqu'un message est posté, les amis qui suivent l'expéditeur du post (message) prennent immédiatement connaissance du message. Chaque membre étant identifié par son nom d'utilisateur, peut à son tour suivre d'autres amis en vue de prendre connaissance des mises à jour disponibles sur leurs pages individuelles. Le membre qui est suivi peut accepter ou non la requête de suivre à son tour, ses amis adeptes qui le suivent. Il existe plusieurs façons de poster un message sur ce réseau social :

- **La notification (mention)**

Lorsqu'un expéditeur mentionne une série de mots clés comme @username dans le message, cela signifie qu'il s'adresse à un groupe spécifique de destinataires. Par conséquent, la plateforme weibo notifie les destinataires concernés avec le message sur leur page d'accueil.

- **La réplique (repost)**

C'est aussi une autre façon de communiquer sur weibo. La réplique concerne tous les amis adeptes d'un utilisateur, autour d'un sujet de prédilection.

- **Le hashtag (#)**

Le message marqué de hashtag indique un sujet spécifique autour duquel l'on souhaite lancer un débat. Si le sujet attire un nombre important de membres, alors celui-ci apparait dans la liste des sujets d'actualité.

1.4 Vecteurs de propagation du pourriel

Le pourriel peut s'attaquer à divers médias électroniques : les courriels, le téléphone fixe ou mobile, les messageries instantanées (SPIM), les forums de discussion, les moteurs de recherche, les wikis, les blogues, etc.

1.4.1 Vecteur par courrier électronique (Courriel)

Le pourriel par courrier électronique est le type de pourriel le plus répandu. Selon les statistiques publiées sur le site de Sophos [7] en 2014, la vélocité des échanges était évaluée à 196 milliards de courriels par jours, dont 95% des courriels reçus étaient des pourriels. Ces statistiques sont confirmées par Trend Micro, une des compagnies de lutte antispam.

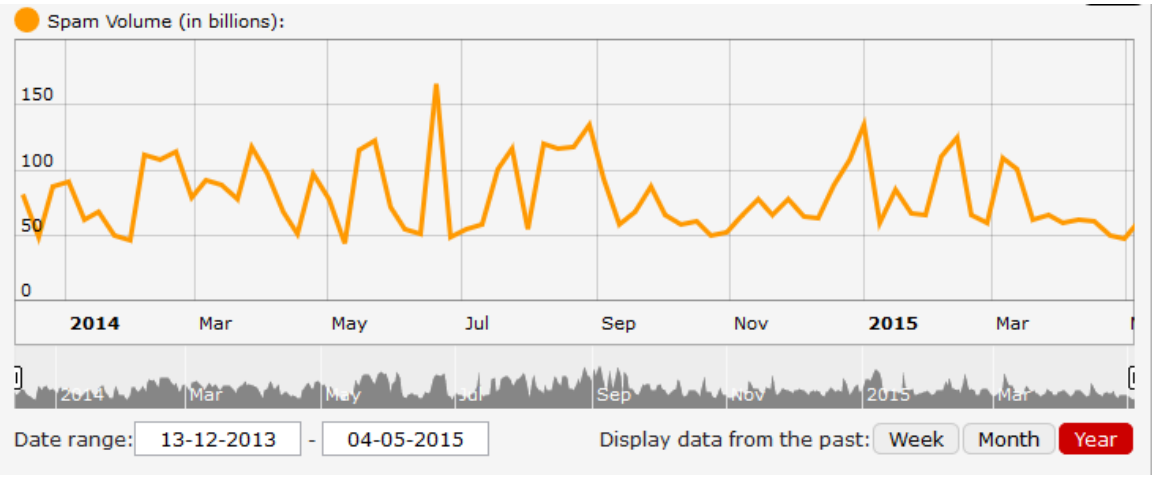

Figure 3: *Volume du pourriel, source http://www.trendmicro.com*

Le coût d'envoi des messages étant dérisoire, il est plus facile de diffuser un message à des millions de destinataires qui assumeront à leur tour les coûts de réception et de stockage. Les polluposteurs redoublent d'imagination pour masquer leurs activités et restent habiles, souvent en falsifiant les adresses d'expéditeur ou en déployant des serveurs SMTP (Simple Mail Transfer Protocol) qui permettent des envois anonymes. Généralement, les polluposteurs utilisent des robots d'indexation pour collecter des adresses victimes auxquelles le pourriel est envoyé. Selon Wikipédia, il existerait des marchés où les polluposteurs peuvent se procurer des listes d'adresses sous forme de CD-ROM, ce qui renforce le phénomène du pourriel.

1.4.2 Vecteur par SMS (Short Message Service)

Depuis quelques années, une nouvelle forme de pourriels menace les réseaux de téléphonie fixe et mobile. En effet, les messages textes désignés par SMS sont couramment utilisés entre les utilisateurs de téléphone cellulaire, pour substituer les appels vocaux dans des situations où la communication vocale est impossible ou indésirable. Cette façon de communiquer est très répandue puisque le message texte coûte nettement moins cher que de passer un appel téléphonique vers un autre téléphone mobile [4]. Selon les études menées par l'Union Internationale de Télécommunication (UIT), le SMS est devenu une véritable industrie commerciale, avec un volume de trafic qui est passé de 6,9 milliards à la fin 2010 à 8000 milliards la fin 2011. Le revers de la médaille de ce service est que le téléphone cellulaire est devenu la cible des pourriels, avec le nombre croissant de commerces qui ont recours aux SMS pour des raisons de publicités ou pour des communications de masse avec leur clientèle.

La pratique des SMS est bien rare en Amérique du Nord avec un taux estimé à 1% de pourriels de type SMS par jour. Comparativement à la Chine et certains pays de l'Asie, le taux du pourriel a atteint les (20 – 30%) par jour en 2010 soit un taux annuel de 300% [4]. Le prix du SMS étant dérisoire (0,001$) en chine et voire gratuit dans certains pays, ceci aurait sans doute contribué à la croissance du phénomène. Il faut rappeler que les numéros encapsulés dans les SMS sont souvent des numéros audiotels surtaxés. Ce qui rend le phénomène coûteux et ennuyeux pour les consommateurs, d'où la nécessité de s'attaquer à ce fléau.

1.4.3 Vecteur par les blogues et forums

L'un des critères importants de classement dans les moteurs de recherche est la présence de liens vers un site web. Des blogues peuvent alors être créés uniquement dans le but d'augmenter le nombre de liens pointant vers des sites particuliers. On peut tout de même faire usage des blogues existants en ajoutant dans des commentaires, des liens vers un ou plusieurs sites web à promouvoir. Pour empêcher aux robots de diffuser automatiquement ce genre de pollution, la plupart des logiciels de blogue ont intégré dans leur système des contrôles (principe de Captcha) [2]. Dans les forums, certains groupes de discussion ne reçoivent pratiquement que du pourriel. C'est d'ailleurs la raison pour laquelle de nombreux forums sont modérés, c'est-à-dire surveillés par des êtres humains ou un robot qui effectue un tri parmi les articles proposés.

1.5 Motivation des polluposteurs

La première étape vers la conception de mesures efficaces pour s'attaquer au pourriel social, nécessite une bonne compréhension des motivations des polluposteurs. Selon les expériences réalisées par [12], la motivation axée sur le gain financier a été privilégiée. La visite d'un site web par plusieurs utilisateurs génère certainement des revenus aux polluposteurs. Évidemment, ceux-ci s'appuient sur le message électronique, la manipulation des moteurs de recherches et les campagnes publicitaires pour attirer les utilisateurs vers leurs cibles.

Les infrastructures de courriels et les moteurs de recherches parviennent presque à évincer les polluposteurs avec les techniques de listes noires et les algorithmes de classification actuels. En revanche, les réseaux sociaux restent encore une cible d'opportunité pour la publicité surtout que leur usage fait presque partie du quotidien des individus. La masse des utilisateurs et le coût réduit pour placer de la publicité constituent des facteurs attractifs des polluposteurs. En plus, le polluposteur n'a besoin ni de la puissance de calcul de leur ordinateur, ni de la bande passante payante ou encore de l'espace de stockage, puisqu'il s'appuie souvent frauduleusement sur les ressources informatiques d'autrui. Comment faire de l'argent en ciblant les systèmes des réseaux sociaux est sa seule préoccupation. C'est ainsi que le pourriel continue d'exister et de prospérer grâce aux revenus qu'il peut engendrer. Ceci explique la motivation de certains commanditaires car près de 11 % d'internautes admettent avoir acheté un produit à la suite de la réception d'un pourriel publicitaire [7].

1.6 Le pourriel social

C'est une menace dans les réseaux sociaux qui se présente sous deux formes principales [1] :

- sous forme de lien (une image, lien hypertexte, affiche publicitaire),
- sous forme textuelle (ex. un tweet, requête d'ajout à la liste des amis).

1.6.1 Le pourriel de liens

Il consiste à proposer des liens cliquables. L'objectif étant de rediriger les utilisateurs vers une campagne publicitaire ou vers un fichier exécutable, permettant d'attaquer la machine cible. L'intérêt particulier de ce type de pourriel réside dans le référencement abusif qui consiste à tromper les moteurs de recherches sur la qualité de la page ou du site web. Cette technique

permet d'obtenir selon les mots clés indexés, un bon classement dans les résultats des moteurs de recherche. Habituellement les utilisateurs préfèrent visiter la première page, et donc les polluposteurs visent le premier rang de classement. Toutefois, le pourriel peut aussi se présenter sous forme de ferme de liens (link farm), c'est-à-dire un site web hébergeant des listes de liens vers d'autres sites qui sont sous le contrôle des polluposteurs. C'est également une technique qui vise l'amélioration de classement des sites web.

À l'heure actuelle, il y'a une avancée significative dans le filtrage des pourriels de courriels grâce aux systèmes de listes noires des URL et du filtrage de contenu. Mais les techniques pour détecter, classifier et supprimer efficacement les URL à l'intérieur des messages des réseaux sociaux présentent encore des défis de taille. À cet effet, plusieurs organisations passent en revue manuellement tous les messages postés par les utilisateurs, une opération très coûteuse en termes de temps, tout en comptant avec les erreurs humaines.

1.6.2 Pourriel textuel

Il se présente sous deux formes : la création de comptes factice et le piratage informatique.

1.6.2.1 Création de comptes factice

Il s'agit de créer des comptes factices sur les réseaux sociaux pour ajouter des amis afin d'extraire des renseignements personnels maladroitement publiés sur leurs pages d'accueil. Ces comptes sont souvent dotés « d'images chaudes ou coquettes » ou des textes de contenu carrément captivant pour attirer le maximum d'utilisateurs. L'un des exemples les plus marquants est la chaîne de lettre. Elle consiste à relayer un message, sous peine d'un quelconque malheur en cas de refus de diffusion, ou simplement diffuser au maximum possible un message qui nécessite des actions humanitaires. On se souvient des comptes factices repérés sur Facebook, qui collectaient frauduleusement des dons prétendument destinés au tremblement de terre en Haïti.

1.6.2.2 Piratage informatique

Dans cette catégorie, il s'agit de manipuler malicieusement le compte de la victime pour en faire de ce dernier un polluposteur. Cette méthode semble la plus efficace jusqu'à date, puisque tous les contacts d'un carnet d'adresses croient avoir affaire à leur ami en qui ils ont confiance. Ainsi,

les liens postés par ce compte sont visités sans inquiétude. Généralement, il est question d'activer une extension (plug-in) à partir de votre navigateur. Ensuite, cette opération diffuse automatiquement à votre insu le message malveillant, y compris vos identifiants de connexion.

Le pourriel textuel peut aussi servir d'une arnaque de type hameçonnage (Phishing), où le destinataire est invité à entrer ses identifiants sur un compte qui est sous le contrôle des polluposteurs. Par exemple, lors d'un achat en ligne sur un site malveillant. L'utilisateur est alors dupé de manière subtile afin de lui soustraire des renseignements personnels qui peuvent ensuite être utilisés pour accéder à d'autres comptes, tels que les comptes bancaires. C'est donc de l'ingénierie sociale qu'exploitent les polluposteurs pour gagner la confiance des utilisateurs sur les réseaux sociaux.

Par ailleurs, différentes techniques de diffusion de messages malveillants sont utilisées sur les réseaux sociaux par le biais des robots (spamboats). Leurs activités semblent plus ou moins complexes, mais leur objectif principal est d'échapper au contrôle des administrateurs réseaux, tout en diffusant des pourriels.

1.7 Les modes opératoires des polluposteurs dans les réseaux sociaux.

Les spambots sont des robots polluposteurs dans les réseaux sociaux. Selon les différentes stratégies de pollution adoptées, on distingue quatre formes de robots [2]:

«Displayer» : ce type de robot ne diffuse pas de pourriel, mais les affiches plutôt sur son propre profil. Pour afficher le contenu du pourriel, il faut visiter le profil. On retrouve très souvent ce type de robot sur MySpace et Facebook, et il reste le moins agressif et le moins efficace, car il ne touche qu'un nombre très réduit de personnes.

«Bragger» : ce robot se présente sous différentes formes selon le réseau : sous forme de lien pour effectuer par exemple la mise à jour de votre statut sur Facebook, et sous forme d'un tweet sur Twitter. Les liens n'étant pas visibles aux visiteurs et amis des victimes, seuls les utilisateurs directement connectés au robot sont concernés par la campagne du pourriel. Ce type de robot est plus présent sur Facebook et sur Twitter.

«Poster» : c'est le robot le plus virulent de tous. Selon le type de réseau, par exemple sur Facebook, il envoie le pourriel directement sur les murs de chaque victime. Le pourriel est

ensuite immédiatement diffusé à toute la liste des amis des victimes y compris les visiteurs du profil, ce qui élargit le nombre de personnes touchées.

«Whisperer» : est un spambot qui diffuse des courriels ou des messages instantanés privés à leurs victimes. À la différence de Poster, seule la victime est capable de consulter le message pourriel. Ce type de robot joue en particulier sur la spontanéité et la naïveté des utilisateurs, et il est le plus souvent repéré sur le réseau Twitter.

D'autre part, les activités des robots peuvent se résumer en deux grandes catégories: les **«greedys»** et les **«stealthy».** Dans la première catégorie, les robots encapsulent dans un message un lien pointant directement vers un site de campagne de pourriel, tandis que le second robot diffuse des messages d'apparence légitime, mais injecte sporadiquement les messages malicieux. Dans les activités de ce dernier, on remarque souvent des messages avec des images mises en pièces jointes et qui ne font apparaître aucun lien. Cet aperçu du phénomène de pourriel social témoigne que le pourriel est réellement une menace et ne doit pas laisser indifférente la communauté des chercheurs.

1.8 Contribution du projet

Dans ce projet, nous abordons la classification de pourriels textuels dans les sites de réseaux sociaux. Nous utilisons des classificateurs supervisés à la fois probabilistes et discriminants pour filtrer le pourriel. Nous utiliserons des corpus publics existant pour l'apprentissage. Nous adoptons l'approche de sac à mots pour extraire les caractéristiques sémantiques qui permettront de distinguer les pourriels des messages légitimes. Après la phase d'apprentissage, nous validons notre méthodologie sur un ensemble de messages.

