DÉPARTEMENT GÉNIE INDUSTRIEL

PROJET BIBLIOGRAPHIQUE

Détection des spams e-mail

Réalisé par :
AYMEN HAJJEJ
EYA JLASSI
Classe: 1ATA

Encadrés par :
MME MEYSSA BEN SAAD

Année universitaire:
2022/2023

Remerciements

En premier lieu, nous remercions notre enseignante et encadrante Mme. Meyssa ben Saad pour toutes les connaissances qu'elle nous a communiquées tout au long de l'élaboration de ce projet, pour sa grande disponibilité, et ses précieux conseils. Nous remercions également Mme Hedia Chaker, responsable de la filière Techniques Avancées, son sérieux remarquable et ses encouragements, qui nous ont été d'une très grande utilité

Nous remercions également, toutes les personnes qui nous ont conseillées et relues lors de la rédaction de ce projet, notamment nos camarades de la 1re année Techniques Avancées.

Enfin, nous adressons une pensée affective à nos parents, nos familles et nos amis pour leurs supports tout au long de ce projet.

Table des matières

Table des figures

Introduction

La notion d'e-mail est apparue pour la première fois en 1965 à l'institut de technologie du Massachusetts sous le nom de programme « Mailbox ». Elle était utilisée dans le but de laisser des messages sur l'ordinateur de l'université pour que d'autres puissent les consulter. Cependant, même si cette innovation a permis la diffusion des informations entre les étudiants, elle était incapable de les faire circuler entre les ordinateurs, néanmoins hors de l'université. Ce n'est qu'en 1971 que la messagerie électronique a connu une révolution grâce à l'ingénieur Américain Ray Tomlinson et au développement du réseau ARPANET [1]. En effet, la conception de deux programmes SNDMSG [2] et CPYNET [3] et leur association par Tomlinson ont donné naissance au premier prototype de l'email tel que l'on connaît aujourd'hui. Et voila plus de 50 ans après, le courriel électronique n'a pas pris une seule ride mais au contraire, il a révolutionné la manière de communication entre les individus en général, et dans les entreprises en particulier. Aujourd'hui, même avec l'apparition d'autres canaux modernes pour communiquer, le mailing reste le plus sollicité. En effet, en 2020, le nombre global des utilisateurs d'e-mail s'élève à 4 milliards de la population mondiale et il est estimé à devenir 4.6 milliards en 2025. Il convient de signaler que cette valeur représente presque le nombre des utilisateurs actifs des réseaux sociaux en 2020 qui est estimé à 4.48 milliards d'individu. De surcroît, il serait utile d'attirer l'attention sur le fait qu'environ 306.4 millions de mails sont envoyés dans le monde par jour, en autres termes, environ 3 millions d'e-mails sont expédiés chaque seconde.

Ce taux important rend le processus de l'acheminement du courriel dans les serveurs coûteux en temps. Le problème, qui peut aggraver la situation , est le fait qu'environ 65% des e-mails envoyés sont des courriels indésirables ou des spam. Il peut paraître difficile de prendre conscience de la gravité du sujet. C'est pourquoi nous souhaitons dans ce travail tenter d'expliquer la notion de spams et mettre en valeurs ses coûts. De surcroît, il serait utile d'attirer

1. ARPANET (advanced research projects agency network) était l'un des premiers réseaux informatiques, construit en 1969 comme un support robuste pour transmettre des données militaires sensibles et pour relier des groupes à la pointe de la recherche à travers le territoire des États-Unis

2. SNDMSG ou Send Message est un programme concu par Ray Tomlinson qui permet à deux utilisateurs connectés sur le même ordinateur de se laisser mutuellement des messages.

3. CPYNET,programme concu par Ray Tomlinson dont le nom est inspiré du nom anglais"copy net", peut envoyer des fichiers sur l'un ou l'autre des ordinateurs reliés par ARPANET

l'attention sur le fait que la circulation de ces pourriels peut encombrer inutilement le trafic réseau et rendre la réception et l'envoi des e-mails utiles très lents. Pour lutter contre ce problème, plusieurs sociétés ont commencé à implanter des filtres antispams dans leurs systèmes qui permettent d'éliminer la grande majorité des courriels électroniques non désirables.

Cependant, cette technique, même si elle paraît bénéfique, peut commettre des fautes. Par exemple, classer le courriel d'un employé comme un faux positif [4] peut causer des pertes financières à la société. Plus généralement, dans l'ère du marketing digital, une entreprise, légitime à communiquer avec ses clients et abonnés, peut se trouver incapable de les atteindre. À titre comparatif, seuls 79% des e-mails de marketing authentiques atteignent les boîtes de réception des abonnés. Plus sévèrement, tomber dans les pièges des filtres antispams peut conduire à la perte des clients puisque ces derniers peuvent perdre confiance en cette société si leurs e-mails sont considérés comme du spam, ce qui peut entraîner une baisse de la fidélité des clients et de la vente. De surcroît, l'entreprise peut également subir une réputation négative et une perte de clients potentiels et de prospect si son courrier électronique est considéré comme du spam.[30]

Dans ce contexte, ce travail a pour but d'aider le lecteur à ne pas tomber dans ces pièges en essayant de répondre à cette question :

Quelles sont les principales techniques utilisées dans les filtres antispam ?

Ceci étant dit, nous nous intéresserons dans une première partie à définir le courrier électronique et à expliquer son cheminement à travers le réseau informatique du point de vue technique. Dans la deuxième partie, nous tenterons d'introduire au lecteur la notion de spam et d'expliquer ses objectifs et les nuisances qu'il peut causer. Enfin, dans le troisième chapitre, nous exposerons les principales techniques de détection des courriels non sollicités.

4. Un faux positif est un e-mail légitime identifié à tort comme courrier indésirable par spamblocker

Chapitre 1

Description générale d'un e-mail et de son fonctionnement

Lorsqu'il s'agit d'aborder ce sujet, il est important de disposer de certains prérequis pour une analyse complète et rigoureuse. Dans cette section, nous allons établir les bases nécessaires pour une approche technique de ce sujet. Pour cette raison, cette partie sera consacrée à clarifier certaines notions pour le lecteur et à expliquer l'acheminement des courriels depuis son ordinateur jusqu'aux destinataires finaux, en signalant à chaque étape l'intervention des filtres anti-spam. En nous concentrant sur les prérequis pertinents, nous pourrons ultérieurement développer une compréhension complète des techniques de détection des spams et fournir des informations utiles aux parties prenantes concernées.

1.1 Description d'un courrier électronique :

1.1.1 Définition d'un e-mail :

Un courrier électronique, également nommé courriel, mail, ou e-mail (de l'anglais), est un message écrit envoyé électroniquement via un réseau informatique d'un expéditeur ou émetteur à un destinataire. Chaque mail doit respecter des règles normalisées afin d'autoriser sa transmission via la messagerie électronique et son dépôt dans la boîte aux lettres électronique du destinataire. Ces conditions seront explicitées dans des parties ultérieures. Tout d'abord, pour recevoir et consulter un courriel, l'utilisateur doit disposer d'un compte de mail électronique accessible à travers une adresse électronique unique et un mot de passe.

1.1.2 Définition d'une adresse électronique :

Une adresse électronique, adresse courriel ou adresse e-mail est une chaîne de caractères permettant d'acheminer du courrier électronique dans une boîte aux lettres informatique. Elle rend l'utilisateur identifiable et accessible dans le réseau des serveurs. Les adresses de courrier électronique utilisées sur Internet sont codées dans un nombre très limité de caractères, sous-ensemble de l'ASCII [1]. Un codage spécial appelé UTF-74 [2], surtout utilisé en Asie, permet néanmoins de représenter tous les caractères Unicode [3]en utilisant uniquement les caractères autorisés.[26]

Une adresse électronique est tout d'abord constituée des trois éléments suivants, dans cet ordre :

— Une partie locale, identifiant généralement une personne (lucas, Jean.Dupont, joe123) ou un nom de service (info, vente, postmaster) ;
— Le caractère séparateur @ (arobase), signifiant at (« à » ou « chez ») en anglais ;
— L'adresse du serveur, généralement un nom de domaine identifiant l'organisme (entreprise, association, mairie, université, voire individu) hébergeant la boîte électronique (exemple.net, exemple.com, exemple.org).

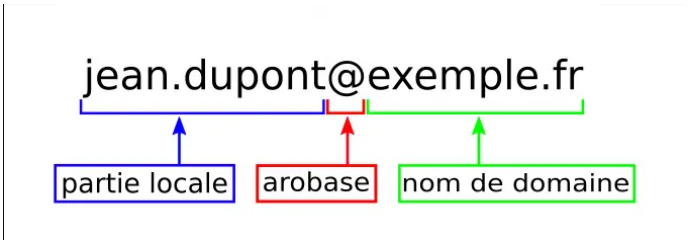

FIGURE 1.1 – Adresse électronique
[2]

1.1.3 Composition :

Un courrier électronique est composé de deux parties : l'entête et le corps du message, séparés par une ligne vide.

1. L'American Standard Code for Information Interchange ,plus connu sous l'acronyme ASCII , est une norme informatique de codage de caractères apparue dans les années 1960
2. abréviation de l'anglais Universal Character Set Transformation Format1 - 8 bits, est un codage de caractères informatiques conçu pour coder l'ensemble des caractères du « répertoire universel de caractères codés »,
3. Unicode est un standard informatique qui permet des échanges de textes dans différentes langues, à un niveau mondial.