Ce rapport est organisé comme suit : dans le deuxième chapitre, nous présentons l'état de l'art relatif aux différentes approches (statistiques, dynamiques et heuristiques) mises au point pour le filtrage du pourriel en général, et du pourriel social en particulier. Dans le troisième chapitre, nous abordons les méthodes d'apprentissage automatique et leurs avantages dans la détection de pourriels. Dans le quatrième chapitre, nous présentons notre contribution en termes de résultats obtenus après avoir effectué le prétraitement et la représentation vectorielle du corpus de données textuelles collectées. Enfin, nous terminons le rapport par une conclusion générale.

Chapitre 2- État de l'art

2.1 Introduction

Plusieurs publications récentes démontrent l'intérêt d'empêcher la pollution numérique dans les réseaux sociaux. Plusieurs chercheurs ont examiné l'usage des méthodes classiques pour filtrer le pourriel social. D'autres ont exploré de nouvelles techniques de filtrage à partir des techniques d'apprentissage automatique. Dans la réalité, le pourriel social a été une opportunité pour tester, comparer et valider plusieurs méthodes de classification automatique.

2.2 Pourriel de courriels et du web

Les premières publications concernant la classification du courrier électronique basées sur l'apprentissage automatique datent de 1996 [13]. Rocchio [14] s'est intéressé à la reconnaissance des pourriels par la classification thématique de messages. Il représente le message à classifier sous la forme d'un vecteur. Ensuite il évalue la distance (généralement euclidienne ou de Manhattan définissant la pertinence d'un message) entre ce vecteur et les vecteurs similaires de chaque classe, associant le message à celle dont le vecteur prototype est le plus proche [15].

Cohen [13] effectue la comparaison de Ripper avec la méthode de Rocchio. Ripper est un classificateur capable de définir le contexte d'un mot. Ce classificateur utilise un ensemble de règles (présence ou absence des mots du dictionnaire), qui est automatiquement constitué pendant la phase d'apprentissage. Ripper effectue aussi une phase d'optimisation afin de réajuster les ensembles de règles. Les deux méthodes ont présenté des résultats similaires, mais on constate une amélioration lorsque l'apprentissage est fait individuellement pour chaque utilisateur plutôt que collectivement pour tous les utilisateurs.

Pantel [16] et Sahami et al. [17] proposent l'utilisation d'un classificateur de type Bayes naïf pour le filtrage de pourriels. Pour la réalisation de son expérience, [17] sélectionne 500 caractéristiques de chaque message, et constate que la classification binaire (ham, pourriel) était plus efficace que la classification multi-classes basée sur le genre du message (érotisme, hameçonnage, etc.).

Drucker et al. [15] proposent l'usage des machines à vecteurs de support (SVM) pour le filtrage du pourriel. D'une part, ils comparent les performances d'un classificateur SVM linéaire avec les classificateurs basés sur les règles (Rocchio, Ripper) et d'autre part avec les arbres « Boosted trees ». Ils ont abouti à plusieurs configurations expérimentales, en occurrence sur les modes d'apprentissage, la mise au point et la sélection d'attributs. Plusieurs de leurs conclusions publiées demeurent valides jusqu'à date :

- Les méthodes basées sur des règles ne sont pas performantes pour le filtrage de pourriel,
- Les performances des SVM et les arbres de type « boosted trees » sont comparables, mais le taux de faux positifs enregistré par SVM est plus faible,
- L'apprentissage des arbres de type « boosted trees » est excessivement lent,
- Pour ce qui concerne les méthodes SVM, les attributs binaires qui indiquent la présence ou l'absence de termes donnent de meilleurs résultats, alors que les attributs multinomiaux, c'est-à-dire le nombre d'occurrences des termes, sont à privilégier pour les « boosted trees ».
- Les procédures de sélection d'attributs sont d'une grande complexité et le mieux est de les intégrer dans l'apprentissage si l'on souhaite les utiliser,
- Enfin, il faut aussi tenir compte des termes neutres dits « stop words ».

Dans la littérature, plusieurs travaux de recherche ont confirmé l'efficacité des approches SVM pour filtrer le pourriel. Kolcz [18] s'est intéressé aux erreurs de classification spécifiques à chaque classe, pendant qu'Islam [19] publie sa méthode de sélection d'attributs afin d'améliorer la classification du pourriel SMTP. Joachims [20] à son tour a amélioré les méthodes SVM linéaires, par des implémentations plus adaptées à un contexte de classification généralisé. Sculley [21] a évité de travailler sur des données synthétiques et effectué des tests plus réalistes dans l'apprentissage de messages et le filtrage en ligne de pourriels. À cet effet, il a déployé une méthode dite ROSVM (Relaxed Online SVM) qui a amélioré légèrement l'efficacité de la classification des pourriels surtout en optimisant la complexité de l'algorithme.

Blanzieri [22] offre un aperçu complet des techniques d'apprentissage automatique qui peuvent être appliquées au filtrage de courriels. Hao et al. [23] décrivent un moteur de réputation basée sur des caractéristiques légères, telles que la distance géographique entre l'émetteur et le récepteur et des attributs diurnes, c'est-à-dire des attributs liés au temps de la journée. Bien que

la cible soit le pourriel classique, le processus dans le suivi de réputation de l'expéditeur en utilisant des attributs similaires (par exemple, la période à laquelle les messages ont été envoyés au cours de la journée), semble aussi applicable au pourriel social.

Whittaker et al. [24] décrivent leur méthode d'apprentissage évolutive de détection d'hameçonnage et des systèmes de listes noires. Étant donné qu'une quantité considérable de pourriels des médias sociaux comprend des liens vers des sites malveillants, leur détection reste toujours un défi. Dans le même ordre d'idées, Thomas et al. [25] proposent Monarch, un système évolutif qui permet la détection en temps réel des URL qui pointent vers des pages web de pourriels tels que déterminés par les caractéristiques de l'URL, le contenu de la page et les propriétés d'hébergement du domaine cible. Ils montrent que la durée de la campagne publicitaire du pourriel sur Twitter était plus longue que dans les courriels classiques

Les pourriels dans les commentaires des blogues ont également attiré l'attention des chercheurs qui ont déployé des méthodes d'apprentissage automatiques [26, 27], à savoir les techniques SVM et les classificateurs bayésiens pour leur détection. Mishne et al. [28] ont utilisé la modélisation linguistique pour trouver des divergences sémantiques entre les blogues sur lesquels sont affichés des liens de commentaires de pourriels et les sites cibles vers lesquels ils pointent (qui pourraient, par exemple, contenir des contenus pour adultes). De même, Markines et al. [8] ont appliqué des techniques similaires (SVM et AdaBoost) pour la détection de pourriels dans les réseaux sociaux.

Les travaux précédents ont permis de diminuer le risque du pourriel SMTP qui continue de menacer sous d'autres formes (ex. le pourriel social). Les technologies éprouvées dans le filtrage des pourriels SMTP peuvent être utiles dans la lutte contre le pourriel social, étant donné que le pourriel en général exploite les mêmes vecteurs de propagation, à savoir le texte, les images, la vidéo, l'audio, etc. Ainsi, plusieurs chercheurs se sont inspirés de méthodes classiques de filtrage pour mettre au point de nouvelles méthodes de détection des activités des polluposteurs sur les sites de réseaux sociaux.

2.3 Caractérisation du pourriel social

Les sondages effectués par Heymann et al. [29] sur les plateformes des réseaux sociaux ont débouché sur certaines caractéristiques communes de pourriels. Ces caractéristiques sont basées

sur l'identité des utilisateurs, le classement selon le contenu et le comportement à travers les affiches. On se sert des caractéristiques de l'identité signalées comme pourriel par les autres utilisateurs et par des modérateurs de confiance pour ensuite entrainer les classificateurs. D'autres chercheurs vont se focaliser sur la collecte, l'identification des attributs et la classification des différents types de pourriels des réseaux sociaux.

Zinman et Donath [30] ont utilisé les caractéristiques des profils et des commentaires du réseau MySpace. Mais les performances relativement faibles de leurs travaux démontrent de la difficulté à procéder à la classification manuelle du pourriel social. Plusieurs autres études [2, 31] vont adopter l'approche des «Honeypot» qui consiste à créer des profils d'attraction avec la seule intention d'observer les activités des polluposteurs. Les données collectées serviront ensuite à entrainer des classificateurs. Les caractéristiques les plus pertinentes sont le nombre de requêtes d'ajout d'amis, le ratio des URL dans le texte, etc. Webb et al. [31] vont aussi utiliser la technique des «Honeypots» pour révéler différents types de polluposteurs de réseaux sociaux, la démographie typique de leurs profils et les pages web qui ont tendance à faire de la publicité.

Benevenuto et al. [32] identifient les attributs sociaux des pourriels et des messages légitimes (ham) à partir d'une analyse faite sur les vidéos YouTube. Ces attributs concernent le nombre de fois que la vidéo est lue, le nombre de commentaires émis sur une vidéo et les attributs publics d'un profil. Ils ont ensuite utilisé les machines à vecteurs de support (SVM) pour la classification, et obtiennent 96% de précision dans la détection des utilisateurs opportunistes (promoteurs), avec seulement 57% de polluposteurs potentiels.

Les contenus indésirables des réseaux sociaux ne sont pas nécessairement des pourriels, ou relatifs à l'escroquerie. Plusieurs utilisateurs témoignent aussi de comportements inappropriés de la communauté, où les utilisateurs postent des contenus offensifs et de harcèlements. Yin et al. [33] ont combiné l'analyse des sentiments avec des listes de mots de blasphème et des fonctionnalités contextuelles pour identifier le harcèlement à partir des corpus de données de Slashdot et MySpace. D'autres travaux sur les plates-formes des réseaux sociaux concernent la collecte des renseignements personnels qui facilite les vols d'identité ou des attaques directes des ordinateurs des utilisateurs, compromettant un grand nombre de comptes [27, 34].

2.3.1 Détection du pourriel social

SocialSpamGuard [35] est un système de détection de pourriel de médias sociaux qui analyse les caractéristiques des messages textuels et des images des médias sociaux. Pour effectuer des tests, le système utilise un cluster «General Activity Detection - GAD» pour l'échantillonnage de pourriel et des messages légitimes, puis entraine un classificateur avec les caractéristiques analysées. Cependant, le système est construit à partir de quelques caractéristiques des utilisateurs de Facebook accessibles au public (ex. le profil). L'accès à un nombre très réduit de caractéristiques ne contribue pas aussi pour accroitre l'efficacité de ce système.

2.3.2 Identification des comptes pourriels

En 2010, Stringhini et al. [2] se sont intéressés spécifiquement aux réseaux sociaux Facebook, MySpace et Twitter. À cet effet, six caractéristiques sont mises en évidence pour évaluer des comportements propres aux polluposteurs :

$$\mathcal{R} = \frac{\text{Following}}{\text{Followers}} \qquad (1)$$

Le ratio $\mathcal{R}$ renseigne si la plupart des requêtes d'ajout d'amis sont rejetées ou acceptées. Si $\mathcal{R}$ est élevé, alors on déduit qu'il s'agit d'un profil de polluposteur du fait du nombre faible d'amis.
Le ratio $\mathcal{U}$ renseigne de la proportion de messages, qui encapsule des URL dans le texte.

$$\mathcal{U} = \frac{\text{messages contenant des urls}}{\text{total des messages}} \qquad (2)$$

Où $\mathcal{S}$ est le paramètre qui renseigne sur le degré de similarité des messages envoyés par un utilisateur, sachant que les robots diffusent des messages souvent similaires :

$$\mathcal{S} = \frac{\sum_{\mathfrak{p}\in\mathcal{P}} c(\mathfrak{p})}{l_a l_{\mathfrak{p}}} \qquad (3)$$

Où $\mathcal{P}$ est l'ensemble des combinaisons possibles de couples de messages parmi tous les messages enregistrés pour un certain compte, $\mathfrak{p}$ est une paire de messages, $c(\mathfrak{p})$ est le nombre de mots partagés entre deux messages, l_a est la longueur moyenne de messages postés par cet utilisateur et $l_{\mathfrak{p}}$ est le nombre de combinaisons de message. L'idée derrière cette formule est que le profil qui envoie des messages similaires aura une faible valeur de $\mathcal{S}$.

La métrique $\mathcal{F}$ (friend choice) renseigne si un profil se sert d'une liste de noms pour choisir ses amis ou pas. Elle est formulée comme suit :

$$\mathcal{F} = \frac{\mathcal{T}_n}{\mathfrak{D}_n} \qquad (4)$$

Où $\mathcal{T}_n$ exprime le nombre total de noms dans un carnet d'adresses d'un utilisateur et $\mathfrak{D}_n$ exprime le nombre de prénoms distincts. La valeur $\mathcal{F}$ des profils légitimes tend vers 1, alors que celle des profils des polluposteurs peut atteindre 2 et plus.

$$M = nombre\ de\ messages\ envoyés\ par\ un\ profil \qquad (5)$$

Selon leurs expériences, les robots envoient moins de 20 messages pendant leur durée de vie pour éviter de se faire repérer, par contre les profils légitimes peuvent atteindre des centaines de messages.

$$FN = le\ nombre\ d'amis\ dont\ dispose\ un\ profil \qquad (6)$$

L'idée est que les profils légitimes ont de fortes chances de disposer de plus de contacts que le profil d'un polluposteur.

L'ensemble des caractéristiques mises en évidence va servir au développement d'un programme de détection des profils de polluposteurs sur Facebook et Twitter. En revanche, certaines caractéristiques sont légèrement modifiées, afin de les adapter aux plateformes des deux systèmes qui sont différentes. Le classificateur utilisé est l'algorithme des forêts d'arbres décisionnels (Random Forest Algorithm), implémenté dans le logiciel Weka [2]. Le classificateur est alors entrainé sur un corpus de données locales (Réseau de Los Angeles) composées de 1000 profils (173 pourriels et 827 légitimes). Par validation croisée de 10 partitions d'entrainement et de test sur le réseau Facebook, seulement 2% de faux positif et 1% de faux négatifs ont été enregistrés. Pour 500 profils sur Twitter, seulement 2,5% de faux positif et 3% de faux négatifs ont été mal classés. Enfin, pour évaluer la méthode de façon réelle, le programme va être proposé à Twitter pour un essai en ligne d'une durée de 3 mois (mars à juin 2010). Pour 135834 profiles analysés, 15932 ont été reportés comme polluposteurs avec seulement 75 faux positifs d'après Twitter.

D'après les expériences de Stringhini et al. [2], lorsque le corpus est formé des données des réseaux géographiques (par exemple, le réseau de Los Angeles et celui de New York), les

performances du classificateur se dégradent. L'hypothèse en est que les « honeynet account » utilisés par les différentes institutions n'ont pas une façon identique d'attraction. Cela veut dire qu'a priori, aucun compte expérimental ne dispose d'une garantie qu'il est en procession d'un échantillon hétérogène des types de pourriels émis partout dans le monde. À ce sujet, la comparaison d'études d'origines diverses, réalisées à la même période a révélé des différences notables.

2.4 Construction d'un modèle hybride universel

De Wang et al. [36] proposent un environnement ouvert de détection de pourriel pour les réseaux sociaux. Cette approche de plusieurs volets évite à chaque réseau de développer une solution spécifique à sa plateforme, et se prête donc pour une classification de type associative. Leur modèle définit alors trois objets standards des réseaux sociaux: le profil, le modèle de message et le modèle de page web. Ensuite, les attributs les plus communs relatifs aux profils des utilisateurs des réseaux sociaux d'une part, et autres attributs relatifs aux modèles de messages d'autre part, seront sélectionnés. Les attributs de messages sont traités sur la base des informations contenues dans l'en-tête du message, à savoir : les champs de l'expéditeur et du récepteur (signé par, envoyé par), l'horaire de l'activité, le sujet et le contenu. À cette liste s'ajoutent certaines informations sur l'adresse IP de l'expéditeur. Et pour ce qui concerne l'authentification des pages web, on retient les attributs communs se trouvant dans l'en-tête de la session HTTP, à savoir : la connexion, la taille du message, le serveur et le statut. Leur approche va également s'appuyer sur la définition de listes noires, des listes de réputation d'adresses IP des émetteurs, et le calcul de similarité pour classifier les messages.

2.4.1 Combinaison de règles

Sachant qu'un seul critère ne suffit pas pour atteindre un niveau d'efficacité de filtrage de pourriels multimédias, plusieurs règles seront combinées. Ainsi, pour procéder à la classification des messages, les auteurs associent à chaque modèle un classificateur, et combine 4 niveaux de règles (ou critères) de contrôle: la règle ET, la règle OU, le vote majoritaire et le modèle bayésien.

- La règle ET classifie un objet comme pourriel si tout classificateur associé à chaque modèle l'a classifié comme pourriel.

- La règle OU classifie chaque objet comme pourriel si au moins un classificateur associé à chaque modèle l'a détecté comme pourriel,
- Le vote majoritaire identifie un objet comme pourriel si la majorité des classificateurs l'a détecté comme pourriel,
- Le modèle bayésien va classifier le message selon l'équation (7): soit l'objet x associé à la classe $y \in Y = \{ham, spam\}$. Si l'on suppose une variable cachée Z d'un évènement à classifier (ex. éducation, santé, etc.), et $p(y|x)$ la probabilité d'appartenir à la classe y sachant x, la valeur $p(y|x)$ est donnée par la probabilité marginale de y et de la probabilité jointe de la variable Z et de y:

$$p(y|x) = \sum_i p(y|Z_i, x)p(Z_i|x), \qquad (7)$$

Conjointement à cette approche, De Wang et al. [36] ont utilisé un apprentissage incrémental pour éviter la perte d'efficacité dans le temps d'un filtre dont l'apprentissage n'est pas mise à jour, afin d'être supportable dans certaines limites.

Bien que ces approches permettent de réduire les flots de messages pour améliorer la classification des messages (ham, pourriel), leurs inconvénients majeurs résident dans les coûts trop élevés pour entraîner les modèles afin de détecter de nouveaux pourriels. De plus, les différentes caractéristiques des réseaux sociaux, par exemple, la longueur des messages dans Facebook par rapport à celle dans Twitter (140 octets ou 165 caractères maximum), peuvent réduire les avantages de la mise en collaboration des corpus de pourriels à travers différentes plateformes.

Dans la même perspective, Yardi et al.[11] effectuent une étude sur un échantillon de polluposteurs prélevé sur le réseau Twitter. Cette étude révèle que le comportement des polluposteurs est bien différent des utilisateurs légitimes, surtout par le mode d'affichage des tweets, les attitudes des adeptes et les fans de leurs pages. Stringhini et al.[11] procèdent par la création d'un certain nombre de profils d'attraction sur Facebook, Twitter et Myspace. Ils identifient cinq communes caractéristiques potentielles pour la détection de polluposteurs, entre « Followee-To-Follower », le ratio des URL, la similarité des messages, les messages envoyés, et le nombre d'amis. À l'analyse de ces deux approches expérimentales, elles sont bien

convaincantes pour la détection de polluposteurs, mais elles manquent des détails de spécifications et d'évaluation de prototype.

Wang [37] va poursuivre sur la même lancée en proposant un algorithme de classification basé sur Naïve Bayes. Il effectue des tests de discrimination des comportements suspects des normaux à partir des données de Twitter. Il parvient aux résultats de 89% de précision (vrais positifs, faux positifs). Gao et al. [38] ont exploré les messages sur les murs Facebook, en analysant les propriétés temporelles des profils, les caractéristiques des URL, le ratio des postes, et autres caractéristiques des comptes malveillants. Ils soulignent également que les diverses campagnes de pourriels se font par annonce de produits dans un laps de temps. Ils précisent que le pourriel sur Facebook se présente souvent sous forme d'image attrayante, ou de message très soigné. D'après leur approche, les résultats par campagne sont meilleurs par rapport à un examen individuel de pourriels. Ces deux autres approches manquent également assez de précision dans la discrimination des classes.

Facebook [39] donne un aperçu de leur système immunitaire, un système de défense contre l'hameçonnage, la fraude et le pourriel. Le système est composé de trois parties : la première partie est composée de classificateurs, et la seconde composée d'une implémentation logicielle dérivée de «Feature Extraction Language» ou (FXL). Cette seconde partie extrait et prépare les caractéristiques pour la classification. La troisième partie est un moteur d'analyse des politiques qui prend des mesures concernant des messages ayant une connotation suspecte. Ce service est de très haut niveau, car il prend en compte quelques détails importants des activités indésirables sur le site Facebook à savoir les faux profils, le harcèlement, les comptes compromis, les logiciels malveillants et le pourriel.

Contrairement à la recherche qui se concentre sur la détection dynamique du pourriel sur la base de l'activité de l'utilisateur, Irani et al. [40] montrent que les caractéristiques statiques associées à l'authentification de l'utilisateur sur MySpace sont suffisantes pour entraîner un classificateur effectif contre le pourriel social. Ils précisent que les algorithmes d'arbres de décision C4.5 offrent de meilleures performances que Naïve Bayes dans ce cas. Comme dans tous travaux de recherches, cet examen n'est qu'un travail partiel et ne concerne que les informations de profils recueillies par attraction des Honeypots. Ainsi, un apport complémentaire des données privées

collectées sur les utilisateurs à savoir les attributs de furetage, les adresses IP et l'emplacement géographique viendraient améliorer sensiblement les performances du classificateur. Benevenuto et al. [41] procèdent à la mise en place d'un corpus de données collectées sur Twitter. En s'appuyant sur 62 attributs extraits, ils analysent les contenus des tweets pour faire un parallèle entre les comportements des utilisateurs légitimes et des polluposteurs. Ces attributs seront ensuite entrainés dans différents modèles pour la classification des utilisateurs.

Bosma et al. [42] se servent des rapports générés suite aux signalements des utilisateurs sur le pourriel, comme un outil pour la construction d'un environnement de détection non supervisé de pourriels des réseaux sociaux. Leur approche consiste à compter le nombre de rapports émis contre un utilisateur suspect, et ensuite pondérer ces rapports basés sur la réputation de l'utilisateur. La détermination de la réputation et la fiabilité des utilisateurs dans les réseaux sociaux ont été bien étudiées [43] et semblent être des facteurs prometteurs à la classification du pourriel social. Leur modèle utilise un classificateur bayésien et les liens des messages avec un contenu similaire, sans tenir compte d'autres caractéristiques. Cette étude est l'une des rares qui a servi d'un modèle de test sur les données non publiques, y compris les messages privés, les rapports de signalement de pourriel et des profils d'utilisateurs d'un grand site de réseautage social néerlandais.

2.5 Analyse des réseaux sociaux RenRen et Sino Wei Bo

En 2012, Zhu et al. [44] estiment que les activités observées sur des profils des réseaux sociaux sont très éparses. Ainsi, la définition des modèles sur la base des activités du passé pourrait conduire à des échecs de classification des profils. C'est ainsi qu'il va proposer un modèle de classification de pourriel basé sur la factorisation matricielle des activités des utilisateurs. À cet effet, une matrice d'activité A sera factorisée en deux matrices latentes U et V respectivement la matrice des utilisateurs et celle de leurs activités.

$$\min f(U,V) = \sum_{a_{ij} \in A} I_{ij}\left(a_{ij} - U_i V_j\right)^2 + \frac{\lambda f}{2}\left(\|U\|^2{}_F + \|V\|^2{}_F\right) \qquad (8)$$

Où I_{ij} est une fonction indicatrice qui est égale à 1 si l'élément correspondant dans la matrice existe et 0 sinon.

$$I_{ij} = \begin{cases} 1 \text{ si } U_i, V_j \text{ existe,} \\ 0 \text{ sinon} \end{cases} \quad (9)$$

Le premier terme de la relation (8) est appelé le coût d'approximation. Les trois termes au carré appelés norme de Frobenius, ou norme euclidienne, représentent la partie de régularisation pour éviter le surajustement de la factorisation. Le coefficient de régularisation λf est le compromis entre la perte d'approximation et les termes de régularisation. Cette méthode permet d'induire de façon collaborative un ensemble succinct d'attributs latents. À cet effet, ils vont exploiter les données du réseau RenRen (www.renren.com), pour construire une base de données composée de plus de 1000 attributs les plus communs qui renseignent sur des liens de parenté que chaque individu entretient avec les différents cercles d'amis existants sur les réseaux. Cependant, ces deux approches sont basées sur une grande quantité de caractéristiques sélectionnées qui pourrait coûter en termes de puissance de calcul, puis va engendrer une complexité trop élevée dans l'entraînement des modèles.

Suite à une étude effectue par Liu et al. [45] sur le réseau social Sina Weibo, on parvient à dégager trois caractéristiques inhérentes aux habitudes de vie des polluposteurs. Il s'agit de leur agressivité dans la publicité, la duplication des contenus des affiches et le comportement agressif des adeptes. Ce qui permet de mettre en évidence trois ensembles de caractéristiques. Dans leur démarche, l'équipe adopte une approche particulière, basée sur la combinaison de modèles pour construire des groupes de dispositifs de classification. Les différents groupes de classificateurs formés sont ensuite testés sur chacune des trois caractéristiques générées, et fonctionnent conjointement comme un classificateur anti-spam unique. L'objectif visé par cette combinaison de classificateurs est d'améliorer les performances du filtre anti-spam. Cependant, la précision de classification obtenue est seulement de 82,06%. L'hypothèse plausible de ce résultat en dessous des performances escomptées serait le faible nombre d'attributs (8 au maximum) que contient chaque sous-ensemble de chaque ensemble de caractéristiques.

Xianghan et al. [11] ont réutilisé les concepts de Liu et al. [45], avec des idées complémentaires, et obtient des résultats les plus performants de l'état de l'art, avec un taux de précision de 99%. Pour parvenir à cette performance, d'abord l'équipe propose un modèle de classification de type SVM (les machines à vecteurs de support), en portant le nombre de caractéristiques de 8 à 18. Ensuite, la pertinence de chaque caractéristique sélectionnée est évaluée à travers l'application

Java, Weka. L'usage combiné de ces caractéristiques a été un facteur important pour atteindre les performances escomptées.

2.6 Avantages des modèles de classification hybrides

On remarque à travers la littérature que la combinaison de plusieurs modèles de classification améliore nettement les performances de filtrage de pourriel social. En effet dans un processus de classification, chaque modèle dispose de ses forces et de ses faiblesses pour classifier de nouvelles données. En plus, compte tenu de la diversité des formes d'entrées (valeurs booléennes, discrètes, continues), un modèle peut bien classifier un exemple de données et se tromper significativement lors de la classification d'un autre. Ainsi, la combinaison des modèles permet de pallier les erreurs de classification des uns et des autres vis-à-vis de cette diversité. Bien entendu, si les modèles font les mêmes erreurs lors de la classification, alors le système perd toute sa fiabilité. Cette combinaison complémentaire des modèles, qualifiée de modèles hybrides, permet de combiner les prédictions de façon quelconque en vue d'améliorer les résultats de classification [46].

Lors de la construction des modèles hybrides, la plus simple façon de procéder est d'entrainer les classificateurs de manière indépendante, puis de les faire voter. L'ensemble qui va former le comité du modèle hybride sera la classe qui revient le plus souvent. Des tests ont prouvé que la performance moyenne d'un modèle hybride est supérieure à la moyenne des performances de l'ensemble des modèles composant celui hybride [47].

D'autres méthodes qualifiées de « stacking » apportent également une amélioration aux performances des modèles de classification des messages électroniques. Le processus consiste à entrainer plusieurs modèles de façon séquentielle de telle sorte que le nième modèle dans son apprentissage tienne compte de l'erreur de généralisation de ceux qui le précèdent [48]. À défaut de filtres spécifiques pour détecter le pourriel social, l'approche de combinaison de modèles semble pour l'instant une solution contre cette forme de menace. Elle permet à la communauté des chercheurs d'atteindre des taux d'efficacité de plus en plus élevés.

2.7 Autres méthodes de filtrage de pourriels

2.7.1 Système immunitaire artificiel

Afin de lutter efficacement contre le pourriel, Terri et al. [49] ont fait le parallèle entre le pourriel et les agents pathogènes qui sont combattus par le système immunitaire humain. À partir d'une base de données de gènes, le système génère aléatoirement des anticorps et crée le lymphocyte correspondant afin d'être à même de détecter tout corps étranger entré dans le système. Ensuite, on procède à l'entrainement des lymphocytes sur un corpus de données composées de messages légitimes et de pourriels. Les lymphocytes qui se révèlent inefficaces sont supprimés, et on attribue une date d'expiration à chacun d'eux en fonction de ses performances. À chaque message filtré par un lymphocyte, sa date d'expiration est incrémentée. Dès que les lymphocytes arrivent à expiration, ils meurent et sont remplacés par d'autres qui vont redémarrent un nouveau cycle.

2.7.2 Signature des messages

Un système antispam est efficace dans la lutte des envois anormaux que lorsqu'il a une vision globale des flots de messages qui transitent sur le réseau. C'est pourquoi il doit être positionné au niveau du service d'envoi de message. Cependant, les polluposteurs rendent les envois massifs plus difficiles à détecter. Ils insèrent ou suppriment des séquences aléatoires dans les courriels afin que chaque campagne de pourriel soit unique. Ne pouvant pas se baser uniquement sur un checksum pour identifier les courriels identiques, les techniques de détection doivent se baser sur une autre forme de signature moins sensible à l'insertion et à la suppression de termes. C'est le cas de l'algorithme I-Match [50]. L'algorithme I-Match s'appuie sur l'ensemble des termes uniques du courriel et sur un lexique préalablement établi pour produire la signature du message. Cette signature est alors associée à un unique cluster, ce qui permet d'en déduire la classe du message. C'est une technique qui détecte des envois massif, mais sensible à des modifications aléatoires du corps des messages.

2.7.3 Les contributions de Facebook

Des travaux plus récents tentent d'améliorer l'identification et le filtrage des pourriels dans les réseaux sociaux à partir des méthodes probabilistes [51]. Le concept consiste à utiliser la relation

sociale établie entre l'émetteur et le récepteur afin de déterminer leur proximité, et en déduire la valeur de confiance. Cette valeur est un facteur déterminant dans le calcul de probabilités, qui servira d'indicateur pour la classification.

En industrie, Facebook va proposer l'algorithme de EdgeRank [52] qui attribue à chaque affiche, un score en fonction de certaines caractéristiques, telles que le nombre de « j'aime » ou « likes », le nombre de commentaires et le nombre de réactions postés par un utilisateur. Par conséquent, pour un score d'EdgeRank élevé, moins la chance d'être un polluposteur. L'inconvénient de cette approche est que les polluposteurs pourraient aussi se constituer en un réseau et poster des commentaires entre eux afin d'obtenir un score élevé d'EdgeRank.