1.1.3.1 L'en-tête :

Les champs d'entête sont composés de plusieurs paramètres, certains essentiels pour le courrier électronique :

- From : c'est l'adresse électronique de l'expéditeur ou de l'émetteur du message. Ex : expediteur@expediteur.com
- To : adresse électronique du destinataire, autrement dit l'adresse à qui le message est envoyé. Ex : destinataire@destinataire.com
- Date : c'est la date d'envoi du courrier. Dans le cas où aucune date n'est indiquée, le serveur de messagerie se chargera de l'ajouter.
- Received : il s'agit des informations sur tous les serveurs de messagerie traversés par le mail, serveur expéditeur, serveurs intermédiaires et serveur destinataire. Ce champ comprend les noms, adresses IP et date du courrier électronique traité par tous les serveurs.
- Reply-To : adresse alternative à celle du champ From pour recevoir une réponse.
- Subject : le sujet du message.
- Message-ID : référence qui permet d'identifier de manière unique le message.

Une partie de ces informations est générée de manière automatique par le logiciel de messagerie MUA et d'autres sont ajoutées directement par le serveur de messagerie ou les relais de messagerie intermédiaire lors de l'acheminement de l'e-mail.

1.1.3.2 Corps du courriel :

En dessous de cet en-tête se trouve le corps du message. Celui-ci comporte le contenu de l'email, soit le texte mais aussi tous les documents attachés tels que des images, des fichiers Office et autres.

1.2 Acheminement d'un courriel :

Cette partie expliquera les étapes techniques par lesquels un courriel doit passer pour arriver sur l'écran du destinataire. En effet, techniquement parlant, le processus de l'envoi de mail se décompose en quatre étapes :

1. L'envoi de mail par un MUA.
2. Contrôle d'un mail via un MTA.
3. Processus de l'envoi jusqu'au MTA destinataire.
4. Contrôle au niveau de MTA cible et réception de mail.

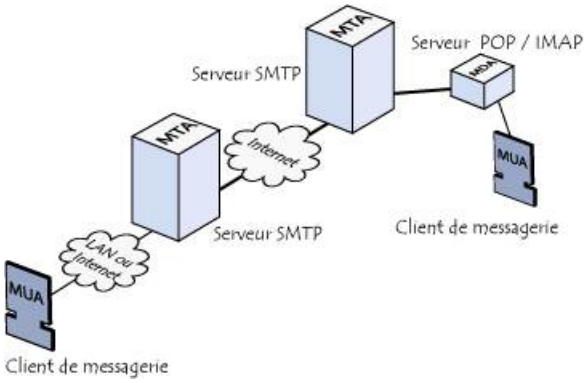

FIGURE 1.2 – Acheminement d'un mail

1.2.1 Envoi d'un email par un MUA

1.2.1.1 Définition d'un Mail User Agent

Un MUA, où littéralement Mail User Agent, est un logiciel qui permet à l'utilisateur d'accéder à sa ou ses boîtes de courrier électronique, lui donnant ainsi la possibilité d'écrire, envoyer et lire des courriels. Il peut être soit sous la forme d'un logiciel appelé client de messagerie (à l'instar de Mozilla Thunderbird, Microsoft Outlook, Eudora Mail ou Lotus Notes), soit sous la forme d'un site du type Webmail (à l'instar de Gmail ou Yahoo), permettant l'accès aux messages depuis n'importe quelle connexion Internet via un navigateur web.[4]

1.2.1.2 Envoi d'un email par un MUA

La première étape dans l'envoi de mail réside dans l'accès à un MUA. Ainsi, l'utilisateur a la possibilité de choisir son destinataire, l'objet de son courriel et de saisir le corps de son message. Une fois la saisie est achevée, et juste avoir la pression du bouton « envoyer », le processus d'expédition est déclenché. Après avoir envoyé l'e-mail, le client de messagerie convertit le message par la suite en deux catégories, à savoir l'en-tête et le corps de l'e-mail et le charge sur le serveur de messagerie sortant : un serveur SMTP sous la forme d'une session SMTP.

1.2.2 Contrôle d'un email via un MTA

1.2.2.1 Définition d'un Simple Mail Transfer Protocol SMTP

Simple Mail Transfer Protocol ou littéralement » protocole simple de transfert de courrier » est un protocole de communication utilisé pour transférer les courriers électroniques vers les serveurs de messagerie électronique. Il s'agit d'un protocole simple se basant sur ces trois étapes : la première étape réside dans la spécification de l'expéditeur du message, la deuxième

dans la spécification des destinataires. Et une fois leurs existences sont vérifiées, la troisième étape se conclut dans le transfert du corps de l'e-mail.

1.2.2.2 Définition d'un serveur SMTP

Le serveur SMTP est tout simplement le serveur qui sert à envoyer les mails de l'expéditeur vers un serveur destinataire. Chaque serveur SMTP admet une adresse IP. En effet, une adresse IP où littéralement adresse d'Internet Protocole est un numéro d'identification qui est attribué de façon unique soit de manière permanente ou provisoire à chaque périphérique relié à un réseau informatique qui utilise l'Internet Protocole. L'adresse IP est à la base du système d'acheminement des paquets de données sur Internet. Elle a un impact direct sur la délivrabilité des emails puisque c'est à partir d'elle que les fournisseurs d'accès d'Internet vont identifier et déterminer la réputation de l'envoyeur.

1.2.2.3 Définition d'un MTA

MTA ou également Message Transfer Agent (agent de transfert de message) est un logiciel pour serveur de transmission de courriels. Son rôle consiste à la réception, habituellement par le protocole SMTP, des emails envoyés soit par des clients de messagerie électronique (MUA), soit par d'autres MTA et de les redistribuer à des Mail Delivery Agent et/ou d'autres MTA.

1.2.2.4 Définition d'un serveur DNS

Les serveurs de noms (DNS)[4] jouent un rôle primordial dans la direction du trafic sur Internet. Pour commencer, il faut bien signaler que les équipements (ou hôtes) connectés à un réseau IP, comme Internet, possèdent une adresse IP qui les identifie sur le réseau. Ces adresses sont numériques afin de faciliter leur traitement par les machines.Pour faciliter l'accès aux hôtes sur un réseau IP, un mécanisme a été mis en place pour associer un nom à une adresse IP. Ce nom, plus simple à retenir qu'une suite de chiffres, est appelé « nom de domaine ».
La tâche des serveurs DNS consiste à la résolution d'un nom, en autres termes à trouver l'adresse IP qui lui est associée.
À titre explicatif, en tapant l'URL d'un site web (texte.com) dans la barre d'adresse et en appuyant sur la touche Entrer, le navigateur envoie une requête aux serveurs de noms de ce domaine que le répond avec l'adresse IP du serveur du ce site web. Ainsi,le navigateur demande le contenu du site Web auprès du serveur associé, le récupère et l'affiche sur l'écran de l'utilisateur.[13]

4. Domain Name System

Cependant, les serveurs DNS font bien plus que convertir un nom de domaine en adresse IP. Sur demande, ils fournissent diverses informations nécessaires à l'échange simple et sécurisé de données sur Internet. Ces informations sont stockées dans ce qu'on appelle des enregistrements de ressources DNS.

1.2.2.5 Fonctionnement

L'interaction entre le MUA et le serveur SMTP se fait par l'intermédiaire d'une session SMTP qui contient toutes les informations contenues dans le courriel.

```
telnet smtp.----.---- 25
Connected to smtp.----.----.
220 smtp.----.---- SMTP Ready
HELO client
250-smtp.----.----
250-PIPELINING
250 8BITMIME
MAIL FROM: <auteur@yyyy.yyyy>
250 Sender ok
RCPT TO: <destinataire@----.---->
250 Recipient ok.
DATA
354 Enter mail, end with "." on a line by itself
Subject: Test

Corps du texte
.
250 Ok
QUIT
221 Closing connection
Connection closed by foreign host.
```

FIGURE 1.3 – Session SMTP
[3]

Un serveur de mail dispose d'un programme constamment actif dans le but de récupérer les emails et de les envoyer : le Mail Transfer Agent ou MTA. Ce MTA représente la base de logiciel d'un serveur de messagerie. Il commence par contrôler si une adresse de messagerie est correcte (si elle respecte la forme énoncée ci-dessus). Puis, le MTA recherche le serveur de messagerie

du destinataire. Pour cela, il effectue une requête dans le système de noms de domaines sur la partie « nom de domaine ». Cette requête permet de récupérer l'adresse IP du serveur de messagerie du domaine. À ce moment, il ouvre une connexion TCP avec ce serveur distant, connexion sur laquelle on transmet l'adresse de l'expéditeur, celle du destinataire et le message. Le serveur distant n'est pas forcément la destination finale, il peut être un simple relais ou une passerelle. Le serveur qui a accepté un message doit le délivrer ou bien prévenir l'expéditeur en cas d'échéance. Une fois, le serveur correspondant au domaine est identifié, il faudra vérifier que la partie locale de l'adresse existe bel et bien. Si l'adresse email est introuvable que ce soit pour des raisons de fautes de frappe ou autres, l'email est renvoyé par le MTA à l'expéditeur avec une annotation. Si tout est en ordre, le MTA de l'expéditeur transfère le mail au MTA du destinataire. Mais, avant que le MTA du fournisseur de messagerie n'envoie l'email, la taille maximale autorisée de l'email est contrôlée. Si les pièces jointes sont trop lourdes, l'expéditeur en sera informé et l'envoi sera alors interrompu. La plupart des fournisseurs de messagerie passent les messages au crible à la recherche de spams et de logiciels malveillants tels que des virus ou des chevaux de Troie avant de les envoyer. Si le mail ne contient ni spam, ni logiciel malveillant et si sa taille est conforme, le MTA enregistre alors le message.[25]

1.2.3 Le processus d'envoi jusqu'au destinataire MTA

Une fois toutes les vérifications sont effectuées avec succès, le MTA de l'expéditeur transfère le mail au MTA du destinataire. Pour se faire, le message est scindé en plusieurs parties, à l'instar d'un transfert de données lambda sur le Web. L'envoi de paquets de données (dont la taille maximale est de 64 KB) présente plusieurs avantages. D'un côté, de petits envois peuvent facilement être réalisés. D'un autre côté, les parties d'un email peuvent utiliser plusieurs chemins de données afin d'atteindre le serveur de mail cible. Les paquets ont toujours recours aux chemins d'accès où le trafic est faible. Le trafic entre les serveurs de messagerie se base sur l'utilisation de nœuds d'échange Internet. Les paquets d'un email atterrissent via ces nœuds de distribution au serveur de messagerie du destinataire et seront ensuite assemblés.