2.7.4 Les modèles heuristiques

L'analyse heuristique se base sur des critères permettant d'examiner l'en-tête et le corps d'un message afin de déceler des caractéristiques qui discriminent un message légitime du pourriel. Habituellement, un moteur d'analyse heuristique s'appuie sur plusieurs centaines de critères représentés sous forme d'expressions relationnelles ou régulières. Une expression régulière est un motif que l'on peut appliquer à une chaîne afin de voir si ladite chaîne correspond au motif.

L'utilisation des expressions relationnelles permet de trouver les variations de mots sensibles, ce qui augmente les chances de détecter du pourriel. Par exemple, le moteur heuristique analyse la proportion de code HTML, d'images, des références liées à la pornographie, des termes liés à l'acquisition facile d'argent. Il vérifie également si le sujet du message est vide ou non, ou si l'identificateur du message contient des signes comme « $ » souvent utilisé par robots polluposteurs. Suite à cette analyse, il en résulte un score comparé à un seuil fixé par l'administrateur réseau afin de déduire la classe du message. Cette technique souffre deux inconvénients majeurs : une mise à jour constante des critères pour faire face aux perpétuelles astuces des polluposteurs, ensuite l'analyse est fastidieuse face à un grand volume de données.

2.7.5 Réputation de l'émetteur sur les réseaux sociaux

Selon les spécialistes de la sécurité informatique, la meilleure technique pour trier ses messages reste celle des listes blanches. Elle consiste à établir manuellement ou semi automatiquement une liste de contact en qui l'on a confiance et dont on sait que les messages sont légitimes. En

pratique, les correspondances envoyées par les personnes de confiance sont classées dans la boîte de réception et le reste des messages sont considérés comme des pourriels. Mais cette technique souffre de deux problèmes majeurs. D'une part, l'utilisateur doit lui-même assurer la maintenance de la liste blanche ce qui peut dans certains cas, représenter une quantité énorme de travail. D'autre part, elle se base seulement sur des contacts connus.

C'est ainsi que les correspondances en provenance de personnes ou d'organismes non encore répertoriés dans la liste blanche seront confondues aux pourriels. Pour éviter de tels scénarios, on utilise les réseaux de réputation [53], qui consiste à noter (entre 0 et 10) chaque émetteur en fonction de son réseau de connaissances. Cela permet d'attribuer à chaque utilisateur un indice de confiance qui est un indicateur pour mieux classifier les messages. D'autre part, la confiance attribuée aux correspondances envoyées par des émetteurs inconnus pourra être inférée pour peu que les émetteurs soient connus par une ou plusieurs personnes du réseau social de l'utilisateur.

2.8 Sommaire

On remarque qu'il existe de nombreuses mesures techniques pour filtrer le pourriel. Ces mesures peuvent être regroupées en deux grandes catégories : les modèles basés sur le filtrage de contenu et les modèles basés sur l'identité. Dans le premier modèle dit statistique, une série de méthodes d'apprentissage automatique et heuristique sont mises en œuvre pour analyser le contenu selon les mots-clés et selon des motifs qui sont potentiellement liés aux pourriels [12]. On procède également à la comparaison de signature des messages associée aux envois massifs et anormaux, l'analyse des fichiers logs et des signalements des objets collectivement rapportés par les utilisateurs, afin de classifier les messages électroniques. Dans le second modèle (dynamique) basé sur l'identité, l'approche la plus couramment utilisée consiste à maintenir à jour, une liste blanche et une liste noire d'adresses de courriels qui devraient être filtrées ou non par un mécanisme antispam [54].

Les deux approches en haut ont chacune ses faibles : la convergence des modèles dynamique est lente avec la taille explosive de données, alors que les modèles statistiques ont besoin de relancer constamment la phase d'apprentissage afin de s'adapter aux nouvelles données. Dans le chapitre suivant, nous exposons les différentes méthodes de classification automatique de pourriels, suivit d'une analyse des problèmes de filtres anti-spam.

Chapitre 3- Méthodes d'apprentissage et détection de pourriel

3.1 Introduction

Dans ce chapitre, nous présenterons quelques algorithmes de filtrage de pourriels basés sur l'apprentissage automatique et domaines connexes. Ces algorithmes sont issus des travaux menés par la communauté de recherche scientifique, et avec le concours des spécialistes de la sécurité informatiques en milieu industriel.

3.2 Les méthodes non supervisées

3.2.1 Principe

Les méthodes de classification non supervisées peuvent être réparties en deux grands groupes : les méthodes de partitionnement et les méthodes hiérarchiques. En général, toutes ces méthodes s'appuient sur la notion de similarité intraclasse et de dissimilarité interclasse, puis sur une représentation vectorielle des objets. Contrairement à l'apprentissage supervisé qui fournit les valeurs réelles de sorties pour les données d'entrées, l'apprentissage non supervisé ne fournit pas ces sorties. Soit n le nombre de données, et $X^{(i)}$ un vecteur de $\mathcal{D}$ dimensions (caractéristiques), et les données d'apprentissage représentées par: $\mathbb{D}_n = \{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$. Cette approche est le plus souvent utilisée pour des applications de regroupement (clustering), et le but est d'utiliser l'ensemble $\mathbb{D}_n$ pour construire des groupes homogènes, représentants des classes de données. Au cours de ce processus, il revient aux algorithmes de découvrir par eux-mêmes les structures plus ou moins cachées des données afin de former des groupes d'individus dont les caractéristiques sont communes [32]. L'apprentissage non supervisé est appliqué dans plusieurs secteurs d'activités tels que le traitement d'images, la reconnaissance de la voix humaine, la recherche documentaire, la recherche pathologique (médical), etc.

3.2.2 Les méthodes de partitionnement

Elles consistent à grouper des données selon leurs degrés de similarité. L'algorithme le plus connu de cette catégorie est celui des K-moyennes. Il permet de partitionner automatiquement un ensemble de données en K groupes ou « clusters ». Son principe consiste à classifier les points

sur la base du critère du plus proche voisin. Il s'agit d'abord de calculer les K points qui représentent les centres de gravité des groupes à former, ensuite affecter les autres points aux centres les plus proches. L'affectation est basée sur le calcul de distance entre les points et les centres de gravité. Plusieurs distances peuvent être définies telles que la distance euclidienne ou la distance de Manhattan. La suite du processus consiste à raffiner des regroupements de manière itérative. À chaque itération, on recalcule les centres de gravité des groupes et des points sont réaffectés aux nouveaux groupes formés. L'algorithme des K-moyennes se résume comme suit :

Soit un ensemble de n données $\mathbb{D}_n = \{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$, et chaque donnée $x^{(i)}$possède $\mathcal{D}$ dimensions réelles $\left({x_1}^{(i)}, {x_2}^{(i)}, \dots, {x_{\mathcal{D}}}^{(i)}\right)$. Supposons que les données appartiennent à K différents groupes $G = \{G_1, G_2, \dots G_K\}$ $(K \leq n)$. Pour chaque donnée $x^{(i)}$, on associe une variable indicatrice $r_{ik} \in \{0,1\}$, qui prendra sa valeur comme suit :

$$r_{ik} = \begin{cases} 1 & si \quad x^{(i)} \in G_k \\ 0 & si \quad x^{(i)} \notin G_k \end{cases} \qquad (10)$$

On peut définir une fonction objective (somme des carrés résiduels: *SCR*) à minimiser pour réaliser un groupement optimal selon la relation (11) :

$$\text{SCR} = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \left\| x^{(i)} - \mu^{(k)} \right\|^2 \qquad (11)$$

Où $\mu^{(k)}$ est la moyenne des données du groupe G_k, il revient à trouver les r_{ik} et les $\mu^{(k)}$ pour minimiser la fonction SCR.

3.2.3 Les méthodes hiérarchiques

Elles se présentent sous deux formes: la méthode descendante et la méthode ascendante.

3.2.3.1 Méthode hiérarchique descendante

À l'inverse d'une fusion progressive, la classification descendante est une analyse d'un ensemble de données qui vise à obtenir des singletons par décomposition des éléments. Généralement, cette méthode procède par dichotomie. Par exemple, pour un ensemble de groupe G de données, on le divise en 2 classes G_0 et G_1, ensuite les classes G_0 et G_1 sont divisées en 2, respectivement en G_{00}, G_{01} et G_{10}, G_{11} et ainsi de suite jusqu'à obtenir un singleton.

3.2.3.2 Méthode hiérarchique ascendante

La classification ascendante dite « bottom-up » est la plus connue des méthodes de classification automatique. Elle consiste à définir une matrice de similarité à partir de l'échelon le plus fin, qui sert à consolider progressivement et effectuer la synthèse des données. C'est l'exemple de classifications opérées par un arbre hiérarchique ou dendrogramme. L'obtention d'une partition se fait par coupure de l'arbre à un seuil donné. En général, cette méthode conduit à une catégorisation polythétique (classification des données ayant un grand nombre d'attributs en commun). La méthode hiérarchique ascendante se présente sous deux autres formes : la méthode du lien simple, et la méthode du diamètre ou du lien complet.

A. Méthode du simple lien (Single Link)

La méthode du simple lien consiste à fusionner à chaque itération, la paire d'objets la plus proche (les points séparés par la plus petite distance) entre chacun des deux groupes. Cette méthode est adaptée aux classes de longue étendue, des chaînes, et des ellipsoïdes. La distance entre les points est exprimée selon l'équation (12) :

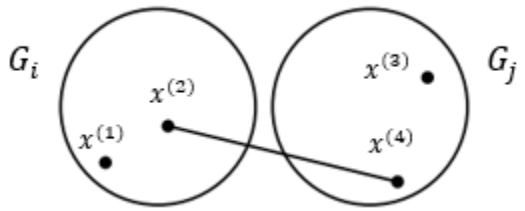

Figure 4: *Groupement par le simple lien*

$$dist_{min}(G_i, G_j) = \min_{\substack{1 \le i \le n \\ 1 \le j \le m}} \left\| x^{(i)} - x^{(j)} \right\| ;\ x^{(i)} \in G_i\ , x^{(j)} \in\ G_j \qquad (12)$$

Où n *et* m *désignent respectivement les cordonnées de* G_i *et de* G_j

B. Méthode du diamètre ou lien complet (Complete Link)

La méthode du diamètre consiste à unir la paire d'objets la moins similaire (les points les plus éloignés) entre chacun des deux groupes. Sachant que toutes les entrées dans une classe sont liées à une autre avec la similarité minimum, ainsi on peut dans certains cas obtenir de petites classes fortement liées.

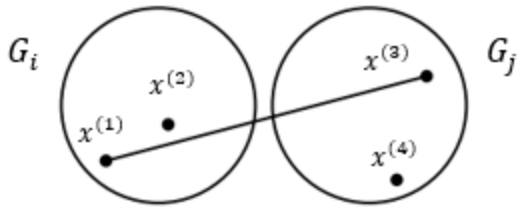

Figure 5: *Groupement par le lien complet*

La distance entre les points est exprimée selon l'équation (13) :

$$dist_{max}(G_i, G_j) = \max_{\substack{1\leq i\leq n \\ 1\leq j\leq m}} \left\| x^{(i)} - x^{(j)} \right\| ;\ x^{(i)} \in G_i\ , x^{(j)} \in\ G_j \qquad (13)$$

3.3 Les méthodes supervisées

3.3.1 Principe

En ce qui concerne les algorithmes d'apprentissage supervisé, on fournit à l'algorithme des données d'entrainement $\mathbb{D}_n = \left\{\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)\right\}$, $x^{(i)}$constitue les données d'entrée et $y^{(i)}$ la cible. On dispose donc d'une sélection initiale de données dont la classification est connue. Ces données sont supposées indépendantes et identiquement distribuées, et serviront à l'apprentissage de l'algorithme. Enfin, la classification se fait par l'algorithme selon le modèle appris. Il existe plusieurs algorithmes d'apprentissage supervisé, parmi lesquels les arbres décisions, les réseaux de neurones, Naïve Bayes, SVM, etc.

3.3.2 Arbres de décision (datamining)

L'autre grande technique de classification est celle des algorithmes à base d'arbres de décision tels que: ID3, C4.5, SLIQ ou CART. Les messages préalablement étiquetés manuellement sont découpés selon leurs attributs pertinents et insérés dans la base de données. Ensuite l'algorithme d'inférence va générer un arbre de décision à partir duquel l'ensemble de règles sera extrait, puis directement intégré aux systèmes de filtrage anti-spam. La précision de la classification dépendra de la qualité de la préparation des données et des caractéristiques qui serviront à l'entraînement.

3.3.3 La méthode par réseaux de neurones

Les réseaux de neurones forment une structure composée d'une succession de couches de nœuds, qui définissent une fonction de transformation non linéaire des vecteurs d'entrées aux sorties. Ces vecteurs représentent les caractéristiques des messages pouvant servir, après apprentissage, à

reproduire une forme de raisonnement humain. Ces caractéristiques permettent d'ajuster les coefficients synaptiques du réseau de neurones durant la phase d'apprentissage. L'apprentissage se fait à partir d'une collection de messages (ham, pourriel) préalablement triés manuellement, et peut éventuellement être incrémental afin d'être le mieux adapté possible aux nouvelles formes de pourriels qui peuvent surgir. Une fois l'apprentissage réalisé, la structure fonctionne comme un filtre antispam classique très efficace en fonction des infrastructures. Le nombre de couches déployées et la disposition des neurones dans le réseau influencent le résultat de classification.
Comparés aux autres méthodes de classification supervisée, les réseaux de neurones sont rapides et permettent de régler le taux de mauvaise classification (Faux Positif) en ajustant le seuil de sensibilité, mais nécessitent également un long entraînement et régulier pour faire face aux nouvelles données de pourriels.

3.3.4 Les méthodes bayésiennes

Le classificateur de Bayes est basé sur le Théorème de Bayes, et s'appuie sur les probabilités jointes des termes et des catégories (classes) pour estimer la probabilité d'une catégorie sachant un message à classifier. Le problème général de classification peut-être posé comme le choix de la meilleure hypothèse associée à un objet, après observation d'un ensemble d'exemples d'apprentissage. Pour définir la meilleure hypothèse, on considère celle qui est la plus probable (celle dont la probabilité d'erreur est la plus faible).

Considérons $p(y),\ y \in Y = \{ham, spam\}$ la probabilité associée à une hypothèse de classification avant que l'objet à classifier (soit un message) ne soit observé. Il s'agit de la probabilité à priori de l'hypothèse. Après avoir observé un message reçu $m \in \mathcal{M}$, la probabilité a posteriori est évaluée selon la règle de Bayes pour chaque classe, selon l'équation (14):

$$p(y|m) = \frac{p(\mathcal{M} = m|Y = y).p(Y = y)}{p(\mathcal{M} = m)} \qquad (14)$$

Où $\mathcal{M} = \{m^{(1)}, m^{(2)}, \dots, m^{(n)}\}$ représente une séquence de messages dans leur format original.
Pour comparer des probabilités a posteriori des classes candidates, la règle de décision optimale est basée sur le choix de la classe qui maximise cette probabilité:

$$\hat{y} = \arg\max_{y \in Y} \frac{p(\mathcal{M}|y).p(y)}{p(\mathcal{M})} \qquad (15)$$
$$\propto \arg\max_{y \in Y} p(\mathcal{M}|y).p(y)$$

Le dénominateur $p(\mathcal{M})$ est une valeur constante pour toutes les classes, donc négligeable. L'estimation des probabilités a priori des classes $p(y)$ est moins complexe par rapport à celle des probabilités a posteriori. Dans le premier cas, un nombre réduit d'hypothèses permet d'obtenir une précision suffisante. Par contre, l'estimation des probabilités a posteriori $p(\mathcal{M}\ |y)$ est un problème plus complexe à résoudre. Cette complexité est liée au nombre parfois explosif des paramètres devant être estimés.

Pour exprimer l'indépendance des attributs, on adopte une alternative de solution à caractère naïve, d'où le nom Naïve Bayes. L'hypothèse d'indépendance signifie que la probabilité conditionnelle d'un terme sachant une classe est supposée indépendante des probabilités conditionnelles des autres termes sachant la même classe. Autrement dit, les probabilités associées à chaque terme peuvent être estimées individuellement. Les modèles de classification de naïve Bayes tirent leur efficacité de cette hypothèse. Plusieurs autres classificateurs dérivés de ce modèle sont utilisés pour la génération de documents textuels et le filtrage du pourriel.

3.3.4.1 Modèle évènementiel

Dans ce modèle on considère que l'objet est généré à partir d'un mélange de n distributions de probabilité, n étant le nombre de classes. Dans un premier temps, le processus consiste à faire un choix aléatoire d'une classe de probabilité à priori $p(y)$, puis on génère ensuite un message selon la distribution des termes spécifique à la classe choisie. Compte tenu de ce principe de fonctionnement, le classificateur de Naïve Bayes est qualifié de classificateur génératif.

Soit $\mathbb{D}_n = \{(\mathfrak{m}^{(1)}, y^{(1)}), \dots, (\mathfrak{m}^{(n)}, y^{(n)})\}$, $(\mathfrak{m}^{(n)}, y^{(n)}) \in \mathcal{M} \times Y$ *un ensemble d'exemple à classifier, et* un échantillon de n exemples de $\mathcal{M}$ utilisés pour l'apprentissage. $\mathcal{V} = \{t^{(1)}, t^{(2)}, \dots, t^{|\mathcal{V}|}\}$ désigne le vocabulaire constitué de tous les termes présents dans tous les objets, et $\mathfrak{m} = (m_1, m_2, \dots, m_{\mathcal{L}})$ un vecteur composé de $\mathcal{L}$ mots d'un message. $\mathbb{I}_c(, \mathfrak{m}^{(i)}, c)$ est la fonction indicatrice de la classe d'appartenance de l'exemple $\mathfrak{m}^{(i)}$.

$$\mathbb{I}_c(\mathfrak{m}^{(i)}, c) = \begin{cases} 1 \text{ si } \mathcal{y}^{(i)} = c, \\ \\ 0 \; sinon \end{cases} \qquad (16)$$

Les expressions $I_p(t, \mathfrak{m})$ et $I_n(t, \mathfrak{m})$ sont des fonctions indiquant respectivement, la présence et le nombre d'occurrences du terme t dans le message $\mathfrak{m}$.

3.3.4.2 Modèle multivarié de Bernoulli

Pour générer un message dans ce modèle, on effectue plusieurs tirages au sort. Chaque terme du vocabulaire résulte d'un tirage, pour décider si le terme en question est présent dans le message. Chaque tirage suit une loi de Bernoulli de probabilité $p(t|\mathcal{y})$. Selon ce modèle, on estime la probabilité, dans chaque classe du message généré par la relation (17):

$$p(\mathcal{M}|\mathcal{y}) = \prod_{t \in \mathcal{V}} \left(p(t|\mathcal{y})^{\mathbb{I}_p(t,\mathcal{M})} (1 - p(t|\mathcal{y}))^{(1-\mathbb{I}_p(t,\mathcal{M}))} \right) \qquad (17)$$

La distribution des termes dans les classes $p(t|\mathcal{y})$ est habituellement évaluée avec le correcteur de Laplace selon l'équation (18) :

$$\hat{p}(t|\mathcal{y}) = \frac{1 + \sum_{\mathfrak{m} \in \mathbb{D}_n} \mathbb{I}_p(t, \mathfrak{m}) \mathbb{I}_c(\mathfrak{m}, \mathcal{y})}{2 + \sum_{\mathfrak{m} \in \mathbb{D}_n} \mathbb{I}_c(\mathfrak{m}, \mathcal{y})}, t \in \mathcal{V}, \mathcal{y} \in Y \qquad (18)$$

3.3.4.3 Modèle Multinomial

Dans ce modèle, la génération du message est liée par un nombre limité t, la longueur du message, des tirages aléatoires avec remise, des termes d'un vocabulaire, chaque terme pouvant apparaitre plus d'une fois. La probabilité du message dans chaque classe est estimée selon l'équation (19) :

$$p(\mathcal{M}|\mathcal{y}) = p(|\mathcal{M}| = \mathcal{L}|\mathcal{y}). \mathcal{L}! \prod_{t \in \mathcal{V}} \frac{p(t|\mathcal{y})^{\mathbb{I}_n(t,\mathfrak{m})}}{\mathbb{I}_n(t, \mathfrak{m})!} \qquad (19)$$

Le correcteur de Laplace permet d'estimer les probabilités de chaque terme dans sa classe:

$$\hat{p}(t|\mathcal{y}) = \frac{1 + \sum_{\mathfrak{m} \in \mathbb{D}_n} \mathbb{I}_n(t, \mathfrak{m}) \mathbb{I}_c(\mathfrak{m}, \mathcal{y})}{|\mathcal{V}| + \sum_{E \in \mathcal{V}} \sum_{\mathfrak{m} \in \mathbb{D}_n} \mathbb{I}_n(E, \mathfrak{m}) \mathbb{I}_c(\mathfrak{m}, \mathcal{y})}, t \in \mathcal{V}, \mathcal{y} \in Y \qquad (20)$$

Où E désigne le vocabulaire du sous-ensemble d'exemple, tiré de l'ensemble de départ $\mathbb{D}_n$.

Contrairement au modèle multivarié, le produit multinomial ne tient compte que des termes présents dans les messages c'est-à-dire les t termes au maximum.

3.3.4.4 Modèle Multinomial avec attributs booléens

Le modèle multinomial ne considère que la présence ou absence des termes :

$$p(\mathcal{M}|\mathcal{y}) = p(|\mathcal{M}| = \mathcal{L}|\mathcal{y}).\mathcal{L}! \prod_{t\in\mathcal{V}} p(t|\mathcal{y})^{\mathbb{I}_p(t,\mathfrak{m})} \qquad (21)$$

Et, on peut estimer la probabilité de chaque terme dans sa classe par l'équation (22).

$$\hat{p}(t|\mathcal{y}) = \frac{1 + \sum_{\mathfrak{m}\in\mathbb{D}_n} \mathbb{I}_p(t,\mathfrak{m})\mathbb{I}_c(\mathfrak{m},\mathcal{y})}{|\mathcal{V}| + \sum_{E\in\mathcal{V}}\sum_{\mathfrak{m}\in\mathbb{D}_n} \mathbb{I}_p(E,\mathfrak{m})\mathbb{I}_c(\mathfrak{m},\mathcal{y})}, t \in \mathcal{V}, \mathcal{y} \in Y \qquad (22)$$

3.3.5 La régression logistique

À l'instar des méthodes bayésiennes, l'algorithme de régression logistique a une interprétation probabiliste puisque sa décision de classification est fondée sur la probabilité a posteriori de la classe. D'une manière générale, le modèle de régression logistique considère que le logarithme de la vraisemblance a posteriori peut être décrit comme une fonction linéaire du vecteur de caractéristiques de l'objet (message) à classifier. Ainsi la probabilité a posteriori est représentée par une fonction sigmoïde agissant sur le vecteur de caractéristiques $X = (x_1, x_2, \dots, x_{\mathcal{D}})$.

$$p(\mathcal{y} = spam|X) = \frac{1}{1 + e^{(-f(x))}} \qquad (23)$$

Où $f(x) = w_0 + \sum_{d=1}^{\mathcal{D}}(w_d x_d) = w_0 + w^T x$, $\mathcal{y} \in Y = \{ham, spam\}$

Cela implique que $(w_0, w_1, \dots, w_{\mathcal{D}})$ représente le vecteur qui définit un hyperplan qui sépare les deux classes (ham, pourriel). Cette formulation de l'hypothèse évite la simulation du processus de génération de l'objet comme le ferait Naïve Bayes et permet une évaluation directe de la probabilité a posteriori de la classe. L'apprentissage de cet algorithme consiste alors à déterminer le paramètre *w* qui doit être estimé à partir des données.

3.3.6 La méthode SVM

Les machines à vecteurs de support (SVM) décrivent une approche de classification supervisée basée sur une interprétation géométrique, en s'appuyant sur la notion de marge maximale. La figure 6 explique le principe général des méthodes SVM.

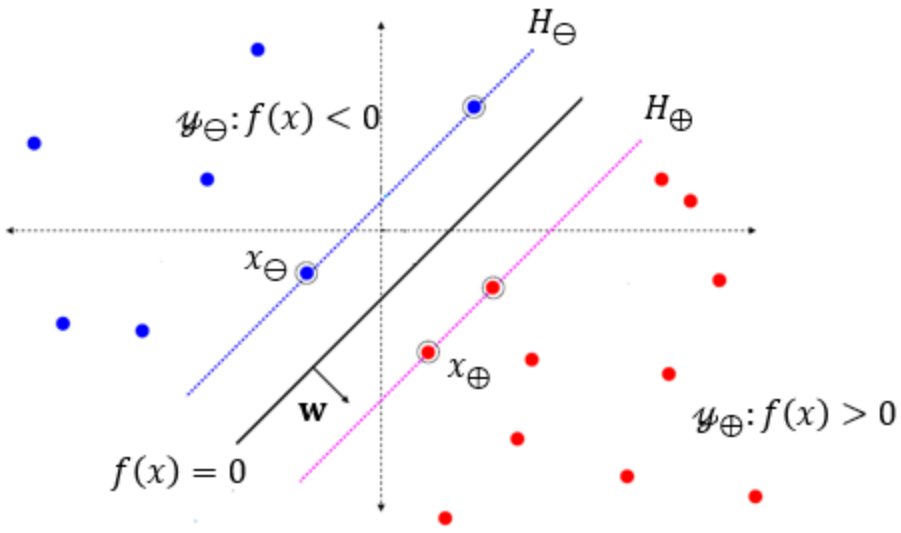

Figure 6: *indique le principe des SVM*

Pour ce qui concerne la classification du pourriel, les étiquettes (ham, pourriel) appartiennent à un ensemble de classe ordonnée $y \in \{+1, -1\}$. La marge est la distance qui sépare la frontière de séparation (un hyperplan $f(x) = 0$) des exemples les plus proches. Les vecteurs définissant la distance entre la frontière et les exemples les plus proches sont des Vecteurs de Support. L'apprentissage consiste à trouver l'hyperplan assurant le principe de maximisation de la marge, à partir de l'ensemble d'exemples, dont la solution est un problème d'optimisation quadratique. On peut toutefois utiliser ce type de classificateur pour résoudre des problèmes non linéaires par projection dans un espace de dimension supérieure.

Si nous cherchons à séparer des ensembles de données en deux classes $y_\oplus$ (classe des positifs) et $y_\ominus$ (classes des négatifs), l'algorithme SVM permet de trouver un hyperplan séparateur des deux groupes. Pour optimiser cette séparation, SVM recherche l'hyperplan pour lequel la distance entre la frontière des deux groupes et les points les plus proches est maximale.

De façon général, soit $H_\oplus$ et $H_\ominus$ les hyperplans contenant les données $x_\oplus$ et $x_\ominus$, respectivement les plus proches dans la classe $y_\oplus$ et $y_\ominus$. Les coefficients w $et\ w_0$ sont choisis tels que:

$$\begin{cases} \mathrm{w}^T x_\oplus + w_0 = +1. \\ w^T x_\ominus + w_0 = -1. \end{cases} \qquad (24)$$

La valeur de la marge est alors donnée géométriquement par l'équation (25):

$$\frac{\mathrm{w}^T}{\|w\|}\left(x_{\oplus}-x_{\ominus}\right)=\frac{w^T\left(x_{\oplus}-x_{\ominus}\right)}{\|w\|}=\frac{2}{\|w\|} \qquad (25)$$

Pour maximiser la marge, il faut minimiser la valeur de $\|w\|$. Ainsi pour classifier une nouvelle donnée x, il faudra juste résoudre l'équation (26) :

$$f(x)=\ \mathrm{w}^{\mathrm{T}}x+w_0\begin{cases}\geq 0,\ \text{classe}\ \mathcal{Y}_{\oplus}\\ <0,\ \text{classe}\ \mathcal{Y}_{\ominus}\end{cases} \qquad (26)$$

Bien que les algorithmes de types SVM soient efficaces à la classification binaire, deux autres approches de ses méthodes présentent des solutions palliatives aux problèmes de classification à plusieurs classes:

- Le modèle One-versus-all vise de construire autant de modèles SVM que de classes. Ceci permettra dans le cadre de classification d'un message textuel par exemple, de retenir uniquement le modèle qui aurait retourné la plus grande marge sur l'ensemble des modèles ayant retourné un résultat positif.
- Le modèle One-Versus-one : si on dispose de k nombre de classes, alors ce modèle vise de construire les $k(k-1)/2$ classificateurs, en regroupant les classes deux à deux. La classe qui sera fréquemment retournée pendant l'exécution de l'ensemble des algorithmes sera par exemple celle à laquelle appartient un message textuel**.**

3.3.7 Erreur et validation d'apprentissage

Généralement au cours de l'apprentissage, les évènements observés sont représentés par un ensemble de données $\mathbb{D}_n=\left\{\mathrm{X}^{(1)},\mathrm{X}^{(2)},\ldots,\mathrm{X}^{(n)}\right\}$ et on suppose que les données sont tirées indépendamment dans une même distribution inconnue $P(\mathrm{X},\mathrm{Y})$. C'est l'hypothèse que les données sont indépendamment et identiquement distribuées. L'ensemble $\mathbb{f}:\mathrm{X}\longmapsto Y,\hbar\in\mathcal{H}$ représente des modèles ou des solutions possibles de l'hypothèse, et $Y=\{C^{(1)},C^{(2)},\ldots,C^{(K)}\}$. Pour un élément $x\in\mathrm{X}^{(i)}$ particulier, on peut calculer l'erreur du modèle $f\in\mathbb{f}$ associée à cet élément à l'aide d'une fonction de coût $\mathcal{H}(x,f)$. Ainsi, on peut déterminer l'erreur moyenne sur l'ensemble de données $\mathbb{D}_n$ par la fonction:

$$\hat{\mathcal{G}}\left(f,\mathbb{D}_n\right)=\frac{1}{n}\sum_{i=1}^{n}\mathcal{H}(x,f) \qquad (27)$$

Où $\hat{\mathcal{G}}$ désigne *le risque empirique*.