1.2.4 Un second contrôle à la cible MTA et réception de l'email

1.2.4.1 Définition d'un MDA

Un agent de distribution du courriel, en anglais Mail Delivery Agent souvent abrégé par son sigle MDA, est le logiciel qui intervient dans la dernière étape du processus de distribution d'un courrier électronique. Il est chargé de la déposition du courriel dans la boite à messages du destinataire. Pour cela, il est souvent considéré comme le point final d'un système de

messagerie.[4]

1.2.4.2 Fonctionnement

L'email se trouve désormais sur le serveur de messagerie du destinataire. Avant que celui-ci ne lui parvienne, le MTA du serveur de la messagerie cible contrôle le message entrant. La taille des pièces jointes est à nouveau contrôlée. Si celle-ci dépasse la capacité maximale du serveur de mail du destinataire, le message ne sera alors pas réceptionné mais renvoyé. Par ailleurs, le processus de vérification des spams aura lieu une seconde fois. Si un filtrage d'enveloppe et de fichier attachées identifie le courriel comme un spam, celui-ci est marqué en conséquence et le transfert du message est entièrement interrompu. Si le mail passe ces contrôles avec succès, il sera enregistré sur le serveur de mail du destinataire qui pourra alors le charger. Pour cela, un MDA envoie le message à la messagerie du destinataire. Lorsque celui-ci le réceptionne, le MUA utilisé recourt au MTA du serveur de messagerie. Son contenu est alors à nouveau contrôlé à la recherche de spams ou de logiciels malveillants, cette fois par le client de messagerie du destinataire. Par la suite, le mail peut être ouvert et lu à partir de la boîte de réception.

Ce chapitre est consacré à la découverte du monde du courrier électronique. Il commence par une description générale de l'e-mail et de sa composition afin de fournir les connaissances nécessaires pour comprendre le fonctionnement des filtres anti-spam. Nous avons également expliqué dans ce chapitre le parcours que suit un courriel électronique pour atteindre sa destination et précisé à quelles étapes interviennent les filtres anti-spam. Cependant, la répétition du mot "spam" peut amener à se demander sa définition exacte et ses impacts. Dans le chapitre suivant, nous tenterons de répondre à ces questions.

Chapitre 2

Définition et caractéristiques d'un Spam

Le courrier électronique est à coup sûr l'un des moyens de communication les plus puissants sur Internet et les plus faciles à manipuler également. Cependant, il n'a pas duré longtemps pourque des personnes malveillantes en fassent mauvais usage et créent ce que l'on nomme aujourd'hui le « Spam ».

Ce chapitre va être consacré à la définition du spam ainsi qu'à citer ses différents types, ses principaux objectifs et finalement à donner son impact sur les entreprises aussi bien que sur le niveau individuel.

2.1 Définition générale

Selon la CNIL [1], le spam est l'envoi massif de courriers électroniques non sollicités à des personnes avec lesquelles l'expéditeur n'a jamais eu contact et dont il a obtenu l'adresse électronique d'une façon souvent illégale.

2.2 Contexte historique

Le mot *Spam* est la contraction de l'expression *Spiced Ham* qui veut dire en français *Jambon épicé* introduite à la suite du concours organisé par la société américaine *Hormel Foods* [2] en 1937. Le terme « Spam » tel qu'on l'emploie aujourd'hui désignant un courrier indésirable provient d'un sketch intitulé « Spam » de Monty Python en 1970. Sa première utilisation sur Internet a vu le jour en 1994 lorsqu'un internaute a critiqué la pratique du Spam. La première

1. La Commission nationale de l'informatique et des libertés
2. Hormel Foods Corporation est une entreprise agro-alimentaire américaine basée à Austin dans le sud-est du Minnesota et connue pour ses produits de masse et surtout la marque de viande précuite SPAM commercialisée depuis 1926

pratique du spam était en 1978 par un certain Gary Thuerk, marketeur chez DEC [3], sur le réseau ARPANET [4], et ce en envoyant son message à plus de 600 personnes les invitant à découvrir sa nouvelle machine le 2020. Eventuellement,le gouvernement américain a jugé cet usage comme étant une violation des règles de l'utilisation d'ARPANET.[5, 16, 31]

2.3 Les objectifs et les différents types de Spam

Initialement, les messages Spam étaient uniquement des spams publicitaires dont le seul but était de promouvoir les produits d'une entreprise quelconque et d'en faire la publicité. Au fil du temps, le pourriel [5] s'est amplement développé, et ce pour atteindre des fins de plus en plus hostiles. Voici donc une liste non exhaustive de ses différents types :

2.3.1 L'hameçonnage (Phishing en anglais)

Il consiste à tromper le destinataire en se faisant passer pour des organismes connus par lui, l'objectif dans ce cas est de dérober les données personnelles de l'utilisateur notamment les mots de passe, les numéros des cartes bancaires, etc. C'est le type de spam le plus fréquent.[5, 34]

2.3.2 La publicité

Il s'agit dans ce cas de faire l'éloge d'un produit donné. Ce dernier peut-être un produit cosmétique, pharmaceutique, un logiciel etc.[5, 34]

2.3.3 Le Scam (L'escroquerie)

Ce type est appelé aussi « fraude 419 [6] ». Dans ce genre d'emails l'expéditeur,par exemple, prétend qu'il possède une grosse somme d'argent et indique qu'il a besoin de transférer cette somme en utilisant un compte existant tout en promettant le destinataire de toucher une part de l'argent, ce qui n'est jamais le cas. Un exemple très connu est celui de « The Nigerian Scam ».[5, 34]

3. Société à responsabilité limitée, active dans le secteur d'activité de la fabrication de matériel de distribution et de commande électrique

4. ARPANET (advanced research projects agency network) était l'un des premiers réseaux informatiques, construit en 1969 comme un support robuste pour transmettre des données militaires sensibles et pour relier des groupes à la pointe de la recherche à travers le territoire des États-Unis

5. Contraction de pourri et de courriel, ce terme désigne les spams arrivant dans les boîtes mails en grande quantité

6. La dénomination 419 vient du numéro de l'article du code nigérian sanctionnant ce type de fraude.

2.3.4 Le canular

L'objectif est de diffuser et faire véhiculer une information de nature urgente comme des fausses alertes par l'intermédiaire d'un lien hypertexte. En effet, les spammeurs mettent la compassion et l'empathie du destinataire en jeu afin d'en tirer profit.[5, 34]

2.4 Le coût du Spam

Le Spam étant un phénomène mondial et massif, il est de plus en plus difficile de lui échapper. En fait, les dangers du pourriel concernent les entreprises aussi bien que les particuliers, c'est pour cela qu'on a dédié ce paragraphe pour reconnaitre leurs effets sur ces derniers.

2.4.1 Le coût du spam pour le spammeur

En ce qui concerne l'émetteur, le spam n'entraine aucun coût. En fait, les seuls coûts engendrés par l'envoi du spam sont l'entretien de répertoire d'émission et la formulation du contenu du message. De plus, à fin d'envoyer les messages spams, l'expéditeur emploie des logiciels d'envoi disponibles à des prix accessibles. [18]

2.4.2 Le coût du spam pour les organisations

Selon une étude faite par Ironport Systems [7] en 2007, un salarié dépenserait 5 à 10 minutes à administrer son spam. Ce phénomène qui ne cesse de s'accentuer dissipe vertigineusement de ressources FAI [8].[21]

2.4.2.1 Les coûts directs

En premier lieu, ils sont étroitement liés à la réception (largeur de la bande passante, taux de saturation) et à la sauvegarde du spam sur les serveurs. En effet, dans le cas général chaque message reçu est stocké sans que l'on sache ou bien sans que l'on ait besoin d'avoir recours à lui, ce qui peut être très embêtant et qui peut également nuire à la sauvegarde des données au sein de l'entreprise. En fait, l'encombrement anormal des boîtes de réception induit dans certains cas limites l'indisponibilité du service de messagerie et qui peut, par conséquent, se traduire par des pertes affreuses pour l'entreprise en question.

En second lieu, le pourriel réduit la productivité et la concentration des salariés. En effet, les employés passent leurs temps à gérer les courriers indésirables au lieu de se focaliser sur

7. Ronport Systems, dont le siège est à San Bruno, en Californie, était une société qui concevait et vendait des produits et services destinés à protéger les entreprises contre les menaces Internet
8. Fournisseur d'Accès à Internet

leurs tâches normales et donc c'est du temps perdu qui entraine sans l'ombre d'un doute des pertes atroces pour l'entreprise. D'après une étude faite par NucleusReasearch [9] et KnowledgeStorm [10] en 2007, le Spam engendre des pertes qui valent aux alentours de 712 dollars par an et par employé pour les sociétés américaines.[21]

2.4.2.2 Les coûts indirects

Les pourriels ont également un coût social se traduisant premièrement par une baisse de la volonté des utilisateurs vu que la suppression des spams constitue une tâche souvent pénible pour eux. Deuxièmement, c'est la réputation de l'entreprise qui peut être endommagée puisqu'elle a une grande responsabilité envers la défense de l'individu contre les messageries à contenus choquants.[21]

Nous avons vu tout au long du chapitre ce que c'est le spam ainsi que ses différents types et objectifs. Nous avons pu aussi en tirer les dangers que le spam pose sur les entreprises, les organisations et les particuliers. Dans le chapitre suivant, nous allons nous concentrer sur les techniques de filtrages du spam et les principaux filtres antispams.