Lorsqu'un exemple est aléatoirement tiré de la distribution $P(\mathrm{X})$, un modèle f permet de mesurer *l'erreur espérée* ou *erreur de généralisation* $\mathcal{G}(f)$. Elle est exprimée selon l'équation suivante :

$$\mathcal{G}(f) = \int \mathcal{H}(x, f)P(x)dx \qquad (28)$$

Le problème de l'apprentissage est de trouver $f \in \mathbb{f}$ qui minimise l'erreur de généralisation $\mathcal{G}(f)$. De manière analytique, le calcul de cette valeur n'est pas trivial, et l'on ne peut que l'estimer à partir du risque empirique $\hat{\mathcal{G}}$. Par conséquent, il revient de trouver une fonction f_ε qui minimise le risque empirique pour l'ensemble de données $\mathbb{D}_n$ par la relation :

$$f_\varepsilon(\mathbb{D}_n) = \underset{f \in \mathbb{f}}{\operatorname{argmin}}\, \hat{\mathcal{G}}\,(f, \mathbb{D}_n) \qquad (29)$$

On pourra ensuite mesurer l'erreur d'apprentissage qui est la plus petite perte moyenne sur l'ensemble $\mathbb{D}_n$: $\min_{f \in \mathbb{f}} \hat{\mathcal{G}}\,(f, \mathbb{D}_n)$.

Pendant le processus d'apprentissage, on remarque aussi que la fonction f_ε qui minimise le mieux l'erreur de généralisation sur $\mathbb{D}_n$ n'est pas toujours la meilleure fonction sur un autre exemple $\mathbb{D}_k$ aléatoirement tiré de la distribution $P(\mathrm{X})$. Alors cette petite perte moyenne est un estimé biaisé de $\mathcal{G}(f_\varepsilon(\mathbb{D}_n))$. Bien évidemment la fonction f_ε apprise sur l'ensemble $\mathbb{D}_n$ notée $f_\varepsilon(\mathbb{D}_n)$ et apprise sur l'ensemble $\mathbb{D}_k$ notée $f_\varepsilon(\mathbb{D}_k)$ ne sont pas censés être correctes à chaque endroit où elles sont différentes l'une de l'autre. Cela implique également que ces fonctions sont différentes de la fonction qui minimise le mieux la fonction de généralisation dans l'ensemble $\mathbb{f}$. Par contre, si on effectue des tests de validation sur l'ensemble $\mathbb{D}'_k$, tiré aléatoirement de l'ensemble de départ $\mathbb{D}_n$ on remarque que l'erreur d'apprentissage estimée sur $\mathbb{D}_n$ et validée sur $\mathbb{D}'_k$ notée $\hat{\mathcal{G}}(f_\varepsilon(\mathbb{D}_n),\ \mathbb{D}'_k)$ devient un estimé non biaisé de l'erreur de généralisation $\mathcal{G}(f_\varepsilon(\mathbb{D}_n))$.

Pour cette raison, l'ensemble des données est souvent subdivisé en deux sous-ensembles de validation et d'entrainement. Les données d'apprentissage serviront à l'entrainement du modèle à utiliser, puis l'erreur de généralisation sera estimée à partir de l'erreur empirique mesurée sur l'ensemble de validation, et exprimée selon la relation suivante:

$$\mathcal{G}(f) \approx \hat{\mathcal{G}}(f_{\varepsilon}(\mathbb{D}_{e}), \mathbb{D}_{v}) \qquad (30)$$

Où $\mathbb{D}_{e}$ désignant les données d'entrainement, et $\mathbb{D}_{v}$ les données de validation. Il arrive que la taille des données soit faible pour les diviser en deux sous-ensembles de validation et d'entrainement. Dans ce contexte, les données sont subdivisées en k partitions de taille presque égale, et généralement, dix est la valeur la plus usuelle de k. Le modèle sera donc entrainé sur les k-1 partitions de façon itérative, et à chaque fois, on mesure l'erreur empirique sur l'unique ensemble qui est laissé à l'écart. Enfin, la moyenne des erreurs empiriques obtenue lors des k cycles d'entrainements servira d'estimés non biaisés de l'erreur de généralisation. Le processus d'entraînement comporte trois phases principales: l'initialisation des données, l'apprentissage et la validation.

La phase d'initialisation consiste à trouver un échantillon de messages types. Cet échantillon doit être représentatif des messages à classifier. Les messages servant à l'entraînement doivent fortement ressembler aux messages que le classificateur aura à traiter. Il faut ensuite procéder à la catégorisation (ham, pourriel) des messages. Puis, selon la méthode d'estimation de l'erreur de généralisation adoptée, on séparer de façon aléatoire les ensembles d'entraînement et de validation.

La phase d'entrainement permet à l'algorithme d'apprendre l'existence des différentes catégories de messages et leur association à leurs classes d'appartenance. Un prétraitement des messages est nécessaire afin de transformer le texte libre sous forme de vecteur d'attributs compréhensible pour les algorithmes de classification. Ainsi pour K classes $Y = \{C^{(1)}, C^{(2)}, \dots, C^{(K)}\}$ on peut définir une fonction de coût possible pour $x \in C^{(k)}$ telle que :

$$\mathcal{H}(f,x)=\begin{cases} 0\ si\ f(x)=C^{(k)} \\ \\ 1\ \ sinon \end{cases} \qquad (31)$$

Et l'erreur de généralisation devient :

$$\hat{\mathcal{G}}(f_\varepsilon(\mathbb{D}_e),\mathbb{D}_v) = \left(\frac{nombre\ de\ messages\ mal\ classés\ de\ \mathbb{D}_v}{nombre\ total\ de\ messages\ traité\ de\ \mathbb{D}_v}\right) \qquad (32)$$

À la phase de validation, le classificateur essaye d'attribuer correctement des étiquettes (ham, pourriel) à tous les exemples de messages appartenant à l'ensemble de validation. Le résultat de la classification est ensuite comparé à l'erreur de généralisation. Si les résultats sont satisfaisants, le classificateur est utilisé pour filtrer de nouveaux messages entrants sur les plateformes.

3.4 Analyse et filtrage du pourriel social

L'apprentissage est le processus de construction des modèles à partir d'un ensemble d'exemples de données. Un exemple représente un couple (message, étiquette). L'objectif est de construire un modèle permettant d'associer une étiquette à un message dont l'identité est ignorée pendant l'apprentissage. L'on remarque que ce processus est fortement lié au type d'algorithme de classification, et peut donc prendre des formes différentes. Par exemple, le classificateur SVM consiste à déterminer l'équation d'un hyperplan séparant deux classes alors que naïve Bayes permet de compter pour chaque classe et pour chaque terme du dictionnaire, le nombre de documents où le terme est présent.

Au-delà des deux formes d'apprentissage supervisé et non supervisé, on parle dans ce même processus de filtrage du pourriel social, de deux autres formes d'apprentissage : *l'apprentissage en* ligne et *l'apprentissage hors ligne*. Dans l'apprentissage hors ligne (dite apprentissage en batch), les exemples sont entièrement traités dès le départ avant toute opération de classification. Dans l'apprentissage en ligne les exemples sont des objets réels à classifier et l'apprentissage se fait, au fur et à mesure, grâce au retour d'information concernant les classifications préalablement effectuées. Ce type d'apprentissage a deux caractéristiques particulières qui le

distinguent de l'apprentissage hors ligne: l'ordre dans lequel sont présentés les exemples est important (ex. l'ordre chronologique). Ensuite, il n'est pas nécessaire de borner le nombre d'exemples utilisés pour l'apprentissage. Enfin ils sont intégrés à un dispositif permettant d'ignorer automatiquement les exemples très anciens et devenus inutiles. Compte tenu du caractère évolutif des messages électroniques, l'apprentissage d'un filtre anti-spam relève typiquement de l'apprentissage en ligne.

3.5 Prétraitement et réduction de dimension des messages

L'un des problèmes de classification du pourriel est le nombre des attributs à sélectionner face à leur nombre explosif. Par exemple, pour un nombre $\mathcal{D}$ d'attributs au départ, il existera $(2^{\mathcal{D}} - 1)$ sous-ensemble d'attributs à sélectionner. Il arrive de se confronter à un nombre d'attributs extraits des différents exemples pouvant atteindre un ordre très élevé (10^7), et dont très peu ont un pouvoir discriminant. À partir d'une fonction d'erreur existante, on ne peut tester rapidement toutes les combinaisons que si $\mathcal{D}$ est relativement petit. Et lorsque ce problème est légèrement traité, deux pièges classiques de l'apprentissage sont tendus:

- La complexité algorithmique (le temps de traitement et éventuellement des problèmes de stockage),
- Le surajustement (overfitting), qui consiste à attribuer à un problème, une dimension plus importante qu'il ne le faut. Dans le cas précis de filtrage du pourriel social, ce phénomène diminue la capacité de généralisation des algorithmes de classification, et donc une perte d'efficacité.

Pour pallier à ce phénomène, la réduction de dimension de représentation des messages est indispensable.

3.5.1 Élimination des mots vides de sens

Le filtrage de pourriels aura lieu qu'après avoir effectué un prétraitement spécifique notamment, la suppression des mots vides de sens (stop word), généralement dépendants de la langue et qui désignent par exemple :

- En anglais: if, why, me, then, with, I, so, when, etc.

- En français, il s'agit des articles (le, la, les, un, aux, l', …), des prépositions (à, de, pour, par, sans, …), des adverbes (ici, là, si, aussi, ça, …), les opérateurs de conjonctions (ou, et, donc, or, ni, car, …), et des pronominaux (il, elle, eux, …).

De même, la réduction de dimension nécessite souvent la suppression des mots très fréquents, puisqu'ils sont présents partout et donc ne portent pas d'information pertinente sur la discrimination des classes. Il en est de même pour les mots très rares, car leur faible fréquence ne permet pas de construire des règles stables. Dans la plupart des cas, ces deux catégories de mots accroissent utilement le temps d'apprentissage et la complexité des algorithmes [4]. Malgré la suppression de ces deux catégories de mots, les candidats restent encore très élevés et il faut utiliser les méthodes statistiques pour choisir les mots les plus pertinents.

3.5.2 Représentation des messages avec des lemmes et des stemms

Dans ce même processus de réduction de dimensions des messages, les méthodes d'extraction des attributs peuvent s'appuyer sur des traitements morphologiques des termes. Ces traitements visent à regrouper les attributs (mots ou termes) d'une même famille à partir d'un dictionnaire. Les approches les plus adoptées sont :

- La lemmatisation qui consiste à ramener sous une forme canonique les différentes formes que peut revêtir un mot. Ainsi on ramène à la forme «appel» les termes «appel, appels, appelle, appelles » et à la forme « appeler » les termes «appelé, appelaient, appellera, appelleront». La lemmatisation cherche à regrouper les racines lexicales en réduisant un nom à sa forme la plus simple au masculin singulier et remplace les formes verbales conjuguées à l'infinitif. Généralement, un lemme correspond systématiquement à un mot réel de la langue.
- Racinisation (angl. Stemming) : similaire à la lemmatisation, ce procédé repose sur des contraintes linguistiques et consiste à extraire le radical ou racine (stemme) à partir des flexions des mots.

Contrairement au lemme qui correspond à un mot réel de la langue, la racine ne l'est pas forcément. Par exemple, le mot « aime » a pour radical « aim » qui ne correspond pas à un mot réel. Par contre dans l'exemple de « porteur », le radical est « port » qui lui l'est. Ce processus vise donc à réduire le nombre des attributs en faisant abstraction des pluriels et des multiples façons d'exprimer la même pensée.

3.5.3 Normalisation de la longueur des messages

Pour normaliser la fréquence du terme, on divise le nombre d'occurrences de ce terme par la valeur de la fréquence la plus haute d'un terme quelconque présent dans ce message. Cette opération vise à ramener les attributs des messages à ceux d'un message de longueur fixe, longueur adoptée comme une référence. Des expériences ont montré que les polluposteurs font usage des messages courts, dont la taille est souvent inférieure à 20 Ko. Salton et al [55] ont montré à partir des corpus de TREC que les messages plus longs ont plus de chance de satisfaire les critères d'une requête que les messages courts. Pour éviter que les métriques attribuent des poids à certains termes proportionnellement à la taille des messages dans lesquels ils sont présents, nous avons jugé nécessaire de procéder à la normalisation de la longueur des messages à une taille égale. Plusieurs approches existent pour la normalisation, et les plus usuelles sont :

- La méthode de Graham qui recommande l'utilisation des $(\mathcal{D} = 15)$ termes les plus pertinents,
- La méthode de Cormack [56] qui recommande l'utilisation des n-grams de taille fixe pour représenter un message.

Pour cette même opération de normalisation des messages, le choix des termes pertinents peut se faire par le calcul de l'information mutuelle, le gain d'information, et autres méthodes utilisant les fréquences d'apparition des termes.

- **Gain d'information**

Lorsque le vocabulaire véhiculé par les messages du corpus est connu, on peut utiliser la technique des associations par des probabilités telle que le gain d'information pour extraire les termes ayant de grandes valeurs discriminatives. Par exemple, la présence ou l'absence d'un terme t_i peut être évaluée par la mesure du gain d'information. Dans ce contexte précis de filtrage de pourriel, pour un terme t_i apparaissant dans un exemple de message, le gain d'information mesure l'entropie ou la diminution d'incertitude moyenne d'appartenance de ce terme à l'ensemble $Y = \{ham, spam\}$. Le gain d'information se mesure par la relation suivante :

$$IG(t_i) = H(Y) - H(Y|t_i) \qquad (33)$$

Où :

$$IG(t_i) = -\sum_{y \in Y} p(y) log_2 p(y) + \sum_{\substack{y \in Y \\ t \in T_i}} p(t) p(y|t) \log_2 p(y|t)$$

Et $T_i = \{t_i \ \overline{t}_i\}$ indique la présence ou l'absence du terme dans le document.

- **Information mutuelle**

Une autre technique intéressante des associations de probabilité est l'information mutuelle qui est un critère permettant de maximiser les contraintes d'association des objets entre classes. Cette approche exprime le rapport d'association entre le terme t_i et une classe donnée $y \in Y$.

$$\mathrm{I}(t_i, y) = \log_2 \frac{p(t_i, y)}{p(t_i)p(y)} = \log_2 \frac{p(t_i|y)}{p(t_i)} = \log_2 \frac{p(y|t_i)}{p(y)} \qquad (34)$$

En ce qui concerne le filtrage du pourriel, cette valeur $\mathrm{I}(t_i, y)$ peut être estimée à partir du tableau des données de l'attribut versus la classe.

3.5.4 Les modèles de représentation des messages

En général, les algorithmes de classification ne savent pas manipuler les données brutes (non structurées). C'est pourquoi, après la phase de prétraitement, il faut représenter les données sous forme de matrice. Le plus souvent, cette matrice sera représentée par un tableau de données, dont les cases sont des entiers définis dans un même système de mesures et d'unités homogènes [26]. D'autres types de tableaux existent, et peuvent servir à la classification, tels que les tableaux de mesures ou les tableaux logiques de modularité 0 ou 1. Certaines fois, il arrive qu'un tableau de fréquence soit assimilé à un tableau de données des k (i, j) observations d'un évènement (i, j), mais le vecteur sera la représentation la plus courante où chaque dimension correspond à un attribut du message.

3.5.4.1 Modèle de sac à mot

Au cours de notre expérimentation, nous avons utilisé le modèle de sac à mot pour représenter le corpus. Généralement, dans les processus de traitement statistique des textes, le modèle basé sur la présence de mots d'un certain vocabulaire communément appelé modèle de sac à mots (bag of Words) est le plus utilisé. On peut le considérer comme étant une suite de caractères appartenant

à un dictionnaire, et le message électronique représente l'expérience de tirer avec replacement des mots dans un sac. L'ensemble des messages du corpus sera représenté par un vecteur de la même taille que le dictionnaire, dont la composante *k* indique le nombre d'occurrences du *k-ème* mot du dictionnaire dans le document. L'idée est de transformer les messages en vecteurs dont chaque composante représente un mot. Cette représentation des messages exclut toute analyse grammaticale et toute notion de distance entre les termes, d'où l'appellation « sac de mots ».