9. Nucleus Research est une entreprise qui fournit des services de recherche et de conseil technologiques axés sur le retour sur investissement à l'échelle mondiale

10. KnowledgeStormest une société qui fournit une ressource de recherche en ligne pour des solutions et des informations technologiques.

Chapitre 3

Les filtres anti-spams

Dans cette section, nous allons explorer les techniques de détection de spam les plus importantes. Comme nous l'avons expliqué précédemment dans le premier chapitre, un courriel électronique est constitué de deux parties distinctes : l'en-tête et le corps. Ceci étant dit, les développeurs ont essayé d'explorer cette répartition et ont mis en place deux types de filtrage : le filtrage d'en-tête et le filtrage de contenu.

Dans la première partie de cette section, nous allons nous concentrer sur les différentes techniques de filtrage d'en-tête. Dans la seconde partie, nous nous pencherons sur les techniques de filtrage de contenu. Ces deux types de filtrage sont essentiels pour détecter les spams, et nous allons détailler chaque technique pour fournir une compréhension complète de leur fonctionnement.

3.1 Filtrage d'enveloppe

Le filtrage de contenu s'applique uniquement à l'entête du message, qui contient souvent assez d'informations pour pouvoir distinguer un spam. En effet, les informations contenues dans l'entête sont très importantes pour la simple raison que l'envoi de corps mail n'est effectué que si l'en-tête a été reçue et acceptée. Le logiciel de détection de spam peut dès le début détecter les mails indésirables et les bloquer avant même que leur corps ne soit envoyé. Ce qui a pour bénéfice de diminuer considérablement le trafic sur la passerelle SMTP et gagner du temps. De surcroît, le taux de faux positifs dans ce type de filtrage est quasiment nul. En effet, lorsqu'un filtre d'enveloppe a identifié un courriel comme du spam, il se trompe rarement. Dans ce but, ce type de filtre repose sur plusieurs techniques.

3.1.1 Liste blanche

3.1.1.1 Définition

Une liste blanche définit en général un ensemble d'entités (personnes, domaines, machines, etc.) auxquels on attribue un niveau de liberté ou de confiance maximum dans un système particulier.

3.1.1.2 principe

L'idée de ce type de filtrage repose sur le fait qu'il est plus simple de cibler les personnes à qui on veut donner la permission de nous envoyer des courriels que d'énumérer ceux dont on ne veut pas en recevoir. Pour cette raison, la technique la plus radicale pour trier ses courriels est encore celle des listes blanches (white-list). En effet, cette méthode consiste à établir manuellement ou semi-automatiquement une liste de contact (adresses emails, noms des domaines, adresse IP) en qui on a confiance et de qui on souhaite accepter des mails. De ce fait, les courriels envoyés par les personnes de confiance sont classés dans la boîte de réception et les autres sont envoyés dans un sous-répertoire pour les courriels indésirables. Cette liste blanche peut-être mise à jour soit manuellement en ajoutant les informations de l'expéditeur, soit automatiquement puisque tout destinataire à qui on envoie un mail sera systématiquement ajouté à cette liste.

3.1.1.3 Efficacité

En considérant le fait que la liste blanche est générée de façon manuelle ou semi-manuelle, cette méthode semble gourmande en terme de base de données. En effet, cette technique nécessite un important réseau social pour s'assurer qu'aucun mail légitime n'est classifié comme spam. De plus, une mise à jour régulière pour ajouter les nouveaux contacts est indispensable et coûteuse en temps. Cependant, cette méthode est très avantageuse si le compte mail est destiné pour communiquer avec un réseau restreint de destinataires. Dans ce cas, elle permet de lutter indirectement contre le taux de mauvaise classification (Faux Positif).

3.1.2 SPF (Sender Policy Framework)

3.1.2.1 Définition

Afin de comprendre cette méthode, il faut bien rappeler que le système de courrier électronique actuel repose sur le protocole simple de transfert de courrier (SMTP). Le fait que le

SMTP, tel qu'il est spécifié dans les RFC 5321 et RFC 5322, n'a pas de mécanisme d'authentification intégré, ce qui permet à tout utilisateur de s'identifier avec le nom de domaine de son choix. Cette propriété indulgente donne aux spammeurs la possibilité d'imiter des adresses électroniques. En effet, 99 % des spams sont émis via des serveurs piratés et/ou non dédiés à l'envoi des courriels. Pour surmonter le problème de l'imitation, l'authentification de l'expéditeur a été proposée dans un certain nombre de méthodes dont la plus utilisée est "Sender Policy Framework". SPF utilise la zone DNS (Domain Name System) d'un domaine pour authentifier les adresses IP d'envoi.

3.1.2.2 Principe

Le détenteur d'un domaine ajoute, dans la zone DNS de ce domaine, un enregistrement du type TXT qui indique quels sont les serveurs autorisés ou non à envoyer du courriel pour ce domaine. Ainsi, si mail.domainea.com est le seul serveur autorisé à envoyer du courriel pour domainea.com, cette information sera spécifié dans l'enregistrement TXT. Le système vérifie que le serveur envoyant le courriel est bien dans la liste des serveurs autorisés. Sinon, il s'agit d'un spam.
Pour fonctionner correctement, le support SPF doit être activé sur le filtre antispam.[29]

3.1.3 Liste noire

3.1.3.1 Definition

Une liste noire (en anglais blacklist), par opposition à une liste blanche, est une liste rassemblant les identifiants virtuels d'individus ou d'entités jugées indésirables, par une personne, un groupe ou une organisation donnée.[5]

3.1.3.2 Principe

Par analogie à la liste blanche, une liste noire est composée d'adresses de messagerie, de noms de domaines ou d'adresses IP à qui on souhaite bloquer l'envoi de mail. En d'autres termes, tous les expéditeurs répertoriés dans cette black List ne peuvent plus envoyer d'autres données auprès du destinataire, que ce soit par email ou via un formulaire de contact puisque tout email qu'il envoie est considéré comme indésirable.

3.1.3.3 Efficacité

Cette technique permet la réception des emails de nouveaux expéditeurs sans avoir l'obligation de les ajouter à la liste blanche ce qui garantit plus d'efficacité dans le travail. Cependant,

à l'instar de la liste blanche, une telle liste nécessite un maintien régulier pour ajouter les nouveaux contacts indésirables.

3.1.4 RBL (Realtime Blackhole List)

3.1.4.1 Definition

La Realtime Blackhole List (RBL) peut être définie comme une grande liste noire généralisée. Elle a pour but de fournir une liste de serveurs réputés comme principaux envoyeurs de spam. Le RBL a été mis en œuvre pour la première fois par Mail Abuse Prevention System (MAPS) [1], qui ont été acquises par Trend Micro [2] en 2005. [27]

3.1.4.2 Principe

Cette liste contient les adresses IP obtenues par diverses méthodes. L'une de deux techniques les plus utilisées est la détection de serveurs ne respectant pas les préconisations de la RFC [3], les serveurs ouverts à tout le monde ou encore des serveurs hébergés dans certains pays et qui sont susceptibles d'être considérés comme spammeurs. L'autre technique consiste à donner un score à chaque serveur. Si un courriel, issu d'un serveur spécifique, était marqué comme spam, le serveur de messagerie le prendrait en compte. Si plusierus personnes font de même, au bout de plusieurs dénonciations, le serveur d'envoi est listé dans RBL.
Cette méthode permet de construire une liste noire généralisée. Lorsqu'un filtre de détection de spam reçoit un courriel, il vérifie si le serveur d'envoi est contenu dans un RBL. Si oui, le courriel est catégorisé comme spam.

3.1.5 La liste grise

3.1.5.1 Définition

Le Greylisting ou l'utilisation de la liste grise est une technique antispam particulièrement efficace qui repose sur le principe RFC 5321. Ce principe postule qu'un serveur de réception de courriel, s'il ne peut pas traiter la réception d'un message, doit retourner un code 421. Ce code d'erreur indique au serveur qui envoie le courriel d'attendre un certain délai et de réessayer l'envoi un peu plus tard.[6]

1. MAPS est une organisation qui fournit un soutien anti-spam en tenant à jour des listes noires.
2. Trend Micro : une société de développement des logiciels de sécurité pour les serveurs et les environnements de cloud computing
3. Request For Comments :ce sont un ensemble de documents qui font référence auprès de la Communauté Internet et qui décrivent, spécifient, aident à l'implémentation, standardisent et débattent de la majorité des normes, standards, technologies et protocoles liés à Internet et aux réseaux en général

3.1.5.2 Principe

Remarquant cette particularité, les experts de la sécurité du courriel ont donc envisagé une technique de détection de spam exploitant ce principe : la liste grise. Cette méthode fonctionne avec une base de données. En effet, chaque enregistrement de la base de données constitue un triplet composé de l'adresse IP du serveur qui envoie le courriel, de l'adresse courriel de l'expéditeur, et de l'adresse courriel du destinataire, formant ainsi une clé unique. Cette clé est ensuite stockée dans la base de données accompagnée avec la date de la première connexion de ce triplet au serveur. Ainsi, lorsqu'un message est reçu par le serveur de courriel du destinataire, ce dernier vérifie dans sa base l'existence du triplet.

• Si le triplet n'est pas dans sa base de données, il l'ajoute avec la date actuelle. Il renvoie ensuite le code d'erreur 421, indiquant au serveur qu'il devra renvoyer le message.

• Si le triplet est déjà dans la base de données, le serveur vérifie le délai entre la date courante et celle stockée dans la base (la date de la première connexion). Si le délai est supérieur ou égal à un délai prédéfini par le destinataire (par exemple, 5 minutes), le message est accepté. Sinon, le serveur retourne un numéro d'erreur 421. Après un certain temps (défini également dans l'enregistrement), l'enregistrement devient inactif et le serveur doit renvoyer un 421. Ainsi, lorsque le MTA envoyeur reçoit le 421, s'il est légitime, il attendra un délai prédéfini dans la configuration du serveur expéditeur du message (ou Mail Transfer Agent) avant de renvoyer le message. Sinon, dans le cas du MTA non légitime (utilisé par les spammeurs), il n'attendra pas et le renvoie beaucoup plus tôt que le délai passe ou pire il ne le renverra pas puisque la très grande majorité des spammeurs préfère sacrifier un courriel et viser d'autres destinataires plutôt que d'attendre et ainsi, diminuer leur performance.

3.1.5.3 Efficacité

Cette méthode est une des méthodes qui permet d'atteindre des taux d'efficacité très élevés. En effet, ce taux était de l'ordre de 99 % quand il a été proposé en 2003. Même si cette efficacité est moins importante due à l'augmentation de l'utilisation des Webmail par les spammeurs, elle reste quand même de l'ordre de 80-90 %. De plus, cette technique a l'avantage d'être facile à implanter et à utiliser à condition que l'administrateur système soit prêt à accepter les délais chaque fois qu'un email d'un nouvel expéditeur est envoyé.

3.1.6 Mail en masse

La détection de spam par l'envoi en masse est une technique couramment utilisée pour détecter les pourriels. Cette technique consiste à surveiller le nombre d'e-mails envoyés par une adresse ou un domaine en particulier, et à déterminer si ces e-mails sont considérés comme

non sollicités en fonction du nombre d'e-mails envoyés en une courte période de temps. Si un nombre élevé d'e-mails est envoyé en peu de temps, il y a de fortes chances que ces e-mails soient du spam. Les fournisseurs de messagerie, les filtres antipollupostage et les logiciels de sécurité utilisent généralement cette technique pour détecter les pourriels. Ils peuvent également surveiller d'autres indicateurs, tels que les expéditeurs connus pour envoyer du spam et les mots-clés qui lui sont associés, pour déterminer si un e-mail est considéré comme indésirable. Si un e-mail correspond à plusieurs critères de détection de spam, il sera souvent classé comme un pourriel et sera généralement déplacé dans un dossier de spam ou bloqué complètement. Il faut bien noter que les serveurs SMTP des Webmail et messagerie classique à l'instar de Gmail, free, Hotmail... ne sont pas adaptés à l'envoi de messages en masse. Leurs adresses IP ne sont pas surveillées. Ils sont partagés avec des spammeurs. Pour cette raison, si des mails sont envoyés en masse depuis des messageries classiques, ils finiront donc immanquablement dans le répertoire spam.

Dans ce chapitre, nous avons examiné les techniques de filtrage des courriels électroniques basées sur l'analyse de l'enveloppe. Cette méthode repose sur le fait que le corps du courriel n'est transmis que si l'adresse IP de l'expéditeur est reçue et acceptée. Cette approche permet d'obtenir des résultats de faux positifs presque déterministes, cependant, une question se pose :

Comment un logiciel peut-il détecter un courriel indésirable si celui-ci parvient à passer le filtrage d'enveloppe avec succès ?

3.2 Filtrage de Contenu

3.2.1 Introduction

Ce type de filtrage, comme son nom l'indique, consiste à analyser le contenu des messages, à titre d'exemples les images, les mots utilisés, les adresses URLs, le code HTML, etc. En effet, ces filtres s'appliquent dans le cas où le filtrage d'enveloppe n'a pas dévoilé des informations suffisantes pour considérer le mail comme spam. De plus, ils sont effectués en complémentarité avec ceux d'enveloppe et permettent en fin de compte de classer les messages non traités. Les filtres de contenu sont beaucoup plus sensibles que les filtres d'enveloppe, et ce grâce à la nature subjective des donnés qu'ils analysent.

Dans cette partie, nous nous focaliserons sur les principaux filtres et techniques de filtrage de contenu tels que le filtrage par mots-clés, le filtrage heuristique, etc. Nous introduirons par la suite la notion d'apprentissage automatique ainsi que les principaux types d'algorithmes d'apprentissage tout en citant les différentes méthodes se basant sur le Machine Learning et tout en indiquant leur importance majeure dans la détection des pourriels.

3.2.2 Filtrage par mots-clés

3.2.2.1 Principe

Cette technique consiste à construire une liste noire de mots-clés qui sont employés dans les spams d'une manière très probable et qu'on souhaite les détecter par la suite. Éventuellement, après avoir associé à chaque mot trouvé dans le courriel, un score est calculé en vue de classer le message en spam ou en ham.[5, 34]

3.2.2.2 Avantages et inconvénients

L'avantage principal de cette méthode est la rapidité de l'analyse réalisée sur la liste .Toutefois, cette technique présente plusieurs inconvénients. En premier lieu, le message doit être balayé plusieurs fois de suite vu le nombre colossal d'expressions disponibles. En deuxième lieu, la liste doit être régulièrement mise à jour étant donné que le spam subit constamment des transformations et ceci est dû principalement au fait que les spammeurs veulent à tout prix plus de profits. Or cette tâche s'avère très compliquée et complexe voire impossible essentiellement à cause de la taille considérable de la liste. En d'autres termes, pour chaque mot-clé, on peut trouver plusieurs variantes c'est-à-dire qu'on peut avoir plusieurs combinaisons possibles de lettres formant le mot considéré sans lui changer la signification et le sens. À titre d'exemple, voici une liste non exhaustive des variantes qu'on peut trouver pour le mot « money » :

— M.O.N.E.Y
— MoNeY
— M*O*O*N*E*E*Y*Y

Enfin, le danger de cette méthode réside dans l'obtention d'un taux considérable de faux positifs. [22, 24, 10]

3.2.3 Filtrage des adresses URLs

3.2.3.1 Principe

Les adresses URL sont fortement présentes dans les spams vu leur rôle central dans la mise en place d'une connexion entre le prédateur et la proie. De plus, elles confèrent à la cible la possibilité de disposer de plus de renseignement sur le produit en question. C'est pour cela qu'il est tout à fait légitime de créer des filtres anti-pourriels qui ciblent les adresses URLs. Le filtrage des liens hypertextes consiste donc à balayer le message au plus une fois et exhiber simulatnément toutes les adresses URL existentes c'est-à-dire une liste noire de liens hypertextes rassemblés par les moteurs de recherche qui leur semblent suspects ou bien dangereux pour enfin de compte mettre à l'épreuve leur existence dans une base de donnée bien connue à l'avance. Par exemple, la base d'URLs utilisée par ze-filter [4] est celle de subl.org. Cette base de données dispose grosso modo de plus de 20000 entrées où le taux de détection est supérieur ou égal à 70 % avec une très faible erreur relative (taux très faible de faux positifs). [17, 19]

3.2.3.2 Les avantages

Le point fort de ce filtre, à l'opposition de ceux des mots-clés et des expressions régulières (ce filtre fait l'objet du paragraphe suivant), réside dans le parcours rapide de la base de données concernée malgré sa taille gigantesque dans la plupart des cas. [17]

3.2.4 Filtrage par les expressions régulières

Les expressions régulières ou rationnelles sont des critères ou des modèles de recherche (en anglais pattern) dans les chaines de caractères. L'objectif d'utilisation de telles expressions est de sélectionner ou de faire des substitutions sur les chaines de caractères. En d'autres termes, l'emploi des expressions régulières permet à partir d'une simple ligne de code de fournir à partir d'un mot et donc d'une chaine de caractères de trouver ses différentes variantes ce qui augmente considérablement la probabilité de détecter les spams. Par exemple, si on considère

4. C'est un logiciel de filtrage de courrier gratuit développé par José Martins da Cruz. Il a été connu sous le nom de J-chkmail. Ce filtre se base sur l'analyse comportementale et l'analyse du contenu.

le mot "Viagra" et si on suppose de plus que le spammeur veut déjouer un filtre par mots-clés en utilisant le mot "Viaaagraa" alors la commande, par exemple sous Linux, qui permet de contourner ce mauvais tour est [17, 19] :

$ grep " ^Vi+a+gra+$"

3.2.5 Filtrage d'image

3.2.5.1 Principe

En général, les pièces jointes présentent un obstacle important pour les filtres anti-spam. Bien que la détection de spam à partir de l'analyse des images contenues dans le courriel soit une tâche presque irréalisable, il existe bien une multitude de techniques permettant de vérifier si les images ont été déjà utilisées dans un spam ou non. On peut donc tirer de ce qui précède que le filtrage d'image réside en fait dans l'analyse des images obtenues dans les messages au niveau des propriétés du fichier d'image (format, taille du fichier, taille de l'image) ainsi que le contenu de l'image (couleurs , taille des pixels). Une autre technique pour déterminer la nature du mail consiste à considérer la répartition et la distribution des fichiers images dans le message textuel et donc on cherche à savoir dans quel contexte elles ont été employées. Par ailleurs, on calcule une somme, dite de contrôle [5] sur l'image et on la compare avec d'autres sommes disponibles sur Internet ce qui permet ensuite au système de vérifier si l'image a déjà été utilisée dans pourriel.[17, 33, 23]

3.2.5.2 Les inconvénients

Cependant, ces différentes techniques ne sont pas infaillibles. En effet, il est possible d'envoyer des images avec des fonds de couleurs variables ou bien changer le bruit de l'image de telle façon que seuls les humains peuvent distinguer le message voulu. De plus, pour combattre les techniques de computer vision (en français vision de l'ordinateur) comme l'OCR [6], les spammeurs utilisent les méthodes CAPTCHA [7]. En outre , les techniques de computer vision nécessitent des ressources CPU énormes ce qui les rendent inefficaces à grande échelle.[33]

5. D'après Wikipédia, une somme de contrôle est séquence de données numériques calculée à partir d'un bloc de données plus important comme un message ou un fichier. Leur précision énorme lorsqu'on veut savoir si ce bloc a été préservé lors d'une opération de copie,stockage ou de transmission, constitue leur intérêt majeur.
6. Optical Character Recognition
7. Completely Automated Public Turing test to tell Computers and Humans Apart