3.5.5 La fonction du filtre

La phase de filtrage est la dernière étape d'un processus de classification. Le filtre représente l'élément central que l'on peut décomposer en trois modules essentiels.

- La représentation de l'information représente le module de traitement où les attributs des messages sont extraits par les méthodes de classification. Ce module prépare les données pour être utilisées par les algorithmes de classification.
- L'algorithme de classification est le cerveau du filtre. En d'autres termes, il représente le module intelligent du filtre. Il est réservé à l'implémentation informatique d'une méthode de classification, telle que la régression logistique, le perceptron ou les machines à vecteurs de support (SVM), etc.
- Les paramètres représentent l'ensemble des variables utilisées par l'algorithme de classification.
-

Enfin le classificateur est l'ensemble formé de paramètres et d'algorithme de classification.

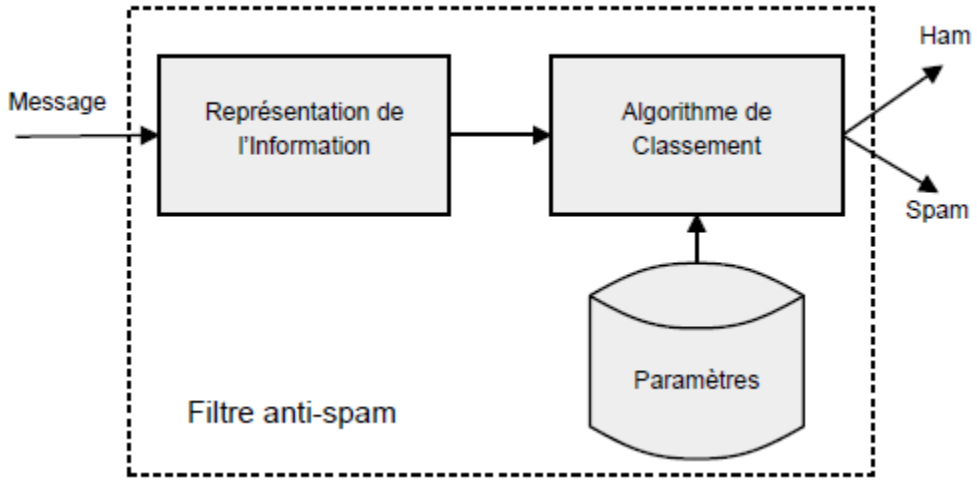

Figure 7: *Principe de fonctionnement du filtre antispam, source : pourriel, classement automatique de messages électroniques de Jose-Marcio Martins Da Cruz.*

3.6 Sommaire

Bien qu'il existe plusieurs scénarios de filtrage de pourriel, la plupart des modèles déployés adopte un processus logique simple : les messages sont soumis au filtre dans l'ordre de leur arrivée. Après traitement, le filtre associe le message à l'une des deux classes (ham, pourriel) indiquant, éventuellement, l'incertitude de la classification à l'aide d'une valeur numérique (par exemple le score). Sur certaines plateformes de réseaux sociaux, l'intervention d'un modérateur pour corriger les erreurs de classification peut s'avérer nécessaire.

Chapitre 4- Expérimentation et résultats

4.1 Aperçu général

Dans ce chapitre, nous présentons les principaux résultats des expériences que nous avons effectuées à partir d'un corpus public. Les corpus sont généralement protégés par les promoteurs de réseaux sociaux pour préserver la vie privée des usagers. C'est pourquoi nous nous sommes orientés vers un corpus public « **FINAL »** mis au point à des fins expérimentales par Tiago et al. [4]. Ce corpus est composé de 5574 messages réels. De plus, nous avons opté d'effectuer notre expérimentation sur ce corpus du fait que le contenu des messages soit similaire à celui des tweets : la taille maximale d'un message est environ de 140 octets, et comportes des abréviations et certaines fois des émoticônes que l'on retrouve dans le corps des messages des réseaux sociaux. Nous avons donc réalisé quatre différents classificateurs cités dans l'état de l'art pour leur efficacité dans le filtrage du pourriel en général et du pourriel social en particulier. Il s'agit de naïves Bayes, de la régression linéaire, des arbres de décision, et des machines à vecteurs de support (SVM).

4.2 Corpus de données et prétraitement

Le corpus de données dénommé **FINAL** est un corpus de messages réels et publics collectés à partir de plusieurs sources à savoir :

- **GrumbleTex**t, un site web britannique où les utilisateurs de téléphones mobiles réalisent des signalements lorsqu'ils reçoivent un appel ou un message douteux, l'incitant par exemple à texter ou à rappeler un numéro surtaxé. À partir de ce site, 425 messages pourriels de type SMS ont été collectés.
- **NUS** (National University of Singapore) est un corpus mis au point par le département informatique de cette université, où 3375 messages ont été tirés au hasard sur un total de 10,000 messages légitimes.
- Le corpus **INIT-v.0.1** a mis en public 1002 messages légitimes et 322 pourriels.
- Le corpus linguistique mise au point par Caroline TAGG de l'université de Birmingham, composé de 425 messages ont également servi à la mise au point du corpus **FINAL**.

C’est donc la combinaison de ces ensembles de données qui donnera naissance au corpus « **FINAL** » composé de 4827 messages légitimes et de 747 pourriels, soit un total de 5574 messages, dont voici un extrait de contenu de ce corpus.

Tableau 1: *contenu des messages du corpus FINAL*

text3 <5572x2 cell>

	1	2
1	0	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got am...
2	0	Ok lar... Joking wif u oni...
3	1	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive en...
4	0	U dun say so early hor... U c already then say...
5	0	Nah I don't think he goes to usf, he lives around here though
6	1	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for...
7	0	Even my brother is not like to speak with me. They treat me like aids patent.
8	0	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your ...
9	1	WINNER!! As a valued network customer you have been selected to receivea Â£900 prize reward! ...
10	1	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with ca...
11	0	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried e...
12	1	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/d...
13	1	URGENT! You have won a 1 week FREE membership in our Â£100,000 Prize Jackpot! Txt the word:...
14	0	I've been searching for the right words to thank you for this breather. I promise i wont take your ...
15	0	I HAVE A DATE ON SUNDAY WITH WILL!!

4.2.1 Segmentation des messages

Dans notre démarche d’évaluation des classificateurs, nous avons au préalable effectué une segmentation des messages, c’est-à-dire brisé le flux de texte en « mots ». Cette segmentation considère comme mot les séquences de caractères séparés par des espaces, les tabulations, les retours chariots, les points, les virgules, les deux points et les tirets. Ce processus a permis de décomposer naturellement les messages et obtenir des attributs sous forme de vecteurs de mots (tableau 2). On désigne par *forme* ou *forme graphique* tous les attributs extraits de cette façon. Habituellement, ces attributs ne correspondent pas forcément à des mots ayant le sens linguistique courant. Pour cette raison, on préfère le mot *forme* pour designer l’unité de segmentation.

Dans ce même processus, nous avons jugé important de convertir certaines séquences courantes telles que les numéros de téléphone ou les montants en dollars en une représentation de format

NNNNN au lieu de 06.12.14.17.09. Cette représentation vise à réduire le nombre d'entrées du vocabulaire et surtout à retenir que la forme en question fait référence à un numéro de téléphone plutôt qu'à sa valeur précise. Enfin, il était utile de représenter systématiquement tous les messages vers une casse de caractères uniques : les minuscules.

Tableau 2: *prétraitement des messages*

temp <5572x1 cell>

	1
1	go','until','jurong','point','crazy','available','only','in','bugis','n','great','world','la','e','buffet','cine','there','got','a...
2	ok','lar','joking','wif','u','oni
3	free','entry','in','N','a','wkly','comp','to','win','fa','cup','final','tkts','NNst','may','NNNN','text','fa','to','NNNNN','t...
4	u','dun','say','so','early','hor','u','c','already','then','say
5	nah','i','don','t','think','he','goes','to','usf','he','lives','around','here','though
6	freemsg','hey','there','darling','it','s','been','N','week','s','now','and','no','word','back','i','d','like','some','fun','you...
7	even','my','brother','is','not','like','to','speak','with','me','they','treat','me','like','aids','patent
8	as','per','your','request','melle','melle','oru','minnaminunginte','nurungu','vettam','has','been','set','as','your','cal...
9	winner','as','a','valued','network','customer','you','have','been','selected','to','receivea','NNN','prize','reward','to',...
10	had','your','mobile','NN','months','or','more','u','r','entitled','to','update','to','the','latest','colour','mobiles','with'...
11	i','m','gonna','be','home','soon','and','i','don','t','want','to','talk','about','this','stuff','anymore','tonight','k','i','ve',...
12	six','chances','to','win','cash','from','NNN','to','NN','NNN','pounds','txt','cshNN','and','send','to','NNNNN','cost',...
13	urgent','you','have','won','a','N','week','free','membership','in','our','NNN','NNN','prize','jackpot','txt','the','word...
14	i','ve','been','searching','for','the','right','words','to','thank','you','for','this','breather','i','promise','i','wont','take','...
15	i','have','a','date','on','sunday','with','will

Les statistiques de base du corpus **FINAL** compte au total 81175 tokens ou mots, composé de 63632 mots extraits des messages légitimes, et de 17547 mots extraits des messages pourriels (tableau 3 et 4).

Tableau 3: *Les statistiques de base*

Messages	Proportion (nombre)	Proportion (%)
Hams	4827	86.6
Pourriels	747	13.40
Total	5574	100.00

Tableau 4: *Statistiques des occurrences de mots*

Hams	63,632
Pourriels	17,543
Total	81,175
Moy par Message	14.56

Moy par Hams	13.18
Moy par Spams	23.48

4.2.2 Normalisation du corpus

Pour préserver la sémantique des messages, aucun prétraitement spécifique à une langue n'est effectué (pas d'élimination des mots vides de sens, ni recherches de radicaux des mots) étant donné que certains chercheurs pensent que ces méthodes affectent la performance et la précision du filtrage des pourriels [56, 57]. Seules les formes redondantes ont été supprimées. Pour normaliser le tableau des fréquences des formes graphiques, toutes les entrées non nulles ont été sommées. Ce qui a permis de réduire le nombre de caractéristiques à 4875 au lieu de 118619 entrées. Chaque message est enfin transformé sous forme d'un vecteur de mots (tokens).

4.2.3 Protocole expérimental

Nous avons adopté les terminologies suivantes dans ce chapitre:

- Ham pour désigner les messages légitimes (non-pourriel)
- Spam pour désigner le pourriel

Ensuite des notions spécifiques sont utilisées pour juger des performances d'un classificateur:

Vrai positif (VP): "hams" correctement classés.

Faux négatif (FN): "hams" ayant été étiquetés "spam" par le classificateur.

Vrai négatif (VN): "spams" correctement classés.

Faux positif (FP): "spams" ayant été étiquetés "ham" par le classificateur.

En dehors du ratio global de bonne classification (Vrai positif + Vrai négatif), c'est le ratio de Faux Positif qui va nous intéresser le plus car il faut éviter de perdre un message important à cause d'une erreur de classification. Cela pourrait avoir des conséquences désastreuses surtout pour une entreprise. Les performances de chaque algorithme seront donc évaluées selon deux aspects: le ratio des faux positifs et le taux d'erreur.

$$taux\ d'erreur = \frac{FP + FN}{VP + FP + VN + FN} \qquad (35)$$

$$ratio\ de\ faux\ positifs = \frac{FP}{VN + FP} \qquad (36)$$

Enfin, nous donnerons nos résultats sous forme de matrice de confusion assimilable au tableau de contingence suivant:

Tableau 5: *tableau de contingence de classification des messages*

	Réalités	
	Ham	Spam
Ham	VP	FN
Spam	FP	VN

Avant de passer à la phase de classification, nous devrions vérifier si un terme présent dans un message était présent ou non dans le dictionnaire. Cela nous permettra d'établir la matrice de fréquences des termes (vocabulaire), qui servira à prédire la classe d'appartenance d'un message. La présence ou l'absence des termes est respectivement marquée par (1) et (0) dans le tableau 6.

Tableau 6: *Matrice de fréquences des termes par messages*

data <5572x118619 double>

	1	2	3	4	5	6	7	8	9
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	1	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	1
9	0	1	0	0	0	0	0	0	1
10	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0
13	0	1	0	0	0	0	0	0	0
14	0	1	0	0	0	0	0	0	0
15	0	1	0	0	0	0	0	0	0

4.2.4 La phase de classification

La matrice de fréquences représentant les vecteurs de caractéristiques a été partitionnée en deux sous sous-ensembles: des données d'entrainement et des données de validation tout en préservant le ratio des messages pourriels et ham dans ces sous-ensembles. Ensuite nous avons effectué la validation croisée c'est-à-dire pour chaque sous-ensemble, nous avons créé 10 paires de partitions d'entrainement et de validation pour examiner la performance comme une moyenne

sur 10 itérations. Chaque catégorie de messages est divisée en des ensembles de taille presque égale (558 messages par itération).

Pour définir l'appartenance à la classe ham/spam, (spamcité) nous estimons la probabilité de chaque mot comme suit:

$P(t|S) =$ la probabilité qu'un terme spécifique apparaisse dans les messages spams,

$P(t|H) =$ la probabilité qu'un terme spécifique apparaisse dans les messages ham.

Et enfin nous définissons la probabilité a postériori du terme t comme suit:

$$P(S|t) = \left(\frac{P(t|S)P(S)}{P(t|S)P(S) + P(t|H)P(H)}\right) \qquad (37)$$

Chaque probabilité est estimée en utilisant une proportion de messages dans l'ensemble d'entrainement. Donc on évalue $P(t|S)$ uniquement sur la fraction de messages spams contenant le terme t et $P(t|H)$ est également évaluée sur une proportion de messages légitimes contenant le terme t.

Soit m le nombre de messages dans l'ensemble d'entrainement, et le i-ème message contient n_i mots. Soit le dictionnaire contenant $|N| = 4875\ mots$. $P(t|y = 1)$ est le paramètre qui estime la probabilité qu'un mot particulier d'un message spam, soit le kieme mot dans le dictionnaire. $P(t|y = 0)$ est le paramètre qui estime la probabilité qu'un mot particulier du message ham, est le kieme mot dans le dictionnaire. L'expression $P(y)$ est le paramètre qui estime qu'un message quelconque soit un spam.