3.2.6 Analyse heuristique

L'analyse heuristique est une technique de filtrage fondée sur l'analyse du contenu des messages. En effet, cette technique analyse, teste et note la présence des formes et des codes spécifiques[9]. Parmi les différents types de tests, on peut citer à titre d'exemple, la vérification si l'objet est vide ou non ou bien le calcul du pourcentage du contenu du message visant l'acquisition facile d'argent par rapport à la partie restante du message

Après avoir soumis les courriels à plusieurs vérifications agissant sur l'ensemble du contenu textuel du message et son en-tête. Un score, appelé *Spam Score*, est ensuite calculé à partir des différents résultats issus de l'ensemble des tests appliqués. Ce score sert à la fin à classer la nature du message, et ce en le comparant avec un seuil prédéfini par l'utilisateur et réalisant le meilleur balancement entre le taux des faux positifs et négatifs. Cependant, il existe des exceptions à cette règle :

— Si le mail est reçu à partir des expéditeurs connus par le destinataire ou bien à qui il a accordé son consentement alors dans ce cas, il est hors de question que le message soit considéré comme non sollicité (Opt-In).
— En revanche, si le destinataire n'a pas octroyé son accord alors le message est systématiquement considéré comme du spam (Opt-Out).
— De surcroît, il est possible pour l'utilisateur de définir des réglementations propres à lui et dans ce cas, les messages ne vont subir aucun test et vont être considérer comme légitimes. [17, 31, 19]

3.2.7 Filtrage basé sur le Machine Learning

3.2.7.1 Définition et introduction au Machine Learning

Selon *Arthur Samuel*[8], Le *Machine Learning*[9] est le domaine d'étude qui permet aux ordinateurs d'apprendre sans être explicitement programmés. L'apprentissage automatique est une forme d'intelligence artificielle qui se base sur la création de systèmes ou de modèles qui apprennent ou améliorent leurs performances en fonction des données qu'ils traitent. La création et le développement du modèle se font toujours en passant par plusieurs étapes. Ces étapes peuvent être résumées en quatre étapes principales :

1. Sélectionner et préparer un ensemble de données d'entrainement ou *Dataset*. Cet ensemble, qui doit être bien préparé, sert à nourrir le modèle de Machine Learning.
2. Sélectionner l'(les) algorithme(s) à exécuter.

8. Arthur Samuel, né le 5 décembre 1901 et mort le 29 juillet 1990, est un pionnier américain du jeu sur ordinateur, de l'intelligence artificielle et de l'apprentissage automatique
9. En français : apprentissage automatique

Le choix de l'algorithme dépend du type et du volume des données d'entrainement.

3. On passe ensuite à la phase d'entrainement de l'algorithme

   Il s'agit d'un processus itératif dans lequel les résultats fournis par l'algorithme sont comparés à ceux prédéfinis. Ce processus se répète plusieurs fois jusqu'à ce que les résultats fournis soient corrects la plupart du temps.

4. Enfin, il faut utiliser et améliorer le modèle, en le testant sur des nouvelles données.

Le Machine Learning se fonde sur les algorithmes. En effet, on distingue deux principaux types d'algorithmes d'apprentissage : l'apprentissage supervisé et l'apprentissage non supervisé. [28]

3.2.7.2 Les méthodes d'apprentissage automatique

L'apprentissage supervisé

En apprentissage supervisé, on présente aux algorithmes les entrées ainsi que les sorties voulues, et ce en vue de determiner une corrélation entre les deux. En apprentissage supervisé, l'algorithme apprend grâce à un jeu de données comportant les solutions voulues appelées « étiquettes ». Dans la pratique, les algorithmes de Machine Learning supervisés sont les plus couramment utilisés.[34]

En apprentissage supervisé, on distingue les problèmes dits de *Classification* et ceux de *Regréssion*. Les algorithmes de classification et de régréssion sont utilisés pour faire des prédictions. En effet, si le but du problème est de prédire des valeurs continues comme les prix, les salaires, l'âge, etc ... alors c'est un problème de régression. Si l'objectif du problème est de prédire des valeurs discrètes comme mâle ou femelle, vrai ou faux, etc ... alors c'est un problème de classification. Ce dernier s'avère très utile pour traiter les problèmes de détection des spams. Voici une liste non exhaustive des principaux algorithmes de classification :

— Classifieur naïf de Bayes
— Les machines à vecteurs de support ou SVM
— L'arbre de décision
— Les k-plus proches voisions
— Réseau neuronal artificiel

L'apprentissage non supervisé

En apprentissage non supervisé, les modèles sont entrainés en utilisant des bases d'apprentissage non étiquetées. Les algorithmes non supervisés reconnaissent les motifs et les corrélations entre les relations.

La différence entre les deux types d'apprentissage réside dans la méthode employée dans le traitement des données afin de faire des prédictions.[8]

3.2.8 Les principaux algorithmes de classification

3.2.8.1 Le classifieur naïf de Bayes

Le classifieur de bayésien naïf est une méthode d'apprentissage qui repose sur une hypothèse simplificatrice forte :

Les variables X_i sont mutuellement et conditionnellement indépendantes par rapport à C.

Où :

— X_i est une variable explicative, $i = 1 \dots n$

— C est une classe du problème de classifcation [15, 7]

Sous cette hypothèse *naïve* l'interaction entre les différentes variables peut-être négligée. En effet, cela a peu d'importance.

Ce classifieur repose sur la formule de Bayes :

$$P(C \backslash X) = \frac{P(C)P(X \backslash C)}{P(X)}$$

avec :

— $P(C)$ et $P(X)$ sont respectivement les probabilités de C et X

— $P(X \backslash C)$ et $P(C \backslash X)$ sont respectivement les probabilités de X sachant C et de C sachant X

Écrivons ce théorème d'une autre manière :

$$P(C \backslash X_1 \cap \dots \cap X_n) = \frac{P(C)P(X_1 \cap \dots \cap X_n \backslash C)}{P(X_1 \cap \dots \cap X_n)}$$

En langage courant, cela signifie :

$$\text{postérieure} = \frac{\text{antérieure} \times \text{vraisemblance}}{\text{evidence}}$$

La probabilité $P(X \backslash C)$ est très difficile à calculer, c'est là où intervient l'hypothèse *naïve*. Dans ce cas, la probabilité [10] de la k$^{\text{ème}}$ classe C_k devient :

$$P(C_k \backslash X) = \frac{P(C_k) \prod_i P(X_i \backslash C_k)}{\sum_{j=1}^{k} P(C_j) \times \prod_i P(X_i \backslash C_j)}$$

10. La démonstration de ce théorème peut être trouvée dans Wikipédia

3.2.8.2 Application du classifieur bayésien naïf dans la détection des spams : Le filtre bayésien

Ce classifieur linéaire, introduit par *M.Paul Graham* en 2002, permet d'avoir un taux de détection très élevé dépassant les 99% et avec une erreur relative minime. En réalité, les probabilités d'apparition dans un pourriel diffèrent d'un mot à un autre. À titre d'exemple, il est plus probable de croiser les mots *Argent* et *Viagra* dans un spam que dans un courrier légitime. [11]

Comme tout algorithme d'apprentissage automatique, on a recours à une phase d'entrainement pour évaluer les probabilités antérieures et postérieures. En fait, on divise l'ensemble des données disponibles en deux sous-ensembles. Le premier set sert essentiellement à améliorer l'algorithme et on doit indiquer manuellement pour chaque message contenu dans ce groupe s'il est légitime ou non tout en ajustant simultanément les probabilités c'est-à-dire on les calcule de nouveau après l'ajout et la sauvegarde de chaque nouveau message dans la base de données.

Après cette phase, on teste l'algorithme entrainé et c'est là où le deuxième set intervient. On sait bien au préalable que la détection du spam est un problème de classification à deux classes : Spam « S » et Ham « H » et donc notre objectif est de calculer la probabilité qu'un message donné soit un spam en sachant qu'il contient un mot donné « M ». Cette probabilité s'obtient par le biais du théorème de Bayes :

$$P(S\backslash M) = \frac{P(S)P(M\backslash S)}{P(S)P(M\backslash S) + P(H)P(M\backslash H)}$$

Si on part de l'intuition, il n'y a aucun prétexte pour lequel la probabilité qu'un message quelconque est considéré comme spam est différente de celle d'avoir un message légitime et donc :

$$P(H) = P(S) = 0.5$$

Pourtant des statistiques de Symantec[20, 22] ont montré que la probabilité qu'un message quelconque soit un spam est considérable et vaut au minimum 80% et donc :

$$P(S) = 0.8$$
$$P(H) = 0.2$$

Bien entendu, savoir si un message est un spam ou non en se basant uniquement sur la présence d'un seul mot peut conduire évidemment à l'erreur. Pour remédier à ce problème, on considère

plusieurs mots puis on combine leurs spamicité [11] par l'intermédiaire de la formule ci-dessous :

$$P = \frac{p_1 p_2 \dots p_n}{p_1 p_2 \dots p_n + (1-p_1)(1-p_2)\dots(1-p_n)}$$

avec :

— $p_i = P(S \backslash M_i)$

— P = La probabilité que le message soit un spam

— p_i = La probabilité que le message soit un spam sachant qu'il contient le mot M_i

En guise de conclusion, bien que les hypothèses des classifieurs naïfs bayésiens soient simples, ces derniers ont fait preuve d'une efficacité plus que suffisantes dans beaucoup de situations.