Nous calculons les trois paramètres $P(t|y = 1), P(t|y = 0)$, et $P(y)$ à partir des données d'entrainement. Ensuite pour faire la prédiction sur des messages tests, nous utilisons les paramètres calculés pour effectuer les comparaisons avec $P(t|y = 1) * P(y = 1)$ et $P(t|y = 0) * P(y = 0)$. Au lieu de procéder à une comparaison des valeurs de paramètres directement avec les probabilités (méthode non triviale), on préfère plutôt comparer leurs logarithmes. D'où un message sera classé pourriel si:

$$\log P(t|y = 1) + \log P(y = 1) > \log P(t|y = 0) + \log P(y = 0) \qquad (38)$$

4.3 Les classificateurs déployés et résultats

4.3.1 Classificateur Naïve Bayes

L'idée principale de l'algorithme de classification Naïve Bayes est que l'on peut calculer la probabilité qu'un mot appartienne à une certaine classe en appliquant le théorème de Bayes. Ce théorème suppose que la probabilité qu'un attribut apparaisse dans un document est indépendante de la probabilité de tout autre attribut apparaissant aussi dans ce document. L'efficacité d'un classificateur est exprimée par le ratio de bonnes prédictions sur le nombre total de messages. Initialement, nous avons observé les effets du prétraitement du corpus avec le classificateur de Bayes. Pour l'ensemble des messages, nous avons obtenu les résultats suivants:

Tableau 7: *Naïve Bayes avant la validation croisée*

	Réalités	
	Ham	Spam
Ham	4794	31
Spam	37	710

Le taux d'erreur pour cette classification : 1.22%

Le ratio de faux positifs est de 0.64%.

Après validation croisée, la meilleure classification obtenue est la suivante:

Tableau 8: *performance des classificateurs naïfs Bayes*

	Réalités	
	Ham	Spam
Ham	481	8
Spam	6	63

Le meilleur taux d'erreur de 2.51%

4.3.2 La régression logistique

Pour ce qui concerne la classification des messages textuels, y représente l'étiquette du message (ham, spam) et noté sous forme binaire ordonnée $y \in Y = \{0, 1\}$, tandis que, le vecteur de caractéristiques X représente l'ensemble du vocabulaire des messages.

Nous avons obtenu les résultats suivants après entrainement du classificateur:

Tableau 9: *performance de régression logistique*

	Réalités	
	Ham	Spam
Ham	373	97
Spam	19	68

Soit un taux d'erreur de 20.82%

4.3.3 Les arbres de décision

Tableau 10: *performance des arbres de décision*

	Réalités	
	Ham	Spam
Ham	479	10
Spam	7	62

Soit un taux d'erreur de 3.04%

4.3.4 Les machines à vecteurs de support

Tableau 11: *performance des SVM*

	Réalités	
	Ham	Spam
Ham	485	9
Spam	7	57

Soit un taux d'erreur de 2.87 %

Il faut préciser qu'en matière de classification de spams, il souhaitable d'évaluer la performance des classificateurs par la moyenne harmonique ($F-mesure$) entre la précision (P_r) et le rappel (R). Cette métrique raffine les résultats puisqu'elle représente une synthèse issue de la combinaison de la précision et du rappel.

$$F-mesure = \frac{2P_r R}{P_r + R} = \frac{2 \times TP}{2 \times TP + FP + FN} \qquad (39)$$

$$\text{avec} \quad P_r = \frac{TP}{TP+FP} \quad \text{et} \quad R = \frac{TP}{TP+FN}$$

Ainsi, le tableau 12 présente le résumé de meilleures performances obtenues pour chaque classificateur pour l'ensemble des messages du corpus FINAL.

Tableau 12: *résumé des performances des classificateurs*

Classificateurs	***Précision***	***Rappel***	***F-mesure***
NAIVE BAYES	***91.8***	***98.37***	***98.57***
SVM	***89.49***	***98.18***	***98.38***
Arbre de décision	***88.53***	***97.96***	***98.26***
Régression logistique	***84.59***	***79.36***	***86.54***

Selon notre expérimentation, Naïves Bayes et SVM performent mieux que les arbres de décision et encore bien loin de la régression logistique. Ensuite la durée maximale de chaque cycle de d'exécution sur un ordinateur i7- 4500U, 2.4Ghz et 8Go de RAM est de :

- 15ms pour Naïve Bayes,
- 4 minutes pour SVM
- 20 minutes pour les arbres de décision

Conclusion générale

Dans cet essai, nous avons présenté un aperçu des activités des polluposteurs sur les réseaux sociaux (exemple : Facebook, Twitter, Myspace et Sino Weibo). Nous avons ensuite examiné un certain nombre d'attributs que partagent les polluposteurs et qui les différencient des usagers légitimes. Nous avons enfin entrainé et évalué différents classificateurs pour la détection de pourriels textuels. Nous avons obtenu des résultats très probants, qui démontrent la fiabilité des méthodes d'apprentissage pour détecter le pourriel social.

Bien que les résultats témoignent de la précision des méthodes d'apprentissage pour filtrer le pourriel textuel, nous envisageons dans les travaux futurs étendre notre analyse à l'ensemble des pourriels multimédias (image, vidéo et l'analyse des liens entre pages Web). Dans cette même optique, nous tenterons d'effectuer un filtrage en ligne du pourriel multimédia. Cela nous permettra de tenir compte de l'aspect évolutif du pourriel afin de rendre notre système fonctionnel sur une longue période avant de nécessiter une mise à jour.

Bibliographie

[1] Nexgate. (2015, 2 mai). *2013 state of social media pourriel*. Available: http://nexgate.com/wp-content/uploads/2013/09/Nexgate-2013-State-of-Social-Media-Spam-Research-Report.pdf

[2] C. K. Gianluca Stringhini, Giovanni Vigna, "Detecting spammers on social networks," in *Proceedings of the 26th Annual Computer Security Applications Conference ACSAC '10*, New York, NY, USA, 2010, pp. 1-9.

[3] L. G. Enhua Tan, A Songqing Chen, Xiaodong Zhang, Yihong Zhao, "UNIK: unsupervised social network spam detection," pp. 479-488, 2013.

[4] J. M. G. H. Tiago A. Almeida , Akebo Yamakami, "Contributions to the study of SMS spam filtering: new collection and results," 2011.

[5] L. Y. Huan Liu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,* vol. 17 N0. 4, Avril 2005.

[6] P. Guillon, "Etat de l'art du pourriel, solutions et recommandations " 10 décembre 2008.

[7] Sophos. (2015, 2 mai). *SPAMIONSHIP des douze pays émettant le plus de pourriels – quel rôle *VOUS* pouvez jouer* Available: https://www.sophos.com/fr-fr/press-office/press-releases/2014/10/dirty-dozen-spampionship.aspx

[8] B. Markines, C. Cattuto, and F. Menczer, "Social Spam Detection," presented at the 18th International World Wide Web Conference (W3C 2009), 2009.

[9] Statista. (2015). Available: http://www.statista.com/topics/1164/social-networks/

[10] Facebook. (2015, 1 mai). Available: http://newsroom.fb.com/

[11] Z. Z. Xianghan Zheng, Zheyi Chen, Yuanlong Yu, Chunming Rong, "Detecting Spammers on Social Networks," *Neurocomputing,* 8 February 2015.

[12] B. C. Markines, Ciro, Menczer, Filippo, "Social Spam Detection," presented at the 18th International World Wide Web Conference (W3C 2009), 2009.

[13] W. W. Cohen, "Learning rules that classify e-mail," presented at the In Papers from the AAAI Spring Symposium on Machine Learning in Information Access, AAAI Press, 1996.

[14] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys (CSUR),* vol. 34 Issue 1, pp. 1- 47, March 2002

[15] D. W. H. Drucker, and V. N. Vapnik, "Support vector machines for spam categorization," *Neural Networks, IEEE Transactions on 10* vol. no. 5, pp. 1048–1054, 1999.

[16] P. P. a. D. Lin, "Spamcop : A spam classification and organization program Learning for Text Categorization," presented at the AAAI Technical Report, (Madison, Wisconsin), 1998.

[17] S. D. Mehran Sahami, David Heckerman, and Eric Horvitz "A bayesian approach to filtering junk E-mail, Learning for Text Categorization," presented at the AAAI Technical Report, Papers from the 1998 Workshop (Madison, Wisconsin), 1998.

[18] A. K. a. J. Alspector, "SVM-based filtering of E-mail spam with content-specific misclassification costs," *TextDM 2001 (IEEE ICDM-2001 Workshop on Text Mining)* 2001.

[19] M. U. C. Md. Rafiqul Islam, and Wanlei Zhou, "An innovative spam filtering model based on support vector machine," in *CIMCA '05 : Proceedings of the International Conference on Computational Intelligence for Modelling Control and Automation*, Washington, DC, USA), 2005, pp. 348–353.

[20] T. Joachims, *Making large-scale support vector machine learning practical* 1999.

[21] D. S. a. G. M. Wachman, "Relaxed online SVMs for spam filtering," in *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA 2007, pp. 415–422.

[22] E. B. a. A. Bryl., "A survey of learning-based techniques of email spam filtering. Artificial Intelligence," pp. Rev.29:63-92, March 2008.

[23] N. A. S. S. Hao, N. Feamster, A. G. Gray, and S. Krasser, "Detecting spammers with snare: Spatio-temporal network-level automatic reputation engine," *In USENIX Security Symposium,* pp. 101-118, 2009.

[24] B. R. C. Whittaker, and M. Nazif, "Large-scale automatic classication of phishing pages," *In NDSS,* 2010.

[25] C. G. K. Thomas, J. Ma, V. Paxson, and D. Song, "Design and evaluation of a real-time url spam filtering service," in *In Proceedings of IEEE Symposium on Security and Privacy*, 2011.

[26] A. J. P. Kolari, T. Finin, T. Oates, and A. Joshi. , "Detecting spam blogs: A machine learning approach," *In AAAI,* 2006.

[27] B. C. D. D. Nagamalai, and J.-K. Lee. , "Bayesian based comment spam defending tool," *CoRR,* p. 1011.3279, 2010.

[28] D. C. G. Mishne, and R. Lempel, " Blocking blog spam with language model disagreement," 2005.

[29] G. K. P. Heymann, and H. Garcia-Molina, "Fighting pourriel on social web sites: A survey of approaches and future challenges. Internet Computing," *IEEE , Cloud Computing and Big Data (CloudCom-Asia),* pp. 11(6):36-45, nov.-dec 2007.

[30] A. Z. a. J. S. Donath, "Is britney spears spam?," *In CEAS,* 2007.

[31] J. C. K. Lee, and S. Webb, "Uncovering social spammers: social honeypots + machine learning," *In SIGIR,* pp. 435-442, 2010.

[32] T. R. Fabrício Benevenuto , Virgílio Almeida, Jussara Almeida, Marcos Gonçalves, "Detecting spammers and content promoters in online video social networks," *SIGIR '09 Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* pp. 620-627 2009.

[33] B. D. D. D. Yin, Z. Xue, L. Hong, A. Kontostathis, and L. Edwards "Detection of Harassment on Web 2.0. In Content analysis in the WEB 2.0," *(CAW2.0),* 20019.

[34] M. M. M. Huber, E. Weippl, G. Kitzler, and S. Goluch, "Friend-in-the-middle attacks: Exploiting social networking sites for spam.," *IEEE Internet Computing,* pp. 15-28, May 2011.

[35] C. X. L. X. Jin, J. Luo, and J. Han, "Socialspamguard: A data mining-based spam detection system for social media networks," *PVLDB,* vol. 4(12), pp. 1458-1461, 2011.

[36] D. I. De Wang, Calton Pu, "SPADE: a social-spam analytics and detection framework," 2014.

[37] A. H. Wang, "Don't follow me: spam detection in Twitter, security and cryptography," *IEEE, (SECRYPT), in Proceedings of the 2010 international conference on* . pp. 1 - 10, 2010.

[38] H. Gao, "Towards Online Heterogeneous Spam Detection and Mitigation for Online Social Networks," *ProQuest Dissertations and Theses,* 2013.

[39] E. C. T. Stein, and K. Mangla, "Facebook immune system.," in *In Proceedings of the 4th Workshop on Social Network Systems, SNS '11*, New York, NY, USA, 2011, pp. 8:1-8:8.

[40] S. W. D. Irani, and C. Pu, "Study of static classication of social spam proles in myspace," 2010.

[41] T. R. Fabricio Benevenuto, Adriano Veloso, Jussara Almeida, Marcos Goncalves, Virgilio Almeida, "Practical Detection of Spammers and Content Promoters in Online Video Sharing Systems," *IEEE Systems, Man, and Cybernetics Society,* vol. 42, pp. 688 - 701 30 November 2011 2012.

[42] E. M. M. Bosma, and W. Weerkamp, "A framework for unsupervised spam detection in social networking sites," presented at the In ECIR, 34th European Conference on Information Retrieval, Barcelona, 2012.

[43] C. E. Antonio Lupher, Reynold Xin, "Feature Selection and Classification of Spam on Social Networking Sites, Berkeley University, USA."

[44] X. W. Yin Zhu, Erheng Zhong Nanthan N. Liu, He Li, Q. Yang, "Discovering Spammers in Social Networks," *Association for the Advancement of Artificial Intelligence (www.aaai.org),* 2012.

[45] J. T. X. Hu, Y.Zhang, H. Liu, "Social spammer detection in microblogging," *ACM, Proceedings of the twenty-third Internation Joint Conference on Artificial Intelligence,* pp. 2633 - 2639, 2012.

[46] D. D. Lewis, "An evaluation of phrasal and clustered representations on a text categorization task," in *Proceedings of the 15thannual international ACM SIGIR conference on Research and development in information retrieval*, Copenhagen, Danemark, 1992, pp. 37 - 50.

[47] A. K. e. J. Vedelsby, " Neural network ensembles, cross validation, and active learning," in *Advances in Neural Information Processing Systems 7*, ed: MIT Press, 1995, pp. 231 - 238.

[48] G. F. L. e. W. A. Stubblefield, *Artificial intelligence: structures and strategies for complex problem solving*: Addison-Wesley, 1998.

[49] T. O. a. T. White, "Immunity from spam : An analysis of an artificial immune system for junk email detection," in *Proceedings, 4th International Conference, ICARIS*, Banff, Alberta, Canada, 2005, pp. 276-289.

[50] A. C. A. Kołcz, "Lexicon randomization for near-duplicate detection with I-Match," *Supercomput Springer Science+Business Media, LLC 2008,* pp. 45: 255–276, 26 January 2008.

[51] X. Z. Lingshan Xu, Chunming Rong, "Trust Evaluation Based Content Filtering in Social Interactive Data," *IEEE ,Cloud Computing and Big Data (CloudCom-Asia), International Conference on 16-19 Dec,* pp. 538 - 542 2013.

[52] T. S. Leyla Bilge, Davide Balzarotti, Engin Kirda, "All your contacts are belong to us: automated identity theft attacks on social networks," *ACM, WWW '09 Proceedings of the 18th international conference on World wide web* pp. Pages 551-560 2009.

[53] J. H. Jennifer Golbeck, "Reputation Network Analysis for Email Filtering," *CEAS,* 2004.

[54] E. S. Jaeyeon Jung, "An Empirical Study of Spam Traffic and the Use of DNS Black Lists," *MIT Computer Science and Artificial Intelligence Laboratory , 32 Vassar Street, Cambridge, MA 02139, USA,* 2004.

[55] J. A. a. A. S. Gerard Salton, "Automatic Text Decomposition and Structuring," *IPM - Information Processing & Management,* vol. 32 pp. 127--138, 1996.

[56] G. V. Cormack, "Email Spam Filtering: A Systematic Review," *Foundations and Trends in Information Retrieval,* vol. 1 Issue 4, April 2008.

[57] J. Z. Le Zhang, Tianshun Yao, "An evaluation of statistical spam filtering techniques," *ACM Transactions on Asian Language Information Processing (TALIP),* vol. 3 Issue 4, pp. 243-269 December 2004.