3.2.8.3 Algorithme de K-NN

L'algorithme des *K-plus proches voisins* ou *K-Nearest Neighbours (KNN)* est un algorithme de Machine Learning qui appartient à la classe des algorithmes d'apprentissage supervisé. On suppose que l'ensemble E contient n données labellisées et u une autre donnée n'appartenant pas à E qui ne possède pas de label. On souhaite connaître la classe de u. Pour cela, on fixe une fonction d qui renvoie la distance (qui reste à choisir) entre la donnée u et une donnée quelconque appartenant à E et un entier k inférieur ou égal à n.
Le principe de l'algorithme de k-plus proches voisins repose sur les étapes suivantes :

1. On calcule les distances entre la donnée u et chaque donnée appartenant à E à l'aide de la fonction d.
2. On retient les k données du jeu de données E les plus proches de u.
3. On attribue à u la classe qui est la plus fréquente parmi les k données les plus proches.

Le principe de cet algorithme justifie l'importance du choix de fonction distance. En effet, ce choix dépend essentiellement des types de données qu'on manipule. À titre d'illustration pour les données quantitatives (exemple : poids, salaires, taille, montant de panier électronique etc...) et du même type, la distance euclidienne est un bon candidat. Quant à la distance de Manhattan, elle est une bonne mesure à utiliser quand les données (input variable) ne sont pas du même type (exemple : âge, sexe, longueur, poids etc...). Finalement, pour la détermination des similarités entre deux séquences de symboles, la distance de Hamming est un bon choix.
Rappel : Normes de Hölder
Pour un entier $p \geq 1$, les normes de Hölder sont définies par :

11. La probabilité qu'un mot ou qu'un ensemble de mots, appartienne à un spam.

$$\|x\|_p = (|x_1| + \cdots + |x_n|)^{\frac{1}{p}} \; ; \forall x \in \mathbb{R}^n$$

Pout tout x et y de $\mathbb{R}^n$, On définit la distance entre x et y associée à la norme $\|.\|$ par :

$$d(x,y) = \|x - y\|$$

Par définition, la distance euclidienne est la distance associée à la norme 2 et la distance de Manhattan est celle associée à la norme 1.
Pour $x = (x_i)_{0 \leq i \leq n-1}$ *et* $y = (y_i)_{0 \leq i \leq n-1}$, la distance de Hamming $d_H(x,y)$ entre x et y est :

$$d_H(x,y) = Card(\{i \; tq \; x_i \neq y_i\})$$

Où : Card(E) désigne le cardinal de l'ensemble E

Le choix de k est important. En effet, la détermination de la classe dominante dépend du nombre des k plus proches voisins. Pour cela plus k est petite, plus le taux d'erreur est élevé. Par ailleurs, plus on utilise de voisins (un nombre K grand), plus on sera fiable dans notre prédiction. Toutefois, si on utilise un nombre de k voisins avec $k = N$ et N étant le nombre d'observations, on risque d'avoir du overfitting et par conséquent un modèle qui se généralise mal sur des observations qu'il n'a pas encore vu.

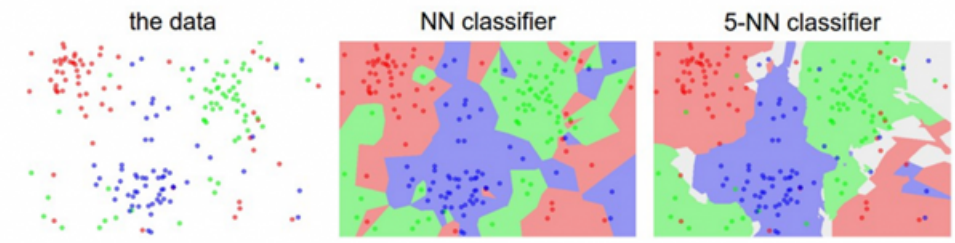

FIGURE 3.1 – Comparaison entre les classifieurs N-NN et 5-NN

3.2.8.4 Application dans la détection des spams de l'algorithme K-NN

Le classifieur *KNN* est considéré comme un classifieur qui se base sur l'exemple c'est-à-dire que les données d'apprentissage sont utilisées pour la comparaison. Quand un nouveau message doit être classifié les k messages les plus proches (voisins) sont trouvés et si une proportion suffisante a été classifiée dans une certaine catégorie alors le nouveau document est lui aussi attribué à cette catégorie. Pour décider que le message est légitime ou non, on considère la catégorie de messages les plus proches de lui.
Voici l'algorithme des k voisins les plus proches :

1. L'apprentissage

   On stocke les données d'apprentissage

2. Filtrage

   Étant donné un message X, on determine les k voisins les plus proches dans les données d'apprentissage. Si la proportion des spam dans les k voisins est largement suffisante alors le message est un spam.

L'un des inconvénients de cette méthode est qu'il n'y a pas de paramètres imposés pour contrôler le nombre de faux positifs. Ce problème est facilement résolu si on change le critère de classification au critère suivant :

Si, pour un l donné, on a l ou plus messages spams dans les k plus proches voisins de X alors ce dernier est considéré comme un spam, sinon X est un ham

Comme on l'a indiqué au paragraphe précédent, le choix du réel k est de la plus haute importance. En effet, l'inconvénient majeur de cette méthode est que la précision des résultats dépend essentiellement du choix de k. Ce choix affecte directement les taux des faux positifs ou des faux négatifs. Pour contrer ce problème, on utilise un autre classifieur appelé *SNN* qui va ajuster automatiquement la valeur de k. L'avantage substantiel que ce dernier présente est qu'il permet de trouver la véritable valeur de k et donc il assure une classification beaucoup plus précise. [23, 1]

3.2.8.5 Machines à Vecteurs de Support : SVM

L'algorithme des machines à vecteurs de support ou SVM [12], introduit par *M.Vapnik* en 1995, est une méthode de classification par apprentissage supervisé. Le principe fondamental des SVM consiste à ramener le problème de discrimination à celui de la recherche d'un hyperplan optimal qui va séparer deux ensembles de points (données) en se basant sur l'emploi de fonctionnelles dites *Noyau* (*Kernel* en anglais). Les points les plus proches de cet hyperplan sont appelés *Vecteurs de support*. On appelle *La marge*, la distance minimale entre l'hyperplan et l'ensemble d'apprentissage. On peut donc dire que l'hyperplan séparateur optimal est celui qui maximise la marge afin de bien classer les nouveaux éléments.
À fin d'atteindre cet objectif, deux astuces ont été introduites :

1. La première consiste à définir l'hyperplan comme solution d'un problème d'optimisation sous contraintes dont la fonction objective ne s'exprime qu'à l'aide de produit scalaire entre vecteurs.

12. Support-Vector Machine

2. Le passage à la recherche des surfaces séparatrices non linéaires est obtenu par l'introduction d'une fonction noyau dans le produit scalaire induisant implicitement une transformation non linéaire des données vers un espace intermédiaire. [31, 34]

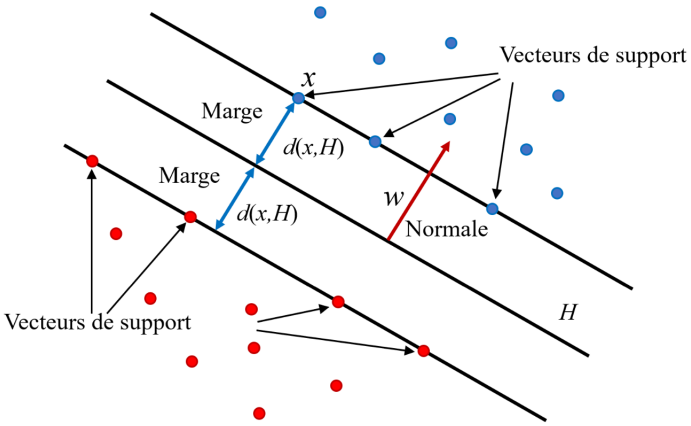

FIGURE 3.2 – Hyperplan, marge et vecteurs de support

3.2.8.6 Utilisation des SVM pour classifier le spam

Dans la pratique, l'une des applications du SVM est qu'on le combine avec les classifieurs bayésiens naïfs [5]. On cherche à exploiter les points forts des classifieurs bayésiens et des SVM dans la détection des spams et le text-mining. Cette méthode a été présentée par Chui-Yu Chiu et Yuan-Ting Huan. Tout d'abord, on doit faire la sélection des caractéristiques car leur nombre peut être élevé donc il est avantageux de considérer seulement celles qui représentent les messages le mieux avant de passer à la phase d'apprentissage du filtre. [13] Pour cela, on utilise la méthode TF-IDF. Cette méthode consiste à pondérer les différents termes d'un document par suite évaluer l'importance de chacun d'entre eux dans le document, relativement à un corpus donné.[14] On définit la fréquence tf_m d'un mot m comme étant le nombre d'occurences de ce mot dans le texte. On introduit également la fréquence inverse de document idf_m pour modéliser le poids du mot m dans le corpus. Elle peut être déterminée par le biais de la formule :

$$idf_m = \log \frac{|D|}{|E|}$$

Où :

— $E = \{d\,;\, m \in d\}$ et d signifie un document

13. La selection des caractéristiques ou Feature Selection consiste à choisir un sous-ensemble de taille M caractéristiques à partir d'un ensemble original de manière à les réduire d'une façon optimale.

— $|D|$ = Le nombre total de documents du corpus
— $|E|$ = Le nombre de document où le mot m apprait

Le poids du mot m dans le corpus est le produit de tf_m et idf_m :

$$p_m = tf_m \times idf_m$$

Ensuite, on détermine les antérieures du classifieur bayésien naïf puis on passe aux probabilités à posteriori. Par la suite, on se sert des résultats fournis par la méthode TF-IDF et des probabilités conditionnelles du classifieur déja calculées en vue d'entrainer l'algorithme des machines à vecteurs de support pour enfin se servir de ce dernier à fin d'effectuer la classification des données. [31]

Dans cette partie, nous nous sommes penchés sur la deuxième phase de filtrage d'un e-mail. En effet, une fois qu'un courriel passe la phase de filtrage d'enveloppe avec succès, son contenu est analysé à travers des filtres superficiels et des filtres plus approfondis.

Le filtrage superficiel se base sur la présence de particularités propres aux spams au niveau des mots, des images, des adresses URL, etc. Ce type de filtrage est assez simple et ne requiert pas beaucoup de ressources. Cependant, il peut être moins efficace que le filtrage approfondi.

Le filtrage approfondi, quant à lui, utilise des algorithmes de machine learning tels que le classifieur bayésien naïf, la méthode des KNN et enfin SVM pour déterminer la classe de l'e-mail reçu. Ce type de filtrage est plus complexe et plus précis que le filtrage superficiel. Il permet de réduire les faux positifs et d'augmenter la précision de la détection de spams.

Ceci étant dit, il est légitime de se questionner sur l'importance du filtrage d'enveloppe. Toutefois, il convient de rappeler que le corps d'un e-mail n'est transféré que si l'en-tête est accepté. Pour cette raison, le filtrage d'enveloppe permet de détecter certains spams à l'avance en se basant sur des critères tels que l'adresse IP ou l'adresse de l'expéditeur et de les bloquer dés le début. Ceci permet de réduire le temps de traitement en évitant de dépenser du temps et de l'énergie sur le filtrage de contenu si l'on connaît d'avance la réponse. De plus, transférer le corps d'un spam, que l'on peut détecter à travers l'en-tête, ne fait qu'alourdir le trafic du réseau et ralentir le processus d'envoi des courriels légitimes.

En somme, le filtrage d'enveloppe est une première ligne de défense importante contre les pourriels et les cyberattaques qui peuvent être véhiculées via le courrier électronique. Combinés avec un filtrage de contenu, ces deux types de filtrage permettent d'améliorer les performances globales du système de messagerie et de réduire le taux de faux positifs.

Conclusion générale

Le but de cette thèse était d'introduire les principales techniques de détection et de filtrage des spams e-mail dans un cadre général afin d'aider les lecteurs à comprendre la complexité du mécanisme qui se cache derrière l'étiquette "spam" et d'améliorer les chances que leurs courriels parviennent à destination. Comme présenté précédemment, le spam est l'envoi massif de courriels électroniques non sollicités à des personnes avec lesquelles l'expéditeur n'a jamais eu de contact et dont il a souvent obtenu l'adresse électronique illégalement. Sans filtres anti-spam, la gestion des pourriels peut coûter aux employés 2 minutes par e-mail et 28 jours par an, ce qui semble coûteux en temps et en mémoire du système.

Dans une première partie de ce projet, le but est d'expliquer certaines notions techniques pour aider le lecteur à explorer et à comprendre le monde de l'échange de courriers électroniques. Pour cela, nous nous sommes concentrés sur les différentes composantes de l'e-mail et sur son acheminement de l'écran jusqu'à sa destination finale. Nous nous sommes ensuite attaqués à la clarification du terme "spam" et des dangers qu'il pose.

En troisième lieu, cette dissertation a pour vocation d'illustrer les différentes techniques de filtrage des spams e-mail, d'expliquer leur fonctionnement et d'évaluer leur efficacité. Nous avons découvert dans cette partie que le filtrage se déroule en deux phases, l'une portant sur les informations fournies par l'en-tête et l'autre sur le contenu du courriel.

Après avoir exploré les différentes techniques, on ne peut qu'affirmer qu'aucune méthode n'a réussi jusqu'à aujourd'hui à réaliser un taux 100% de faux négatif et 0% de faux positif. En effet, toute nouvelle méthode ne s'attarde pas à être contourné par les spammeurs. Ce qui nous pousse à demander :

Arrivera-t-il un jour où on développe un filtre déterministe idéale ?

Bibliographie

[1] Khorsi AHMED : An overview of content-based spam filtering techniques. *Informatica*, 31(3), mai 2007.

[2] ALTOSPAM : Fonctionnement de la messagerie électronique, janvier 2019. *URL : https ://www.altospam.com/actualite/2019/01/comment-fonctionne-lemail/.*

[3] ALTOSPAM : Format et structure d'un courrier électronique, février 2016. *URL : https ://www.altospam.com/actualite/2016/02/format-et-structure-dun-email.*

[4] Schäfer CARLO : Detection of compromised email accounts used for spamming in correlation with mail user agent access activities extracted from metadata. *In 2015 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, pages 1–6, 2015.

[5] Gherabi CHARAF EDDINE : *Détection des spams se basant sur les techniques de classification*. Thèse de doctorat, Université Mohamed Boudiaf-M'sila faculté des mathématiques et de L'informatique, 2018.

[6] Pin-Ren CHIOU, Po-Ching LIN et Chun-Ta LI : Blocking spam sessions with greylisting and block listing based on client behavior. *In 2013 15th International Conference on Advanced Communications Technology (ICACT)*, pages 184–189, 2013.

[7] Salperwyck CHRISTOPHE, Lemaire VINCENT et Hue CARINE : Classifieur naïf de bayes pondéré pour flux de données. *In EGC*, pages 275–286, 2014.

[8] IBM Cloud EDUCATION : What is unsupervised Machine Learning ? *URL :https ://www.ibm.com/cloud/learn/unsupervised-learning*, septembre 2020.

[9] Passigue EKPAO ANANI : *Analyse et détection de pourriels textuels dans les réseaux sociaux par apprentissage*. Thèse de doctorat, Université du Québec en Outaouais, août 2015.

[10] Kaspersky ENCYCLOPEDIA : Techniques de protection contre les spams. *URL : https ://encyclopedia.kaspersky.fr/knowledge/spam-protection-technologies/*, juillet 2018.

[11] Rossi FABRICE : Filtrage de spam par méthodes probabilistes. *Multi-System and Internet Security Cookbook*, 8:74–80, juillet 2003.

[12] Calas GUILLAUME : Les techniques de classification de courriels. *EPITA*, 2009.

[13] Jung JAEYEON, Sit EMIL, Balakrishnan HARI et Morris ROBERT : Dns performance and the effectiveness of caching. *In Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*, pages 153–167, 2001.

[14] Menghour KAMILIA et Souici-Meslati LABIBA : Sélection de caractéristiques pour le filtrage de spams. *In CORIA*, pages 349–360, 2010.

[15] Wikipédia l'encyclopédie LIBRE : La classification naïve bayésienne.
URL : https ://fr.wikipedia.org/wiki/Classification_naïve_bayésienne, décembre 2022.

[16] Wikipédia l'encyclopédie LIBRE : Le spam.
URL : https ://fr.wikipedia.org/wiki/Spam, novembre 2022.

[17] Wikipédia l'encyclopédie LIBRE : La lutte anti-pollupostage.
URL : https ://fr.wikipedia.org/wiki/Lutte_anti-spam, janvier 2023.

[18] Stephane MANHES : Comment déterminer le coût d'un spam ? Altospam.
URL : https ://www.altospam.com/actualite/2010/02/comment-determiner-le-cout-d'un-spam/, février 2010.

[19] Da Cruz MARTINS : Filtrage de messagerie sur des gros serveurs. *Ecole des Mines de Paris*, janvier 2005.

[20] Dylan MORS et Dermot HARNETT : State of spam : a monthly report, septembre 2009.
URL :https ://eval.symantec.com/mktginfo/enterprise/other_resources/b-state_of_spam_report_09-2009.en-us.pdf.

[21] Dagorn NATHALIE : Sensibilisation aux coûts et conséquences du spam. *Terminal. Technologie de l'information, culture et société*, (105), 2010.

[22] Maghrebi Nour el HOUDA : La detection des spams. Mémoire de D.E.A., Faculté des sciences de l'ingénieur , Université Djili Liabes sidi Bel Abbes, 2014.

[23] Saad OMAR M., Darwish ASHRAF et Faraj RAMADAN : A survey of machine learning techniques for spam filtering. *IJCSNS International Journal of Computer Science and Network Security*, février 2012.

[24] Lagadec PHILIPPE : Filtrage de messagerie et analyse de contenu. *SSTIC*, juin 2004.

[25] Jonathan POSTEL : Rfc0821 : Simple mail transfer protocol, 1982.

[26] Têtue ROMY : Qu'est-ce qu'une adresse e-mail ? 1998.

[27] Nazirova SAADAT : Improvement of anti spam technology with the help of an estimation of reliability of the sender. 2008.

[28] Léa SCHAPIRA : Machine learning : Définition, fonctionnement, novembre 2021. *URL : https ://datascientest.com/machine-learning-tout-savoir*.

[29] Devrim SIPAHI, Gökhan DALKLÇ et Mehmet Hilal ÖZCANHAN : Detecting spam through their sender policy framework records. *Security and Communication Networks*, 8(18): 3555–3563, 2015.

[30] STATISTA : Number of e-mail users worldwide from 2017 to 2025, 2022. *URL :https ://www.statista.com/statistics/255080/number-of-e-mail-users-worldwide/*.

[31] Ousman YASSINE et PARAKH : *Une nouvelle approche pour la détection des spams se basant sur un traitement de données catégorielles*. Bibliothèque et Archives Canada, Ottawa, juillet 2013.

[32] Siddique ZEESHAN BIN, Khan MUDASSAR ALI, Din IKRAM UD, Almogren AHMAD, Mohiuddin IRFAN et Nazir SHAH : Machine learning-based detection of spam emails. *Scientific Programming*, 2021.

[33] Wang ZHE, Josephson WILLIAM K, LV QIN, Charikar MOSES et Li KAI : Filtering image spam with near-duplicate detection. *In CEAS*, 2007.

[34] Bahouche ZIANE, Hamza L et Belhadad HICHAM : *Détection des Spams basé sur le Machine Learning*. Thèse de doctorat, Université A. Mira-Bjaia, 2021.

[35] Comment ça MARCHE : Fonctionnement de la messagerie électronique, décembre 2022. *URL :https ://web.maths.unsw.edu.au/lafaye/CCM/courrier-electronique/fonctionnement-mta-mua.htm*.